Bayesian spatio–temporal modelling of rainfall through non–homogeneous hidden Markov models

Sarah Elizabeth Germain

Thesis submitted for the degree of
Doctor of Philosophy

School of Mathematics & Statistics
Newcastle University
Newcastle upon Tyne
United Kingdom

November 2010

I dedicate this thesis to Chris, who kept me going with his reassurance and love. I am also grateful to my Mum and Dad for their unwavering support and tolerance and to my brother, Peter, for his advice, encouragement and hugs.

Acknowledgements

I would like to thank my supervisors, Malcolm Farrow and Richard Boys, for their patience, guidance and kindness over the course of my postgraduate education. I would also like to extend my gratitude to Gavin Shaddick for giving me the confidence to complete this thesis. Further, I give my thanks to friends in the School of Mathematics & Statistics for their encouragement and support.

I am grateful to the Engineering and Physical Sciences Research Council for the funding which made it possible to conduct this research.

Abstract

Multi–site statistical models for daily rainfall should account for spatial and temporal dependence amongst measurements and also allow for the event of no rain. Recent research into climate change and variability has sparked interest in the relationship between rainfall and climate, stimulating the development of statistical models that relate large–scale atmospheric variables to local precipitation. Although modelling daily rainfall presents a challenging and topical problem, there have been few attempts taking a subjective Bayesian approach.

This thesis is concerned with developing hidden Markov models (HMMs) for the spatio–temporal analysis of rainfall data, within a Bayesian framework. In these models, daily rainfall patterns are driven by a finite number of unobserved states, interpreted as weather states, that evolve in time as a first order Markov chain. The weather states explain space time structure in the data so that reasonably simple models can be adopted within states. Throughout this thesis, the models and procedures are illustrated using data from a small dense network of six sites situated in Yorkshire, UK.

First we study a simple (homogeneous) HMM in which rainfall occurrences and amounts, given occurrences, are conditionally independent in space and time, given the weather state, and have Bernoulli and gamma distributions, respectively. We compare methods for approximating the posterior distribution for the number of weather states.

This simple model does not incorporate atmospheric information and appears not to capture the observed spatio–temporal structure. We therefore investigate two non–homogeneous hidden Markov models (NHMMs) in which we allow the transition probabilities between weather states to depend on time–varying atmospheric variables and successively relax the conditional independence assumptions. The first NHMM retains the simple conditional model for non–zero rainfall amounts but allows occurrences to form a Markov chain of autologistic models, given the weather state. The second introduces latent multivariate normal random variables to form a hierarchical NHMM in which neither rainfall occurrences nor non–zero amounts are conditionally spatially or temporally independent, given the weather state.

Throughout this thesis, we emphasise the elicitation of prior distributions that convey genuine initial beliefs. For each hidden Markov model studied we demonstrate techniques to assist in this task.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction and objectives of the thesis

Stochastic models of precipitation are either constructed to help planning and decision making or to enhance our understanding of the characteristics and dynamics of the rainfall process. These objectives need not be mutually exclusive. Typically, the more practically motivated models provide a statistical description of the data, rather than the underlying rainfall generating mechanism. However, by identifying relationships and trends within the data and summarising its salient features, they can be helpful in achieving improved insight. In terms of assisting in planning and decision making, stochastic models motivated by *either* objective are useful for generating precipitation series with similar statistical properties to the data used to fit the model. Realistic sequences of rainfall data play an important role in "synthetic" hydrology, ecology and agriculture where they constitute inputs in many models, such as models of flooding, runoff and crop growth. This is particularly pertinent when historical records are of insufficient spatial and/or temporal coverage to evaluate important characteristics of the rainfall process.

In response to questions regarding the potential effects of climate variability and change, recent years have seen an increasing interest in the relationship between rainfall and climate. This has led to the development of stochastic models which provide a link between synoptic (large scale) atmospheric variables and small scale precipitation fields. These models can then, for example, be used to associate particular synoptic atmospheric patterns with the frequencies and characteristics of periods of flooding or drought. This information can give insight into how a water resource system might handle different climatic conditions. The development of *general circulation models* (GCMs) has further stimulated research into these (non-stationary) stochastic models which are capable of predicting rainfall under conditions of altered climate.

GCMs are deterministic mathematical models of the general circulation of the atmosphere. They are based on large systems of differential equations, which are solved numerically and chosen to provide a realistic representation of the physical processes involved. For example, fundamental laws of physics are incorporated, such as conservation of energy, momentum and mass. Given a particular set of initial conditions, the numerical solution of these equations allows simulations of the Earth's atmosphere to be performed on a resolution which is typically constrained to a scale

of approximately 2–5° of longitude and latitude. However, the questions concerning rainfall that are normally of interest in hydrology, agriculture and other scientific fields usually concern local patterns of rainfall over a much finer spatial scale. This has led to the so–called *downscaling* problem in which the idea is to turn predictions over large spatial scales into predictions over smaller spatial scales. There are essentially two ways of addressing this problem. The first is a dynamic model approach, in which regional climate models use the GCM output as initial and lateral boundary conditions to produce simulations of the climate on a finer resolution. Such techniques are highly computationally intensive. The other alternative, in which our interest lies, is *statistical downscaling.*

One class of statistical downscaling models are *weather state models*, first introduced by Hay *et al.* (1991) and also studied by, for example, Bardossy & Plate (1992) and Fowler *et al.* (2000). Using the word "day" as a generic description of a unit of time, weather state models deterministically classify each day into one of a small number of weather states based on the observed atmospheric information, then model precipitation conditionally on the weather state. The effect of climate variability on local precipitation processes can be assessed by first using historical data to make inference about the parameters of the stochastic model. The weather state model is then used to downscale repeated GCM sequences of atmospheric data produced under current climate conditions. Taking a more speculative approach, weather state models can also be used to study the effect of altered climate by using GCM output obtained under altered climate scenarios as input into the stochastic model. However, predictions obtained in this way are based on an implicit assumption that the relationship between the atmosphere and rainfall remains constant under the altered climate conditions. This might be unrealistic in practice.

Hughes & Guttorp (1994*a*) proposed a broad class of spatio–temporal models, referred to as *non–homogeneous hidden Markov models* (NHMMs), linking local precipitation to atmospheric circulation patterns, and showed that these models included the weather state model as a special case. The NHMM differs from the classic weather state model in that the weather states are not determined *a priori*, instead being inferred from the data. Although they are just artefacts of the statistical model, the weather states represent clusterings of distinct precipitation patterns that are likely to be associated with particular atmospheric conditions. The role of the atmospheric data is to influence the temporal evolution of the weather state.

The main objective of this project is to develop homogeneous and non–homogeneous hidden Markov models to model jointly the processes of rainfall occurrence and amount, within a Bayesian framework. The secondary objective is to develop and demonstrate the use of techniques to assist in the task of *elicitation.* This was defined by Garthwaite *et al.* (2005) as the process of formulating an individual's knowledge and beliefs about one or more uncertain quantities into a (joint) probability distribution. In the context of this thesis, this arises in the specification of prior distributions for the model parameters. With regards to the secondary objective, in addition to the work contained in subsequent chapters, we also conducted a thorough investigation to develop prior distributions for the variance matrix of a multivariate normal distribution that were capable of conveying genuine initial beliefs. Although space does not permit full details to be provided in this thesis, the work is available in a technical report, Germain *et al.* (2010*b*).

1.2 Outline of the thesis

The remainder of this thesis is organised as follows. Chapter 2 contains an exploratory examination of the dataset, comprising rainfall measurements at a small network of sites in Yorkshire, that is analysed in subsequent chapters. This includes an introduction to the atmospheric variables on which later models are conditioned, as well as a summary of investigations into the spatial and temporal characteristics of the data. Chapter 3 introduces hidden Markov models and the philosophy and principles of Bayesian statistics, including the problem of model choice/averaging. Inference in hidden Markov models is then formulated in a Bayesian framework and we discuss, comparatively, methods of approximating posterior model probabilities. In the context of hidden Markov models, such techniques can be used to approximate the posterior distribution for the number of hidden states.

In Chapter 4 we present a simple (homogeneous) hidden Markov model for rainfall which assumes that rainfall occurrences and rainfall amounts, given occurrences, are conditionally independent in space and time, given the weather state. Occurrences and non–zero amounts are then modelled as Bernoulli and gamma random variables, respectively, with site (and state) specific parameters. Various within model simulation techniques are available for approximating the posterior distribution of the number of weather states, which we regard as unknown. Following the discussion in Chapter 3, we perform a simulation experiment to compare the performance of a number of these methods, including the recently proposed power posterior approach (Friel & Pettitt, 2008). The model is applied to the Yorkshire dataset, introduced in Chapter 2, and we assess the fit of the model by comparing observed statistics to their posterior predictive distributions.

Chapter 5 presents an NHMM which extends the model from Chapter 4 to allow atmospheric information to influence the probabilities of transition between weather states. Additionally, the assumptions of conditional spatial and temporal independence between rainfall occurrences are relaxed, modelling them as a Markov chain of autologistic models, given the weather state. We conclude by applying the model to the Yorkshire dataset and assessing the fit.

In Chapter 6 we introduce a hierarchical NHMM based on the incorporation of latent multivariate normal random variables. The model is constructed in such a way that the assumptions of conditional independence in space and time, given the weather state, can be relaxed for both rainfall occurrences and non–zero rainfall amounts. Treating latent variables equally as parameters, this is an example of a model for which the set of values with non–zero density in the prior and posterior do not coincide. We explain that for this kind of model, the power posterior approximation of the marginal likelihood requires a correction term. We close the chapter by applying the model to the Yorkshire dataset and comparing the fit with that obtained under the earlier models. Note that for reference in the "modelling" chapters (4, 5 and 6), Appendix E details the main variables and parameters that are used.

Chapter 7 summarises our conclusions and suggests topics for future work. We also provide an overview of our findings when the three hidden Markov models studied in this thesis are applied to a larger, more spatially diffuse network of sites in the UK.

Chapter 2

Exploratory data analysis

2.1 Background

In this chapter, we explore the set of precipitation data which will be analysed in the subsequent chapters of this thesis. This dataset is from a small, dense network of six sites situated in Yorkshire. Data on daily precipitation are available for the winters (December to February) for a period of 30 years (that is, 2707 days). The sites were chosen from a larger dataset and we followed Fowler *et al.* (2000) by selecting six sites which are evenly distributed over the region and give data over the entire 30 year period (although there are small pockets of missing data at two of the sites) with good spatial and altitudinal cover. The mean distance between sites in this network is 82.8 km, with the minimum and maximum being 39.8 km and 133.2 km, respectively. Additional information is available in the form of atmospheric data, comprising a classification of days according to the objective Lamb weather type (LWT) scheme (Jenkinson & Collinson, 1977), which will be explained in Section 2.2, and used as covariate information in the remainder of this thesis.

The locations of the sites in the Yorkshire network can be seen in Figure 2.1. This map also shows the locations of twelve sites from a spatially diffuse network located throughout the entire UK. Further discussion of this larger dataset is deferred until Chapter 7.

According to the American Meteorological Society, in British climatology a *rain day* is defined as a 24 hour period in which at least 0.01 inches or 0.2 mm of precipitation is recorded (Glickman, 2000). Hence 0.2 mm was used as the cutoff for classifying days as wet or dry. For the Yorkshire network, Table 2.1 shows summaries of the proportion of wet days, precipitation on those wet days together with the proportion of missing values and altitude. There are no missing values for four of the six sites. In the remaining two, the proportion of missing values is very low (2.1%). Typically, just over 50% of the days are wet with higher proportions often being associated with sites at higher altitude. Similar patterns are observed for mean daily precipitation on wet days, where typically values are in the range 3.2–3.6 mm with the values observed at higher altitude sites being markedly higher, an extreme case being 10.8 mm for Moorland cottage (site 3) which is at an altitude of 343 m. This is one of two Yorkshire sites located in the Pennines, the other being Great Walden Edge (site 5) at an altitude of 346 m. Standard deviations are generally

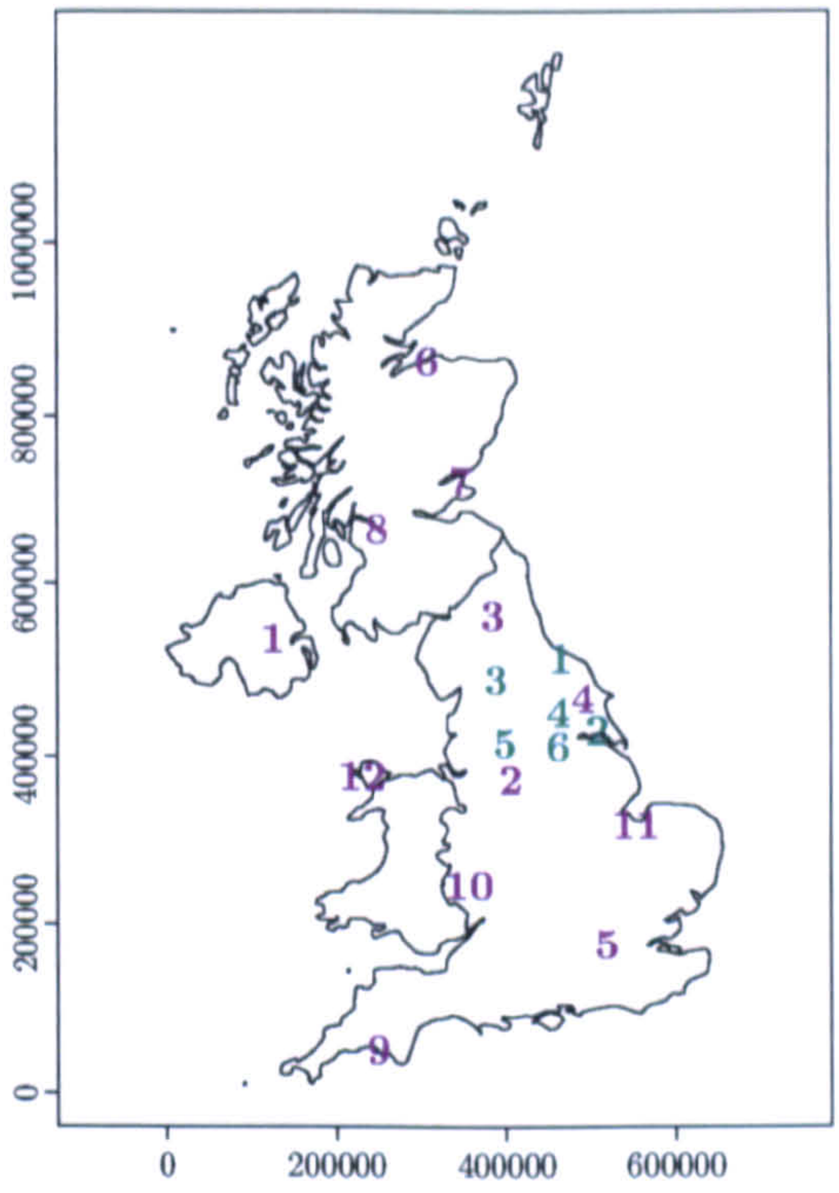


Figure 2.1: Locations of sites within Yorkshire (1–6) and UK (1–12) networks. Axes denote metres from the south eastern point of the British National Grid coordinate system (latitude 49° north, longitude 2° west).

slightly elevated compared to the means, reflected by the coefficients of variation being slightly greater than one.

The data analysis begins with an exploratory look at the mean wet day precipitation and proportion of wet days by site, averaged over the whole period, with regard to the Lamb weather type classification (Section 2.2). This is followed by an examination of the periods of missing values (Section 2.3). Brief details of our exploratory analysis are then provided, firstly considering the possibility of spatial structure and short or long term temporal patterns in the data (Section 2.4). Next we give an overview of some formal time series modelling of the data from the individual sites (Section 2.5). In this final section, we work in a Bayesian framework and use simple Markov and autoregressive models to describe the temporal autocorrelation between rainfall occurrences and between non–zero rainfall amounts.

Although all measurements refer to precipitation, the terms rainfall and precipitation will be used synonymously throughout the remainder of this thesis.

2.2 Atmospheric data: Lamb weather types

Atmospheric data are available in the form of objective Lamb weather types (LWTs). Lamb (1972) developed a subjective weather type classification scheme based on daily synoptic charts

	Site	Altitude (m)	Proportion wet days (%)	Mean daily precip. (mm)	Coefficient of variation	Proportion missing (%)
1	Lockwood Reservoir	193	59.0	3.585	1.166	0.0
2	Pearson Park, Hull	2	55.3	3.243	1.284	2.1
3	Moorland Cottage	343	57.2	10.849	1.266	0.0
4	The Retreat, York	18	49.4	3.426	1.228	0.0
5	Great Walden Edge	346	67.7	5.824	1.197	0.0
6	Kirk Bramwith	7	42.3	3.565	1.189	2.1

Table 2.1: Summary of the data from the Yorkshire network for winter periods between 1961/2 and 1990/1. The mean and coefficient of variation for daily precipitation are based only on wet days.

which depict the state of atmospheric flow over the British Isles at surface level and at a specified height in the atmosphere. Under this scheme an expert analyst can use his or her judgement to determine the weather type on any day in order to give an indication of the daily steering of circulation systems. Jenkinson & Collinson (1977) developed an automated (sometimes called "objective") method for identifying these LWTs using daily gridded mean-sea-level pressure charts. From these data it is possible to calculate estimates of the dominant direction and speed of the flow, as well as its vorticity, which is related to whether a cyclone or anticyclone is present. Particular values of these measures are then associated with specific LWTs so that the classification provides a categorisation of the direction and synoptic type of the surface flow over the British Isles on any particular day.

The Jenkinson classification scheme contains eight main directional types: north (N), north-east (NE), east (E), south-east (SE), south (S), south-west (SW), west (W) and north-west (NW); and three main non-directional types: anticyclonic (A), cyclonic (C) and unclassifiable (U). A further 16 hybrid types, which combine the eight main directional types with the anticyclonic or cyclonic non-directional type, are also recognised. This gives 27 possible objective LWTs, which are shown in Table 2.2. Days on which the vorticity is low and the flow is from the west, for example, will be classified as westerly types, whilst days on which the vorticity is

Label	Objective LWT	Label	Objective LWT	Label	Objective LWT
1	A	27	U	18	C
2	ANE	10	NE	19	CNE
3	AE	11	E	20	CE
4	ASE	12	SE	21	CSE
5	AS	13	S	22	CS
6	ASW	14	SW	23	CSW
7	AW	15	W	24	CW
8	ANW	16	NW	25	CNW
9	AN	17	N	26	CN

Table 2.2: Labelling of the objective Lamb weather types.

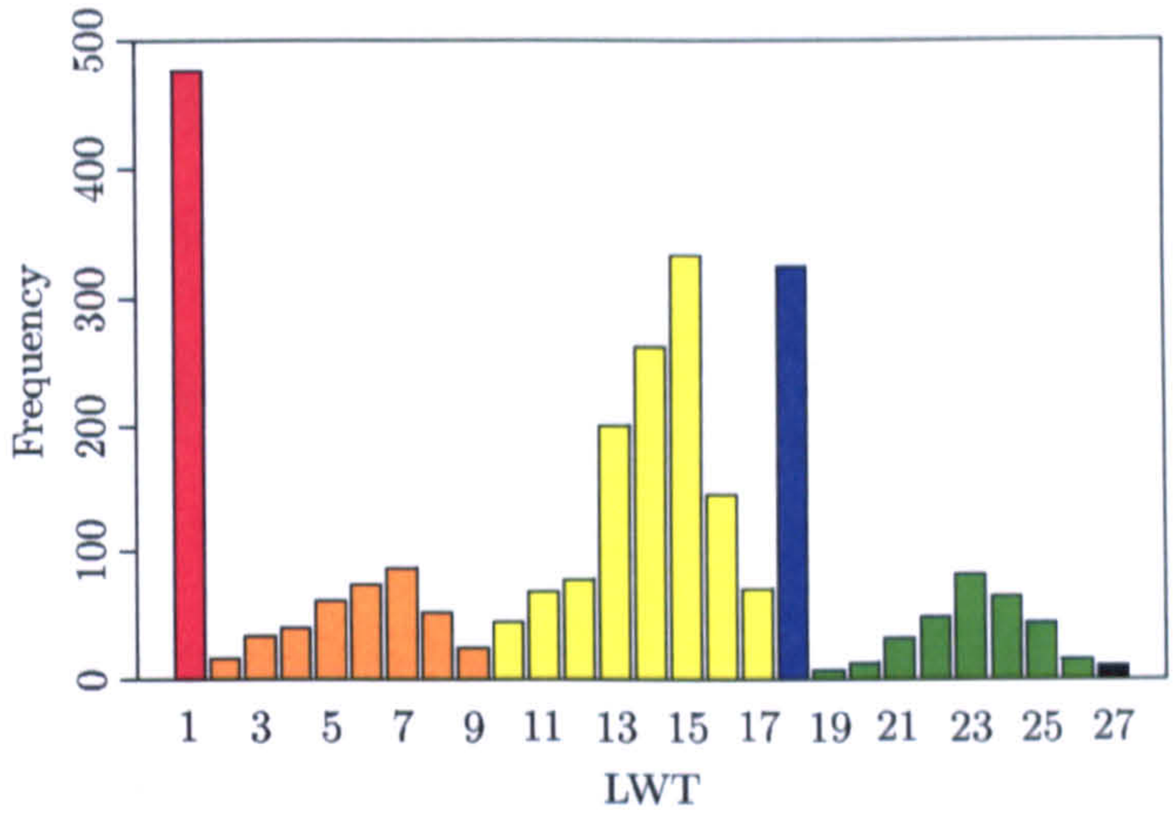


Figure 2.2: Frequencies of occurrence of Lamb weather types within the winters 1961/2–1990/1. Types 1 (and 8–9) are anticyclonic (hybrids), 10–17 are pure directional types, 18 (and 19–26) are cyclonic (hybrids) and 27 is unclassified. See Table 2.2 for further details.

strongly positive or negative will be categorised as cyclonic or anticyclonic, respectively. When the vorticity is only moderately positive or negative, the direction of air flow is also used to provide the classification into one of the hybrid types, for example, a day on which the vorticity is moderately positive and the flow is from the west would be classified as cyclonic westerly (CW). The unclassifiable type is provided to categorise the surface flow on days during which the atmospheric circulation is too complex for it to fall into any of the other types.

The objective Lamb classification scheme has been used to classify the weather type over the British Isles for every day in the period from 1880 to the present day; there are *no missing data* in this time series. The frequencies of their occurrence over the period 1961/2-1990/1 can be seen in Figure 2.2, with the most commonly occurring LWTs being pure anticyclonic (type 1), pure westerly (type 15) and pure cyclonic (type 18). Within the directional classifications, similar patterns in the occurrence of the individual LWTs are observed for both the pure and hybrid groupings.

The proportion of wet days by LWT can be seen in Figure 2.3(a) with the corresponding plots for the mean daily precipitation on wet days shown in Figure 2.3(b). Clear patterns can be seen in the proportion of wet days, with lower proportions being associated with anticyclonic types (1–9) and higher proportions with cyclonic types (18–26). This pattern is also seen to a lesser extent in the variation in mean wet day precipitation amounts across LWTs, where high (low) amounts are typically associated with cyclonic (anticyclonic) types. It also appears clear that for the two Pennine sites, Great Walden Edge and, in particular, Moorland Cottage, higher wet day precipitation amounts are associated with LWTs 5–8, 12–16 and 22–25, with the majority of these being westerly types. This observation was also noted by Fowler *et al.* (2000) who additionally state that northerly and easterly weather types are the main precipitation bearers for the eastern parts of Yorkshire, although this is not immediately apparent from this simple analysis.

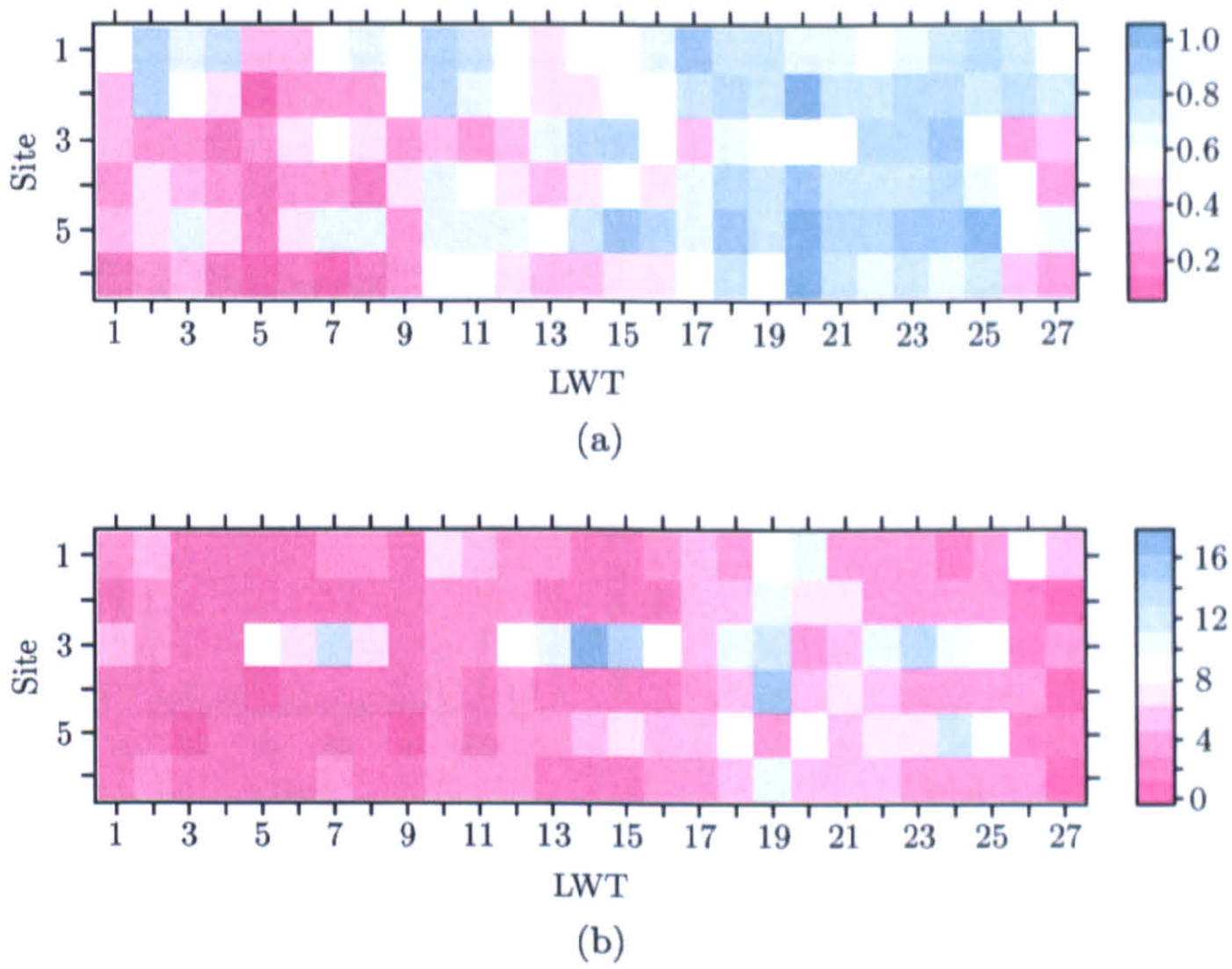


Figure 2.3: (a) Proportion of wet days and (b) mean wet day daily precipitation by Lamb weather type for sites in the Yorkshire network. See Table 2.2 for details.

2.3 Missing values

The periods of missing values in the Yorkshire data occur in non–overlapping sequences at two of the sites, Pearson Park (from 2nd January to 28th February 1991) and Kirk Bramwith (from 3rd January to 28th February 1981). Figure 2.4 shows the sequence of precipitation at all of the sites over the period of missing values at Pearson Park together with measurements from the previous ten days. The distance between Pearson Park and the other sites is also given. The corresponding plot for Kirk Bramwith was similar and is omitted. Neither the preceding values at either of the sites in question, nor the values at the other sites suggest that the presence of missing values could be attributed to abnormally high precipitation, which might have caused the rain gauge to malfunction. Indeed, given the observations at other sites it seems plausible that the missingness was not related to the amount of precipitation during this period.

2.4 Exploratory spatial and temporal analysis

We considered two measures of spatial similarity, one to assess the spatial autocorrelation in rainfall occurrences, and the other to assess spatial autocorrelation in rainfall amounts, given occurrence. For amounts, we use the Spearman's rank correlation coefficient (chosen to avoid having to make likely untenable assumptions about underlying normality), and for occurrences we use the log odds ratio, defined as follows. Consider two sites, i and j. Denote by n_{11} and n_{00} the numbers of days when it is wet and dry at both sites, respectively, by n_{01} the number of days at which it is dry at site i and wet at site j and by n_{10} the number of days at which it

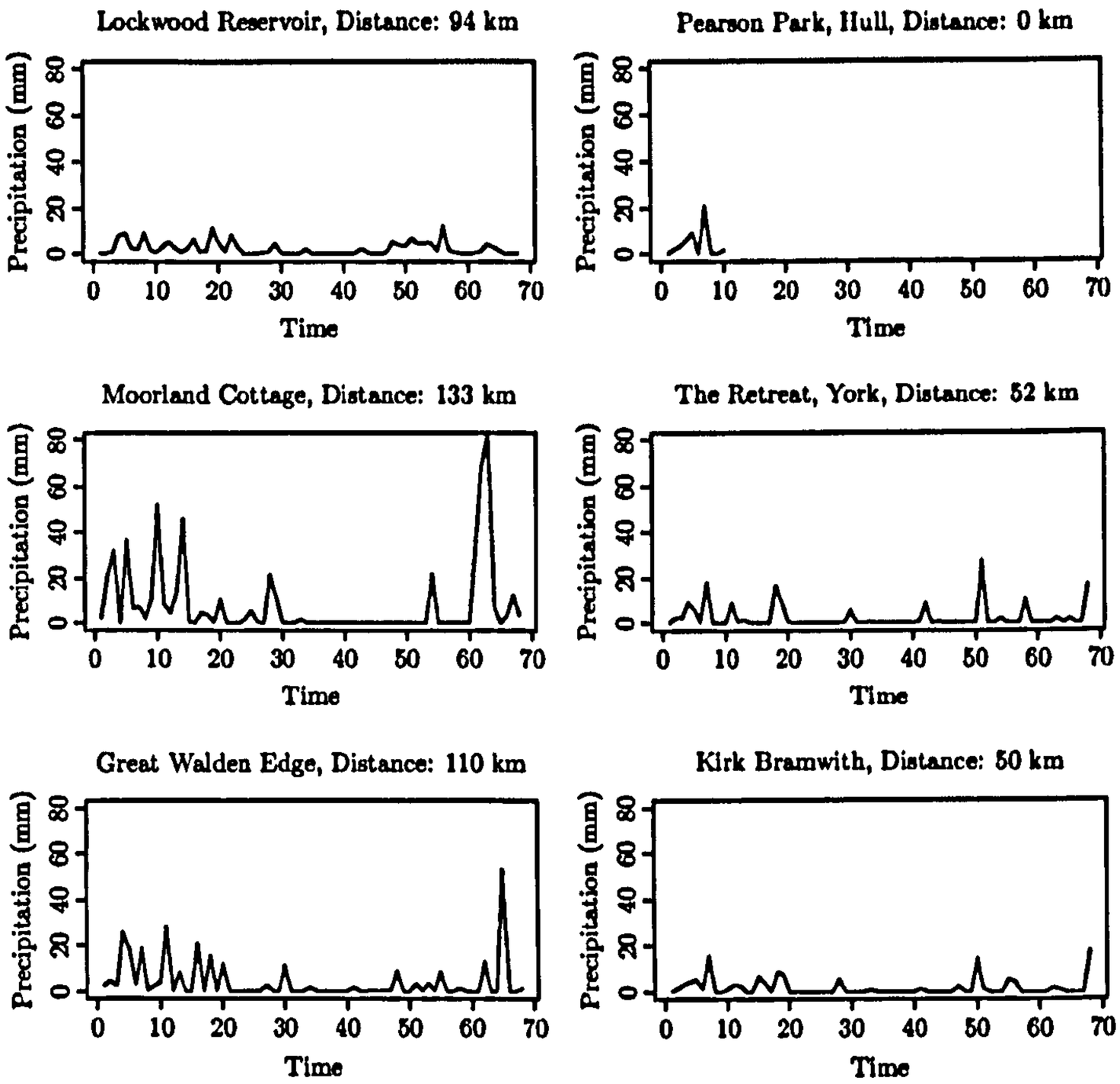


Figure 2.4: Pearson Park: precipitation behaviour at other sites during the period of missing data.

is wet at site i and dry at site j. The odds ratio for rain at site i given rain at site j, against rain at site i given dry at site j is

$$\frac{n_{11}}{n_{01}} \Big/ \frac{n_{10}}{n_{00}} = \frac{n_{00}\, n_{11}}{n_{01}\, n_{10}}$$

which would be the same is we exchanged sites i and j. Therefore the log odds ratio is given by

$$\log\left(\frac{n_{00}\, n_{11}}{n_{01}\, n_{10}}\right).$$

This measure takes values on $\mathbb{R}$ with the sign (positive or negative) indicating the direction of association. Large positive (negative) values indicate strong positive (negative) association and near zero values indicate little or no association.

For the Yorkshire dataset, most log odds ratios and correlations were greater than zero suggesting positive association between rainfall occurrences and non–zero rainfall amounts at most pairs of sites. Plots of the log odds ratios and correlations against the distance between sites showed a clear decrease, implying greater similarity in precipitation characteristics between sites in close proximity. This suggested that there is spatial structure in the data.

By aggregating the data to the yearly level, we considered whether there are any long term trends over the thirty year period. Examination of the proportion of wet days and the mean wet

day precipitation by year for each site showed no obvious long term trend. In terms of short term patterns, time series plots of daily precipitation did not show any within season trend over the thirty year period. This is as expected because seasonal patterns are unlikely to be discernible when only considering calendar winters.

2.5 Simple time series models

To investigate the nature and strength of the temporal autocorrelation in precipitation series we fitted simple time series models to the rainfall data at each site. Classifying each day as wet or dry (as defined in Section 2.1) led to a sequence of binary variables for each site over the entire period. For every site separately, we modelled the data from each winter season as an independent realisation from a two state Markov chain of unknown order q, where $q \in \{0, 1, \ldots, q_{\mathbf{max}}\}$, and conditioned on the first $q_{\mathbf{max}}$ observations in each winter period. We assigned a conjugate beta prior distribution to the Markov transition probabilities for models of each order and a discrete uniform prior to the order of Markovian dependence q. This allowed the posterior mass function for q to be computed in closed form; see Boys & Henderson (2004) for further details. Taking the upper limit to be $q_{\mathbf{max}} = 6$, we computed the posterior mass function for q at each site in the Yorkshire network. In all cases there was negligible posterior support for $q = 0$ which suggests that Markovian dependence is a more tenable assumption than independence. For the majority of sites the Markov models with the most posterior support were of order one or two, although the Pennine site Moorland Cottage was an exception, with the posterior mode for q lying at three.

For rainfall amounts, we define a wet spell of length k as a run of k consecutive wet days which is preceded and followed by a dry day. For each site separately, we transformed the data in wet spells by taking logarithms and then subtracting the mean for that site. The transformed data were then modelled by assuming that each wet spell was an independent realisation from an autoregressive model of order p, where $p \in \{0, 1, \ldots, p_{\mathbf{max}}\}$. Suppose that there are K wet spells. For simplicity, we conditioned on the first $\min\{p_{\mathbf{max}}, T_k\}$ observations in the k–th wet spell ($k = 1, \ldots, K$) where T_k is the length of the k–th wet spell. We adopted the fully conjugate normal inverse–gamma prior distribution for the autoregressive coefficients and conditional variances in the autoregressive models of each order and a discrete uniform prior for the order p. This facilitated analytic computation of the posterior distribution for p and for the model parameters, conditional on each value of p. For more details concerning these calculations, see, for example, Chapter 2 of Denison *et al.* (2002). For each site, independently, we computed the posterior mass function for p, taking $p_{\mathbf{max}} = 2$ to avoid losing too much information due to our conditional model specification. For most sites, the posterior tended to offer more support to AR(1) models than to AR(0) (independence) or AR(2) models. Exceptions were the two Pennine sites where the AR(2) models had the most posterior support. Examination of the marginal posterior distributions for the autoregressive coefficients in the AR(1) model for each site provided evidence of positive autocorrelation within wet spells and of stronger autocorrelations in wet spells at the two Pennine sites. It appears that for sites at higher elevation, the temporal structure in both the rainfall occurrence and amounts processes is such that observations at greater lags continue to influence the distributions of rainfall occurrence and amount.

In fitting these time series models we made a number of simplifications in the modelling. These were largely commensurate with performing simple exploratory analyses. For example, when assigning prior distributions to the model parameters, we chose conjugate priors for convenience. In subsequent chapters, we consider in detail prior elicitation and the challenge of expressing substantive initial information. Also, we made no effort to model the spatial structure. In later chapters we develop models to capture dependence in both space and time.

Chapter 3

Bayesian analysis of hidden Markov models

3.1 Introduction

In this chapter, we introduce hidden Markov models as a means of modelling time series data that exhibit dependence over time. In brief, this is achieved through the introduction of an underlying discrete and unobserved process on which the observed process is modelled conditionally. Each observation in the time series has an associated unobserved or *hidden* state which provides a classification of the data into distinct groups, themselves modelled heterogeneously but within which data are modelled homogeneously. In a finite mixture model the unobserved states, called indicator variables in this context, would usually be modelled as independent and identically distributed (*iid*) random quantities. Hidden Markov models provide a generalisation in which temporal dependence is induced in the observed process by modelling the hidden states as a Markov chain. These concepts are formally introduced in Section 3.2.

This thesis is presented entirely from a Bayesian perspective. To this end, Section 3.3 introduces the Bayesian philosophy and brief details of the principles of Bayesian inference. We then apply these basic principles by formulating the problem of inference for hidden Markov models in a Bayesian framework. The Bayesian approach provides a natural framework for incorporating uncertainty not just in the model parameters, but in the models themselves. In the context of hidden Markov models, given a particular within–state model, this will arise when the number of hidden states is unknown. This is a classic example of a problem in Bayesian model choice (or averaging) which is reviewed in Section 3.4. Two types of approach are available for computing the posterior model probabilities when they are not available analytically; within and across model simulation (Section 3.5). Details of available within model simulation techniques are provided in Section 3.5.1 together with a critical appraisal of their potential performance in analyses involving hidden Markov models. Across model simulation techniques are then briefly discussed in Section 3.5.2.

Note that this chapter is an abridged version of Germain *et al.* (2010*a*) and we occasionally refer to this technical report for more details.

3.2 Hidden Markov models

A hidden Markov model, often abbreviated to HMM, is a bivariate discrete–time stochastic process $\{(\mathbf{Y}_t, S_t) : t = 1, \ldots, T\}$ consisting of an observed process $\{\mathbf{Y}_t : t = 1, \ldots, T\}$ and a hidden process $\{S_t : t = 1, \ldots, T\}$. The latter is often termed the *state* or *regime* and is usually assumed to have a discrete and finite state space. The former can be discrete or continuous, univariate or multivariate. In the standard theory of hidden Markov models this unobserved process is a first order Markov chain and, conditional on these hidden states, the observable random quantities, $\{\mathbf{Y}_t : t = 1, \ldots, T\}$, form a conditionally independent sequence in which the conditional distribution of $\mathbf{Y}_t$ depends only on S_t. In the standard case, the hidden Markov model is assumed to be *homogeneous* in the sense that the Markov chain is homogeneous, that is, the transition probabilities are constant over time, as is the conditional distribution of $\mathbf{Y}_t$ given S_t. Through the observed process $\{\mathbf{Y}_t : t = 1, \ldots, T\}$, inference can be made about *both* the model parameters and the hidden process which is often of interest in its own right. The inclusion of the unobserved random variables means that hidden Markov models may also be regarded as *missing data models* or *latent variable models*. Chapter 4 contains an example of a "standard" hidden Markov model.

Various generalisations of the standard hidden Markov model have been proposed in the literature. For example, the hidden state sequence can be of order greater than one, in which case the conditional distribution of S_t depends only on the preceding d values $\{S_i : i = t - d, \ldots, t - 1\}$. Note that the simple case $d = 0$ corresponds to a finite mixture model with independent mixture component indicators. Another generalisation of the hidden Markov model allows *non-homogeneous* transition probabilities in the hidden process or non–homogeneous within–state (conditional) distributions. In other words, the distribution of S_t given S_{t-1} or of $\mathbf{Y}_t$ given S_t depends in some way on the time index. A non–homogeneous hidden Markov model (NHMM) in which the evolution of the hidden process depends on time through conditioning on time–varying explanatory variables will be introduced in Chapter 5. Finally, another generalisation involves allowing the conditional distribution of $\mathbf{Y}_t$, given the history of the observed process up to and including time $t-1$ (and the whole hidden process) to depend only on the previous q observed values $\{\mathbf{Y}_i : i = t-q, \ldots, t-1\}$ and S_t. These are sometimes called *Markov switching models* and in Chapter 5 we introduce a NHMM in which $q = 1$. Hidden Markov models are now used in a wide variety of applications in areas such as communications engineering, bioinformatics, finance, medicine and meteorology. MacDonald & Zucchini (1997), Cappé *et al.* (2005) and Frühwirth-Schnatter (2006) provide references, examples and a comprehensive treatment of inference for hidden Markov models.

3.2.1 Assumptions

A particular hidden Markov model is partly defined by the choice of conditional independence assumptions which govern the factorisation of the joint distribution of $\{(\mathbf{Y}_t, S_t) : t = 1, \ldots, T\}$. Let θ denote the parameters of the model and introduce the notation $\mathbf{x}_{i:j}$ to denote the sequence $\mathbf{x}_i, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_j$. The "standard" hidden Markov model is then based on the following assumptions:

A1. The first addresses the hidden process and can be summarised as

$$\Pr(S_t = k|S_{1:t-1}, \theta) = \Pr(S_t = k|S_{t-1} = j, \Lambda) = \lambda_{jk}, \qquad j, k \in S_r = \{1, \ldots, r\}$$

for $t = 2, \ldots, T$. This is just the Markov assumption which asserts that, given the hidden state at the previous time point, the current state does not depend on any past states except that which directly precedes it. Moreover, the transition probabilities are constant over time. It follows that the states $\{S_t : t = 1, \ldots, T\}$ are a first order homogeneous r–state Markov chain with transition matrix $\Lambda = (\lambda_{jk})$, where each row, $\lambda_j = (\lambda_{j1}, \ldots, \lambda_{jr})$, is defined on the r–dimensional simplex, $\mathcal{S}_r = \{(x_1, \ldots, x_r) : x_i \geq 0 \, \forall i, \sum x_i = 1\}$.

A2. The second assumption addresses the observed process and can be summarised as

$$(\mathbf{Y}_t|\mathbf{Y}_{1:t-1}, S_{1:T}, \theta) \equiv (\mathbf{Y}_t|S_t = k, \theta) \sim \mathcal{F}(\theta_{\text{obs},k})$$

for $t = 2, \ldots, T$ and $(\mathbf{Y}_1|S_{1:T}, \theta) \equiv (\mathbf{Y}_1|S_1 = k, \theta) \sim \mathcal{F}(\theta_{\text{obs},k})$. Here $\mathcal{F}(\cdot)$ is some parametric distribution family defined over the sampling space $\mathcal{Y}$ such that $\mathcal{F}(\theta_{\text{obs},k})$ has density $p(\mathbf{y}_t|S_t = k, \theta) \equiv p(\mathbf{y}_t \mid \theta_{\text{obs},k})$ indexed by the parameter $\theta_{\text{obs},k}$. This assumption states that $\mathbf{Y}_1, \ldots, \mathbf{Y}_T$ are conditionally independent given the hidden states $(S_1, \ldots, S_T)$.

To complete the joint model for $\{(\mathbf{Y}_t, S_t) : t = 1, \ldots, T\}$, a marginal distribution for S_1 needs to be specified, $\Pr(S_1 = k|\nu) = \nu_k$, where $\nu = (\nu_1, \ldots, \nu_r) \in \mathcal{S}_r$. There are a number of possible choices for ν, but first we review some basic Markov chain theory needed for their description.

A Markov chain is *irreducible* if each state can be reached from any other state in a finite number of moves, i.e. for all j, k and finite h, $\Pr(S_{t+h} = k|S_t = j) > 0$. A particular state, say j, is *periodic* if, starting from that state, the chain returns to it in a fixed number of steps, d_j, or a multiple of d_j, i.e.

$$d_j = \gcd\{h : \Pr(S_{t+h} = j, S_{t+h-1} \neq j, \ldots, S_{t+1} \neq j|S_t = j) > 0\},$$

where gcd is the greatest common divisor operator. Further, state j is *aperiodic* if it is not periodic, i.e. if $d_j = 1$. The Markov chain is aperiodic if this is true of all of its states. If a chain is both irreducible and aperiodic then it has a unique *stationary distribution* defined as the solution to the matrix equation, $\delta = \delta\Lambda$ where δ is a row vector. In addition, regardless of the initial state, S_1, the Markov chain converges to this distribution. Note that a finite state Markov chain will be *ergodic* if it is irreducible and at least one state is aperiodic.

Returning to the choice of the initial distribution ν, if the hidden chain is assumed to be irreducible and aperiodic then a sensible choice might be the stationary distribution, that is, $\nu = \delta$. Although this means the chain starts and hence remains in its stationary distribution, because δ is a function of Λ, this is at the cost of a more complicated joint density for the hidden states, given the model parameters. Aside from this there may be reasons for preferring a non–ergodic hidden chain, for example, left–to–right hidden Markov models in which the Markov chain starts in a particular state before traversing a number of others (without going backwards) and terminating in a fixed state, for example, in a change point problem. In cases where we do not wish to initialise at the stationary distribution, other choices for the initial distribution, ν, are:

1. S_1 is assumed fixed and known, in which case ν is degenerate with all its mass at a single value. This might be a reasonable choice in, for example, a left–right hidden Markov model.

2. S_1 is assumed random and its distribution ν does not depend on Λ, so either:

   (a) ν is fixed, for example, the discrete uniform distribution on $\mathcal{S}_r$, or

   (b) ν is unknown and therefore regarded as another parameter about which inference is to be made.

Choice 2(b) might be sensible if the data comprise several independent realisations from the same hidden Markov model. In such cases, there will be more than one (albeit hidden) "observation" from which to learn about ν. If the data comprise a single realisation from a hidden Markov model then it seems there is nothing to be gained by making ν unknown rather than fixed. The initial distributions in 2(a) and 2(b) both lead to a slight loss of information about the transition probabilities if the chain is irreducible and aperiodic. This is because the first observation made on a stationary Markov process contains some information about the transition probabilities as it is drawn from the stationary distribution of the chain, which itself is dependent on the transition probabilities. Therefore in a Bayesian analysis, a simulation driven possibility for initialising an irreducible and aperiodic Markov chain is to approximate the stationary distribution by inserting a reasonably small number of missing states before the start of the sequence.

For the remainder of this chapter, we denote the model parameters by $\theta = (\theta_{\text{hid}}, \theta_{\text{obs}})$, where $\theta_{\text{hid}} = (\Lambda, \nu)$ and $\theta_{\text{obs}} = (\theta_{\text{obs},1}, \ldots, \theta_{\text{obs},r})$ are the parameters of the hidden and observed processes, respectively. Note that if $\nu = \delta$ then it is enough to write $\theta_{\text{hid}} = \Lambda$. To illustrate the theory and methodology of hidden Markov models, we will use a simple example of a model satisfying assumptions A1 and A2 in which the distribution of $\mathbf{Y}_t = Y_t$ arises from one out of r Bernoulli distributions, depending on the state S_t,

$$(Y_t|Y_{1:t-1}, S_{1:T}, \theta) \equiv (Y_t|S_t = k, \theta) \sim \text{Bern}(p_k) \qquad \text{for } t = 1, \ldots, T.$$

In this case the within–state distributions are indexed by a single parameter, $\theta_{\text{obs},k} = p_k$.

3.2.2 Directed acyclic graphs

In hidden Markov models, directed acyclic graphs (DAGs) are a particularly convenient way of representing the factorisation of the joint probability density function of $\{(\mathbf{Y}_t, S_t) : t = 1, \ldots, T\}$. In general, a DAG represents a factorisation of a joint probability (density) into one or more marginal probabilities and a sequence of conditional probabilities. Each node represents a variable and arrows (directed edges) between nodes represent the presence or otherwise of a direct relationship. Absence of an edge between two nodes implies conditional independence between the corresponding variables. Directed acyclic graphs have no cycles meaning that by following the arrows it is not possible to return to a node after leaving it.

A node V is said to be a parent of another node W if there is an arrow from V to W. Equally, W is then a child of V. The descendants of V are those nodes that can be reached from V by following the direction of the arrows. If the parents of a node V are known then only the

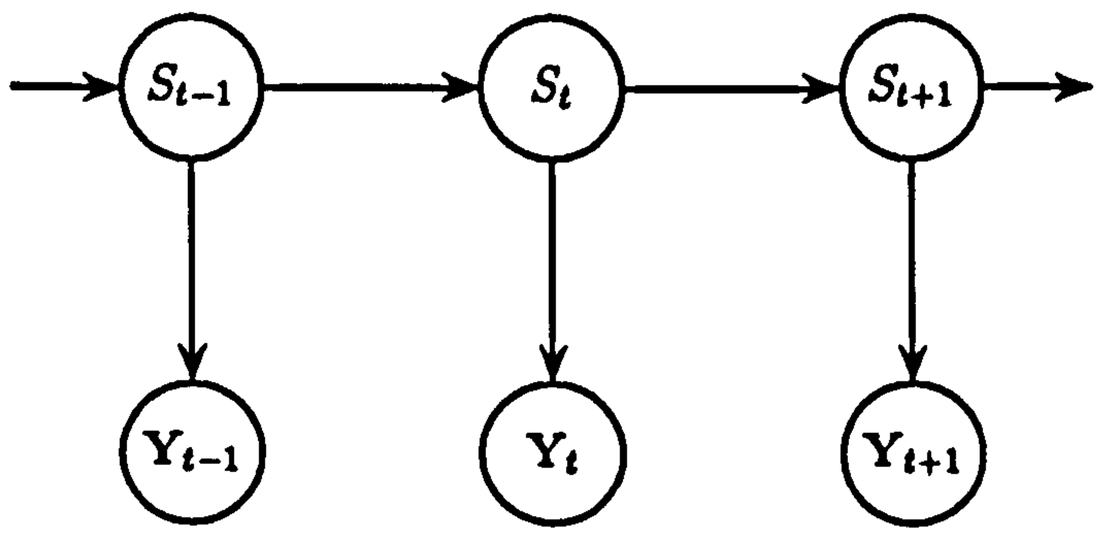


Figure 3.1: A DAG for the hidden Markov model described by assumptions A1 and A2.

descendants of V provide further information about V. Formally, V is conditionally independent of all its non–descendants, given the parents of V. This implies a particular factorisation of the joint distribution of the variables associated with the various nodes $(V_1, \ldots, V_n)$ given by

$$p(V_1, \ldots, V_n) = \prod_{i=1}^{n} p(V_i | \text{parents of } V_i).$$

More details on DAGs and other graphical representations of the local relationships between variables can be found in Lauritzen *et al.* (1990) or Whittaker (1990).

For the hidden Markov model defined by assumptions A1 and A2, a DAG is provided in Figure 3.1. We see, for example, that $\mathbf{Y}_t$ has a single parent, S_t, and no descendants. Therefore, given S_t, $\mathbf{Y}_t$ is conditionally independent of all other hidden states and observations. Dropping notational dependence on the parameters, the implied factorisation of the joint distribution of $\{(\mathbf{Y}_t, S_t) : t = 1, \ldots, T\}$ is

$$p(\mathbf{y}, \mathbf{s}) = \prod_{t=1}^{T} p(\mathbf{y}_t \mid S_t = s_t) \times \Pr(S_1 = s_1) \prod_{t=2}^{T} \Pr(S_t = s_t \mid S_{t-1} = s_{t-1}).$$

3.2.3 General properties and applications of hidden Markov models

An important property of hidden Markov models is their capacity to capture temporal autocorrelation. Consider a hidden Markov model satisfying assumptions A1 and A2. From the DAG in Figure 3.1 we can deduce that if we marginalise over the hidden states, the conditional distribution of $\mathbf{Y}_t$, given the whole history of the observed process up to and including time $t-1$, depends on all the conditioning variables. Therefore, although the hidden process is first order Markov, marginally, the observed process is not a Markov chain of any (finite) order and so does not have the loss of memory property of finite order Markov chains.

It is straightforward to derive the autocorrelation function (ACF), $\rho_{Y_t}(h|\theta) = \text{Corr}(Y_t, Y_{t+h}|\theta)$, for a hidden Markov model satisfying assumptions A1 and A2 when Y_t is univariate and $\{S_t : t = 1, \ldots, T\}$ is irreducible, aperiodic and initialised in its stationary distribution, $\boldsymbol{\nu} = \boldsymbol{\delta}$. For

example, consider the simple Bernoulli model introduced earlier and assume there to be $r = 2$ states. For this model, Frühwirth-Schnatter (2006) shows that the ACF is equal to

$$\rho_{Y_t}(h|\theta) = \frac{\delta_1\delta_2(p_1 - p_2)^2}{\text{Var}(Y_t|\theta)}(\lambda_{11} - \lambda_{21})^h \quad \text{where} \quad \text{Var}(Y_t|\theta) = \sum_{k=1}^{2}\delta_k p_k - \left(\sum_{k=1}^{2}\delta_k p_k\right)^2.$$

In this case, autocorrelation will be present whenever $p_1 \neq p_2$ and $\lambda_{11} \neq \lambda_{21}$. It will be positive at lag one if $\lambda_{11} > \lambda_{21}$ and negative otherwise.

In general, given that the Markov chain is in state j, it is easily shown that the sojourn time in that state is a geometric random variable with parameter λ_{jj}. As such, the expected sojourn time in state j is $1/(1-\lambda_{jj})$, and so, the larger the on–diagonal element, λ_{jj} the more persistent the state. In our simple example above, the greater the persistence probabilities, λ_{11} and $\lambda_{22} = 1 - \lambda_{21}$, the larger the difference $\lambda_{11} - \lambda_{21}$ and hence the stronger the (positive) autocorrelation at lag one, with the opposite conditions leading to strong negative autocorrelation.

The ability of hidden Markov models to capture temporal dependence means that they can be used to model data that exhibit dependence over time. In addition, specific characteristics of these models make them particularly amenable in certain settings. Models satisfying assumptions A1 and A2 are just generalisations of mixture models and so also allow for overdispersion, skewness and multi–modality. This makes hidden Markov models useful when the marginal distributions of observables exhibit these traits. If time series comprise (multivariate) observations on a spatio–temporal process, then hidden Markov models provide a means of simplifying the multi–faceted dependence structure within the data (see, for example, Zucchini & Guttorp, 1991; Hughes *et al.*, 1999, and Chapters 4, 5 and 6). Alternatively, in situations where ordered data are believed to have arisen in distinct segments, hidden Markov models can provide a means of capturing any underlying heterogeneity. Examples include DNA sequence data (Boys & Henderson, 2004) and series with "change points" (Chib, 1998). Finally, if a temporal process can only be observed in noise, hidden Markov models provide a means of extracting the signal, for example, in digital communications (Cover & Thomas, 1991) and speech recognition (Rabiner, 1989).

Although the hidden states are simply statistical devices for introducing dependence in the observed process, there are situations in which the states have a physical interpretation, and are of interest in their own right. Sansom (1998) considers a hidden Markov model for breakpoint rainfall data in which the states are interpreted as states of the rainfall generating mechanism. Learning about the hidden states in such situations may improve understanding of the process which generated the data.

3.3 Bayesian implementation of hidden Markov models

This section begins by providing a brief introduction to the philosophy and principles of Bayesian inference (3.3.1). We then describe the implementation of hidden Markov models within a Bayesian framework. This starts with a discussion of the choice of priors (3.3.2) with particular reference to issues that arise in analyses involving hidden Markov models. This is followed by details of the complete and observed data likelihoods (3.3.3) and the use of MCMC for

posterior inference (3.3.4). Further issues that arise in Bayesian implementations of hidden Markov models are then discussed, including non-identifiability and label switching (3.3.5) and dealing with missing data (3.3.6).

3.3.1 Principles of Bayesian Inference

In the Bayesian framework all unknown quantities including parameters, missing observations and latent variables are treated equally as random variables. The relationships between all unknown quantities and the data are then described by a joint probability distribution. Conditioning on the observed data, the resulting conditional distribution, called the posterior distribution, is used to make inferences about the unknown quantities. In this section, we provide a very brief introduction to Bayesian inference, and refer to the technical report Germain *et al.* (2010*a*) for further details.

Let $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ denote the data and let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ denote the unknown quantities about which we would like to make inference. The key principle of the Bayesian approach is to define a prior distribution $\pi(\boldsymbol{\theta})$, which summarises our prior belief about the unknown quantities, and then to use data to update the prior and construct a posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$, which provides a complete and coherent summary of our post–data uncertainty about $\boldsymbol{\theta}$. The information in the data is contained in the likelihood $L(\boldsymbol{\theta}|\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta})$ and the rule for updating the prior is called Bayes theorem,

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})}{p(\mathbf{y})}, \tag{3.1}$$

or simply, $\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})$. In (3.1), the normalising constant $p(\mathbf{y}) = \int \pi(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}$ is called the marginal likelihood. For future reference, note that the distribution $\pi(\theta_i \mid \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d, \mathbf{y})$ is called the *full conditional* distribution of θ_i, that is, the distribution of θ_i conditional on the data $\mathbf{y}$ and the other parameters $(\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d)$.

Subjective Bayesians regard the prior distribution $\pi(\boldsymbol{\theta})$ as a wholly subjective description of initial uncertainty and choose it to be informative. This is the view taken in this thesis. For a review of methods for eliciting prior distributions, see, for example Garthwaite *et al.* (2005). An alternative approach is to use so–called "non–informative" or vague priors with the intention of conveying little or no prior information, for example, flat priors (in which a parameter is uniformly distributed over some, possibly infinite, range) and reference priors (Berger & Bernardo, 1994). A review of such prior distributions can be found in Kass & Wasserman (1996).

In the majority of cases the normalising constant, $p(\mathbf{y})$, in (3.1) and hence the posterior distribution is not available in closed form, an exception being when the prior distribution is of the same functional form as the likelihood, known as conjugacy. In such cases, Markov chain Monte Carlo (MCMC) methods, for example the Metropolis Hastings algorithm or Gibbs sampling, can be used to generate samples from the posterior distribution. There is a wealth of literature in the area of MCMC, for example, see Gamerman & Lopes (2006), Chib & Greenberg (1995) and Brooks (1998). The basic idea is to set up a Markov chain whose transition probabilities are analytically tractable and which has the required posterior distribution as its stationary

distribution. We can then start from any initial point in the support of the posterior and if the chain is run for long enough we will eventually generate (dependent) samples from the posterior distribution.

There is also a large literature devoted to techniques for diagnosing the convergence of MCMC chains. Some methods are based on numerical checks (see, for example, Geweke, 1992; Gelman & Rubin, 1992) whilst others take the form of visual examination of plots, for instance, trace plots of the MCMC output. Often trace plots from multiple chains, which were initialised at different starting points, are compared to help to detect whether chains have simply become trapped in the region of a local mode, rather than exploring the full posterior. A thorough review of both numerical and graphical convergence diagnostics can be found in Cowles & Carlin (1996). Aside from convergence, another issue with drawing inference from MCMC samples is that of autocorrelation within chains. If there is strong correlation between successive values in the chain, then two consecutive values provide less information about the posterior distribution than if these values were independent. The degree of dependence between successive values can be assessed by computing the autocorrelation function and plotting it against the lag.

In order to assess the fit of a model to the data and to our substantive knowledge, various Bayesian model checking techniques are available. Some methods are based on cross-validation and involve dividing $\mathbf{y}$ into a training set and a validation set. The training set is used to update the prior then the posterior predictive distribution for the validation set is compared with the observed values. For example, if a single observation, y_i, is removed then its conditional predictive density is given by

$$p(y_i|\mathbf{y}_{-i}) = \int p(y_i|\theta, \mathbf{y}_{-i})\pi(\theta|\mathbf{y}_{-i})d\theta,$$

where $\mathbf{y}_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_T)^T$. See Gelfand *et al.* (1992) or Alqallaf & Gustafson (2001) for more details. An alternative approach to model checking is to compare data to hypothetical replicates $\mathbf{y}^{\text{rep}}$ that could have been observed under the model without assuming omission of any members of the sample (Gelman *et al.*, 1995). This avoids the need to recompute the posterior predictive distribution each time a validation set is omitted and is based on the posterior predictive distribution

$$p(\mathbf{y}^{\text{rep}}|\mathbf{y}) = \int p(\mathbf{y}^{\text{rep}}|\theta)\pi(\theta|\mathbf{y})d\theta. \tag{3.2}$$

The basic idea is that if the model fits well then the observed data should look plausible under the posterior predictive distribution. A method based on this principle is discussed further in Chapter 4.

3.3.2 Prior distributions

For hidden Markov models, a standard, and in most cases reasonable, prior assumption is that the parameters of the observed and hidden processes $\theta_{\text{obs}} = (\theta_{\text{obs},1}, \ldots, \theta_{\text{obs},r})$ and $\theta_{\text{hid}} = (\Lambda, \nu)$ are independent *a priori*. The joint prior can then be factorised as

$$\pi(\theta) = \pi(\theta_{\text{obs}}, \theta_{\text{hid}}) = \pi(\theta_{\text{obs}})\pi(\theta_{\text{hid}}). \tag{3.3}$$

For the parameters of the observed process, if learning the value of one state specific parameter $\theta_{\text{obs},k}$ would not cause us to revise our beliefs about the values of any of the others, then it will be reasonable to assume *a priori* independence between $\theta_{\text{obs},1}, \ldots, \theta_{\text{obs},r}$. This will often be the case if states are intended to provide an (autocorrelated) classification of the observed quantities into distinct groups. Under this commonly adopted assumption, we can express the joint prior distribution of θ_{obs} as

$$\pi(\theta_{\text{obs}}) = \prod_{k=1}^{r} \pi(\theta_{\text{obs},k}). \tag{3.4}$$

For the parameters of the hidden process, it is usual to assume *a priori* independence between $\boldsymbol{\nu}$ and $\boldsymbol{\Lambda}$ (unless $\boldsymbol{\nu} = \boldsymbol{\delta}$) and between the rows of the transition matrix $\boldsymbol{\Lambda}$. These are convenient choices which allow the prior to be factorised into a product of functions of $(r+1)$ variables

$$\pi(\theta_{\text{hid}}) = \pi(\boldsymbol{\nu}) \prod_{k=1}^{r} \pi(\boldsymbol{\lambda}_k), \tag{3.5}$$

which mimics the factorisation of the conditional density $\pi(\mathbf{s} \mid \theta_{\text{hid}})$. Such assumptions can simplify subsequent posterior inference and are often reasonable.

When prior beliefs about the parameters of the observed process can be adequately captured by a prior which is conjugate to the conditional density $\pi(\mathbf{y} \mid \mathbf{s}, \theta_{\text{obs}})$ then this will be a convenient choice if posterior inference is via an MCMC scheme which samples the hidden states; see Section 3.3.4. Improper priors, whose density does not integrate to one over the parameter space, are sometimes used as "non–informative" priors. However, for hidden Markov models, it is particularly important that they are not used because they can lead to improper posterior densities $\pi(\theta \mid \mathbf{y})$. See Roeder & Wasserman (1997) for a proof in the case of mixture models.

Consider the parameters of the hidden process, $\theta_{\text{hid}} = (\boldsymbol{\nu}, \boldsymbol{\Lambda})$, and for the purposes of this general section, suppose that the prior for θ_{hid} is factorised according to (3.5). The densities $\pi(s_1 \mid \boldsymbol{\nu})$ and $\pi(\mathbf{s}_{-1} \mid \boldsymbol{\Lambda}, s_1)$ are both of multinomial form, where the notation $\mathbf{s}_{-k} = (s_1, \ldots, s_{k-1}, s_{k+1}, \ldots, s_T)$ denotes omission of the k–th component. Therefore when inference is via an MCMC scheme which samples the hidden states, the Dirichlet distribution will be a conjugate prior for each row of $\boldsymbol{\Lambda}$ and for $\boldsymbol{\nu}$ (if $\boldsymbol{\nu} = (\Pr(S_1 = 1 \mid \boldsymbol{\nu}), \ldots, \Pr(S_1 = r \mid \boldsymbol{\nu}))$ is assumed variable). Consider the d–dimensional Dirichlet distribution $\mathscr{D}_d(\mathbf{a})$ with $\mathbf{a} = (a_1, \ldots, a_d) \in \mathbb{R}_+^d$ and $A = \sum_{i=1}^{d} a_i \in \mathbb{R}^+$ whose density is given in Appendix E. For our purposes it will be convenient to reparameterise this distribution in terms of its mean $\tilde{\mathbf{a}} = \mathbf{a}/A$ and the parameter A, which can be interpreted as the *information content* of the prior, in terms of the size of a hypothetical prior sample; see Germain *et al.* (2010*a*) or Dickey (1982) for further details. We can then write $\mathbf{X} \sim \mathscr{D}_d(A\tilde{\mathbf{a}})$, $\tilde{\mathbf{a}} \in \mathscr{S}_d$, $A \in \mathbb{R}^+$, and the means, variances and covariances are given by

$$\mathrm{E}(\mathbf{X}) = \tilde{\mathbf{a}}, \quad \mathrm{Var}(X_i) = \frac{\tilde{a}_i(1-\tilde{a}_i)}{A+1} \quad \text{and} \quad \mathrm{Cov}(X_i, X_j) = \frac{-\tilde{a}_i\tilde{a}_j}{A+1}. \tag{3.6}$$

It is now clear that once the prior mean has been specified there is only one degree of freedom left to set all the variances and covariances. This makes it impossible to elicit different degrees of belief in the prior estimate (the mean) for different components of $\mathbf{X}$. Moreover all the correlations are negative so it is not possible to express prior belief in positive dependence amongst any of the components. Thus, although it is a convenient choice, the Dirichlet distribution may be unable to provide a true representation of prior belief about transitions in the hidden process.

A more flexible prior on the d–dimensional simplex is the additive logistic normal distribution, which is obtained by a logistic transformation of the $(d-1)$–variate normal distribution (Aitchison, 1986). Having $(d-1)(d+2)/2$ parameters allows the specification of a more flexible prior variance and dependence structure than is possible with the Dirichlet distribution but this prior is not conjugate to the form of a multinomial likelihood function. Furthermore, simple closed form moment expressions are not analytically available and this is an obstacle to prior elicitation.

In general, parameters $(\theta_1, \ldots, \theta_J)$ are said to be exchangeable in their joint distribution if $\pi(\theta_1, \ldots, \theta_J)$ is invariant to permutations of the indices; see, for example, Gelman *et al.* (1995). Therefore exchangeable priors are a common means of representing symmetry in belief. Unless information is available *a priori* to distinguish between the states, it is common to adopt a prior in which $(\theta_{\text{obs},1}, \ldots, \theta_{\text{obs},r})$, are exchangeable and similarly for Λ and ν. For Λ, an exchangeable prior has the property that the joint distribution is unaffected when a particular permutation is applied to the rows as long as the same permutation is applied to the columns or *vice versa*. Under the assumption of *a priori* independence expressed in (3.4), the prior for θ_{obs} will additionally be exchangeable if each term $\pi(\theta_{\text{obs},k})$ belongs to the same distribution family and is parameterised by the same fixed hyperparameters. If the prior for Λ comprises independent Dirichlet distributions for each row, then exchangeability can be achieved by specifying

$$\lambda_k \sim \mathscr{D}_r(E\,\mathbf{e}_k), \qquad \text{independently for } k = 1, \ldots, r \tag{3.7}$$

where the mean hyperparameters $\mathbf{e}_k = (e_{k1}, \ldots, e_{kr})$ are such that $e_{kk} = \alpha \in [0, 1]$ for all $k \in \mathcal{S}_r$ and $e_{k\ell} = (1-\alpha)/(r-1)$ for all $(k, \ell) \in \mathcal{S}_r^2$ such that $k \neq \ell$. Note that the information content parameters E are the same for each row. Finally, assuming the initial distribution ν is independent of Λ, an exchangeable "prior" is given by

$$\nu = \mathrm{U}\{1, \ldots, r\},$$

if ν is fixed, that is, $\Pr(S_1 = 1 \mid \nu) = \cdots = \Pr(S_1 = r \mid \nu) = 1/r$. Note that fixing ν means that we do not learn about the distribution of S_1. Alternatively, if ν is variable and assigned a Dirichlet prior, an exchangeable distribution results from specifying

$$\nu = \Big(\Pr(S_1 = 1 \mid \nu), \ldots, \Pr(S_1 = r \mid \nu)\Big) \sim \mathscr{D}_r(G\,\mathbf{g}), \qquad \mathbf{g} = (1/r, \ldots, 1/r). \tag{3.8}$$

3.3.3 Likelihood

The complete data likelihood function $p(\mathbf{y}, \mathbf{s} \mid \theta)$ is equal to the joint density of the data and the hidden states (the "complete data"), given the model parameters, and can be expressed as

$$p(\mathbf{y}, \mathbf{s} \mid \theta) = p(\mathbf{y} \mid \mathbf{s}, \theta) p(\mathbf{s} \mid \theta) = p(\mathbf{y} \mid \mathbf{s}, \theta_{\text{obs}}) p(\mathbf{s} \mid \theta_{\text{hid}}),$$

where the mass function $p(\mathbf{s} \mid \theta_{\text{hid}})$ is sometimes termed the "prior" for $\mathbf{s}$.

The observed data likelihood function $p(\mathbf{y} \mid \theta)$ is equal to the density of the data given the model parameters and can be represented as a sum of the complete data likelihood over all r^T possible realisations of $\mathbf{s}$, that is,

$$p(\mathbf{y} \mid \theta) = \sum_{\mathbf{s}} p(\mathbf{y}, \mathbf{s} \mid \theta).$$

For even moderately large T, direct computation of this sum becomes infeasible. Fortunately the likelihood function can be computed more easily using a forward recursion such as that described in MacDonald & Zucchini (1997) for a hidden Markov model satisfying assumptions A1 and A2. A different algorithm from the literature, which we make use of in this thesis, arises by writing

$$p(\mathbf{y} \mid \boldsymbol{\theta}) = \prod_{t=1}^{T} p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta}).$$

In classical analyses of hidden Markov models, the terms $p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta})$ are called *one-step ahead predictive densities* and they can be computed in a forward recursion derived in the following way,

$$\begin{aligned} p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta}) &= \sum_{k=1}^{r} p(\mathbf{y}_t, S_t = k \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta}) \\ &= \sum_{k=1}^{r} \Pr(S_t = k \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta}) p(\mathbf{y}_t \mid S_t = k, \mathbf{y}_{1:t-1}, \boldsymbol{\theta}), \end{aligned} \tag{3.9}$$

where $\Pr(S_t = k \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta})$ are called the one–step ahead predictive or "prior" probabilities of S_t having observed data up to time $t-1$ but not at time t. These "prior" probabilities can be calculated according to

$$\begin{aligned} \Pr(S_t = \ell \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta}) &= \sum_{k=1}^{r} \Pr(S_t = \ell, S_{t-1} = k \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta}) \\ &= \sum_{k=1}^{r} \Pr(S_t = \ell \mid S_{t-1} = k, \mathbf{y}_{1:t-1}, \boldsymbol{\theta}) \Pr(S_{t-1} = k \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta}), \end{aligned} \tag{3.10}$$

where terms of the form $\Pr(S_t = k \mid \mathbf{y}_{1:t}, \boldsymbol{\theta})$ are called *filtered state probabilities*, or posterior probabilities of S_t having now observed data $\mathbf{y}_t$. Given knowledge of the "prior" probabilities $\Pr(S_t = \ell \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta})$ and the "likelihood" $p(\mathbf{y}_t \mid S_t = \ell, \mathbf{y}_{1:t-1}, \boldsymbol{\theta})$, they can be calculated using

$$\Pr(S_t = \ell \mid \mathbf{y}_{1:t}, \boldsymbol{\theta}) = \frac{\Pr(S_t = \ell \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta}) p(\mathbf{y}_t \mid S_t = \ell, \mathbf{y}_{1:t-1}, \boldsymbol{\theta})}{p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \boldsymbol{\theta})} \tag{3.11}$$

where the expression for the normalising constant is given in equation (3.9), and that for the "prior" probability is provided in equation (3.10). Equations (3.9)–(3.11) are inter-related and can be computed at times $t = 1, \ldots, T$ in a single forward recursion (generally referred to as *filtering*) detailed in Algorithm 3.3.1.

Therefore we see that the observed data likelihood is simply the product of the normalising constants (3.13) and (3.16) for $t = 2, \ldots, T$ in the filtered state probabilities. Although this approach of filtering the states was traditionally used as an adaptive inference tool, as we shall see in Section 3.3.4.2, it is also involved in block updating of the hidden states when posterior inference is via MCMC with data augmentation.

For greater generality which will be helpful in subsequent chapters, the equations in Algorithm 3.3.1 have not been simplified by taking advantage of the conditional independence assumptions expressed in A1 and A2. In fact they hold under much more general assumptions

Algorithm 3.3.1 Forward Filtering

1: Initialise the forward recursion at $t = 1$:

$$\Pr(S_1 = \ell \mid \mathbf{y}_1, \theta) = \frac{p(\mathbf{y}_1 \mid S_1 = \ell, \theta) \Pr(S_1 = \ell \mid \theta)}{p(\mathbf{y}_1 \mid \theta)}, \tag{3.12}$$

where

$$p(\mathbf{y}_1 \mid \theta) = \sum_{k=1}^{r} p(\mathbf{y}_1 \mid S_1 = k, \theta) \Pr(S_1 = k \mid \theta). \tag{3.13}$$

2: For $t = 2, \ldots, T$ in a forward recursion

(a) Compute the one–step ahead predictive probabilities

$$\begin{aligned} &\Pr(S_t = \ell \mid \mathbf{y}_{1:t-1}, \theta) \\ &\quad = \sum_{k=1}^{r} \Pr(S_t = \ell \mid S_{t-1} = k, \mathbf{y}_{1:t-1}, \theta) \Pr(S_{t-1} = k \mid \mathbf{y}_{1:t-1}, \theta), \end{aligned} \tag{3.14}$$

for $\ell = 1, \ldots, r$.

(b) Compute the filtered probabilities

$$\Pr(S_t = \ell \mid \mathbf{y}_{1:t}, \theta) = \frac{p(\mathbf{y}_t \mid S_t = \ell, \mathbf{y}_{1:t-1}, \theta) \Pr(S_t = \ell \mid \mathbf{y}_{1:t-1}, \theta)}{p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \theta)}, \tag{3.15}$$

where

$$p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}, \theta) = \sum_{k=1}^{r} p(\mathbf{y}_t \mid S_t = k, \mathbf{y}_{1:t-1}, \theta) \Pr(S_t = k \mid \mathbf{y}_{1:t-1}, \theta). \tag{3.16}$$

---

than these. In particular they hold if $\mathbf{Y}_t$ depends on previous values $\mathbf{Y}_{t-1}, \mathbf{Y}_{t-2}, \ldots$ (but only the current value of S_t) or if $\{S_t : t = 1, \ldots, T\}$ is a non–homogeneous first order Markov chain whose conditional distribution at time t depends on t or on some time varying exogenous variables $\mathbf{x}_t$ in addition to S_{t-1}. However, for a simple hidden Markov model based on assumptions A1 and A2, in equations (3.12) and (3.13) we have

$$\Pr(S_1 = \ell \mid \theta) = \nu_\ell, \qquad p(\mathbf{y}_1 \mid S_1 = \ell, \theta) = p(\mathbf{y}_1 \mid \theta_{\text{obs},\ell}),$$

whilst in equation (3.14), from the DAG in Figure 3.1 it is clear that

$$\Pr(S_t = \ell \mid S_{t-1} = k, \mathbf{y}_{1:t-1}, \theta) = \Pr(S_t = \ell \mid S_{t-1} = k, \Lambda) = \lambda_{k\ell},$$

and finally in equations (3.15) and (3.16)

$$p(\mathbf{y}_t \mid S_t = \ell, \mathbf{y}_{1:t-1}, \theta) = p(\mathbf{y}_t \mid S_t = \ell, \theta) = p(\mathbf{y}_t \mid \theta_{\text{obs},\ell}).$$

Note that to avoid problems of numerical instability it is best to calculate the filtered state probabilities by working on a standardised log scale.

3.3.4 Posterior inference via MCMC

The posterior distribution for the model parameters follows from Bayes Theorem as

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) p(\mathbf{y} \mid \boldsymbol{\theta})$$

where $p(\mathbf{y} \mid \boldsymbol{\theta})$ is the observed data likelihood. The complexity of this posterior distribution precludes a fully analytic treatment so MCMC techniques are used to generate (dependent) samples from the posterior distribution. It is possible to devise MCMC algorithms based on Metropolis Hastings moves for which the state space of the sampler is $\boldsymbol{\theta}$ alone (and not the hidden chain $\mathbf{s}$) or even $\mathbf{s}$ alone (and not the model parameters $\boldsymbol{\theta}$). These marginal updating schemes will be outlined briefly in Section 3.3.4.3. However, the most commonly adopted MCMC method in Bayesian analyses of hidden Markov models makes use of the principle of *data augmentation* (Tanner & Wong, 1987) in which the hidden states are introduced as "missing data" and augmented to the state space of the sampler. This greatly simplifies the process of sampling from the posterior, as will be explained in the following section.

3.3.4.1 Data augmentation

It is often possible to choose a prior for $\boldsymbol{\theta}$ which is conjugate to the form of the complete data likelihood, or at least semi–conjugate in its components, meaning the prior for each component is conjugate when the values of the others are fixed. Therefore sampling from the conditional posterior of the model parameters given the hidden states (called the *complete data posterior*) is often routine. Moreover, because the state space for each S_t is discrete and finite, sampling from the conditional posterior of the hidden states given the model parameters is also straightforward. This means hidden Markov models are particularly amenable to a two stage Gibbs sampling strategy which generates samples from the joint posterior distribution of the model parameters and hidden states

$$\pi(\mathbf{s}, \boldsymbol{\theta} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \mathbf{s}, \boldsymbol{\theta}) p(\mathbf{s} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta})$$

by alternating between drawing $\mathbf{s}$ from the conditional posterior distribution $\pi(\mathbf{s} \mid \boldsymbol{\theta}, \mathbf{y})$ (data augmentation) and drawing $\boldsymbol{\theta}$ from the conditional posterior distribution $\pi(\boldsymbol{\theta} \mid \mathbf{s}, \mathbf{y})$. Effectively data augmentation is just a computational tool which allows posterior inferences for the model parameters $\boldsymbol{\theta}$ to be obtained by averaging over the distribution of the hidden states, which are themselves regarded as "missing data". This approach has been adopted by, for example, Albert & Chib (1993), Robert *et al.* (1993), Robert *et al.* (2000) and Boys & Henderson (2004). When *a priori* independence is assumed between $\boldsymbol{\theta}_{\text{obs}}$ and $\boldsymbol{\theta}_{\text{hid}}$ the complete data posterior can be written as

$$\pi(\boldsymbol{\theta} \mid \mathbf{s}, \mathbf{y}) = \pi(\boldsymbol{\theta}_{\text{obs}} \mid \mathbf{s}, \mathbf{y}) \pi(\boldsymbol{\theta}_{\text{hid}} \mid \mathbf{s}) \tag{3.17}$$

and the general form of the MCMC sampling scheme is outlined in Algorithm 3.3.2.

In the standard Gibbs sampler, deriving from the hidden state sequence, there can either be T blocks $s_1, s_2, \ldots, s_T$ or one block $\mathbf{s}$ depending on whether the hidden states are updated one–at–a–time from their full conditional distributions or whether they are updated in a single block from the joint posterior distribution of $\mathbf{s}$, given the model parameters. The former naive scheme is often referred to as *local updating of the hidden chain*, whilst the latter more sophisticated

Algorithm 3.3.2 Sampling from the joint posterior for $(\mathbf{s}, \theta)$ using Gibbs sampling with data augmentation

Repeat the following steps for N iterations $j = 1, \ldots, N$ beyond convergence:

1: Simulate $\theta^{[j]}$ from $\pi(\theta \mid \mathbf{s}^{[j-1]}, \mathbf{y})$:

   (a) Simulate $\theta^{[j]}_{\text{hid}}$ from $\pi(\theta_{\text{hid}} \mid \mathbf{s}^{[j-1]})$.

   (b) Simulate $\theta^{[j]}_{\text{obs}}$ from $\pi(\theta_{\text{obs}} \mid \mathbf{s}^{[j-1]}, \mathbf{y})$.

2: Simulate $\mathbf{s}^{[j]}$ from $\pi(\mathbf{s} \mid \theta^{[j]}, \mathbf{y})$.

approach can be termed *global updating of the hidden chain.* More details will be provided in Section 3.3.4.2. In cases when the prior for θ is conjugate to the complete data likelihood, sampling θ from the complete data posterior is straightforward. In other situations, decomposing θ into several blocks may lead to closed form full conditional densities, but if this is not the case, Metropolis Hastings steps can be introduced within the Gibbs sampling scheme.

3.3.4.2 Sampling from the posterior for $(\mathbf{s} \mid \theta)$

The simplest way to simulate a hidden state sequence from $\pi(\mathbf{s} \mid \theta, \mathbf{y})$ is to employ a Gibbs sampler with T univariate component blocks, that is, the hidden states are drawn sequentially from their full conditional distributions $\pi(s_t \mid \mathbf{s}_{-t}, \theta, \mathbf{y})$ for $t = 1, 2, \ldots, T$. These distributions are particularly easy to derive for a hidden Markov model satisfying assumptions A1 and A2 and further details can be found in Germain *et al.* (2010*a*) or Frühwirth-Schnatter (2006). In spite of its simplicity, the one–at–a–time Gibbs sampling scheme typically has poor convergence properties due to the high dependence amongst the large number of s–blocks. It has been shown by, for example Henderson (1999), that convergence can be improved by introducing a sampling scheme which simulates all the hidden states from $\pi(\mathbf{s} \mid \theta, \mathbf{y})$ in a single block. Single block sampling can be achieved using a *forward backward* algorithm, based on the factorisation of the joint posterior, $\pi(\mathbf{s} \mid \theta, \mathbf{y})$, as

$$\pi(\mathbf{s} \mid \theta, \mathbf{y}) = \pi(s_T \mid \mathbf{y}, \theta) \prod_{t=1}^{T-1} \pi(s_t \mid s_T, s_{T-1}, \ldots, s_{t+1}, \mathbf{y}, \theta).$$

We can therefore sample a value for S_T from its marginal posterior distribution, then a value for S_{T-1} from the conditional posterior of S_{T-1} given S_T, next a value for S_{T-2} from the conditional posterior of S_{T-2} given S_T and S_{T-1}, and so on. For ease of generalisation in subsequent chapters, the forward backward sampling scheme we derive in this section relies on much weaker conditional independence assumptions than A1 and A2, allowing the same relaxations as those discussed and permitted in the derivation of the filtering algorithm in Section 3.3.3.

From the DAG in Figure 3.1 it is clear that S_t is conditionally independent of $\{S_i : i = t+2, \ldots, T\}$, given S_{t+1}, and this would be true even if $\{S_t : t = 1, \ldots, T\}$ formed a non–homogeneous first order Markov chain. Further, there are no direct linkages between S_t and any

of the variables in $\{\mathbf{Y}_i : i = t+1, \ldots, T\}$ that do not pass through S_{t+1} so that $\mathbf{Y}_{t+1}, \ldots, \mathbf{Y}_T$ are conditionally independent of S_t given S_{t+1}. Again, this would be true even if the first order Markov chain $\{S_t : t = 1, \ldots, T\}$ was non-homogeneous. Additionally, if $\mathbf{Y}_t$ depended on its own history as well as S_t, then this statement would hold if we additionally conditioned on $\mathbf{Y}_{1:t}$. This means that under A1 and A2, or the more general conditions detailed here, we have

$$\pi(s_t \mid s_T, s_{T-1}, \ldots, s_{t+1}, \mathbf{y}, \theta) = \pi(s_t \mid s_{t+1}, \mathbf{y}_{1:t}, \theta)$$

for $t = 1, \ldots, T-1$, and can calculate

$$\Pr(S_t = k \mid S_{t+1} = \ell, \mathbf{y}_{1:t}, \theta) \propto \Pr(S_{t+1} = \ell \mid S_t = k, \mathbf{y}_{1:t}, \theta) \Pr(S_t = k \mid \mathbf{y}_{1:t}, \theta)$$

where $\Pr(S_t = k \mid \mathbf{y}_{1:t}, \theta)$ is the filtered probability at time t defined in Section 3.3.3. Therefore once the filtered probabilities have been calculated in a forward sweep, a hidden state sequence can be simulated in a backward sweep starting with the simulation of a value for S_T from $\pi(s_T \mid \mathbf{y}, \theta)$, which is the filtered probability at time T. Full details of the forward backward algorithm are provided in Algorithm 3.3.3.

Algorithm 3.3.3 Forward backward algorithm

To simulate a hidden state sequence $\mathbf{s}^{[j]}$ at iteration j of the MCMC scheme, whilst holding θ fixed at its current value, the algorithm proceeds as follows:

1: Perform the filtering algorithm (Algorithm 3.3.1) conditional on θ to compute the filtered state probabilities $\Pr(S_t = k \mid \mathbf{y}_{1:t}, \theta)$, $k \in \mathcal{S}_r$, for times $t = 1, \ldots, T$.

2: Simulate a value for S_T according to the filtered state probabilities $\Pr(S_T = k \mid \mathbf{y}_{1:T}, \theta)$, $k \in \mathcal{S}_r$.

3: For $t = T-1, \ldots, 1$ compute the conditional probabilities $\Pr(S_t = k \mid S_{t+1} = s_{t+1}^{[j]}, \mathbf{y}_{1:t}, \theta)$, $k \in \mathcal{S}_r$ given by

$$\begin{aligned}\Pr(S_t = k \mid S_{t+1} = s_{t+1}^{[j]}, \mathbf{y}_{1:t}, \theta) &= \frac{\Pr(S_{t+1} = s_{t+1}^{[j]} \mid S_t = k, \mathbf{y}_{1:t}, \theta) \Pr(S_t = k \mid \mathbf{y}_{1:t}, \theta)}{\sum_{\ell=1}^{r} \Pr(S_{t+1} = s_{t+1}^{[j]} \mid S_t = \ell, \mathbf{y}_{1:t}, \theta) \Pr(S_t = \ell \mid \mathbf{y}_{1:t}, \theta)}, \end{aligned} \tag{3.18}$$

and simulate a value for S_t from the distribution defined by these probabilities.

Note that under assumptions A1 and A2, the expression $\Pr(S_{t+1} = s_{t+1}^{[j]} \mid S_t = k, \mathbf{y}_{1:t}, \theta)$ in equation (3.18) reduces to

$$\Pr(S_{t+1} = s_{t+1}^{[j]} \mid S_t = k, \theta_{\text{hid}}) = \lambda_{k s_{t+1}^{[j]}}.$$

3.3.4.3 Marginal updating schemes

Although appealing in their simplicity, problems can arise with data augmentation schemes when there is strong dependence between $\mathbf{s}$ and θ. This may lead to poor mixing over $\mathbf{s}$, when the sampler cannot escape its attraction to local modes in the posterior distribution (Celeux *et al.*, 2000). In this section we consider briefly two alternative MCMC marginal updating schemes

in which the state space is reduced to comprise only $\boldsymbol{\theta}$ or only $\mathbf{s}$, and which may therefore mix better than the Gibbs sampler with data augmentation.

Since the observed data likelihood $p(\mathbf{y} \mid \boldsymbol{\theta})$ can be computed exactly using a forward recursion (see Section 3.3.3) it is possible to construct an MCMC sampler with the marginal posterior distribution for the model parameters $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ as its stationary distribution. Although the posterior distribution will not be of standard form, it can be sampled using a Metropolis Hastings scheme. For example, Cappé *et al.* (2005) and Boys & Henderson (2003) describe algorithms for hidden Markov models with normal and multinomial within–state distributions, respectively. Similar schemes are used in Celeux *et al.* (2000) and Cappé *et al.* (2003). Particularly when the parameter space is highly dimensional, it may be necessary to first decompose $\boldsymbol{\theta}$ into multiple blocks and then to update them one–at–a–time in a Metropolis within Gibbs scheme.

An alternative sampling scheme involves marginalising over the parameters of the hidden Markov model and setting up an MCMC sampler whose equilibrium distribution is the marginal posterior $\pi(\mathbf{s} \mid \mathbf{y})$. Analytic marginalisation over $\boldsymbol{\theta}$ in the joint posterior $\pi(\boldsymbol{\theta}, \mathbf{s} \mid \mathbf{y})$ is possible whenever the observation density comes from the exponential family (see, for example, McCullagh & Nelder, 1989) and priors conjugate to the complete data likelihood are chosen for the model parameters $\boldsymbol{\theta}$. In this case we can write

$$\pi(\mathbf{s} \mid \mathbf{y}) = \frac{p(\mathbf{y}, \mathbf{s})}{p(\mathbf{y})} \propto p(\mathbf{y}, \mathbf{s}), \tag{3.19}$$

and because the conditional posterior for $\boldsymbol{\theta}$ given $\mathbf{s}$ will be of closed form, we can compute the numerator $p(\mathbf{y}, \mathbf{s})$ using

$$p(\mathbf{y}, \mathbf{s}) = \frac{p(\mathbf{y}, \mathbf{s} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{s})}.$$

Although computation of the denominator in (3.19) is not generally feasible, the ability to compute the numerator is sufficient to facilitate one–at–a–time updating of the hidden states from their full conditional distributions in simple Gibbs or Metropolis within Gibbs steps. For further details, see Frühwirth-Schnatter (2006). Note that the requirement for analytic computation of $p(\mathbf{y}, \mathbf{s})$ limits the scope of the marginal updating scheme to those occasions when fully conjugate priors (which adequately capture prior beliefs) are available.

3.3.5 Non–identifiability and label switching

By writing the observed data likelihood as

$$p(\mathbf{y} \mid \boldsymbol{\theta}) = \sum_{\mathbf{s}} p(\mathbf{y}, \mathbf{s} \mid \boldsymbol{\theta}) = \sum_{\mathbf{s}} \nu_{s_1} p(\mathbf{y}_1 \mid \boldsymbol{\theta}_{\text{obs},s_1}) \prod_{t=2}^{T} \lambda_{s_{t-1}s_t} p(\mathbf{y}_t \mid \boldsymbol{\theta}_{\text{obs},s_t}),$$

where the summation is over *all* possible hidden state sequences, it is clear that the observed data likelihood is invariant under permutations of the state labels. This means

$$p\{\mathbf{y} \mid \boldsymbol{\theta}_{\text{obs},1}, \ldots, \boldsymbol{\theta}_{\text{obs},r}, (\lambda_{k\ell}), (\nu_k)\} = p\{\mathbf{y} \mid \boldsymbol{\theta}_{\text{obs},\sigma(1)}, \ldots, \boldsymbol{\theta}_{\text{obs},\sigma(r)}, (\lambda_{\sigma(k)\sigma(\ell)}), (\nu_{\sigma(k)})\}$$

for any permutation $\sigma(\cdot)$ of the integers $\{1, 2, \ldots, r\}$. In other words, renumbering the states in $\mathcal{S}_r$ and permuting the parameter indices in correspondence leaves the likelihood unchanged.

It follows that if an exchangeable prior is chosen then the resulting posterior distribution will also be exchangeable, with $r!$ symmetric modes corresponding to the $r!$ permutations of the state labels. This has the consequence that the model parameters are non-identifiable in the posterior. Similarly, since the complete data likelihood is also invariant under permutations of the state labels, the same comments apply to the joint posterior distribution $\pi(\mathbf{s}, \boldsymbol{\theta} \mid \mathbf{y})$. Writing $\sigma(\mathbf{s}) = \Big(\sigma(s_1), \ldots, \sigma(s_T)\Big)^T$ and $\sigma(\boldsymbol{\theta})$ when the same permutation is applied to the state labels in $\boldsymbol{\theta}$, we have

$$\pi\{\sigma(\mathbf{s}) \mid \mathbf{y}\} = \int \pi\{\sigma(\mathbf{s}), \sigma(\boldsymbol{\theta})\} d\sigma(\boldsymbol{\theta}) = \int \pi(\mathbf{s}, \boldsymbol{\theta} \mid \mathbf{y}) d\sigma(\boldsymbol{\theta}) = \pi(\mathbf{s} \mid \mathbf{y})$$

because the joint posterior $\pi(\mathbf{s}, \boldsymbol{\theta} \mid \mathbf{y})$ is invariant to relabelling and the order of integration can be interchanged arbitrarily. This means that the posterior for $\mathbf{s}$, like that of $\boldsymbol{\theta}$, will also be exchangeable. The non–identifiability of the parameters in $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ results in marginal posterior distributions for the state specific parameters which are the same for all states. Similarly, the non–identifiability of the hidden states in $\pi(\mathbf{s} \mid \mathbf{y})$ results in marginal posterior classification probabilities (that is, $\Pr(S_t = k \mid \mathbf{y})$, $k \in \mathcal{S}_r$) which are the same for all states. For a formal proof, see Frühwirth-Schnatter (2006).

A practical consequence of the non–identifiability of the posterior distribution is that posterior samples are subject to *label switching*, in which random permutations of the hidden state labels occur over the course of the MCMC run. An illustration where label switching can be seen in the MCMC output is provided in Figure 3.2(a). This shows a portion of the trace plots for the probability parameters p_1 and p_2, obtained by simulating from the posterior distribution for a Bernoulli hidden Markov model with $r = 2$ states. The data used were a simulated sample from a hidden Markov model with Bernoulli within–state distributions.

Given an MCMC sampler which mixed properly over all $r!$ posterior modes, each of the modes would be visited equally often. This means that using the MCMC output to obtain summaries of the marginal posterior distributions for state specific parameters, or marginal posterior classification probabilities, would reflect the theoretical invariance. In practice, however, it is often difficult for an MCMC sampler to escape its attraction to one particular mode, especially if the modes are well separated. Therefore when using standard MCMC samplers such as those discussed in Section 3.3.4, it is often the case that label switching either does not occur at all (Celeux *et al.*, 2000) or occurs only occasionally, in an unbalanced fashion, meaning that the $r!$ posterior modes are not equally represented in the MCMC output.

The way in which the (potential) problem of label switching is handled should be tailored according to the goals of the analysis. If the sole objective is prediction, then because the posterior predictive distribution

$$p(\mathbf{y}^{\text{rep}} \mid \mathbf{y}) = \int p(\mathbf{y}^{\text{rep}} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta}$$

is also invariant under permutations of the state labels, its estimate based on the posterior draws will also be robust against label switching. As such nothing needs to be done to account for any label switching that occurs. Indeed, it is the sentiment of Aitkin (1997) that finding a unique labelling is unimportant as the object of interest should be the predictive distribution. If the purpose of sampling from the posterior is to approximate the marginal likelihood of the

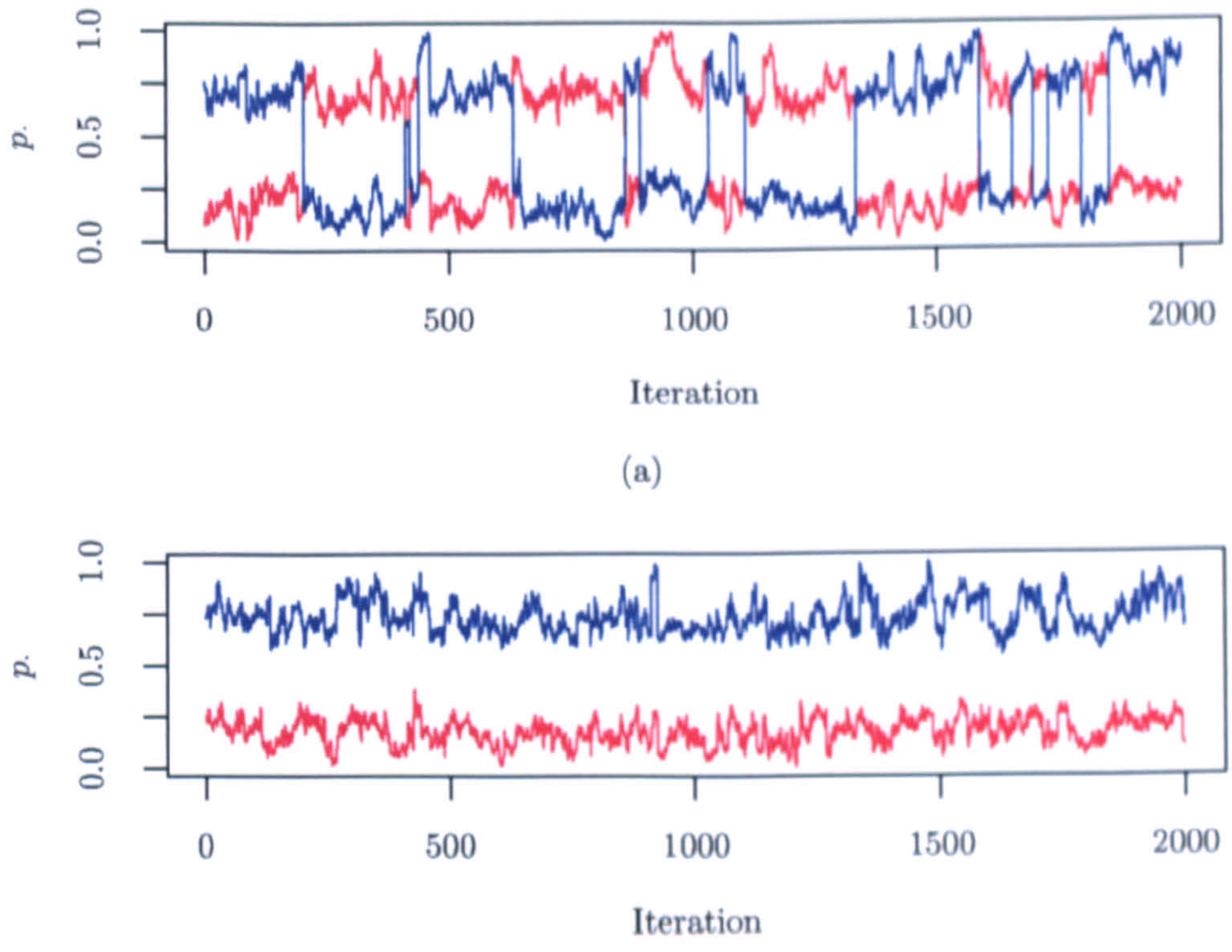


Figure 3.2: Sections of the MCMC output obtained by drawing from the posterior distribution associated with a Bernoulli HMM with $r = 2$ states, based on a simulated dataset. Shown are the trace plots for the probability parameters p_1 and p_2 when the algorithm is employed (a) without and (b) with relabelling.

model, then, as we shall see in Section 3.5.1, it is sometimes necessary to use output based on a sample in which all $r!$ posterior modes have been explored. A simple way of guaranteeing *balanced* label switching between all $r!$ posterior modes is to employ a random permutation sampler (Frühwirth-Schnatter, 2001) in which each draw from the posterior is concluded with a random permutation of the parameter state labels and the hidden state labels if these are also sampled. Celeux *et al.* (2000) comment that although such an approach will give a sampler which visits all the major symmetric modes in the posterior it does not address the root of the problem, namely that the sampler is unable to traverse areas of low posterior density. This means (symmetric) local modes may remain unexplored. Finally, if the goal is classification (for example determining the most probable sequence of hidden states *a posteriori*) or parametric inference then, to be meaningful, one distinct labelling of the states is required. The remainder of this section briefly reviews some of the techniques from the literature for obtaining samples from an *identified* hidden Markov model.

Identifiability can be achieved by imposing an identifiability constraint on one or more of the parameters of the observed or hidden processes, thereby breaking the symmetry in the prior (and thus the posterior) distribution. In our simple Bernoulli example an obvious constraint might be to restrict the probability parameters so that $p_1 < p_2 < \cdots < p_r$ with the consequence that the prior and hence posterior density outside the region where this constraint is satisfied is equal to zero. Although this approach appears to provide a simple solution, in more involved examples, it

can be very difficult to find identifiability constraints which respect the geometry of the posterior sufficiently to induce a unique labelling. In other words, if the posterior modes are not very well separated in the direction of one particular parameter, then using it to define an order is likely to produce an MCMC sample in which draws pertaining to a specific state arise from several of the symmetric modes in the underlying (unconstrained) posterior distribution. As illustrated by, for example, Richardson & Green (1997) or Frühwirth-Schnatter (2001), this means that different identifiability constraints can lead to different marginal posterior distributions for the model parameters. Stephens (1997) provided a formal justification that identifiability constraints can be applied in a post–processing manner, permuting each draw from the posterior to satisfy the constraint. This allows various different constraints to be tested in an effort to find one which isolates a single mode, although there may be no obvious constraint for which this is achievable. When the parameters of the observed process $\boldsymbol{\theta}_{\text{obs},k}$ are themselves multivariate and parameterise multivariate within–state distributions, finding an effective identifiability constraint becomes even more difficult as there may be no natural order to scalar summaries (such as the Euclidean norm) of these multivariate parameters.

An alternative solution is to take the decision theoretic approach of Stephens (2000) or Celeux *et al.* (2000) and make use of relabelling algorithms which aim to find the permutation of the sampled values which minimises the posterior expected loss (Monte Carlo risk) of some chosen loss function. For example, the algorithm might aim to find a permutation of the sampled values which makes the marginal posterior distributions unimodal, taking the form of the natural conjugate family for the parameters. However such approaches involve *two* expensive optimisation steps per draw from the posterior (one to optimise over the hyperparameters in the fitted unimodal marginal posteriors and another to optimise over the permutations). Other algorithms have loss functions which aim to achieve an optimal clustering of observations according to the hidden states. Obviously such algorithms require posterior samples of the hidden state sequences and so are usually used to process the output from MCMC samplers which use data augmentation. Within this general framework, Boys & Henderson (2002) suggest a relabelling scheme that post–processes sampled values using an algorithm which attempts to find the most likely hidden state at every position in the sequence, that is, the marginal posterior mode (MPM) estimate, $\hat{\mathbf{s}}$, which is conveniently available as a by–product. Computing overheads can be substantially reduced by adopting an on–line version of the algorithm which removes the need to store sampled hidden state sequences. The details are provided in full in Algorithm 3.3.4, but essentially after every draw from the posterior, a scoring criterion is used to find the permutation of the sampled values which is the most consistent with the current MPM estimate, $\hat{\mathbf{s}}$. After relabelling according to this permutation, the MPM estimate is updated. Algorithm 3.3.4 is therefore similar to the on–line $r!$ means–type clustering algorithm proposed by Celeux (1998), except in the latter case the mean of the model parameters in each of the r identified states replaces the MPM estimate $\hat{\mathbf{s}}$ and the simple scoring criterion is replaced by a more sophisticated measure of distance between the successively updated "centres" and the current posterior draw.

Boys & Henderson (2002) advise choosing the starting point in Algorithm 3.3.4 to be $\hat{\mathbf{s}}^{*[1]} = \mathbf{s}^{[1]}$ and relabelling the MCMC output for the first time at iteration $j = 2$. In the context of finite mixture models, Nobile & Fearnside (2007) suggest a very similar relabelling scheme which aims to minimise the sum of all distances between the component indicator sequences. However, this algorithm is applied after MCMC sampling so requires storage of all posterior draws of the hidden state sequences.

Algorithm 3.3.4 Relabelling Algorithm

At iteration j, let the current estimate of $\hat{\mathbf{s}}^*$ be $\hat{\mathbf{s}}^{*[j-1]}$ and let the current draws from the posterior for θ and $\mathbf{s}$ be $\theta^{[j]}$ and $\mathbf{s}^{[j]}$. Then

1: Choose the permutation σ_j of $\{1,\ldots,r\}$ which minimises

$$-\sum_{t=1}^{T}\mathbb{I}(\sigma_j(s_t^{[j]}) = \hat{s}_t^{*[j-1]});$$

2: Permute $\mathbf{s}^{[j]}$ and $\theta^{[j]}$ according to σ_j;
3: For $t = 1, 2, \ldots, T$ update the estimate of $\hat{\mathbf{s}}^*$ by setting

$$\hat{s}_t^{*[j]} = \underset{i\in\mathcal{S}_r}{\operatorname{argmax}}\sum_{k=1}^{j}\mathbb{I}(\sigma_k(s_t^{[k]}) = i).$$

Relabelling algorithms cannot be guaranteed to prevent label switching. However, an illustration of how well Algorithm 3.3.4 appears to work in our experience is displayed in Figure 3.2. Adjusting the reversible jump code involved in the production of Figure 3.2(a) to conclude each posterior draw with a run through the relabelling algorithm, the trace plots for a representative portion of the MCMC output for p_1 and p_2 are shown in Figure 3.2(b). There is now no evidence of label switching.

3.3.6 Missing data

In time series there may be isolated occurrences or periods of missing data, and the precise way that the missing data are modelled will be determined by the assumed missing data mechanism. For a comprehensive analysis of missing data see, for example, Chapter 17 of Gelman *et al.* (1995). The simplest assumption is that the missing data are *missing at random*, that is, the conditional distribution of the missing data mechanism does not depend on the missing values, given observed data, parameters and covariates. Moreover if we are prepared to assume *a priori* independence between the parameters of the missing data mechanism and the model parameters, then the missing data mechanism can be termed *ignorable*. Under these two, often very reasonable, assumptions, inferences can be made without any further modelling of the missing data mechanism.

If the missing data mechanism can be assumed to be ignorable then missing data are easily handled within an MCMC framework. In this case the missing values are simply appended to the set of unknown quantities and drawn, or "imputed", from their full conditional distributions (available directly from the "model" for the observed data) on every sweep through the MCMC scheme. Formally this allows an analysis to be performed in which we integrate out the missing values from the joint density of the observed and missing data. However unless it results in significant complication of the MCMC scheme, if we can marginalise over the missing values analytically then this strategy is generally preferable because it may produce a better mixing MCMC sampler owing to the smaller dimension of its state space.

Under assumptions A1 and A2, we can write the joint density of the hidden states, model parameters and data as

$$\pi(\theta)\Pr(S_1 = s_1 \mid \theta)p(\mathbf{y}_1 \mid S_1 = s_1, \theta)\prod_{t=2}^{T}\Pr(S_t = s_t \mid S_{t-1} = s_{t-1}, \theta)p(\mathbf{y}_t \mid S_t = s_t, \theta).$$

If we simply define

$$p(\mathbf{y}_t \mid S_t = s_t, \theta) = \begin{cases} 1, & \text{if } \mathbf{y}_t \text{ is missing,} \\ p(\mathbf{y}_t \mid \theta_{s_t}), & \text{otherwise} \end{cases} \tag{3.20}$$

then the standard filtering algorithm still applies and this can be used to compute the observed data likelihood as previously. If posterior inference is via MCMC with data augmentation, when simulating from the conditional distributions $\Pr(S_t \mid S_{t+1}, \mathbf{y}_{1:t}, \theta)$ using the forward backward algorithm, the data are only involved in the computation of the filtered probabilities and so no modification is needed to the backward sweep. In the computation of the posterior for $(\theta \mid \mathbf{s})$, redefining $p(\mathbf{y}_t \mid S_t = s_t, \theta)$ as in equation (3.20) has no effect on the structure of the posterior. These arguments would still apply if the hidden chain was allowed to be non-homogeneous. However if assumption A2 was relaxed so that $\mathbf{Y}_t$ was allowed to depend on its own history as well as S_t, further work would be required to analytically marginalise over the missing data.

Note that in some applications the assumption that the missing data mechanism is ignorable may not be tenable. In these cases specific aspects of the missing data mechanism may need to be incorporated into the modelling but this will not be considered further in this thesis.

3.4 Inference for hidden Markov models under model uncertainty: concepts

Section 3.3 focused on the Bayesian analysis of hidden Markov models in which the number of states, r, is assumed known. In reality, however, the number of states is often itself an unknown quantity about which we would like to make inference. In the remainder of this chapter, we consider this issue of model uncertainty, firstly describing the concepts of model selection and averaging (Section 3.4) and then considering computational tools with which to estimate the posterior model probabilities (Section 3.5). In hidden Markov models, inference under model uncertainty is sometimes complicated by the non-identifiability of the parameters in the likelihood function of overfitting models, where the "true" number of hidden states is less than the number in the model being analysed. This can arise if the parameters of the within-state models are the same (and transition probabilities from these states into any particular other are the same), or if one of the states is empty meaning the transition probabilities *into* this state are zero. In a frequentist analysis, these sources of non-identifiability can lead to problems due to the potentially irregular behaviour of the likelihood function. Whilst there is no theoretical problem in a Bayesian analysis via MCMC, it is a property of hidden Markov models that the likelihood for an overfitting model with r states will be the same as that for a model with $r - 1$ states. This means that the likelihood alone cannot distinguish between models.

We set the problem of Bayesian inference for hidden Markov models in the presence of model uncertainty in the form of a Bayesian hierarchical model. Suppose that in our prior beliefs (or preferences) the support of r is restricted to the finite countable set $\{1, 2, \ldots, r_{\max}\}$ and that we have assigned a prior probability distribution to the number of states (or more generally model indices)

$$\pi_r(r), \qquad r \in \{1, \ldots, r_{\max}\}.$$

We then regard the hidden Markov model with $1, 2, \ldots, r_{\max}$ states as a different model, parameterised by θ_r, to which we assign a prior

$$\pi(\theta_r \mid r), \qquad \text{for } r \in \{1, \ldots, r_{\max}\}.$$

Finally, we complete our description of the joint distribution for the model index, model parameters and data through the specification of a likelihood function for each hidden Markov model, $p(\mathbf{y} \mid \theta_r, r)$. Note that the dimension of the parameter vector, $\dim(\theta_r) = n_r$, varies between models.

3.4.1 Non–identifiability due to overfitting

In a Bayesian analysis, it is possible to bound the posterior away from non–identifiability through the prior. For example, by choosing a Dirichlet prior distribution for the transition probabilities in which all the hyperparameters are substantially greater than one, the marginal distribution for the transition probability into any state has zero density at zero and little density in the vicinity. This forces the posterior away from a distribution in which the transition probabilities into any state are zero and would discourage non–identifiability due to the existence of one or more empty states. To discourage non–identifiability due to "matching" states, for any particular within–state parameter, say $\phi_{r,j} \in \theta_{r,\text{obs},j}$, $j \in \mathcal{S}_r$, given a particular ordering constraint $\phi_{r,1} < \cdots < \phi_{r,r}$ one could adopt a prior on the distance between the parameters which encouraged separation between them; see Viallefont *et al.* (2002) for an example involving mixtures of Poisson distributions. However, given an overfitting model, even if no effort is made to discourage the occurrence of empty or matching states, the resulting non–identifiability of the posterior presents no theoretical problems in a Bayesian analysis via MCMC. The draws would still constitute a Markov chain with the posterior of the overfitting model as the stationary distribution. The posterior distribution simply averages over the non–identifiable parameters. From a practical point of view, however, as the posterior approaches non–identifiability due to overfitting, label switching is likely to occur *more often* if the posteriors for the within–state parameters of two or more states begin to overlap meaning some of the $r!$ symmetric modes will be situated close together. This also makes label switching increasingly difficult to remedy by any identifiability constraint or relabelling algorithm. In terms of prediction, because the posterior predictive distribution is invariant under permutation of the state labels, the increased prevalence of label switching is unimportant. However if interest lies in learning the properties of the states identified by the model, it is important that we can identify a unique labelling. For more details in the context of finite mixture models, see Frühwirth-Schnatter (2006).

Whilst non–identifiability due to overfitting is not a theoretical problem in a Bayesian analysis, pragmatically, we prefer to report parsimonious, interpretable models. As explained previously there is nothing in the likelihood function to distinguish an overfitting model with r states from

the corresponding reduced model with $r-1$ states. Therefore, when considering hidden Markov models with different numbers of states, it falls to the joint prior for (r, θ_r) to penalise overfitting. The interplay between the conditional prior $\pi(\theta_r \mid r)$ and the likelihood function $p(\mathbf{y} \mid \theta_r, r)$ in Bayesian model selection is borne out through the marginal likelihood.

3.4.2 Defining the marginal likelihood and posterior model probabilities

Regarding the number of states r as a model indicator, the task of discriminating between different values of r (that is, different models) is essentially a problem in Bayesian model selection; see Congdon (2006) for an introduction and Kass & Raftery (1995) or Chipman *et al.* (2001) for a review. Having observed data $\mathbf{y}$, applying Bayes Theorem gives the posterior probability distribution for r over $\{1, 2, \ldots, r_{\max}\}$ as

$$\pi_r(r \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid r)\pi_r(r)}{\sum_{k=1}^{r_{\max}} p(\mathbf{y} \mid k)\pi_r(k)}, \tag{3.21}$$

where the *marginal likelihood*, $p(\mathbf{y} \mid r)$, is obtained by marginalising the joint conditional distribution of $(\mathbf{y}, \theta_r)$, given r, over θ_r,

$$p(\mathbf{y} \mid r) = \int p(\mathbf{y} \mid \theta_r, r)\pi(\theta_r \mid r)d\,\theta_r. \tag{3.22}$$

It is clear from (3.22) that the marginal likelihood can be regarded as the prior predictive density of the data given a hidden Markov model with r hidden states, and is equal to the normalising constant in the conditional posterior density of the model parameters θ_r, given r. The posterior distribution (3.21) adjusts the prior probabilities, $\pi_r(r)$, in light of their relative support from the data, $p(\mathbf{y} \mid r)$. Let us look briefly at the two components in the kernel of the posterior distribution for r. A simple and common choice of prior for the number of hidden states in a hidden Markov model (see, for example, Robert *et al.* (2000)) is the discrete uniform distribution

$$\pi_r(r) = \frac{1}{r_{\max}}, \qquad r \in \{1, 2, \ldots, r_{\max}\},$$

which might be used in an effort to express prior indifference with regards to the number of states, in the sense of favouring all values of r equally. Given the comments in the concluding paragraph of Section 3.4.1, the prior for r is perhaps better viewed, not as a quantification of prior *belief* about the number of hidden states, but as an expression of prior *preference* for more parsimonious models. In other words the prior for r can be used directly to penalise overfitting models, or even just complex models, by choosing a prior which ultimately decays with increasing r. For example, a Poisson distribution truncated to the set $\{1, 2, \ldots, r_{\max}\}$ could be used (see, for example, Boys & Henderson, 2004). However, the marginal likelihood is generally believed to provide a trade off between model fit and model complexity, meaning there should not be any particular *need* to choose $\pi_r(r)$ so that it favours simple models. Congdon (2006) explains this trade–off as follows.

By rearranging Bayes Theorem, the log marginal likelihood can be expressed as

$$\log\{p(\mathbf{y} \mid r)\} = \log\{p(\mathbf{y} \mid \theta_r, r)\} + \log\{\pi(\theta_r \mid r)\} - \log\{\pi(\theta_r \mid \mathbf{y}, r)\}$$

for any θ_r with non–zero density/mass in the posterior. Congdon argues that the term $\log\{\pi(\theta_r \mid r)\} - \log\{\pi(\theta_r \mid \mathbf{y}, r)\}$ acts as penalty to favour more parsimonious models, whilst a more complex model nearly always leads to a higher loglikelihood, $\log\{p(\mathbf{y} \mid \theta_r, r)\}$. However, the exact nature of this trade–off is complicated and depends critically on the choice of priors for the parameters in the different models.

3.4.3 Sensitivity of the marginal likelihood to the prior distribution

From (3.22) it is clear that the marginal likelihood is actually the expectation of the observed data likelihood with respect to the prior distribution $\pi(\theta_r \mid r)$. Therefore when comparing models by their marginal likelihoods, we are actually comparing model–prior combinations. As such, it is inevitable that the marginal likelihood will be sensitive to the prior specification. For example, if $\pi(\theta_r \mid r)$ is chosen to be overly diffuse for any particular r, then there will be little prior support in the region where the likelihood is substantial, which may result in $p(\mathbf{y} \mid r)$ and hence $\pi(r \mid \mathbf{y})$ being downweighted.

In the case of univariate normal mixture models, Frühwirth-Schnatter (2006) and Jennison (1997) provide results which show that a problem similar to Lindley's paradox (Lindley, 1957) can arise, in which increasing the prior variance leads to increasing evidence in favour of a model with only one mixture component. Therefore to avoid overpenalisation of more complex models, with parameter spaces of higher dimension, it is important not to specify excessively diffuse priors. Equally, we want to avoid setting the prior dispersion to be so small as to give the prior an inappropriately high influence. For example, if the spread of the prior is small, this could lead to more complex models being favoured as the prior becomes increasingly amenable to multiple similar states. This effect was observed by Richardson & Green (1997) in studies of prior sensitivity for mixture models with a variable number of components. As the prior variance becomes smaller still, unless the prior is centred exactly in the region of highest likelihood, more simple models are likely to be favoured once more because the likelihood associated with simpler models will be less peaked and so potentially higher in the narrow region of high prior density. In an effort to make the posterior more robust to the choice of prior in mixture and hidden Markov models, some authors, for example, Richardson & Green (1997) and Robert *et al.* (2000) use hierarchical prior specifications for the parameters of the observed process.

Ideally, in any transdimensional analysis, we wish to specify priors that allow the accumulation of posterior probability around models that could provide an appropriate simplification of the data generating mechanism. Especially in cases when we can offer a physical interpretation to the hidden states, it is likely that we will be seeking a model in which the hidden states are well defined and well differentiated. The prior for r, thought of as a penalty for more complex models, is too blunt a tool to express the intricacies of this statement of prior preference. In Section 3.4.1, choices of prior designed to bound the posterior away from non–identifiability were discussed. However, even amongst models which are not overfitting, we might prefer those with fewer states if they better satisfy the latter criteria. A prior which penalises similarity of states would assist in increasing posterior support for models with well differentiated hidden states. This idea was commented upon in the discussion of Richardson & Green (1997) where Lawson & Clark (1997) suggested combating over similarity of states through the use of *inhibition priors*, more traditionally used in cluster modelling. This would involve specifying a particular form

of joint prior for the number of states and, say, the component mean parameters; see Lawson & Denison (2002) for more details. A prior which penalises states which occur infrequently would assist in increasing posterior support for models with well defined hidden states. For example, for a hidden Markov model with r states this might be achieved by choosing the prior for the transition matrix so that it offered very little support in the region where all the inward transition probabilities $(\lambda_{1k}, \ldots, \lambda_{k-1,k}, \lambda_{k+1,k}, \ldots, \lambda_{rk})$, $k \in \mathcal{S}_r$, are small. To make such prior statements we would need to introduce *a priori* dependence between the off-diagonal elements in any column of the transition matrix, although exactly how this might be accomplished sensibly may be a difficult problem.

3.4.4 Model selection versus model averaging

In applications involving hidden Markov models, interest often lies in predicting hypothetical data $\mathbf{Y}^f$ that could have been observed under the model. The overall posterior predictive distribution $p(\mathbf{y}^f \mid \mathbf{y})$ is given by marginalising both $\boldsymbol{\theta}_r$ and r from the joint posterior distribution $\pi(\mathbf{y}^f, \boldsymbol{\theta}_r, r \mid \mathbf{y})$ to obtain

$$p(\mathbf{y}^f \mid \mathbf{y}) = \sum_{r=1}^{r_{\max}} \pi_r(r \mid \mathbf{y}) \int p(\mathbf{y}^f \mid \boldsymbol{\theta}_r, r) \pi(\boldsymbol{\theta}_r \mid \mathbf{y}, r) d\boldsymbol{\theta}_r.$$

This mixture representation is called *model averaging* because instead of basing predictions on a single candidate model, a composite model is created by "averaging" over all competing models. By averaging over the unknown number of states r, the posterior predictive distribution $p(\mathbf{y}^f \mid \mathbf{y})$ properly incorporates our uncertainty about the value of r. Indeed if the only objective of the modelling process is prediction then there is no need to choose any particular model and the problem of model selection is replaced by model averaging. However there are occasions when we might wish to choose a particular value for r where, for example, analysis of one particular hidden Markov model might provide scientific insight into the underlying physical process which generated the data; see, for example, Sansom (1998). Alternatively, we may actually be comparing hidden Markov models with *different* within-state distributions and wish to advise on the "best" model.

Pragmatically, basing predictions on a single chosen model will be less costly in terms of computing time and effort than averaging over all competing models. If the problem is one of model selection and we need to choose a particular value for r then a very simple strategy is to choose the posterior mode, in other words the value for r for which $\pi_r(r \mid \mathbf{y})$ is the largest. In a more formal setting, model selection can be based in a decision theoretic framework (see, for example, DeGroot, 2004; Smith, 1988) in which the ultimate goal is to minimise the expected loss (the "risk") associated with choosing a particular model.

3.5 Inference for hidden Markov models under model uncertainty: computational tools

The posterior distribution $\pi_r(r \mid \mathbf{y})$ provides a complete post-data summary of our uncertainty about the number of hidden states, but the integral (3.22) cannot be evaluated in closed form

meaning the posterior distribution for r must be computed using numerical techniques. There are essentially two computational methods for extracting this posterior information. The first approach, considered in Section 3.5.1, is to use *within model simulation*, in which we first approximate the marginal likelihood for each model and then compute the posterior for r through application of Bayes Theorem. The second approach, outlined briefly in Section 3.5.2, is to make use of so–called *across model simulation*, or transdimensional MCMC methods, of which the most well known is the reversible jump MCMC (RJMCMC) algorithm (Green, 1995).

3.5.1 Within model simulation

The computation of the marginal likelihood for hidden Markov models is a non–trivial integration problem. Various methods have been proposed for approximating the marginal likelihood; see Congdon (2006) for a thorough introduction, Bos (2002) for a brief comparative study and Frühwirth-Schnatter (2006) for a comparison between the more commonly used techniques for finite mixture models. Two features of hidden Markov models make approximation of the marginal likelihood a particularly difficult problem. The first is linked to the irregular asymptotic behaviour of the posterior distribution in models which are overfitting. The second is the existence of $r!$ symmetric (major) modes in the posterior distribution $\pi(\theta_r \mid \mathbf{y}, r)$ when an exchangeable prior is used. Moreover, there may additionally be (symmetric) local modes present when there are several r–state models competing to provide an explanation of the data; see, for example, Celeux *et al.* (2000). Failure to take account of the multiple modes in the posterior distribution can lead to bias or inefficiency in the approximations.

In Sections 3.5.1.1–3.5.1.4 we present details of various techniques for approximating the marginal likelihood. This is followed in Section 3.5.1.5 with a discussion outlining their benefits and drawbacks in the general context of hidden Markov models. In Chapter 4 we consider the relative merits of the different techniques in applications involving the specific types of hidden Markov model in which we have interest.

3.5.1.1 Laplace approximation

The Laplace approximation to the marginal likelihood is given by expanding the natural logarithm of the posterior kernel, $h(\theta_r) = \log\{p(\mathbf{y} \mid \theta_r, r)\pi(\theta_r \mid r)\}$, as a quadratic about its mode $\tilde{\theta}_r$. Exponentiating then gives an approximation to the posterior kernel which has the form of a normal density with mean $\tilde{\theta}_r$ and covariance matrix $H(\tilde{\theta}_r)$, where $H(\tilde{\theta}_r)$ is minus the inverse Hessian matrix of $h(\theta_r)$ evaluated at $\tilde{\theta}_r$. Integrating the approximation of $\exp\{h(\theta_r)\}$ with respect to θ_r gives

$$p(\mathbf{y} \mid r) \simeq (2\pi)^{n_r/2}|H(\tilde{\theta}_r)|^{1/2}p(\mathbf{y} \mid \tilde{\theta}_r, r)\pi(\tilde{\theta}_r \mid r).$$

The asymptotic justification of the Laplace approximation is obtained under the same regularity conditions that guarantee asymptotic normality of the posterior density. The approximation error is therefore small when the posterior, $\pi(\theta_r \mid \mathbf{y}, r)$, is approximately multivariate normal.

3.5.1.2 Monte Carlo simulation techniques

Approximations of the marginal likelihood using *importance sampling* are based on the identity

$$p(\mathbf{y} \mid r) = \int \frac{p(\mathbf{y} \mid \boldsymbol{\theta}_r, r)\pi(\boldsymbol{\theta}_r \mid r)}{q(\boldsymbol{\theta}_r)} q(\boldsymbol{\theta}_r) d\boldsymbol{\theta}_r = \mathrm{E}_q \left\{ \frac{p(\mathbf{y} \mid \boldsymbol{\theta}_r, r)\pi(\boldsymbol{\theta}_r \mid r)}{q(\boldsymbol{\theta}_r)} \right\}, \tag{3.23}$$

where $\mathrm{E}_q(\cdot)$ denotes expectation with respect to the density $q(\boldsymbol{\theta}_r)$, a suitable *importance density* which should be chosen to sample from the more "important" parts of the space of integration. Using Monte Carlo integration, an estimate of $p(\mathbf{y} \mid r)$ based on an independent and identically distributed (*iid*) sample from $q(\boldsymbol{\theta}_r)$ is given by

$$\hat{p}_{\mathrm{IS}}(\mathbf{y} \mid r) = \frac{1}{L} \sum_{\ell=1}^{L} \frac{p(\mathbf{y} \mid \boldsymbol{\theta}_r^{[\ell]}, r)\pi(\boldsymbol{\theta}_r^{[\ell]} \mid r)}{q(\boldsymbol{\theta}_r^{[\ell]})} \quad \text{where} \quad \boldsymbol{\theta}_r^{[\ell]} \overset{iid}{\sim} q(\boldsymbol{\theta}_r) \text{ for } \ell = 1, \ldots, L. \tag{3.24}$$

A sufficient but not necessary condition for the variance of this estimator to be finite (Frühwirth-Schnatter, 2006) is that the ratio $p(\mathbf{y} \mid \boldsymbol{\theta}_r, r)\pi(\boldsymbol{\theta}_r \mid r)/q(\boldsymbol{\theta}_r)$ is bounded so $q(\boldsymbol{\theta}_r)$ should have heavier tails than the unnormalised posterior $p(\mathbf{y} \mid \boldsymbol{\theta}_r, r)\pi(\boldsymbol{\theta}_r \mid r)$ as well as being a good approximation to $\pi(\boldsymbol{\theta}_r \mid \mathbf{y}, r)$. By taking $q(\boldsymbol{\theta}_r) = \pi(\boldsymbol{\theta}_r \mid r)$ in the integrand of (3.23), we recover the *Monte Carlo approximation*

$$\hat{p}_{\mathrm{MC}}(\mathbf{y} \mid r) = \frac{1}{L} \sum_{\ell=1}^{L} p(\mathbf{y} \mid \boldsymbol{\theta}_r^{[\ell]}, r), \qquad \boldsymbol{\theta}_r^{[\ell]} \overset{iid}{\sim} \pi(\boldsymbol{\theta}_r \mid r), \tag{3.25}$$

as a special case of the importance sampling estimator.

A related approximation, the *reciprocal importance sampling* estimate, is based on the identity

$$1 = \int q(\boldsymbol{\theta}_r) \frac{p(\mathbf{y} \mid r)\pi(\boldsymbol{\theta}_r \mid \mathbf{y}, r)}{p(\mathbf{y} \mid \boldsymbol{\theta}_r, r)\pi(\boldsymbol{\theta}_r \mid r)} d\boldsymbol{\theta}_r, \tag{3.26}$$

from which we obtain

$$p(\mathbf{y} \mid r) = \left\{ \int \frac{q(\boldsymbol{\theta}_r)}{p(\mathbf{y} \mid \boldsymbol{\theta}_r, r)\pi(\boldsymbol{\theta}_r \mid r)} \pi(\boldsymbol{\theta}_r \mid \mathbf{y}, r) d\boldsymbol{\theta}_r \right\}^{-1} = \left[\mathrm{E}_\pi \left\{ \frac{q(\boldsymbol{\theta}_r)}{p(\mathbf{y} \mid \boldsymbol{\theta}_r, r)\pi(\boldsymbol{\theta}_r \mid r)} \right\} \right]^{-1}, \tag{3.27}$$

where here $\mathrm{E}_\pi(\cdot)$ denotes expectation with respect to the posterior, $\pi(\boldsymbol{\theta}_r \mid \mathbf{y}, r)$. An estimate of the marginal likelihood based on an MCMC sample from the posterior is then given by

$$\hat{p}_{\mathrm{RI}}(\mathbf{y} \mid r) = \left\{ \frac{1}{M} \sum_{m=1}^{M} \frac{q(\boldsymbol{\theta}_r^{[m]})}{p(\mathbf{y} \mid \boldsymbol{\theta}_r^{[m]}, r)\pi(\boldsymbol{\theta}_r^{[m]} \mid r)} \right\}^{-1} \quad \text{where} \quad \boldsymbol{\theta}_r^{[m]} \sim \pi(\boldsymbol{\theta}_r \mid \mathbf{y}, r) \text{ for } m = 1, \ldots, M. \tag{3.28}$$

A sufficient but not necessary condition for this estimator to have finite variance is that the ratio $q(\boldsymbol{\theta}_r)/p(\mathbf{y} \mid \boldsymbol{\theta}_r, r)\pi(\boldsymbol{\theta}_r \mid r)$ is bounded so $q(\boldsymbol{\theta}_r)$ should have thinner tails than the unnormalised posterior $p(\mathbf{y} \mid \boldsymbol{\theta}_r, r)\pi(\boldsymbol{\theta}_r \mid r)$ and be a good approximation to $\pi(\boldsymbol{\theta}_r \mid \mathbf{y}, r)$. If we take $q(\boldsymbol{\theta}_r) = \pi(\boldsymbol{\theta}_r \mid r)$ in (3.27) we obtain a special case of $\hat{p}_{\mathrm{RI}}$ called the *harmonic mean estimator*

$$\hat{p}_{\mathrm{HM}}(\mathbf{y} \mid r) = \left\{ \frac{1}{M} \sum_{m=1}^{M} \frac{1}{p(\mathbf{y} \mid \boldsymbol{\theta}_r^{[m]}, r)} \right\}^{-1}, \tag{3.29}$$

where $\theta_r^{[1]}, \ldots, \theta_r^{[M]}$ are MCMC draws from the posterior.

Meng & Wong (1996) introduced the *bridge sampling* technique, which is based on the identity

$$1 = \frac{\int \alpha(\theta_r)\pi(\theta_r \mid \mathbf{y}, r)q(\theta_r)d\,\theta_r}{\int \alpha(\theta_r)q(\theta_r)\pi(\theta_r \mid \mathbf{y}, r)d\,\theta_r} = \frac{\mathrm{E}_q\{\alpha(\theta_r)\pi(\theta_r \mid \mathbf{y}, r)\}}{\mathrm{E}_\pi\{\alpha(\theta_r)q(\theta_r)\}}, \tag{3.30}$$

where $q(\theta_r)$ is an importance density approximation to the posterior and $\alpha(\theta_r)$ is the bridge function which satisfies

$$C_\alpha = \int \alpha(\theta_r)\pi(\theta_r \mid \mathbf{y}, r)q(\theta_r)d\,\theta_r > 0. \tag{3.31}$$

Let $\pi^*(\theta_r \mid \mathbf{y}, r) = p(\mathbf{y} \mid \theta_r, r)\pi(\theta_r \mid r)$ denote the unnormalised posterior. Substituting $\pi(\theta_r \mid \mathbf{y}, r) = \pi^*(\theta_r \mid \mathbf{y}, r)/p(\mathbf{y} \mid r)$ into (3.30) and rearranging leads to the bridge sampling estimator of the marginal likelihood,

$$\hat{p}_{\mathrm{BS}}(\mathbf{y} \mid r) = \frac{\frac{1}{L}\sum_{\ell=1}^{L} \alpha(\tilde{\theta}_r^{[\ell]})\pi^*(\tilde{\theta}_r^{[\ell]} \mid \mathbf{y}, r)}{\frac{1}{M}\sum_{m=1}^{M} \alpha(\theta_r^{[m]})q(\theta_r^{[m]})}, \tag{3.32}$$

which is based on an *iid* sample from the importance density, $\tilde{\theta}_r^{[\ell]} \overset{iid}{\sim} q(\theta_r)$ for $\ell = 1, \ldots, L$, as well as an MCMC sample from the posterior, $\theta_r^{[m]} \sim \pi(\theta_r \mid \mathbf{y}, r)$ for $m = 1, \ldots, M$. Note that by taking $\alpha(\theta_r) = 1/q(\theta_r)$ or $\alpha(\theta_r) = 1/\pi^*(\theta_r \mid \mathbf{y}, r)$ we recover the importance or reciprocal importance sampling estimators, respectively. Based on *iid* draws from both $\pi(\theta_r \mid \mathbf{y}, r)$ and $q(\theta_r)$, Meng & Wong (1996) show that an asymptotically optimal choice of $\alpha(\theta_r)$, which minimises the expected relative error of the estimator $\hat{p}_{\mathrm{BS}}(\mathbf{y} \mid r)$ is

$$\alpha(\theta_r) \propto \frac{1}{Lq(\theta_r) + M\pi(\theta_r \mid \mathbf{y}, r)}. \tag{3.33}$$

When the bridge function is chosen to satisfy (3.33), Frühwirth-Schnatter (2004) shows that the relative mean square error of the bridge sampling estimator depends on a sum of ratios, each of which is bounded regardless of the tail behaviour of $q(\theta_r)$. This makes the optimal bridge sampling estimator less sensitive to the tail behaviour of the importance density $q(\theta_r)$ relative to the posterior $\pi(\theta_r \mid \mathbf{y}, r)$ than the importance sampling estimator or the reciprocal importance sampling estimator. The optimal bridge function (3.33) depends on the normalised posterior density, which is unknown. Meng & Wong (1996) suggest an iterative procedure to obtain $\hat{p}_{\mathrm{BS}}(\mathbf{y} \mid r)$ as the limit of a sequence $\hat{p}_{\mathrm{BS},t}(\mathbf{y} \mid r)$ as $t \to \infty$. At each step, $t = 1, 2, \ldots$, the most recent estimate $\hat{p}_{\mathrm{BS},t-1}(\mathbf{y} \mid r)$ is used to normalise the posterior kernel $\pi^*(\theta_r \mid \mathbf{y}, r)$ and produce a new estimate $\hat{p}_{\mathrm{BS},t}(\mathbf{y} \mid r)$ using (3.32) and (3.33) via the recursion

$$\hat{p}_{\mathrm{BS}_r,t} = \frac{\frac{1}{L}\sum_{\ell=1}^{L} \frac{\pi^*(\tilde{\theta}_r^{[\ell]} \mid \mathbf{y}, r)}{Lq(\tilde{\theta}_r^{[\ell]}) + M\pi^*(\tilde{\theta}_r^{[\ell]} \mid \mathbf{y}, r)/\hat{p}_{\mathrm{BS}_r,t-1}}}{\frac{1}{M}\sum_{m=1}^{M} \frac{q(\theta_r^{[m]})}{Lq(\theta_r^{[m]}) + M\pi^*(\theta_r^{[m]} \mid \mathbf{y}, r)/\hat{p}_{\mathrm{BS}_r,t-1}}}. \tag{3.34}$$

The recursion is typically initialised at $t = 0$ by one of the other marginal likelihood approximations and is run until convergence. A simple choice for $q(\theta_r)$ is to take the importance sampling density to be the prior, yielding the estimator

$$\hat{p}_{\mathrm{BSP}_r,t} = \frac{\frac{1}{L}\sum_{\ell=1}^{L} \frac{p(\mathbf{y} \mid \tilde{\theta}_r^{[\ell]}, r)}{L + Mp(\mathbf{y} \mid \tilde{\theta}_r^{[\ell]}, r)/\hat{p}_{\mathrm{BSP}_r,t-1}}}{\frac{1}{M}\sum_{m=1}^{M} \frac{1}{L + Mp(\mathbf{y} \mid \theta_r^{[m]}, r)/\hat{p}_{\mathrm{BSP}_r,t-1}}}. \tag{3.35}$$

Friel & Pettitt (2008) propose a method to compute the marginal likelihood based on samples from the so–called *power posterior*, defined as

$$\pi_t(\theta_r \mid \mathbf{y}, r) \propto p(\mathbf{y} \mid \theta_r, r)^t \pi(\theta_r \mid r) \tag{3.36}$$

where $t \in [0, 1]$ is an auxiliary variable (or "temperature" parameter). Borrowing ideas from path sampling (Gelman & Meng, 1998) allows the log marginal likelihood to be expressed as

$$\log p(\mathbf{y} \mid r) = \int_0^1 \mathrm{E}_{\theta_r|\mathbf{y},r,t}\{\log p(\mathbf{y} \mid \theta_r, r)\} dt, \tag{3.37}$$

where the expectation of the half mean deviance in the integrand is with respect to the power posterior at temperature t.

To estimate the integral in (3.37), Friel & Pettitt (2008) suggest two alternatives: a serial MCMC approach and a population MCMC approach. Here we provide details on the former. Using this approach, the integral is discretised over $t \in [0, 1]$ as $0 = t_0 < t_1 < \ldots t_{n-1} < t_n = 1$, and then approximated by the trapezoidal rule,

$$\log p_{PP}(\mathbf{y} \mid r) = \sum_{i=0}^{n-1} (t_{i+1} - t_i) \frac{\mathrm{E}_{\theta_r|\mathbf{y},r,t_{i+1}}\{\log p(\mathbf{y} \mid \theta_r, r)\} + \mathrm{E}_{\theta_r|\mathbf{y},r,t_i}\{\log p(\mathbf{y} \mid \theta_r, r)\}}{2}, \tag{3.38}$$

where $\log p_{PP}(\mathbf{y} \mid r) \simeq \log p(\mathbf{y} \mid r)$. By separately sampling from the power posterior at each temperature t, the expectations $\mathrm{E}_{\theta_r|\mathbf{y},r,t}\{\log p(\mathbf{y} \mid \theta_r, r)\}$ in (3.38) can be estimated by Monte Carlo integration.

Denote by $\theta_{r,i}^{[j]}$ the j–th iteration of parameters θ_r from an MCMC sampler exploring the power posterior at temperature t_i, $\pi(\theta_r|\mathbf{y}, r, t_i)$. Given a particular choice of n and a particular spacing for the t_i's, the algorithm is presented in Algorithm 3.5.1. Note that for reasonably closely spaced t_i, (3.39) should give a reasonable starting value for the next Markov chain so that little burn–in is required.

To apply the power posterior approach to hidden Markov models, we can either use data augmentation, appending the hidden states to the set of unknowns, or we can marginalise over the hidden states. Friel & Pettitt (2008) use the former approach in computing the marginal likelihood for a hidden Markov random field model, although in their case marginalisation over the hidden variables would be more involved than in a standard hidden Markov model. The data augmentation approach has the advantage that if the conditional density $p(\mathbf{y}|\mathbf{s}, \theta_r, r)$ follows an

Algorithm 3.5.1 Estimating the marginal likelihood via power posteriors

1: Initialise $\theta_{r,0}^{[0]}$, if possible at the prior mean (thereby guaranteeing immediate convergence to the power posterior at temperature 0, i.e. the prior).
2: For $i = 0, \ldots, n$:

(a) Set the temperature parameter t_i

(b) Generate a sample $\{\theta_{r,i}^{[K+1]}, \ldots, \theta_{r,i}^{[R]}\}$ by MCMC sampling from $\pi(\theta_r|\mathbf{y}, r, t_i)$

(c) Estimate the expectation

$$\mathrm{E}_{\theta_r|\mathbf{y},r,t_i}\{\log p(\mathbf{y}|\theta_r, r)\} \simeq \frac{1}{R-K}\sum_{j=K+1}^{R} \log p(\mathbf{y}|\theta_{r,i}^{[j]}, r)$$

(d) While $i < n$ initialise the next chain at an estimate of the mean for $p(\theta_r \mid \mathbf{y}, r, t_i)$

$$\theta_{r,i+1}^{[0]} = \frac{1}{R-K}\sum_{j=K+1}^{R} \theta_{r,i}^{[j]} \tag{3.39}$$

3: Compute $\log p_{PP}(\mathbf{y} \mid r)$ using (3.38)

exponential family, then raising that to a power t gives a distribution belonging to the same exponential family. This means the power posterior could easily be sampled, given a conjugate prior. If the hidden states are integrated out of the joint power posteriors for $(\theta_r, \mathbf{s} \mid r)$, then the approach bears some resemblance to simulated tempering (Celeux *et al.*, 2000), in which t is introduced to allow easier movement around the posterior for θ_r. As with the marginal updating scheme in Section 3.3.4.3, the power posteriors in this case will not be of standard form requiring Metropolis Hastings updates for the parameters θ_r.

3.5.1.3 Chib's method

Chib (1995) proposed a marginal likelihood approximation based on the equation

$$p(\mathbf{y} \mid r) = \frac{p(\mathbf{y} \mid \theta_r, r)\pi(\theta_r \mid r)}{\pi(\theta_r \mid \mathbf{y}, r)} \qquad \text{for all } \theta_r \text{ such that } \pi(\theta_r \mid \mathbf{y}, r) > 0,$$

which is simply a rearrangement of Bayes Theorem. Therefore, for a given θ_r^*, if a good approximation to the posterior $\pi(\theta_r^* \mid \mathbf{y}, r)$ can be constructed then the marginal likelihood can be estimated by

$$\hat{p}_{\mathrm{CM}}(\mathbf{y} \mid r) = \frac{p(\mathbf{y} \mid \theta_r^*, r)\pi(\theta_r^* \mid r)}{\hat{\pi}(\theta_r^* \mid \mathbf{y}, r)}, \tag{3.40}$$

where for greater efficiency in estimation, typically θ_r^* is taken to be a point of high posterior density. The estimate of the posterior ordinate is based on the marginal/conditional decomposition

of $\pi(\boldsymbol{\theta}_r \mid \mathbf{y}, r)$ into $B \leq n_r$ blocks,

$$\pi(\boldsymbol{\theta}_r \mid \mathbf{y}, r) = \prod_{i=1}^{B} \pi(\boldsymbol{\theta}_{r,i} \mid \mathbf{y}, r, \boldsymbol{\theta}_{r,1}, \ldots, \boldsymbol{\theta}_{r,i-1}).$$

Each of the ordinates $\pi(\boldsymbol{\theta}_{r,i} \mid \mathbf{y}, r, \boldsymbol{\theta}_{r,1}, \ldots, \boldsymbol{\theta}_{r,i-1})$ is then estimated in a separate MCMC run, fixing certain parameter blocks at the high density points for those blocks as appropriate. To estimate the ordinates using the techniques described in Chib (1995), all the full conditional distributions in the Gibbs sampler must be known. Chib & Jeliazkov (2001) extended the method to handle situations in which some of the full conditional distributions have unknown normalising constants and the corresponding parameters are updated in Metropolis Hastings steps.

When approximating the marginal likelihood for hidden Markov models and an exchangeable prior is used, application of Chib's method will produce a biased estimate if the label switching problem is not addressed; see, for example, Frühwirth-Schnatter (2004) or Marin & Robert (2008). In brief, if the approximation is based on MCMC output that fails to visit all of the $r!$ modes in the posterior, it cannot provide an accurate estimate of the posterior ordinate or therefore the marginal likelihood. One way of overcoming this problem is to use an MCMC sampler which, by design, only explores one of the $r!$ posterior modes and then to adjust the marginal likelihood appropriately, by multiplying with a factor of $r!$; see Frühwirth-Schnatter (2004) for a full explanation. The problem with this approach is that the "correction" will lead to a biased estimate if the MCMC scheme fails to isolate a single mode of the posterior distribution. Another solution is to adjust the approximation of the posterior ordinate so that it respects the symmetry of the posterior, for example, Frühwirth-Schnatter (2004) suggests basing the approximation on the MCMC output of a random permutation sampler, which forces exploration of all $r!$ posterior modes. Marin & Robert (2008) propose an alternative fix.

3.5.1.4 Marginal posterior methods

In the literature on finite mixture models, other techniques for making inference about the number of components are based on the marginal posterior distribution for the (latent) component indicators, say $\mathbf{s}$, having first integrated the model parameters $\boldsymbol{\theta}_r = (\boldsymbol{\theta}_{r,\text{obs}}, \boldsymbol{\theta}_{r,\text{hid}})$ out of the joint posterior $\pi(\boldsymbol{\theta}_r, \mathbf{s} \mid \mathbf{y}, r)$. These approaches are designed for models where this marginalisation can be performed analytically, which demands the choice of a conjugate prior for $\boldsymbol{\theta}_r$. Nobile & Fearnside (2007) proposed an MCMC sampler, known as the allocation sampler, whose state space consists of the number of components and the component indicators. It can therefore be regarded as analogous to the reversible jump sampler of Richardson & Green (1997), on a reduced state space, in which transitions occur between discrete spaces containing different numbers of elements, as opposed to spaces of variable dimension. The sampler comprises moves which do not affect the number of components and moves which change it.

Having integrated out the model parameters from the joint posterior density $\pi(\boldsymbol{\theta}_r, \mathbf{s} \mid \mathbf{y}, r)$, Steele *et al.* (2006) suggest an importance sampling estimator of the marginal likelihood, expressed in terms of the latent component indicators,

$$p(\mathbf{y} \mid r) = \sum_{\mathbf{s}} p(\mathbf{y} \mid \mathbf{s}, r) p(\mathbf{s} \mid r). \tag{3.41}$$

Here summation is over all possible component indicator combinations. The authors propose using a mixture importance mass function $q(\mathbf{s})$ which respects the symmetry of the posterior. To avoid making $q(\mathbf{s})$ overly concentrated, its first component mass function is the prior, $p(\mathbf{s} \mid r)$. Other components, centred at posterior modes, are then added to $q(\mathbf{s})$ incrementally until it is judged that no important parts of the space of integration in (3.41) have been missed. This includes the $r!$ (major) symmetric modes as well as smaller local modes in the posterior mass function.

3.5.1.5 Discussion: comparing the methods of approximation

Although the Laplace approximation of the marginal likelihood can be very effective when the posterior kernel is sufficiently well–behaved, it is justified using arguments which appeal to the asymptotic normality of the posterior density. However, for overfitting hidden Markov models, asymptotic normality of the posterior may not hold, making the assumptions underlying the approximation untenable.

Of the first three Monte Carlo simulation techniques, the optimal bridge sampling estimator has an advantage over the importance sampling or reciprocal importance sampling estimators of having a bounded variance. However, its computation is more expensive because of the requirement to use a recursive formula together with *both* an MCMC sample from the posterior *and* an *iid* sample from an importance density. More generally, the strengths and weaknesses of these three estimators are intrinsically linked to the choice of importance sampling density. Given a well chosen $q(\boldsymbol{\theta}_r)$ from which it is easy to sample (or evaluate in the case of reciprocal importance sampling), approximation of the marginal likelihood should be reasonably easy and efficient. To work well, the importance sampling density should have most if its density concentrated in "important" parts of the space of integration, in other words its shape should be similar to that of the posterior. The multimodality of the posterior density therefore makes it difficult to find a suitable importance density, and a poorly chosen $q(\boldsymbol{\theta}_r)$ can lead to biased estimators or estimators with high variance. In the following paragraphs we discuss possible choices for $q(\boldsymbol{\theta}_r)$, emphasising their merits and weaknesses.

One simple and automatic choice is to take $q(\boldsymbol{\theta}_r) = \pi(\boldsymbol{\theta}_r \mid r)$ which, assuming an exchangeable prior is selected, respects the symmetry of the posterior by offering equal support to all of its $r!$ symmetric modes. With importance sampling, this choice leads to the Monte Carlo estimator in equation (3.25). However, since the prior is usually flat relative to the posterior, this simple estimator is likely to be inefficient as many draws from the prior will fall in regions of low likelihood. Bos (2002) compared marginal likelihood estimators for a simple regression model and found this estimator to be unstable, with a substantially larger variance than that of any of the other estimators considered. Steele *et al.* (2006) reached similar conclusions with regards to the variance of the Monte Carlo estimator in a comparison involving finite mixture models. However, it was notable that the Monte Carlo estimator was less biased than the more sophisticated approaches considered, with a considerably shorter computing time.

Taking $q(\boldsymbol{\theta}_r) = \pi(\boldsymbol{\theta}_r \mid r)$ in the reciprocal importance sampling estimator yields the harmonic mean estimator in equation (3.29). In computing this estimator, all that is required to obtain the summands is the observed data likelihood, evaluated at the posterior draws. If the MCMC

scheme makes use of the forward backward algorithm to sample the hidden states, these values are available directly as a by-product of forward filtering. This means that the harmonic mean estimator can be computed immediately following MCMC sampling. In spite of its ease of computation, however, the harmonic mean estimator is prone to being unstable if there happens to be a few very small likelihood values in the MCMC output. In a simulation study involving finite mixtures of Poisson distributions, Frühwirth-Schnatter (2006) found the harmonic mean estimator to perform rather poorly, particularly when the differences between the means of the component Poisson distributions were large. This may have been because, when the posterior modes are well separated, the prior (which is unimodal) provides an even poorer approximation to the unequivocally multimodal posterior. In an effort to stabilise the harmonic mean estimator, Newton & Raftery (1994) suggested a hybrid estimator, based on combined samples from the prior and posterior. Starting with the simulation consistent marginal likelihood estimator

$$\frac{\sum_{m=1}^{M} \pi(\theta_r^{[m]} \mid r)/q(\theta_r^{[m]} \mid r) \times p(\mathbf{y} \mid \theta_r^{[m]}, r)}{\sum_{m=1}^{M} \pi(\theta_r^{[m]} \mid r)/q(\theta_r^{[m]} \mid r)}$$

and taking the importance density to be $q(\theta_r) = \delta\pi(\theta_r \mid r) + (1-\delta)\pi(\theta_r \mid \mathbf{y}, r)$ with $0 < \delta < 1$ and δ small, for example $\delta = 0.05$, leads to the estimator

$$\hat{p}_{\mathrm{NRF}}(\mathbf{y} \mid r) = \frac{\sum_{m=1}^{M} \dfrac{p(\mathbf{y} \mid \theta_r^{[m]}, r)}{\delta\hat{p}_{\mathrm{NRF}}(\mathbf{y} \mid r) + (1-\delta)p(\mathbf{y} \mid \theta_r^{[m]}, r)}}{\sum_{m=1}^{M} \{\delta\hat{p}_{\mathrm{NRF}}(\mathbf{y} \mid r) + (1-\delta)p(\mathbf{y} \mid \theta_r^{[m]}, r)\}^{-1}}. \tag{3.42}$$

This can be computed by a standard iterative scheme. To avoid having to simulate from the prior, Newton & Raftery (1994) also suggest an approximation to (3.42) based on a sample of size M from the posterior and a notional sample of size $\delta M/(1-\delta)$ from the prior, such that all likelihood values $p(\mathbf{y} \mid \theta_r^{[m]}, r)$, evaluated at the notional prior draws, are equal to their expected value $p(\mathbf{y} \mid r)$. This yields the approximation

$$\hat{p}_{\mathrm{NRA}}(\mathbf{y} \mid r) = \frac{\dfrac{\delta M}{(1-\delta)} + \sum_{m=1}^{M} \dfrac{p(\mathbf{y} \mid \theta_r^{[m]}, r)}{\delta\hat{p}_{\mathrm{NRA}}(\mathbf{y} \mid r) + (1-\delta)p(\mathbf{y} \mid \theta_r^{[m]}, r)}}{\dfrac{\delta M}{(1-\delta)\hat{p}_{\mathrm{NRA}}(\mathbf{y} \mid r)} + \sum_{m=1}^{M} \{\delta\hat{p}_{\mathrm{NRA}}(\mathbf{y} \mid r) + (1-\delta)p(\mathbf{y} \mid \theta_r^{[m]}, r)\}^{-1}} \tag{3.43}$$

which can be computed by a standard iterative scheme. Green (2003) recommends using the hybrid estimator over the Monte Carlo or harmonic mean estimators.

The prior generally provides a poor approximation to the posterior. Choosing an importance sampling density which bears a closer resemblance to the posterior may improve bridge sampling, importance sampling and reciprocal importance sampling estimates of the marginal likelihood. However, finding such a density is not easy. One possibility might be to base the importance sampling density on an approximation to the posterior, obtained using the output from some

initial MCMC run. Being unimodal, a simple normal approximation is unlikely to work well because of the multimodality of the posterior. An obvious remedy might therefore be to introduce an identifiability constraint in order to focus on one particular mode, then to choose the importance sampling density $q(\theta_r)$ to approximate this mode. This demands an adjustment to the marginal likelihood estimator in which the prior in the summands is normalised over the constrained space by multiplying by $r!$; see Frühwirth-Schnatter (2004) for further explanation. However, the posterior distribution in hidden Markov models often possesses local modes, meaning the (constrained) posterior under the identifiability constraint may itself be multimodal. Again this will be detrimental to the performance of a unimodal importance sampling density like the normal distribution. In response to these problems Frühwirth-Schnatter (2006) proposes taking $q(\theta_r)$ to be an estimate of the (unconstrained) posterior density, based on the expression

$$p(\theta_r \mid \mathbf{y}, r) = \sum_{\mathbf{s}} \pi(\theta_{r,\text{obs}} \mid \mathbf{y}, \mathbf{s}, r) p(\theta_{r,\text{hid}} \mid \mathbf{s}, r) p(\mathbf{s} \mid \mathbf{y}, r) \tag{3.44}$$

where the sum is over all possible hidden states sequences and $\theta_{r,\text{obs}}$ and $\theta_{r,\text{hid}}$ are assumed independent *a priori*. Given MCMC draws $\mathbf{s}^{[1]}, \ldots, \mathbf{s}^{[M]}$ from the posterior for the hidden states, expression (3.44) can be approximated by Rao–Blackwellisation leading to the importance sampling density

$$q(\theta_r) = \frac{1}{M} \sum_{m=1}^{M} \pi(\theta_{r,\text{obs}} \mid \mathbf{y}, \mathbf{s}^{[m]}, r) p(\theta_{r,\text{hid}} \mid \mathbf{s}^{[m]}, r). \tag{3.45}$$

To ensure the importance sampling density captures all the modes of the posterior it is essential that the MCMC sampler forces balanced label switching. The approximation in (3.45) is based on the assumption that the priors for $\theta_{r,\text{obs}}$ and $\theta_{r,\text{hid}}$ are conjugate so that the summands in (3.45) are available in closed form. If the priors for some of the parameters in $\theta_{r,\text{obs}}$ are only semi–conjugate, then Frühwirth-Schnatter (2004) provides a more general importance sampling density, of which (3.45) is a special case. In a simulation experiment involving Poisson mixture models, Frühwirth-Schnatter (2006) found the importance sampling density (3.45) to lead to approximately unbiased marginal likelihood estimates, with the bridge sampling estimator having a smaller relative mean square error than the importance or reciprocal importance sampling estimators. One of the most appealing features of this importance sampling density is that its construction can be incorporated into MCMC sampling. Therefore little extra work is required to approximate the marginal likelihood after obtaining a posterior sample. However, the main drawback is that there is no obvious extension in situations when some of the full conditional densities have unknown normalising constants.

This is not a problem when estimating the marginal likelihood via power posteriors and the method remains applicable when non–conjugate priors are used. Although the same is true of Chib's (extended) method, estimation by power posteriors is more automatic and involves considerably less bookkeeping. The approximation is sensitive to the chosen number of temperatures (n) and to the spacing of the t_i's. Therefore the need to make these choices is a drawback of this approach. Furthermore, with a large number of temperatures, obtaining the estimate can be computationally slow. This is especially noticeable when compared to other techniques, in which the marginal likelihood can be estimated more or less directly from the output of standard MCMC sampling. Finally, the validity of the approximation depends crucially on convergence

of the collection of Markov chains but checking all of them may be impractical. As such, we should bear in mind the pitfalls of unsupervised MCMC.

Chib's method lends itself particularly well to MCMC sampling when data augmentation is used. Given the available corrections, therefore, it is an attractive method if the number of blocks B is not too large and especially if the full conditional distributions are all known. For instance, if the complete data posterior $\pi(\theta_r \mid \mathbf{s}, \mathbf{y}, r)$ is available in closed form, then an estimate of the posterior ordinate is given by

$$\hat{\pi}(\theta_r^* \mid \mathbf{y}, r) = \frac{1}{N} \sum_{i=1}^{N} \pi(\theta_r^* \mid \mathbf{s}^{[i]}, \mathbf{y}, r)$$

where $\mathbf{s}^{[1]}, \ldots, \mathbf{s}^{[N]}$ are the MCMC draws of the hidden states obtained using a random permutation sampler. The summary quantities required to compute this sum can be stored during MCMC sampling, so afterwards very little extra computing time and effort is required to approximate the marginal likelihood. However, for hidden Markov models in which the number of parameter blocks is large, many reduced MCMC runs would be required to compute the estimate. This would require considerable computing time, as well as judicious bookkeeping.

The scope of methods based on the marginal posterior for the latent variables is limited to model/prior combinations in which marginalisation over the parameters can be performed analytically. However, within the limits of their viability, these methods address the specific peculiarities of latent variable models (such as the multimodality of the posterior) and so might be expected to perform well. More significantly, because the parameters are integrated out of the model analytically, the methods of approximation remain essentially the same even for within–state distributions of high dimension.

3.5.2 Across model simulation

In dealing with model uncertainty, an alternative to within model simulation is transdimensional MCMC, also known as across model simulation. These methods involve constructing Markov chains which simultaneously traverse both the parameter and the model space, in exploration of the joint posterior distribution, $\pi(\theta_r, r \mid \mathbf{y})$.

The most widely implemented approach to transdimensional MCMC is reversible jump MCMC introduced by Green (1995) as a generalisation of the Metropolis Hastings algorithm, which additionally includes the model indicator. The reversible jump algorithm allows the construction of an ergodic Markov chain with states of the form (r, θ_r) and the joint posterior distribution of the parameters and the model indicator as its stationary distribution. The sampler "jumps" between models by periodically proposing moves from one model to another, each of which is rejected with a probability which ensures that the chain possesses the correct stationary distribution. Attractive features of RJMCMC include the efficiency associated with simultaneous exploration of the model and parameter space and the potential for improved mixing, especially if posteriors are multimodal. In this case, the possibility of moving between different models can lead to easier passage between local modes than would be possible with standard fixed dimensional samplers. For example, Richardson & Green (1997) make this observation in an

application involving finite mixture models with an unknown number of components. In spite of these benefits, the success of any RJMCMC scheme is largely dependent on the ability to construct efficient proposal distributions for transdimensional moves. Although there have been attempts to provide guidelines (see, for example, Brooks *et al.*, 2003; Godsill, 2001), this is a challenging problem, especially when the difference in dimensionality between different models is large.

Besides RJMCMC, a variety of alternative across model simulation techniques have also been applied to mixture and hidden Markov models with an unknown number of components/states. These include product–space MCMC methods (Carlin & Chib, 1995), continuous time MCMC samplers based on marked point processes (Cappé *et al.*, 2003; Shi *et al.*, 2002) and the saturated state space approach (Brooks *et al.*, 2003). All provide the potential benefits of efficiency and improved mixing but, like RJMCMC, these benefits can only be achieved if the across model sampler promotes good mixing between models. For more details on across model simulation techniques, see the technical report Germain *et al.* (2010*a*), or the reviews in Green (2003) or Sisson (2005).

In general, one of the main factors influencing the choice between within and across model simulation should be the number of models under consideration, in our case, the size of $r_{\max}$. If $r_{\max}$ is large, computing the marginal likelihood of each model separately (within model simulation) would be computationally prohibitive, whereas an across model approach may be viable. Another consideration should be the tenability of designing an across model sampler which mixes well over the joint space of the model indicators and the model parameters. Generally, this is likely to be easiest when models differ only by the presence/absence of a small number of parameters, and is more difficult otherwise. For example, Robert *et al.* (2000) compared the performance of reversible jump samplers when they modelled various sets of data using both mixture models and hidden Markov models, with univariate normal within component/state distributions. For transdimensional moves, acceptance rates were found to be around 20–30% when using mixture models. However, when modelling the same data using hidden Markov models, for which there were much larger differences in dimensionality between parameter spaces for different models, the rates were substantially lower (around 0.3–4.4%). Acceptance rates of a similarly low magnitude were found by Dellaportas & Papageorgiou (2006) when they generalised the reversible jump scheme proposed by Richardson & Green (1997) to model mixtures of multivariate normal distributions.

In this thesis, we model rainfall using hidden Markov models for which the number of states is unknown. However, motivated by physical arguments (see Section 4.7.1.1), we choose to limit $r_{\max}$ so that the number of states is not overly large. In principle, therefore, within model simulation will be feasible. Moreover, because we model rainfall at multiple sites, the within-state distributions will be multivariate and highly parameterised. It is likely, therefore, that building an efficient across model sampler would be very difficult. Consequently, we do not consider transdimensional MCMC further in this thesis and, instead, investigate within model simulation techniques.

Chapter 4

A homogeneous hidden Markov model for rainfall data

4.1 Introduction

Zucchini & Guttorp (1991) pioneered the use of hidden Markov models for describing daily precipitation at multiple sites. In applications of multi-site rainfall modelling, the unobserved states in a hidden Markov model correspond to particular patterns of precipitation at the sites. Although these states might not be identifiable with interpretable weather types, they are generally intended to summarise the meteorological situation and as such can be interpreted as *"weather states"*. In their seminal work, Zucchini & Guttorp (1991) characterised the temporal structure in their hidden Markov model for rainfall occurrence by assuming a homogeneous first order Markov chain for the hidden states and conditional (temporal) independence in the observed process, given the hidden process. That is, the temporal structure could be summarised by assumptions A1 and A2, respectively, from Chapter 3. The spatial structure of their model was then simplified through an assumption of conditional independence between sites given the weather state. Following this initial publication, the basic model has been extended in a variety of ways. In brief, extensions have included allowing the hidden Markov chain to be non-homogeneous with transition probabilities dependent upon observed atmospheric variables (Hughes & Guttorp, 1994*a*); explicitly modelling spatial dependence between rainfall occurrences within weather states (Hughes & Guttorp, 1994*b*; Hughes *et al.*, 1999); and additionally modelling precipitation amounts, summarising the spatial structure both with (Bellone *et al.*, 2000; Betro *et al.*, 2008) and without (Thompson *et al.*, 2007; Ailliot *et al.*, 2009) an assumption of conditional independence given occurrence and the weather state. Data from Australia (Hughes *et al.*, 1999; Charles *et al.*, 2004), New Zealand (Thompson *et al.*, 2007; Ailliot *et al.*, 2009), North America (Hughes & Guttorp, 1994*a*; Bellone *et al.*, 2000), South America (Robertson *et al.*, 2004), Africa (MacDonald & Zucchini, 1997) and the Mediterranean (Betro *et al.*, 2008) have been analysed, but to our knowledge, hidden Markov models have not found application in analyses of UK data.

More significantly in terms of this thesis, in all of these studies, the problem has been formulated

in a frequentist framework using standard inferential and computational techniques such as maximum likelihood and the EM algorithm, respectively. The Bayesian approach which we present is therefore novel in this aspect, and through it we can provide a complete and coherent summary of all post data uncertainty, including that surrounding the number of hidden states. An additional benefit of modelling within the Bayesian paradigm is the facility to incorporate prior knowledge. We show how, with some consideration, prior beliefs about the rainfall process can easily be encapsulated in probabilistic form.

The remainder of this chapter is organised as follows. Section 4.2 describes a hidden Markov model for precipitation, including a discussion of the underlying assumptions, specific parameterisations that have been used in this implementation and an exploration of the spatio-temporal dependence structure. This is followed in Section 4.3 by an explanation of the chosen prior distribution and details of how prior information can be incorporated. A description of the likelihood is given in Section 4.4, followed by details of the MCMC scheme used for posterior inference in Section 4.5, including details of the conditional posterior distribution of the parameters, given the weather states. The section which follows considers the problem of inference about the number of weather states. Following the introduction to within model simulation in Chapter 3, Section 4.6 presents details of a simulation experiment, in which several methods for estimating the marginal likelihood are compared. Section 4.7 applies the model and inferential procedures to the Yorkshire dataset, and include details of the prior specification, the resulting posteriors and the use of the posterior predictive distribution for model checking.

4.2 Description of the hidden Markov model

Following the introduction to hidden Markov models presented in Chapter 3, suppose there exists a *hidden* or unobservable discrete-valued stochastic process which we interpret as the *weather state*. We denote by S_t the weather state at time t, $t = 1, 2, \ldots, T$, and by $\mathcal{S}_r = \{1, 2, \ldots, r\}$ its state space. For example, if we assumed there to be just two weather states, $r = 2$, then broadly speaking we might expect one to be associated with wet weather conditions and the other to be associated with dry weather conditions. Our interest lies in modelling daily rainfall data and so a day represents one time unit, although the theory would remain applicable for different (discrete) units of time.

Consider a network comprising n sites at which rain is measured. Let $\mathbf{D}_t = (D_t^1, D_t^2, \ldots, D_t^n)^T$ be an n-dimensional random vector for the process of rainfall occurrence defined so that

$$D_t^i = \begin{cases} 1, & \text{if there is greater than or equal to } c\text{mm rain on day } t \text{ at site } i, \\ 0, & \text{otherwise,} \end{cases}$$

for some suitable cut-off $c$mm. According to the American Meteorological Society, in British climatology a *rain day* is defined as a 24 hour period in which at least 0.01 in. or 0.2 mm of precipitation is recorded (Glickman, 2000), so we use the cut-off $c = 0.2$mm. Let $\mathbf{W}_t = (W_t^1, W_t^2, \ldots, W_t^n)^T$ be an n-dimensional random vector for the process of rainfall amount defined so that W_t^i is the amount of rain on day t at site i, taken to be 0 if $D_t^i = 0$. Note that the weather state on any particular day is common to all sites in the network.

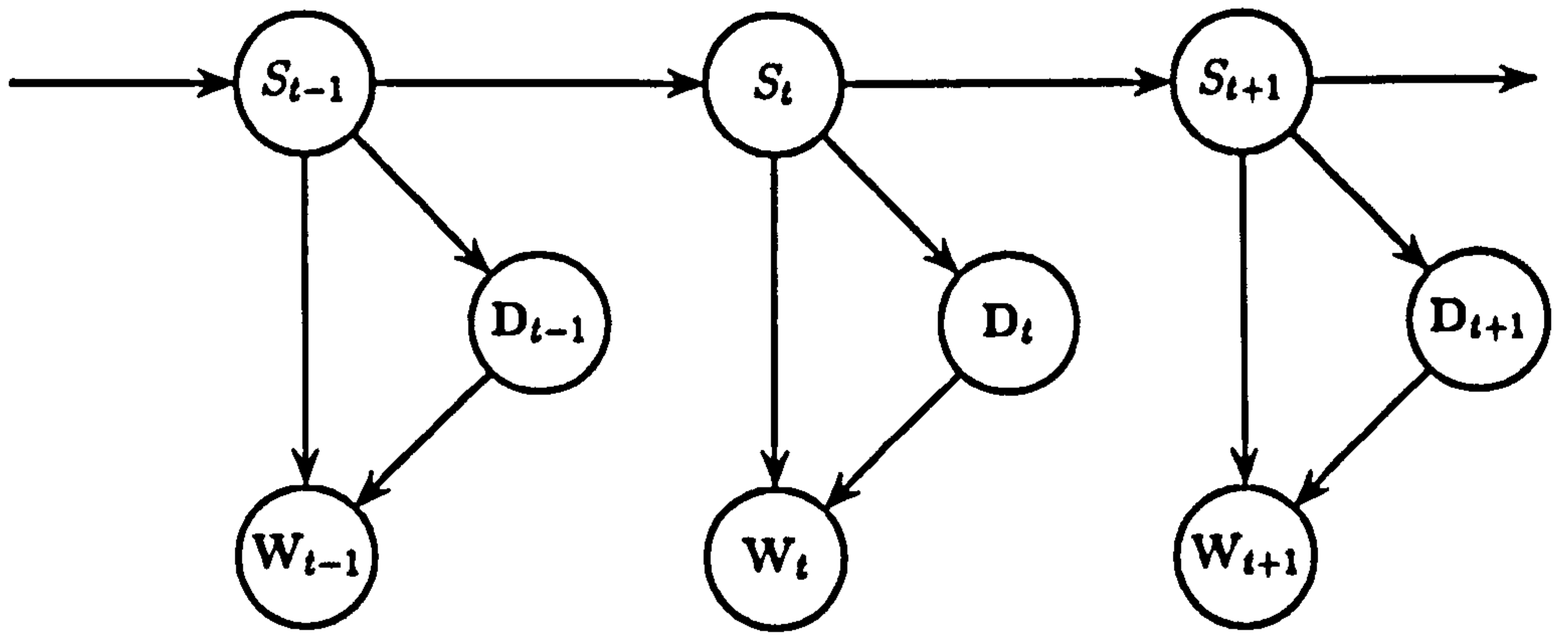


Figure 4.1: A DAG showing the (temporal) dependence structure in the class of HMMs described by assumptions A1 and A2 and the factorisation of the joint mixed density and mass function $p(\mathbf{w}_t, \mathbf{d}_t \mid S_t, \theta_{\text{obs}})$ given in equation (4.1).

4.2.1 Assumptions of the hidden Markov model

In this chapter we focus on a class of hidden Markov models for rainfall occurrence and amount whose temporal structure is defined by the "standard" assumptions from Chapter 3, namely

A1. $\Pr(S_t \mid S_{1:t-1}, \theta) = \Pr(S_t = k \mid S_{t-1} = j, \theta_{\text{hid}}) = \lambda_{jk}, \qquad j, k, \in S_r$
for $t = 2, \ldots, T$ with $\Pr(S_1 \mid \theta) = \Pr(S_1 = k \mid \theta_{\text{hid}}) = \nu_k$.

A2. $p(\mathbf{w}_t, \mathbf{d}_t \mid \mathbf{w}_{1:t-1}, \mathbf{d}_{1:t-1}, S_{1:T}, \theta) = p(\mathbf{w}_t, \mathbf{d}_t \mid S_t = k, \theta_{\text{obs}})$ for $t = 1, \ldots, T$.

These assumptions assert that the temporal dependence in the weather state process (described by A1) captures all the temporal persistence in the precipitation process. Together A1 and A2 describe a broad class of models for precipitation occurrence and amount. A particular hidden Markov model within this class is defined by the parameterisation chosen for $p(\mathbf{w}_t, \mathbf{d}_t \mid S_t = k, \theta_{\text{obs}})$.

4.2.2 Parameterisation for the precipitation process

We factorise the joint mixed density and mass function $p(\mathbf{w}_t, \mathbf{d}_t \mid S_t = k, \theta_{\text{obs}})$ as

$$p(\mathbf{w}_t, \mathbf{d}_t \mid S_t = k, \theta_{\text{obs}}) = \Pr(\mathbf{D}_t = \mathbf{d}_t \mid S_t = k, \theta_{\text{obs}}) p(\mathbf{w}_t \mid \mathbf{D}_t = \mathbf{d}_t, S_t = k, \theta_{\text{obs}}). \tag{4.1}$$

The resulting factorisation of the joint distribution for $\{(\mathbf{W}_t, \mathbf{D}_t, S_t) : t = 1, 2, \ldots, T\}$ is shown in Figure 4.1. However this DAG gives no indication of the conditional spatial structure at any particular time point. In other words, at a given time, t, it provides no information about the relationships between the variables in the joint conditional distributions $\Pr(D_t^1 = d_t^1, \ldots, D_t^n = d_t^n \mid S_t = k, \theta_{\text{obs}})$ or $p(w_t^1, \ldots, w_t^n \mid \mathbf{D}_t = \mathbf{d}_t, S_t = k, \theta_{\text{obs}})$.

The reasonably simple model for $p(\mathbf{w}_t, \mathbf{d}_t \mid S_t, \theta_{\text{obs}})$, which we will consider in this chapter, assumes conditional spatial independence of rainfall occurrence, given the weather state and conditional spatial independence of rainfall amount, given occurrence and the weather state, so

$$\begin{aligned} p(\mathbf{w}_t, \mathbf{d}_t \mid S_t = k, \theta_{\text{obs}}) &= \Pr(\mathbf{D}_t = \mathbf{d}_t \mid S_t = k, \theta_{\text{obs}}) p(\mathbf{w}_t \mid \mathbf{D}_t = \mathbf{d}_t, S_t = k, \theta_{\text{obs}}) \\ &= \prod_{i=1}^{n} \Pr(D_t^i = d_t^i \mid S_t = k, \theta_{\text{obs}}) p(w_t^i \mid D_t^i = d_t^i, S_t = k, \theta_{\text{obs}}), \end{aligned} \tag{4.2}$$

where

$$\Pr(W_t^i = 0 \mid D_t^i = 0) = 1, \quad (W_t^i \mid D_t^i = 1, S_t = k, \theta_{\text{obs}}) \sim \text{Ga}(\alpha_{ik}, \beta_{ik}), \tag{4.3}$$

$$\text{and} \quad D_t^i \mid S_t = k, \theta_{\text{obs}} \sim \text{Bern}(p_{ik}). \tag{4.4}$$

Here, for site i in weather state k, p_{ik} is the probability of rain whilst α_{ik} and β_{ik} are the shape and scale parameters in the gamma distribution, $\text{Ga}(\alpha_{ik}, \beta_{ik})$, for rainfall amounts on wet days. Bellone *et al.* (2000) use the same parameterisation of the precipitation process in their NHMM for rainfall occurrence and amount. Although the gamma distribution is the most commonly used model in the literature for non–zero rainfall amounts, other distributions have been used, for example, a mixture of exponentials (Woolhiser & Roldán, 1986) or the lognormal distribution (Smith, 1994). Betro *et al.* (2008) fitted a hidden Markov model to precipitation data from Sardinia in which the non–zero rainfall amounts were modelled using a mixture of Weibull distributions with a fixed shape parameter, choosing the number of mixture components using the BIC. For their dataset, it was judged that the more standard distributions did not have sufficiently long tails to capture the extreme rainfall events which often affect the central and southeastern coast of the island. However for daily rainfall data in the UK, the gamma distribution has been shown to provide a good model in most regions and most seasons; see Gregory *et al.* (1993).

The shape and scale parameters of a gamma distribution are not natural quantities about which we can elicit prior opinions. This motivates a reparameterisation which allows the specification of priors at a level of the model that we can interpret. For example, we might reparameterise each gamma $\text{Ga}(\alpha_{ik}, \beta_{ik})$ distribution in terms of its mean α_{ik}/β_{ik} and variance α_{ik}/β_{ik}^2; or in terms of its mean and coefficient of variation, $1/\sqrt{\alpha_{ik}}$, the latter quantity being defined as the ratio of the standard deviation to the mean. We prefer to use the coefficient of variation, rather than the variance, because it corresponds more naturally to the way in which most people would think about their uncertainty regarding, especially positive, random quantities; see, for example, Garthwaite *et al.* (2005). This is because in such cases, it is often easier to express our degree of belief on a multiplicative rather than additive scale, as a point estimate plus or minus some percentage "error". Moreover, in specifying the priors for the parameters of the gamma distribution, it is advantageous if the chosen parameterisation allows us to think about changing one parameter, without having to alter our beliefs about the other. This is particularly true for a problem such as the one at hand, where it is necessary to elicit prior information for multiple sites. Suppose we have already specified the hyperparameters in the prior for a site known to be generally "wet". Now suppose that we wish to think about the prior for another site, known to be generally "dry". It would be natural to think first about how the mean differs between the two sites. For a positive valued quantity, if we think the mean differs, we should also expect the variance to differ, but not necessarily the coefficient of variation. With these arguments in mind

we reparameterise the gamma distribution for non–zero rainfall amounts at site i in weather state k as

$$(W_t^i \mid D_t^i = 1, S_t = k, \theta_{\text{obs}}) \sim \text{Ga}\left(\frac{1}{v_{ik}^2}, \frac{1}{v_{ik}^2 m_{ik}}\right), \quad (i,k) \in \{1,2,\ldots,n\} \times \mathcal{S}_r \tag{4.5}$$

where m_{ik} is the mean and v_{ik} is the coefficient of variation of the gamma distribution. The chosen parameterisation of the precipitation process is therefore

$$p(\mathbf{w}_t, \mathbf{d}_t \mid S_t = k, \theta_{\text{obs}}) = \prod_{i=1}^{n} \left\{ p_{ik} \text{Ga}\left(w_t^i \middle| \frac{1}{v_{ik}^2}, \frac{1}{v_{ik}^2 m_{ik}}\right) \right\}^{d_t^i} (1 - p_{ik})^{1-d_t^i}, \tag{4.6}$$

where $\text{Ga}(w \mid \alpha, \beta)$ denotes the gamma $\text{Ga}(\alpha, \beta)$ density (see Appendix E) evaluated at w.

For notational convenience, we collect the set of unknown rainfall occurrence probabilities into a $n \times r$ matrix $\mathcal{P}$ with (i,k)-th entry p_{ik}. Similarly we collect the mean and coefficient of variation parameters into $n \times r$ matrices $\mathcal{M} = (m_{ik})$ and $\mathcal{V} = (v_{ik})$, respectively. The set of all model parameters is therefore denoted by $\theta = (\theta_{\text{hid}}, \theta_{\text{obs}})$, where

$$\theta_{\text{hid}} = (\Lambda, \nu) \in \mathcal{S}_r^r \times \mathcal{S}_r$$

parameterises the hidden process and

$$\theta_{\text{obs}} = (\mathcal{P}, \mathcal{M}, \mathcal{V}) \in [0,1]^{nr} \times \mathbb{R}_+^{nr} \times \mathbb{R}_+^{nr}$$

parameterises the observed process. The notation $[0,1]^x$ denotes the product of x $[0,1]$ intervals, $\mathbb{R}_+^x$ denotes the product of x $(0,\infty)$ intervals and $\mathcal{S}_r^x$ denotes the product of x unit simplices, each one of dimension r.

4.2.3 Exploring the spatio–temporal dependence

It is important to appreciate that although in this relatively simple model we assume conditional temporal and spatial independence, given the weather state, marginally spatio–temporal dependence is induced by the *common* weather state which evolves in time according to a first order Markov chain. The temporal properties of hidden Markov models were illustrated in Chapter 3, in which we presented the autocorrelation function for a 2–state hidden Markov model with Bernoulli within–state distributions, and showed how positive temporal autocorrelation could be induced. To illustrate the spatial properties of this particular hidden Markov model, we prove the following proposition.

Proposition 4.1. *Denote* $\Pr(S_t = k \mid \theta) = g_k(t; \theta_{hid})$ *for* $k = 1, \ldots, r$, *noting that* $g_k(t; \theta_{hid}) = \delta_k$ *if the hidden Markov chain* $\{S_t : t = 1, \ldots, T\}$ *is irreducible, aperiodic and in its stationary distribution,* $\delta = (\delta_1, \ldots, \delta_r)$. *For any pair of sites* $(i,j) \in \{1, \ldots, n\}^2$, $i \neq j$, *at time* $t = 1, \ldots, T$, *we have*

$$Cov(D_t^i, D_t^j \mid \theta) = Cov_{S_t \mid \theta}(p_{iS_t}, p_{jS_t}) \tag{4.7}$$

and

$$Cov(W_t^i, W_t^j \mid \theta) = Cov_{S_t \mid \theta}\{E(W_t^i \mid S_t, \theta), E(W_t^j \mid S_t, \theta)\} \tag{4.8}$$

where $E(W_t^i \mid S_t = k, \theta) = m_{ik} p_{ik}$.

Proof. Beginning with equation (4.7), by definition

$$\mathrm{Cov}(D_t^i, D_t^j \mid \theta) = \mathrm{E}(D_t^i D_t^j \mid \theta) - \mathrm{E}(D_t^i \mid \theta)\mathrm{E}(D_t^j \mid \theta).$$

Using the Law of Total Expectation, and the conditional independence of D_t^i and D_t^j, given S_t, we have

$$\begin{aligned}\mathrm{Cov}(D_t^i, D_t^j \mid \theta) &= \sum_{k=1}^{r} \mathrm{E}(D_t^i \mid S_t = k, \theta)\mathrm{E}(D_t^j \mid S_t = k, \theta) g_k(t; \theta_{\mathrm{hid}}) \\ &\quad - \left\{\sum_{k=1}^{r} \mathrm{E}(D_t^i \mid S_t = k, \theta) g_k(t; \theta_{\mathrm{hid}})\right\}\left\{\sum_{k=1}^{r} \mathrm{E}(D_t^j \mid S_t = k, \theta) g_k(t; \theta_{\mathrm{hid}})\right\} \\ &= \mathrm{Cov}_{S_t|\theta}\{\mathrm{E}(D_t^i \mid S_t, \theta), \mathrm{E}(D_t^j \mid S_t, \theta)\} \\ &= \mathrm{Cov}_{S_t|\theta}(p_{iS_t}, p_{jS_t}).\end{aligned}$$

After a little algebra, note that this result can be written as

$$\begin{aligned}\mathrm{Cov}(D_t^i, D_t^j \mid \theta) &= \sum_{k=1}^{r} p_{ik}p_{jk} g_k(t; \theta_{\mathrm{hid}}) - \left\{\sum_{k=1}^{r} p_{ik} g_k(t; \theta_{\mathrm{hid}})\right\}\left\{\sum_{k=1}^{r} p_{jk} g_k(t; \theta_{\mathrm{hid}})\right\} \\ &= \sum_{k=1}^{r-1} \sum_{\ell=k+1}^{r} (p_{ik} - p_{i\ell})(p_{jk} - p_{j\ell}) g_k(t; \theta_{\mathrm{hid}}) g_\ell(t; \theta_{\mathrm{hid}}). \qquad (4.9)\end{aligned}$$

Next, since W_t^i and W_t^j are also conditionally independent, given S_t, it follows by analogy with the derivation above that

$$\mathrm{Cov}(W_t^i, W_t^j \mid \theta) = \mathrm{Cov}_{S_t|\theta}\{\mathrm{E}(W_t^i \mid S_t, \theta), \mathrm{E}(W_t^j \mid S_t, \theta)\},$$

where, using the Law of Total Expectation,

$$\mathrm{E}(W_t^i \mid S_t = k, \theta) = \mathrm{E}(W_t^i \mid D_t^i = 0, S_t = k, \theta)(1 - p_{ik}) + \mathrm{E}(W_t^i \mid D_t^i = 1, S_t = k, \theta)p_{ik} = m_{ik}p_{ik}.$$

□

Equation (4.7) shows that the covariance between rainfall occurrences at two sites is simply the covariance over weather states between the conditional probabilities of rain, given the state. Therefore, the weather states can induce *positive* correlation between D_t^i and D_t^j if the conditional probabilities of rain at sites i and j are similar to each other in most states. *Negative* correlation can be induced if most weather states correspond to very different conditional probabilities of rain at the two sites. This conclusion is borne out through equation (4.9) which shows that positive association will arise between D_t^i and D_t^j if either $p_{ik} > p_{i\ell}$ *and* $p_{jk} > p_{j\ell}$ or $p_{ik} < p_{i\ell}$ *and* $p_{jk} < p_{j\ell}$ for most pairs of states, k and ℓ. Conversely, negative association will arise if $(p_{ik} - p_{i\ell})$ and $(p_{jk} - p_{j\ell})$ broadly have opposite signs.

Similarly, from equation (4.8) it is clear that the covariance between rainfall amounts at two sites is simply the covariance over weather states between the conditional mean amounts, given the state. As a result, positive (negative) association will arise between W_t^i and W_t^j if the conditional mean rainfall amounts at sites i and j are similar (different) to each other in most states.

4.3 Prior distribution

Uncertainty about the unknown model parameters, *a priori*, is expressed through a prior distribution of the form

$$\pi(\theta) = \pi(\theta_{\text{hid}})\pi(\theta_{\text{obs}}) = \pi(\Lambda)\pi(\nu)\pi(\mathcal{P})\pi(\mathcal{M})\pi(\mathcal{V}), \tag{4.10}$$

where the density for ν would be omitted if S_1 was given a distribution parameterised by Λ, such as the stationary distribution of the chain. Implicit in equation (4.10), is an assumption of *a priori* independence, not only between the parameters of the observed and hidden processes, but also between the parameter blocks within each of these components of θ. We assume that the prior distribution (4.10) is exchangeable across weather states because we do not wish to distinguish between any of the weather states *a priori*.

The remainder of this section provides a description and justification of the particular prior distributions that we choose for each of $\mathcal{P}$, $\mathcal{M}$, $\mathcal{V}$, Λ and ν. The joint distributions that we choose are based on a set of independence assumptions which may not be truly representative of our prior beliefs. Therefore we also discuss priors which could be used to express belief in more complex relationships between parameters. The section concludes with details of how the hyperparameters can be chosen in order to incorporate our prior knowledge.

For simplicity we assume *a priori* independence between the rainfall probability parameters at each site and in each weather state so that

$$\pi(\mathcal{P}) = \prod_{i=1}^{n}\prod_{k=1}^{r}\pi(p_{ik}).$$

The rainfall occurrence probabilities in $\mathcal{P}$ are each defined on the interval $[0, 1]$ and so a suitable prior distribution for each p_{ik} is the beta distribution,

$$p_{ik} \sim \text{Beta}(a_{1ik}, a_{2ik}), \quad (i,k) \in \{1,\ldots,n\} \times \mathcal{S}_r, \tag{4.11}$$

for fixed hyperparameters a_{1ik} and a_{2ik}. This is a convenient choice because the beta distribution is conjugate to a likelihood function of Bernoulli form.

If it is known before seeing the data that a particular site has a tendency to be generally wet or generally dry, then the assumption of *a priori* independence across weather states may be brought into question. However, if such information is not available then an independence assumption seems reasonable because the probabilities $p_{i1},\ldots,p_{ir}$ at any site, i, are expected to correspond to weather states representing distinct precipitation conditions. Therefore knowledge that, say, p_{i1} was greater than its expected value would not affect our prior beliefs about the expectation of p_{ik} for $k \in \mathcal{S}_r \setminus \{1\}$. For a network of sites which are spatially well separated, the assumption of *a priori* independence across sites within the same weather state may be more reasonable than it would be for a dense network of sites. In the latter case we might believe that each weather state will represent broadly similar precipitation conditions at all sites so that learning say p_{1k} was greater than its expected value would lead to an upward revision of our beliefs about the expectation of p_{ik} for $i \in \{2,\ldots,n\}$. This would mean the probabilities $p_{1k},\ldots,p_{nk}$ for each $k \in \mathcal{S}_r$ are actually positively correlated in our prior beliefs. An example

of a prior which would allow $p_{1k}, \ldots, p_{nk}$ to be correlated *a priori* is the multivariate logit-normal distribution. Joe (1997) defines a random vector $\mathbf{P} = (P_1, \ldots, P_n)^T \in [0,1]^n$ as having a multivariate logit–normal distribution with parameters μ and Σ if

$$\left(\log\left(\frac{P_1}{1-P_1}\right), \ldots, \log\left(\frac{P_n}{1-P_n}\right)\right) \sim \mathrm{N}_n(\mu, \Sigma).$$

Clearly this distribution is not conjugate to a likelihood function of Bernoulli form. Although this is not prohibitive to its use, it makes MCMC more difficult and time consuming, as parameter updates require Metropolis Hastings steps. Moreover the moments of $\mathbf{P}$ are not available in any simple closed form (for the univariate case; see Johnson, 1949) which makes it difficult to elicit prior beliefs, particularly about the dependence between the components of the random vector $\mathbf{P}$. A simpler way of inducing *a priori* correlation, whilst retaining the computational convenience of (semi)-conjugacy, is to adopt a hierarchical beta prior such as

$$p_{ik} \mid p_k \sim \mathrm{Beta}\Big(p_k q_k, (1-p_k)q_k\Big), \quad p_k \sim \mathrm{Beta}(u_{1k}, u_{2k}), \quad (i,k) \in \{1, \ldots, n\} \times \mathcal{S}_r, \tag{4.12}$$

where u_{1k} and u_{2k} are fixed hyperparameters and q_k could either be fixed or given a distribution on $\mathbb{R}^+$. Marginalising over p_k (and q_k if it is not assumed fixed) in the joint distribution of $\Big(p_{1k}, \ldots, p_{nk}, p_k, (q_k)\Big)$ gives a distribution for $(p_{1k}, \ldots, p_{nk})$ on $[0,1]^n$ in which the p_{ik} are correlated.

The mean and coefficient of variation parameters in the gamma distributions for rainfall amounts on wet days will also be assumed independent across weather states and across sites so that

$$\pi(\mathcal{M}) = \prod_{i=1}^{n}\prod_{k=1}^{r} \pi(m_{ik}) \quad \text{and} \quad \pi(\mathcal{V}) = \prod_{i=1}^{n}\prod_{k=1}^{r} \pi(v_{ik}).$$

The validity of each assumption of *a priori* independence was questioned for the rainfall probability parameters and similar physical considerations cause us to question the validity of the assumptions here. Nevertheless, the use of more sophisticated priors which allow *a priori* dependence across sites and/or weather sites will be reserved for later chapters. The parameters in $\mathcal{M}$ and $\mathcal{V}$ are only constrained to lie on the positive real line so any distribution with support on $\mathbb{R}^+$ would be suitable for the individual m_{ik} and v_{ik}. We choose inverse gamma distributions for the elements of the matrix $\mathcal{M}$,

$$m_{ik} \sim \mathrm{IG}(b_{1ik}, b_{2ik}), \quad (i,k) \in \{1, \ldots, n\} \times \mathcal{S}_r, \tag{4.13}$$

and gamma distributions for the coefficient of variation parameters $\mathcal{V}$,

$$v_{ik} \sim \mathrm{Ga}(c_{1ik}, c_{2ik}), \quad (i,k) \in \{1, \ldots, n\} \times \mathcal{S}_r. \tag{4.14}$$

The inverse gamma distribution is chosen for each mean rainfall amount parameter because it is semi–conjugate; see Section 4.5. Choosing the semi–conjugate prior therefore has the computational benefit of allowing the mean parameters to be updated in simple Gibbs steps, without recourse to Metropolis Hastings schemes. There is no conjugate or semi–conjugate prior for the elements of the matrix $\mathcal{V}$, and other distributions such as the lognormal would provide alternatives. In order to represent *a priori* belief in dependence amongst the m_{ik} or v_{ik}, suitable

priors would be the multivariate lognormal distribution with non-diagonal variance matrix, or a hierarchical prior formulated in a similar way to (4.12).

The hyperparameters $\{a_{1ik}, a_{2ik}, b_{1ik}, b_{2ik}, c_{1ik}, c_{2ik} : i = 1, \ldots, n, k \in \mathcal{S}_r\}$ are chosen to reflect prior beliefs concerning the rainfall occurrence and amount processes. This will be discussed in Sections 4.3.1 and 4.3.2.

Possible priors for the transition matrix $\mathbf{\Lambda}$ in a hidden Markov model were discussed in Section 3.3.2. In spite of its rather inflexible dependence structure, we find that our prior beliefs about transitions between weather states can be adequately captured by adopting independent Dirichlet distributions for the rows of $\mathbf{\Lambda}$,

$$\boldsymbol{\lambda}_j \sim \mathscr{D}_r(E_j \mathbf{e}_j), \quad j \in \mathcal{S}_r, \tag{4.15}$$

where $\mathrm{E}(\boldsymbol{\lambda}_j) = \mathbf{e}_j$. Here $\mathbf{e}_j \in \mathscr{S}_r$ and $E_j \in \mathbb{R}^+$ are fixed hyperparameters which are chosen to reflect prior beliefs about the mean sojourn time for any particular weather state. This will be formalised in Section 4.3.3.

Following the discussion in Section 3.2.1 regarding the choice of initial distribution, depending on the particularities and inferential objectives of analyses in subsequent sections within this chapter, we will consider two possibilities for the initial distribution: (i) $\boldsymbol{\nu}$ is a fixed probability distribution and (ii) $\boldsymbol{\nu}$ is variable and assigned a conjugate Dirichlet prior, parameterised by its own hyperparameters, $G \in \mathbb{R}^+$ and $\mathbf{g} \in \mathscr{S}_r$.

4.3.1 Prior beliefs about the probabilities of rainfall

The Beta(a_1, a_2) distribution has mean $a_1/(a_1 + a_2)$ and variance

$$\frac{a_1 a_2}{(a_1 + a_2)^2 (a_1 + a_2 + 1)}.$$

Specifying a mean and a variance gives a pair of simultaneous equations which can be solved for a_1 and a_2. In practice, it might be easier to express our prior beliefs, or at least the degree of belief in our prior point estimate, using the *equivalent prior sample* approach (Garthwaite *et al.*, 2005). For Bernoulli data/beta prior combinations, this involves regarding the prior as containing information equivalent to a hypothetical prior sample of length $(a_1 + a_2)$ days in which there were a_1 wet days and a_2 dry days.

The prior for $\mathcal{P}$ is exchangeable across weather states which means that for each site, $i = 1, 2, \ldots, n$, $a_{1ik} = a_{1i}$ and $a_{2ik} = a_{2i}$ for all $k \in \mathcal{S}_r$. As such the only sensible choice for our prior point estimate at site i, for example, the mean, would seem to be one which reflects our beliefs about the probability of rain on a "typical" day at that site. Based on the equivalent prior sample approach, for each site, i, the prior specification for a particular weather state has an equivalent length of $(a_{1i} + a_{2i})$ days so the total information content of the prior specification is $r(a_{1i} + a_{2i})$ days. The information content should be chosen to be small, that is, large prior variances should be selected to represent our prior belief that any particular weather state might be associated with extremely wet (i.e. $p_{ik} \sim 1$) or extremely dry (i.e. $p_{ik} \sim 0$) conditions at site i, rather than the "typical" conditions quantified by the prior point estimate.

4.3.2 Prior beliefs about the mean and coefficient of variation for non-zero rainfall amounts

As there is no fully conjugate prior for the parameters in $\mathcal{M}$ or $\mathcal{V}$, prior opinion cannot be assessed using the equivalent prior sample approach. Armed with an understanding of the meaning of these parameters, we can choose the hyperparameters in their prior distributions by thinking about the median and another percentile such as the lower or upper 5% point of each distribution rather than the mean and variance. This is the so-called *quantile method* of eliciting priors in a parametric distribution (Garthwaite *et al.*, 2005). Although the quantiles of the gamma and inverse gamma distributions are not available in closed form, computer languages such as R (R Development Core Team, 2008) can easily be used to solve numerically the appropriate pair of simultaneous equations for the hyperparameters. Again, the only sensible choice for the medians of these priors would seem to be values regarded as representative of the mean and coefficient of variation in the rainfall distribution on a "typical" wet day at the site in question. Recalling that *a priori* exchangeability is assumed across weather states, the choices of the lower or upper 5% points for each site should take into consideration the fact that a weather state might represent an atypical or extreme kind of precipitation climate at that site.

4.3.3 Prior beliefs about the weather states

Our decision to make the prior distribution $\pi(\boldsymbol{\Lambda})$ invariant under permutations of the weather state labels demands that we take the information content parameter E_j to be the same for all rows of $\boldsymbol{\Lambda}$, i.e. $E_j = E$ for all $j \in \mathcal{S}_r$, and the mean hyperparameter $\mathbf{e}_j = (e_{j1}, \ldots, e_{jr})$ to contain elements $e_{jj} = \alpha \in [0, 1]$ and $e_{jk} = (1-\alpha)/(r-1)$ if $j \neq k$. To complete the prior specification we need only select the values of two hyperparameters, α and E, which can be chosen to reflect prior beliefs about the mean sojourn time in any particular weather state; see, for example, Boys & Henderson (2004) for an analogous elicitation strategy in the context of DNA sequence segmentation.

Given $\boldsymbol{\Lambda}$, the sojourn time in weather state $j \in \mathcal{S}_r$ follows a geometric distribution with parameter λ_{jj}, so the expected sojourn time in state j is $1/(1-\lambda_{jj})$. Since the marginal prior distribution for λ_{jj} is $\text{Beta}\{E\alpha, E(1-\alpha)\}$ it follows that the prior induced for $1/(1-\lambda_{jj})$ has mean and variance

$$\text{E}\left(\frac{1}{1-\lambda_{jj}}\right) = \frac{E-1}{E(1-\alpha)-1}, \quad \text{Var}\left(\frac{1}{1-\lambda_{jj}}\right) = \frac{(E-1)E\alpha}{\{E(1-\alpha)-1\}^2\{E(1-\alpha)-2\}}.$$

Specifying a value for the mean, ℓ, and for the variance, c, gives the solution

$$\alpha = (\ell-1)\left\{\frac{\ell(\ell-1)+c}{\ell^2(\ell-1)+c(\ell+1)}\right\} \quad \text{and} \quad E = \frac{1}{c}\left\{\ell^2(\ell-1)+c(\ell+1)\right\}.$$

Alternatively we can think about the mean of $1/(1-\lambda_{jj})$ and the equivalent prior sample size, also known as the information content. In the same way that a $\text{Beta}(a_1, a_2)$ prior can be regarded as containing information equivalent to a hypothetical (Bernoulli) prior sample of size $a_1 + a_2$, we can think of a Dirichlet $\mathscr{D}_r(A\tilde{\mathbf{a}})$ prior as containing information equivalent to a hypothetical multinomial prior sample of size A. In $\pi(\boldsymbol{\Lambda})$ each row is equivalent to E transitions giving the

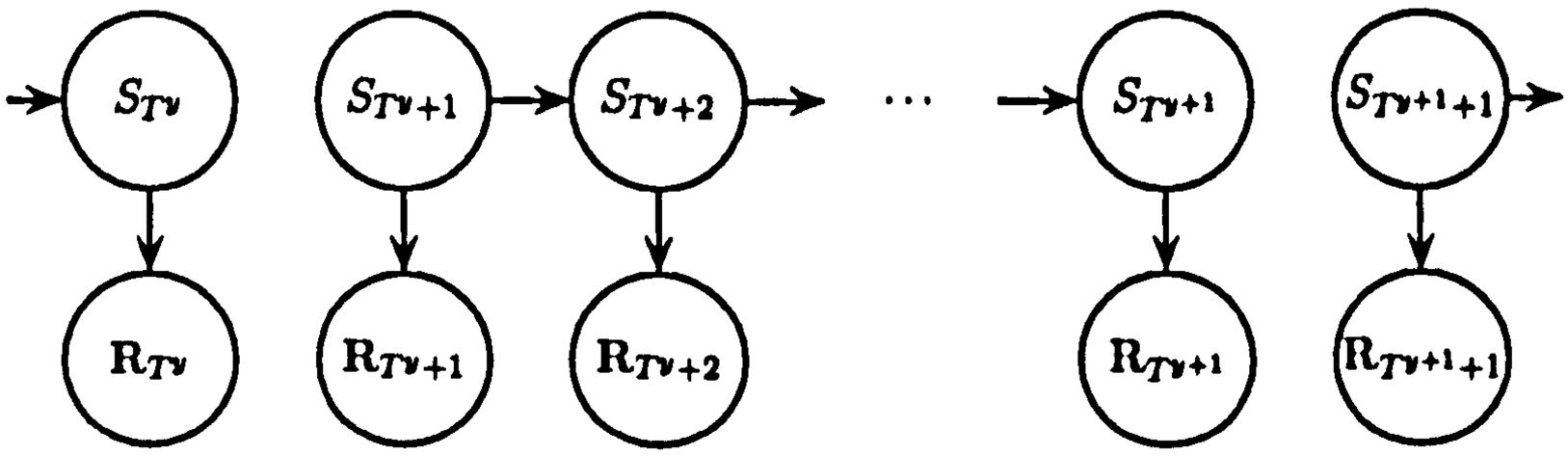


Figure 4.2: A DAG showing the (temporal) dependence structure in independent winter segments.

overall prior specification an equivalent weather state sequence length of $n_\lambda = rE + 1$ days, where the plus one accounts for the existence of n transitions in a sequence of length $n + 1$. Therefore, we can specify a value for n_λ (and hence E) and use this along with the expression for the mean, $\mathrm{E}\{1/(1 - \lambda_{jj})\} = \ell$, to give the solution

$$\alpha = \frac{(E-1)(\ell-1)}{E\ell}.$$

A further alternative would be to fix the values for α and E by thinking about two percentiles in the distribution of $1/(1 - \lambda_{jj})$, for example the median and the lower (or upper) 5% point. From this we can obtain the corresponding percentiles in the distribution of λ_{jj} and then use a computer language such as R to solve the resulting system of two equations numerically for the hyperparameters.

As discussed in Section 3.3.2, the assumption of *a priori* exchangeability across weather states requires that the initial distribution is made equal to the discrete uniform distribution on $\mathcal{S}_r$ if ν is fixed, or that the expectation of ν is $\mathbf{g} = (1/r, \ldots, 1/r)$ if ν is variable and assigned a Dirichlet $\mathscr{D}(G\mathbf{g})$ prior. The information content parameter G can be chosen by making a judgement about the equivalent prior sample size.

4.4 Likelihood

In applications involving hidden Markov models, it is usually assumed that the data comprise one long time series of observations made over consecutive units of time. The observed dataset which we will consider, however, divides naturally into Y sub–series which, together with the weather states, can be modelled as Y independent realisations of the same hidden Markov model. Let the number of days in each sub–series be T_y, $y = 1, \ldots, Y$, and denote the partial sums $T^j = \sum_{y=1}^{j-1} T_y$ with the convention that $T^1 = \sum_{y=1}^{0} T_y = 0$, and clearly $T^{Y+1} = \sum_{y=1}^{Y} T_y = T$. Now we could choose to represent the temporal dependence in the model by the DAG in Figure 4.2, where $\mathbf{R}_t^T = (\mathbf{W}_t^T, \mathbf{D}_t^T)$. To put this into context, for the dataset which we will consider, the y–th sub–series corresponds to the y–th winter (December–February) period so Y will be the number of winters and T_y will be the number of days in the y–th winter. Subsequent analysis will assume that each $\{(\mathbf{w}_{T^y+1}, \mathbf{d}_{T^y+1}, s_{T^y+1}), \ldots, (\mathbf{w}_{T^y+1}, \mathbf{d}_{T^y+1}, s_{T^y+1})\}$ for $y = 1, \ldots, Y$ is an

independent realisation of the hidden Markov model defined by assumptions A1 and A2, with the precipitation process parameterised according to (4.2)–(4.4). Of course a more conventional analysis can be recovered by taking $Y = 1$.

Posterior inference will be via MCMC using data augmentation and so derivation of the full conditional distributions for the model parameters will require the complete data likelihood

$$p(\mathbf{w}, \mathbf{d}, \mathbf{s} \mid \theta) = p(\mathbf{w}, \mathbf{d} \mid \mathbf{s}, \theta_{\text{obs}}) p(\mathbf{s} \mid \theta_{\text{hid}}) \tag{4.16}$$

where

$$\begin{aligned} p(\mathbf{w}, \mathbf{d} \mid \mathbf{s}, \theta_{\text{obs}}) &= \prod_{t=1}^{T} p(\mathbf{w}_t, \mathbf{d}_t \mid S_t = s_t, \theta_{\text{obs}}) \\ &= \prod_{k=1}^{r} \prod_{\{t:s_t=k\}} p(\mathbf{w}_t, \mathbf{d}_t \mid S_t = k, \theta_{\text{obs}}) \\ &= \prod_{k=1}^{r} \prod_{i=1}^{n} p_{ik}^{T_{ik}^1(\mathbf{s})} (1 - p_{ik})^{T_{ik}^0(\mathbf{s})} \times \prod_{k=1}^{r} \prod_{i=1}^{n} \prod_{\substack{\{t:s_t=k,\\ d_t^i=1\}}} \text{Ga}\left(w_t^i \middle| \frac{1}{v_{ik}^2}, \frac{1}{v_{ik}^2 m_{ik}}\right), \end{aligned} \tag{4.17}$$

and

$$\begin{aligned} p(\mathbf{s} \mid \theta_{\text{hid}}) &= \prod_{y=1}^{Y} p(\mathbf{s}_{T^{y}+1:T^{y+1}} \mid \theta_{\text{hid}}) \\ &= \prod_{y=1}^{Y} \left\{ \Pr(S_{T^{y}+1} = s_{T^{y}+1} \mid \nu) \prod_{t=T^{y}+2}^{T^{y+1}} \Pr(S_t = s_t \mid S_{t-1} = s_{t-1}, \Lambda) \right\} \\ &= \prod_{j=1}^{r} \nu_j^{m_j(\mathbf{s})} \times \prod_{j=1}^{r} \prod_{k=1}^{r} \lambda_{jk}^{n_{jk}(\mathbf{s})}, \end{aligned} \tag{4.18}$$

in which

$$\begin{aligned} T_{ik}^{d}(\mathbf{s}) &= \sum_{t=1}^{T} \mathbb{I}(d_t^i = d, s_t = k), \quad m_j(\mathbf{s}) = \sum_{y=1}^{Y} \mathbb{I}(s_{T^{y}+1} = j), \\ n_{jk}(\mathbf{s}) &= \sum_{y=1}^{Y} \sum_{t=T^{y}+2}^{T^{y+1}} \mathbb{I}(s_{t-1} = j, s_t = k) \end{aligned} \tag{4.19}$$

denote the relevant counts and $\mathbb{I}(x)$ is the indicator function.

Note that in the case of a fixed initial distribution for each sub-series, the terms involving ν can be absorbed into a constant of proportionality and omitted from the expression (4.18).

4.5 Posterior inference via MCMC

In Section 3.3.4 we discussed three MCMC techniques for generating (dependent) samples from the posterior distribution of the model parameters in a hidden Markov model, here given by

$$\pi(\theta \mid \mathbf{w}, \mathbf{d}) \propto \pi(\theta) p(\mathbf{w}, \mathbf{d} \mid \theta).$$

These were MCMC with data augmentation, where the state space of the sampler is augmented to include the hidden states (s), and two marginal updating schemes based on Metropolis Hastings updates, where the state space of the sampler is either θ alone or s alone. If the third technique is used, then a sample from the posterior for θ can be obtained from the output $\mathbf{s}^{[j]}$ by drawing one–for–one from the conditional posterior, $\pi(\theta \mid \mathbf{w}, \mathbf{d}, \mathbf{s})$. However, for the hidden Markov model studied in this chapter, this marginal updating scheme can immediately be discarded because the priors for the elements in $\mathcal{M}$ and $\mathcal{V}$ are not fully conjugate and so the parameters cannot be analytically integrated out of the joint posterior distribution $\pi(\theta, \mathbf{s} \mid \mathbf{w}, \mathbf{d})$. In producing an efficient sampler that mixes well, the success of the other marginal updating approach relies on judicious partitioning (or blocking) of θ and laborious tuning of the Metropolis Hastings algorithm. This becomes increasingly difficult as the dimension of θ increases. The main drawback with data augmentation is that mixing over an additional layer can cause convergence difficulties and require longer MCMC runs to obtain a particular number of approximately uncorrelated posterior samples. This problem is particularly acute if there is strong dependence between s and θ, when the sampler can become stuck in local modes. However, unlike the marginal updating scheme for θ, MCMC updating with data augmentation readily scales up to handle multivariate within–state distributions. For a given value of r, the dimension of θ_{obs} increases linearly with the number of sites n, and so, except for small networks, comprising only a few sites, the dimension of the parameter space will be large. Therefore we use MCMC with data augmentation because, compared with marginal updating of θ, it seems better suited to handle the complexity of the within weather state distributions for rainfall occurrence and amount.

Gibbs sampling with data augmentation was outlined generically in Algorithm 3.3.2 and applies as stated because we have assumed *a priori* independence between θ_{obs} and θ_{hid}. Step 1 involves drawing θ from $\pi(\theta \mid \mathbf{s}, \mathbf{w}, \mathbf{d})$ and is described in the following section. Step 2 is the data augmentation step in which s is simulated from $\pi(\mathbf{s} \mid \theta, \mathbf{w}, \mathbf{d})$. We use the generic forward backward scheme detailed in Algorithm 3.3.3, in which we can make use of the simplifications facilitated by assumptions A1 and A2. Since we assume that each sub–series of observations and weather states is an independent realisation from the same hidden Markov model, we can apply the forward backward algorithm separately to each sub–series. Clearly the overall observed data likelihood can then be calculated as the product of the observed data likelihoods for each sub–series.

4.5.1 Sampling from the complete data posterior distribution $\pi(\theta \mid \mathbf{s}, \mathbf{w}, \mathbf{d})$

The complete data posterior distribution is given by Bayes Theorem as

$$\pi(\theta \mid \mathbf{s}, \mathbf{w}, \mathbf{d}) \propto \pi(\theta) p(\mathbf{w}, \mathbf{d}, \mathbf{s} \mid \theta)$$

which can be decomposed as

$$\pi(\theta \mid \mathbf{s}, \mathbf{w}, \mathbf{d}) = \pi(\theta_{\text{obs}} \mid \mathbf{s}, \mathbf{w}, \mathbf{d}) \pi(\theta_{\text{hid}} \mid \mathbf{s})$$

where

$$\pi(\theta_{\text{obs}} \mid \mathbf{s}, \mathbf{w}, \mathbf{d}) \propto \pi(\theta_{\text{obs}}) p(\mathbf{w}, \mathbf{d} \mid \mathbf{s}, \theta_{\text{obs}}) \quad \text{and} \quad \pi(\theta_{\text{hid}} \mid \mathbf{s}) \propto \pi(\theta_{\text{hid}}) p(\mathbf{s} \mid \theta_{\text{hid}}). \tag{4.20}$$

Step 1(a) of Algorithm 3.3.2 involves drawing θ_{hid} from $\pi(\theta_{\text{hid}} \mid \mathbf{s})$. Combining the prior for Λ, (4.15), with the relevant part of the complete data likelihood, (4.18), yields

$$\pi(\Lambda \mid \mathbf{s}) \propto \prod_{j=1}^{r}\prod_{k=1}^{r} \lambda_{jk}^{Ee_{jk}-1} \times \prod_{j=1}^{r}\prod_{k=1}^{r} \lambda_{jk}^{n_{jk}(\mathbf{s})} = \prod_{j=1}^{r}\prod_{k=1}^{r} \lambda_{jk}^{Ee_{jk}+n_{jk}(\mathbf{s})-1},$$

so letting $\mathbf{n}_j(\mathbf{s}) = \Big(n_{j1}(\mathbf{s}), \ldots, n_{jr}(\mathbf{s})\Big)$, we recognise

$$\boldsymbol{\lambda}_j = (\lambda_{j1}, \ldots, \lambda_{jr}) \mid \mathbf{s} \sim \mathscr{D}_r\Big(E\mathbf{c}_j + \mathbf{n}_j(\mathbf{s})\Big) \qquad \text{independently for } j \in \mathcal{S}_r,$$

which can be sampled directly. Similarly if $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_r)$ is variable, assumed independent of Λ *a priori* and assigned the conjugate Dirichlet $\mathscr{D}_r(G\mathbf{g})$ prior then the posterior will be of the same form,

$$\boldsymbol{\nu} \mid \mathbf{s} \equiv \boldsymbol{\nu} \mid s_1 \sim \mathscr{D}_r\Big(G\mathbf{g} + \mathbf{m}(\mathbf{s})\Big), \tag{4.21}$$

where $\mathbf{m}(\mathbf{s}) = \Big(m_1(\mathbf{s}), \ldots, m_r(\mathbf{s})\Big)$, which again can be sampled directly.

Next, step 1(b) of Algorithm 3.3.2 involves sampling θ_{obs} from $\pi(\theta_{\text{obs}} \mid \mathbf{s}, \mathbf{w}, \mathbf{d})$. Starting with the first equation in (4.20) we can then deduce the kernel of the full conditional distribution for any component of θ_{obs} by simply dropping factors independent of that component.

For the rainfall probability parameters, combining the relevant part of the complete data likelihood, (4.17), with the prior, (4.11), is straightforward owing to the conjugacy of the beta distribution to a Bernoulli form of the likelihood expression, and leads to a full conditional distribution for $\mathcal{P}$ such that $p_{11}, \ldots, p_{1r}, p_{21}, \ldots, p_{nr}$ are conditionally independent and

$$p_{ik} \mid \cdots \sim \text{Beta}(a_{1i} + T_{ik}^1(\mathbf{s}), a_{2i} + T_{ik}^0(\mathbf{s})), \qquad (i,k) \in \{1, \ldots, n\} \times \mathcal{S}_r.$$

The notation "$\mid \cdots$" is used to represent conditioning on all other variables, that is, the other model parameters, the weather states $\mathbf{s}$ and the observed data $(\mathbf{w}, \mathbf{d})$, although in this case it is clear that p_{ik} is conditionally independent of $\theta \setminus \{p_{ik}\}$ given $(\mathbf{s}, \mathbf{w}, \mathbf{d})$.

For the mean rainfall amount parameters $\mathcal{M}$, the assumption of *a priori* independence across sites and weather states, along with the semi-conjugacy of the inverse gamma distribution to the so-parameterised gamma form of the likelihood in $p(\mathbf{w}, \mathbf{d} \mid \mathbf{s}, \theta_{\text{obs}})$ leads to a full conditional distribution for $\mathcal{M}$ in which $m_{11}, \ldots, m_{1r}, m_{21}, \ldots, m_{nr}$ are conditionally independent and

$$m_{ik} \mid \cdots \sim \text{IG}\left(b_{1i} + \frac{T_{ik}^1(\mathbf{s})}{v_{ik}^2}, b_{2i} + \frac{T_{ik}^1(\mathbf{s})\bar{w}_{ik}(\mathbf{s})}{v_{ik}^2}\right), \qquad (i,k) \in \{1, \ldots, n\} \times \mathcal{S}_r, \tag{4.22}$$

where

$$\bar{w}_{ik}(\mathbf{s}) = \frac{1}{T_{ik}^1(\mathbf{s})} \sum_{\substack{\{t : s_t = k,\\ d_t^i = 1\}}} w_t^i \tag{4.23}$$

is the arithmetic mean of the rainfall amounts on wet days at site i in weather state k.

Denoting by

$$\bar{w}_{g,ik}(\mathbf{s}) = \left(\prod_{\substack{\{t:s_t=k,\\ d_t^i=1\}}} w_t^i \right)^{1/T_{ik}^1(\mathbf{s})}$$

the geometric mean of the rainfall amounts on wet days at site i in weather state k, the full conditional density for $\mathcal{V}$ is given up to proportionality as

$$\begin{aligned}
\pi(\mathcal{V} \mid \cdots) &\propto \prod_{k=1}^{r}\prod_{i=1}^{n} \mathrm{Ga}(v_{ik} \mid c_{1i}, c_{2i}) \times \prod_{k=1}^{r}\prod_{i=1}^{n} \prod_{\substack{\{t:s_t=k,\\ d_t^i=1\}}} \mathrm{Ga}\left(w_t^i \middle| \frac{1}{v_{ik}^2}, \frac{1}{v_{ik}^2 m_{ik}}\right) \\
&\propto \prod_{k=1}^{r}\prod_{i=1}^{n} \Gamma\left(\frac{1}{v_{ik}^2}\right)^{-T_{ik}^1(\mathbf{s})} v_{ik}^{-\{2T_{ik}^1(\mathbf{s})/v_{ik}^2 - c_{1i}+1\}} m_{ik}^{-T_{ik}^1(\mathbf{s})/v_{ik}^2} \bar{w}_{g,ik}(\mathbf{s})^{T_{ik}^1(\mathbf{s})/v_{ik}^2} \\
&\quad \times \exp\left\{ -\left(c_{2i} v_{ik} + \frac{T_{ik}^1(\mathbf{s}) \bar{w}_{ik}(\mathbf{s})}{v_{ik}^2 m_{ik}} \right) \right\}
\end{aligned}$$

from which we can deduce that the coefficient of variation parameters are conditionally independent across sites and weather states, but their densities are not of standard form, being proportional to

$$\begin{aligned}
\pi(v_{ik} \mid \cdots) \propto \Gamma\left(\frac{1}{v_{ik}^2}\right)^{-T_{ik}^1(\mathbf{s})} & v_{ik}^{-\{2T_{ik}^1(\mathbf{s})/v_{ik}^2 - c_{1i}+1\}} m_{ik}^{-T_{ik}^1(\mathbf{s})/v_{ik}^2} \bar{w}_{g,ik}(\mathbf{s})^{T_{ik}^1(\mathbf{s})/v_{ik}^2} \\
& \times \exp\left\{ -\left(c_{2i} v_{ik} + \frac{T_{ik}^1(\mathbf{s}) \bar{w}_{ik}(\mathbf{s})}{v_{ik}^2 m_{ik}} \right) \right\} \qquad (4.24)
\end{aligned}$$

for $(i,k) \in \{1,2,\ldots,n\} \times \mathcal{S}_r$.

Sampling from the full conditional distributions of $\mathcal{P}$ and $\mathcal{M}$ is standard. In order to sample from the full conditional distribution of each v_{ik}, we introduce a Metropolis Hastings step. Specifically, at each iteration we use a random walk on the gamma scale by generating a proposal, v_{ik}^*, from a gamma distribution whose mean is equal to the current value, v_{ik},

$$v_{ik}^* \mid v_{ik} \sim q(v_{ik}, v_{ik}^*) \equiv \mathrm{Ga}\left(\omega_v^i, \frac{\omega_v^i}{v_{ik}}\right).$$

The term $\omega_v^i \in \mathbb{R}^+$ is a tuning parameter for site i which can be adjusted to control the acceptance rate. This proposal distribution has coefficient of variation $(\omega_v^i)^{-1/2}$ so increasing ω_v^i reduces the coefficient of variation and encourages more moves to be accepted. The acceptance probability of the proposed move is given by

$$\begin{aligned}
\alpha(v_{ik}, v_{ik}^*) &= \min\left\{1, \frac{\pi(v_{ik}^* \mid \cdots) q(v_{ik}^*, v_{ik})}{\pi(v_{ik} \mid \cdots) q(v_{ik}, v_{ik}^*)}\right\} \\
&= \min\{1, A\} \qquad (4.25)
\end{aligned}$$

where

$$A = \frac{(v_{ik}^*)^{-\{2T_{ik}^1(\mathbf{s})/(v_{ik}^*)^2 - c_{1i} + 2\omega_v^i\}} \exp\left[-\left\{c_{2i}v_{ik}^* + \frac{T_{ik}^1(\mathbf{s})\bar{w}_{ik}(\mathbf{s})}{(v_{ik}^*)^2 m_{ik}} + \frac{\omega_v^i v_{ik}}{v_{ik}^*}\right\}\right]}{(v_{ik})^{-\{2T_{ik}^1(\mathbf{s})/(v_{ik})^2 - c_{1i} + 2\omega_v^i\}} \exp\left[-\left\{c_{2i}v_{ik} + \frac{T_{ik}^1(\mathbf{s})\bar{w}_{ik}(\mathbf{s})}{(v_{ik})^2 m_{ik}} + \frac{\omega_v^i v_{ik}^*}{v_{ik}}\right\}\right]}$$
$$\times \frac{\Gamma\left\{\frac{1}{(v_{ik}^*)^2}\right\}^{-T_{ik}^1(\mathbf{s})} \left(\frac{\bar{w}_{g,ik}(\mathbf{s})}{m_{ik}}\right)^{T_{ik}^1(\mathbf{s})/(v_{ik}^*)^2}}{\Gamma\left\{\frac{1}{(v_{ik})^2}\right\}^{-T_{ik}^1(\mathbf{s})} \left(\frac{\bar{w}_{g,ik}(\mathbf{s})}{m_{ik}}\right)^{T_{ik}^1(\mathbf{s})/(v_{ik})^2}}. \tag{4.26}$$

The ratio, A, simplifies, and might be better expressed as

$$\log A = c_{2i}(v_{ik} - v_{ik}^*) + T_{ik}^1(\mathbf{s})\left\{\frac{\bar{w}_{ik}(\mathbf{s})}{m_{ik}} - \log\left(\frac{\bar{w}_{g,ik}(\mathbf{s})}{m_{ik}}\right)\right\}\left\{\frac{1}{(v_{ik})^2} - \frac{1}{(v_{ik}^*)^2}\right\}$$
$$+ (2\omega_v^i - c_{1i})(\log v_{ik} - \log v_{ik}^*) + T_{ik}^1(\mathbf{s})\left[\log\Gamma\left\{\frac{1}{(v_{ik})^2}\right\} - \log\Gamma\left\{\frac{1}{(v_{ik}^*)^2}\right\}\right]$$
$$+ 2T_{ik}^1(\mathbf{s})\left\{\frac{1}{(v_{ik})^2}\log v_{ik} - \frac{1}{(v_{ik}^*)^2}\log v_{ik}^*\right\} + \omega_v^i\left(\frac{v_{ik}^*}{v_{ik}} - \frac{v_{ik}}{v_{ik}^*}\right).$$

4.5.2 Missing data

The Yorkshire dataset that we analyse in Section 4.7 contains some missing values. In the exploratory data analysis in Section 2.3 we examined the periods of missing data and in each case concluded that, given the observations at other sites, it seemed plausible to assume the missingness was not related to the amount of precipitation during the period. Therefore it seems reasonable to assume the missing data are *missing at random*. We further assume *a priori* independence between the parameters of the missing data mechanism and the model parameters. Combining these two assumptions, we take the missing data mechanism to be ignorable and analytically marginalise over the missing values as described in Section 3.3.6.

4.5.3 MCMC scheme

Assuming ν to be variable and independent of Λ, the MCMC scheme can proceed as follows. We initialise the algorithm with a sequence of weather states $\mathbf{s}^{[0]}$ obtained by inputting starting values for the model parameters $\theta^{[0]}$ and applying the forward–backward algorithm as detailed below. Then at each iteration $\ell = 1, 2, \ldots$ we perform a fixed sweep of the following steps:

1. Simulate $\theta^{[\ell]}$ from $\pi(\theta \mid \mathbf{s}^{[\ell-1]}, \mathbf{w}, \mathbf{d})$:
   (a) Simulate θ_{hid} from $\pi(\theta_{\text{hid}} \mid \mathbf{s}^{[\ell-1]})$:
      (i) Simulate $\lambda_j \mid \cdots \sim \mathscr{D}_r\{E_j\mathbf{e}_j + \mathbf{n}_j(\mathbf{s}^{[\ell-1]})\}$ for each $j \in \mathcal{S}_r$.
      (ii) Simulate $\nu \mid \cdots \sim \mathscr{D}_r\{G\mathbf{g} + \mathbf{m}(\mathbf{s}^{[\ell-1]})\}$.
   (b) Simulate θ_{obs} from $\pi(\theta_{\text{obs}} \mid \mathbf{s}^{[\ell-1]}, \mathbf{w}, \mathbf{d})$ by successively passing through the following Gibbs (or Metropolis–within–Gibbs) steps:

(i) Simulate $p_{ik} \mid \cdots \sim \text{Beta}\{a_{1i} + T^1_{ik}(\mathbf{s}^{[\ell-1]}), a_{2i} + T^0_{ik}(\mathbf{s}^{[\ell-1]})\}$ for each pair $(i,k) \in \{1,2,\ldots,n\} \times \mathcal{S}_r$.

(ii) Simulate $m_{ik} \mid \cdots \sim \text{IG}\left\{b_{1i} + \frac{T^1_{ik}(\mathbf{s}^{[\ell-1]})}{v^2_{ik}}, b_{2i} + \frac{T^1_{ik}(\mathbf{s}^{[\ell-1]})\bar{w}_{ik}(\mathbf{s}^{[\ell-1]})}{v^2_{ik}}\right\}$ for each pair $(i,k) \in \{1,2,\ldots,n\} \times \mathcal{S}_r$.

(iii) Perform Metropolis Hastings updates of v_{ik} for each pair $(i,k) \in \{1,2,\ldots,n\} \times \mathcal{S}_r$:

I. Generate a proposal value

$$v^*_{ik} \mid v^{[\ell-1]}_{ik} \sim q(v^{[\ell-1]}_{ik}, v^*_{ik}) \equiv \text{Ga}\left(\omega^i_v, \frac{\omega^i_v}{v^{[\ell-1]}_{ik}}\right).$$

II. Evaluate the acceptance probability of the proposed move, $\alpha(v^{[\ell-1]}_{ik}, v^*_{ik})$, as defined in equations (4.25) and (4.26).

III. Set $v^{[\ell]}_{ik} = v^*_{ik}$ with probability $\alpha(v^{[\ell-1]}_{ik}, v^*_{ik})$ and set $v^{[\ell]}_{ik} = v^{[\ell-1]}_{ik}$ otherwise.

2. Simulate $\mathbf{s}^{[\ell]}$ from $\pi(\mathbf{s} \mid \theta^{[\ell]}, \mathbf{w}, \mathbf{d})$ by applying the forward backward scheme described in Algorithm 3.3.3 separately to each sub–series.

Note that if ν is fixed we simply omit step 1(a)(ii).

4.6 Estimating the marginal likelihood: simulation experiment

In Sections 3.4 and 3.5 we considered Bayesian inference for hidden Markov models when the number of hidden states, r, is treated as an unknown random quantity. We focused primarily on within model simulation techniques and discussed a variety of methods for approximating the marginal likelihood. This section contains a simulation experiment in which the performance of several of these marginal likelihood estimators are compared.

In the remainder of this chapter, r is regarded as a random variable and so, as in Sections 3.4 and 3.5, we distinguish between the parameters and hyperparameters associated with the different hidden Markov models by introducing notational dependence on r. This takes the form of the first subscript in both the model parameters and the hyperparameters in their prior distributions. For example, we denote the parameters of the observed process in an r–state hidden Markov model by $\theta_{r,\text{obs}} = (\mathcal{P}_r, \mathcal{M}_r, \mathcal{V}_r)$ and the hyperparameters in, say, the prior for $(p_{r,ik} \mid r)$ by $a_{r,1i}$ and $a_{r,2i}$.

4.6.1 Background to simulation experiment

Sections 3.5.1.1–3.5.1.4 contained a brief outline of various methods for estimating the marginal likelihood. This was followed in Section 3.5.1.5 by a comparison between these estimators, with emphasis on their application in hidden Markov models. Of the available methods, the Laplace approximation can immediately be discounted because of concerns related to the validity of the approximation in overfitting models. As the prior distribution in our hidden Markov

model for precipitation is not fully conjugate we cannot analytically marginalise the model parameters out of the joint posterior density $\pi(\theta_r, \mathbf{s} \mid \mathbf{w}, \mathbf{d}, r)$ and so the methods based on the marginal posterior for the hidden states are not applicable. Likewise, since the full conditional distributions for the components of $\mathcal{V}$ do not have known normalising constants we cannot construct an importance sampling density using the Rao–Blackwellisation approach suggested by Frühwirth-Schnatter (2004). By the same argument Chib's original method (Chib, 1995) is not applicable although the extension (Chib & Jeliazkov, 2001) which overcomes the problem of intractable full conditional densities is a possibility. However, it is not an attractive option because MCMC updating for our precipitation model involves partitioning θ_r into a number of blocks, nr of which (the v_{ik}) are updated singly in Metropolis Hastings steps. Estimating the marginal likelihood by Chib's extended method would therefore require several reduced MCMC runs and a substantial amount of bookkeeping, which we deem computationally infeasible. This eliminates all the methods considered except the Monte Carlo and harmonic mean estimators, the Newton Raftery hybrid estimator and its approximation, the bridge sampling estimator with importance sampling density equal to the prior, and the estimator based on the power posterior approach. The simulation experiments which follow provides a numerical comparison between these estimators. We consider simplified versions of the hidden Markov model for precipitation in order to allow exact calculation of the marginal likelihood. A consequence is that Chib's original method can be applied, using output from a single MCMC run. Chib's method is generally thought to provide a good approximation to the marginal likelihood of latent variable models (see, for example Frühwirth-Schnatter, 2006; Marin & Robert, 2008) and is included here to provide a benchmark for the performance of the other estimators.

The simulation experiments involve single realisations of a hidden Markov model ($Y = 1$) of length T in networks with $n = 1$ and $n = 4$ sites. We consider models for occurrence alone, i.e. omitting the nodes $\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_T$ from the DAG in Figure 4.1, and for both occurrence and amount. In the latter case, the parameterisation of the precipitation process differs from that in Section 4.2.2 in that non–zero rainfall amounts are modelled using the exponential rather than gamma distribution,

$$(W_t^i \mid D_t^i = 1, S_t = k, \theta_{r,\text{obs}}, r) \sim \text{Exp}(\beta_{r,ik}) \equiv \text{Ga}(1, \beta_{r,ik})$$

in order to facilitate exact computation of the marginal likelihood when (standard) conjugate priors are used. For notational convenience we denote $\mathcal{B}_r = (\beta_{r,ik})$.

4.6.2 Exact computation of the marginal likelihood

The marginal likelihood for the rainfall occurrence model can be expressed as

$$p(\mathbf{d} \mid r) = \sum_{\mathbf{s}} p(\mathbf{d} \mid \mathbf{s}, r) p(\mathbf{s} \mid r), \tag{4.27}$$

whilst for the model which additionally incorporates amounts, it can be written as

$$p(\mathbf{w}, \mathbf{d} \mid r) = \sum_{\mathbf{s}} p(\mathbf{w}, \mathbf{d} \mid \mathbf{s}, r) p(\mathbf{s} \mid r). \tag{4.28}$$

In each case the sum is over all r^T possible weather state sequences of length T. When $\theta_{r,\text{hid}}$ and $\theta_{r,\text{obs}}$ are assumed conditionally independent, given r, and assigned fully conjugate priors,

the summands in (4.27) and (4.28) can be computed exactly. Therefore when T is small it is feasible to compute the marginal likelihood by direct enumeration of these summands over all weather state sequences, allowing the estimates to be compared to the true values.

In these simulation experiments since we only have one sequence ($Y = 1$), for simplicity, we take the initial distribution ν_r to be fixed and equal to the discrete uniform distribution on $\mathcal{S}_r$, i.e. $\Pr(S_1 = k \mid r) = 1/r$ for all $k \in \mathcal{S}_r$. The exchangeable prior distribution described in Section 4.3 must be adjusted as follows to account for the modification to the parameterisation of the precipitation process. Clearly the terms pertaining to the amounts process are dropped in the prior for the occurrence only hidden Markov model. In the model which additionally incorporates rainfall amounts, the prior for $(\mathcal{M}_r, \mathcal{V}_r \mid r)$ is replaced with a prior for $(\mathcal{B}_r \mid r)$, which is exchangeable across weather states and comprises independent Gamma distributions for each $\beta_{r,ik}$,

$$\beta_{r,ik} \mid r \sim \mathrm{Ga}(f_{r,1i}, f_{r,2i}), \quad (i,k) \in \{1,\ldots,n\} \times \mathcal{S}_r.$$

The resulting prior distributions, $\pi(\Lambda_r, \mathcal{P}_r \mid r)$ or $\pi(\Lambda_r, \mathcal{P}_r, \mathcal{B}_r \mid r)$, are fully conjugate, therefore we can calculate

$$\begin{aligned} p(\mathbf{s} \mid r) &= \int p(\mathbf{s} \mid \Lambda_r, r)\pi(\Lambda_r \mid r)d\Lambda_r \\ &= \int \Pr(S_1 = s_1 \mid r) \prod_{t=2}^{T} \Pr(S_t = s_t \mid S_{t-1} = s_{t-1}, \Lambda_r, r)\pi(\Lambda_r \mid r)d\Lambda_r \\ &= \frac{1}{r} \prod_{j=1}^{r} \frac{\Gamma(\sum_{k=1}^{r} E_r e_{r,jk})}{\prod_{k=1}^{r} \Gamma(E_r e_{r,jk})} \int \prod_{k=1}^{r} \lambda_{r,jk}^{E_r e_{r,jk} + n_{jk}(\mathbf{s}) - 1} d\lambda_{r,j}, \end{aligned}$$

and so

$$p(\mathbf{s} \mid r) = \frac{1}{r} \prod_{j=1}^{r} \frac{\Gamma(\sum_{k=1}^{r} E_r e_{r,jk}) \prod_{k=1}^{r} \Gamma\{E_r e_{r,jk} + n_{jk}(\mathbf{s})\}}{\prod_{k=1}^{r} \Gamma(E_r e_{r,jk}) \Gamma\left[\sum_{k=1}^{r} \{E_r e_{r,jk} + n_{jk}(\mathbf{s})\}\right]}$$

where the transition counts $n_{jk}(\mathbf{s})$ associated with a particular weather state sequence $\mathbf{s}$ are defined in equation (4.19) with $Y = 1$.

Similarly, for a rainfall occurrence only hidden Markov model

$$\begin{aligned} p(\mathbf{d} \mid \mathbf{s}, r) &= \int p(\mathbf{d} \mid \mathbf{s}, \mathcal{P}_r, r)\pi(\mathcal{P}_r \mid r)d\mathcal{P}_r \\ &= \int \prod_{t=1}^{T} \Pr(\mathbf{D}_t = \mathbf{d}_t \mid S_t = s_t, \mathcal{P}_r)\pi(\mathcal{P}_r \mid r)d\mathcal{P}_r \\ &= \prod_{k=1}^{r} \prod_{i=1}^{n} \frac{\Gamma(a_{r,1i} + a_{r,2i})}{\Gamma(a_{r,1i})\Gamma(a_{r,2i})} \int p_{r,ik}^{a_{r,1i} + T_{ik}^1(\mathbf{s}) - 1} (1 - p_{r,ik})^{a_{r,2i} + T_{ik}^0(\mathbf{s}) - 1} dp_{r,ik} \\ &= \prod_{k=1}^{r} \prod_{i=1}^{n} \frac{\Gamma(a_{r,1i} + a_{r,2i})\Gamma\{a_{r,1i} + T_{ik}^1(\mathbf{s})\}\Gamma\{a_{r,2i} + T_{ik}^0(\mathbf{s})\}}{\Gamma(a_{r,1i})\Gamma(a_{r,2i})\Gamma\{a_{r,1i} + T_{ik}^1(\mathbf{s}) + a_{r,2i} + T_{ik}^0(\mathbf{s})\}} \\ &= \prod_{i=1}^{n} \left\{ \frac{\Gamma(a_{r,1i} + a_{r,2i})}{\Gamma(a_{r,1i})\Gamma(a_{r,2i})} \right\}^{r} \prod_{k=1}^{r} \frac{\Gamma\{a_{r,1i} + T_{ik}^1(\mathbf{s})\}\Gamma\{a_{r,2i} + T_{ik}^0(\mathbf{s})\}}{\Gamma\{a_{r,1i} + T_{ik}^1(\mathbf{s}) + a_{r,2i} + T_{ik}^0(\mathbf{s})\}} \end{aligned} \tag{4.29}$$

where the counts $T^1_{ik}(\mathbf{s})$ and $T^0_{ik}(\mathbf{s})$ for a particular weather state sequence $\mathbf{s}$ are defined in equation (4.19).

For the hidden Markov model for both occurrence and amount we have

$$p(\mathbf{w}, \mathbf{d} \mid \mathbf{s}, r) = p(\mathbf{d} \mid \mathbf{s}, r) p(\mathbf{w} \mid \mathbf{d}, \mathbf{s}, r) \tag{4.30}$$

where $p(\mathbf{d} \mid \mathbf{s}, r)$ is given in equation (4.29) and

$$\begin{aligned}
p(\mathbf{w} \mid \mathbf{d}, \mathbf{s}, r) &= \int p(\mathbf{w} \mid \mathbf{d}, \mathbf{s}, \mathcal{B}_r, r) \pi(\mathcal{B}_r \mid r) d\,\mathcal{B}_r \\
&= \int \prod_{t=1}^{T} p(\mathbf{w}_t \mid \mathbf{D}_t = \mathbf{d}_t, S_t = s_t, \mathcal{B}_r, r) \pi(\mathcal{B}_r \mid r) d\,\mathcal{B}_r \\
&= \prod_{k=1}^{r} \prod_{i=1}^{n} \frac{f_{r,2i}^{f_{r,1i}}}{\Gamma(f_{r,1i})} \int \beta_{r,ik}^{f_{r,1i}+T^1_{ik}(\mathbf{s})-1} \exp[-\{f_{r,2i} + T^1_{ik}(\mathbf{s}) \bar{w}_{ik}(\mathbf{s})\} \beta_{r,ik}] d\,\beta_{r,ik} \\
&= \prod_{k=1}^{r} \prod_{i=1}^{n} \frac{f_{r,2i}^{f_{r,1i}} \Gamma\{f_{r,1i} + T^1_{ik}(\mathbf{s})\}}{\Gamma(f_{r,1i}) \{f_{r,2i} + T^1_{ik}(\mathbf{s}) \bar{w}_{ik}(\mathbf{s})\}^{\{f_{r,1i}+T^1_{ik}(\mathbf{s})\}}} \\
&= \prod_{i=1}^{n} \left\{ \frac{f_{r,2i}^{f_{r,1i}}}{\Gamma(f_{r,1i})} \right\}^{r} \prod_{k=1}^{r} \frac{\Gamma\{f_{r,1i} + T^1_{ik}(\mathbf{s})\}}{\{f_{r,2i} + T^1_{ik}(\mathbf{s}) \bar{w}_{ik}(\mathbf{s})\}^{\{f_{r,1i}+T^1_{ik}(\mathbf{s})\}}}
\end{aligned} \tag{4.31}$$

with $\bar{w}_{ik}(\mathbf{s})$ defined in equation (4.23) as the arithmetic mean of the non-zero rainfall amounts at site i in weather state k, for a particular weather state sequence, $\mathbf{s}$.

4.6.3 Design of the simulation experiment

Four sets of experiments were performed in which data were simulated from and modelled using hidden Markov models for:

1. Rainfall occurrence only, with $n = 1$ site
2. Rainfall occurrence and amount, with $n = 1$ site
3. Rainfall occurrence only, with $n = 4$ sites
4. Rainfall occurrence and amount, with $n = 4$ sites

For each of these four scenarios the data were generated from hidden Markov models with $r = 2$ states and a transition matrix given by

$$\Lambda_2 = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}.$$

For the first scenario, we simulated 100 replicated datasets, each of length $T = 20$, choosing the parameters in the observed process to be $p_{2,11} = 0.45$ and $p_{2,12} = 0.9$. We then simulated

Scenario	Parameter Set 1	Parameter Set 2	Parameter Set 3
1	$\mathcal{P}_2 = \begin{pmatrix} 0.45 & 0.9 \end{pmatrix}$	$\mathcal{P}_2 = \begin{pmatrix} 0.05 & 0.9 \end{pmatrix}$	$\mathcal{P}_2 = \begin{pmatrix} 0.9 & 0.9 \end{pmatrix}$
2	$\mathcal{P}_2 = \begin{pmatrix} 0.45 & 0.9 \end{pmatrix}$ $\mathcal{B}_2 = \begin{pmatrix} 1.0 & 0.2 \end{pmatrix}$	$\mathcal{P}_2 = \begin{pmatrix} 0.05 & 0.9 \end{pmatrix}$ $\mathcal{B}_2 = \begin{pmatrix} 10.0 & 0.2 \end{pmatrix}$	$\mathcal{P}_2 = \begin{pmatrix} 0.9 & 0.9 \end{pmatrix}$ $\mathcal{B}_2 = \begin{pmatrix} 0.2 & 0.2 \end{pmatrix}$
3	$\mathcal{P}_2 = \begin{pmatrix} 0.45 & 0.9 \\ 0.4 & 0.95 \\ 0.45 & 0.8 \\ 0.5 & 0.85 \end{pmatrix}$	$\mathcal{P}_2 = \begin{pmatrix} 0.05 & 0.9 \\ 0.1 & 0.95 \\ 0.1 & 0.8 \\ 0.05 & 0.85 \end{pmatrix}$	$\mathcal{P}_2 = \begin{pmatrix} 0.9 & 0.9 \\ 0.95 & 0.95 \\ 0.8 & 0.8 \\ 0.85 & 0.85 \end{pmatrix}$
4	$\mathcal{P}_2 = \begin{pmatrix} 0.45 & 0.9 \\ 0.4 & 0.95 \\ 0.45 & 0.8 \\ 0.5 & 0.85 \end{pmatrix}$ $\mathcal{B}_2 = \begin{pmatrix} 1.0 & 0.2 \\ 0.9 & 0.3 \\ 1.1 & 0.3 \\ 1.0 & 0.25 \end{pmatrix}$	$\mathcal{P}_2 = \begin{pmatrix} 0.05 & 0.9 \\ 0.1 & 0.95 \\ 0.1 & 0.8 \\ 0.05 & 0.85 \end{pmatrix}$ $\mathcal{B}_2 = \begin{pmatrix} 10.0 & 0.2 \\ 10.5 & 0.3 \\ 9.5 & 0.3 \\ 9.5 & 0.25 \end{pmatrix}$	$\mathcal{P}_2 = \begin{pmatrix} 0.9 & 0.9 \\ 0.95 & 0.95 \\ 0.8 & 0.8 \\ 0.85 & 0.85 \end{pmatrix}$ $\mathcal{B}_2 = \begin{pmatrix} 0.2 & 0.2 \\ 0.3 & 0.3 \\ 0.3 & 0.3 \\ 0.25 & 0.25 \end{pmatrix}$

Table 4.1: Observed process parameters used to simulate the data in the simulation experiments.

another 100 replicated datasets of length $T = 20$, this time changing the probability of rain in state 1 so that $p_{2,11} = 0.05$ and leaving the probability of rain in state 2 unchanged, i.e. $p_{2,12} = 0.9$. This was repeated a third time, again changing only the probability of rain in state 1 so that in this case $p_{2,11} = p_{2,12} = 0.9$. We proceeded in an analogous fashion for each of the remaining scenarios by simulating 100 replicated datasets of length $T = 20$ from models in which the parameters in $\theta_{2,\text{obs}}$ were fixed at three different sets of values, as given in Table 4.1. Note that within each of the four scenarios, it is only the parameters in the first weather state which change between parameter sets. In addition, within all parameter sets, weather state two (see the second column in the parameter matrices in Table 4.1) is designed to represent a "wet" weather state with high probabilities of rain and large rainfall amounts on wet days. When $\theta_{2,\text{obs}}$ is set according to parameter sets one and two, weather state one is intended to be associated with drier conditions than weather state two, with parameter set two leading to a bigger difference between the states. When data are simulated using parameter set three, a two state hidden Markov model will be overfitting, as the data actually arise from a hidden Markov model in which $r = 1$.

For each parameter set, within each scenario, the performances of nine estimators were compared:

(i) The Monte Carlo estimator

(ii) The Newton Raftery hybrid estimator (and its approximate version) in which the weight assigned to the prior in the mixture importance sampling density is $\delta = 0.01$

(iii) The Newton Raftery hybrid estimator (and its approximate version) in which the weight assigned to the prior in the mixture importance sampling density is $\delta = 0.05$

(iv) The harmonic mean estimator

(v) The bridge sampling estimator with importance sampling density equal to the prior

(vi) The estimator based on the power posterior approach

(vii) Chib's estimator

4.6.4 Implementation

For the purpose of these simulation experiments the hyperparameters in the priors for Λ_2, $\mathcal{P}_2$ and $\mathcal{B}_2$ were chosen to be

$$E_2 = 2, \quad \mathbf{e}_{2,j} = \left(\frac{1}{2}, \frac{1}{2}\right), \quad \text{for } j \in \mathcal{S}_2; \quad a_{2,1i} = a_{2,2i} = 1 \quad \text{for } i \in \{1, \ldots, n\};$$
$$\text{and} \quad f_{2,1i} = f_{2,2i} = 1 \quad \text{for } i \in \{1, \ldots, n\},$$

respectively.

Several of the estimators are based on MCMC samples from the posterior distribution of the model parameters. These were generated by Gibbs sampling with data augmentation. The algorithm used was based on that presented in Section 4.5.3 with a simplification when analysing data generated from the occurrence only model in which steps 1(b)(ii) and 1(b)(iii) were omitted, and a modification when analysing data from the model for both occurrence and amount. This modification was necessary to account for the exchange of the gamma distribution for the exponential distribution and consisted of replacing the aforementioned steps by simulation of $\beta_{2,ik}$ from its full conditional distribution,

$$\beta_{2,ik} \mid \cdots \sim \text{Ga}\left\{f_{2,1i} + T^1_{ik}(\mathbf{s}), f_{2,2i} + T^1_{ik}(\mathbf{s})\bar{w}_{ik}(\mathbf{s})\right\}$$

for each pair $(i,k) \in \{1,2,\ldots,n\} \times \mathcal{S}_2$.

For each set of simulated data, following a burn–in of 10,000 iterations, 50,000 posterior draws were obtained and thinned to every 10–th iterate to reduce computing overheads. This gave a final posterior sample of size 5,000 from which to compute the estimators.

For each set of simulated data, the simple estimators (i)–(v) were constructed as follows. The harmonic mean and approximate versions of the Newton Raftery hybrid estimators were computed directly from the MCMC output. The (original) Newton Raftery hybrid estimators were computed using a mixture of the MCMC draws and draws from the prior. The bridge sampling estimator was computed based on the MCMC sample together with a random sample of size 5,000 from the prior. The Monte Carlo estimator was based purely on a random sample of size 5,000 from the prior. The remainder of this section contains details of the (more involved) computation of Chib's estimator and the estimator based on power posteriors.

In all scenarios, the prior for θ_2 is fully conjugate to the form of the complete data likelihood and so the appropriate complete data posterior, $\pi(\theta_2 \mid \mathbf{s}, \mathbf{d}, r = 2)$ or $\pi(\theta_2 \mid \mathbf{s}, \mathbf{w}, \mathbf{d}, r = 2)$, is

available in closed form. The posterior ordinate in Chib's estimator, (3.40), could therefore be approximated directly from the MCMC output using

$$\begin{aligned}\hat{\pi}(\mathcal{P}_2^*, \Lambda_2^* \mid \mathbf{d}, r=2) &= \frac{1}{N}\sum_{i=1}^{N} \pi(\mathcal{P}_2^*, \Lambda_2^* \mid \mathbf{s}^{[i]}, \mathbf{d}, r=2) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left[\prod_{k=1}^{r} \mathscr{D}_2\{\boldsymbol{\lambda}_{2,k}^* \mid E_2\mathbf{c}_{2,k} + \mathbf{n}_k(\mathbf{s}^{[i]})\}\right. \\ &\qquad \left.\times \prod_{k=1}^{r}\prod_{i=1}^{n} \mathrm{Beta}\{p_{2,ik}^* \mid a_{2,1i} + T_{ik}^1(\mathbf{s}^{[i]}), a_{2,2i} + T_{ik}^0(\mathbf{s}^{[i]})\}\right]\end{aligned}$$

for the hidden Markov models for occurrence only, and

$$\begin{aligned}\hat{\pi}(\mathcal{P}_2^*, \mathcal{B}_2^*, \Lambda_2^* \mid \mathbf{w}, \mathbf{d}, r=2) &= \frac{1}{N}\sum_{i=1}^{N} \pi(\mathcal{P}_2^*, \mathcal{B}_2^*, \Lambda_2^* \mid \mathbf{s}^{[i]}, \mathbf{w}, \mathbf{d}, r=2) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left[\prod_{k=1}^{r} \mathscr{D}_2\{\boldsymbol{\lambda}_{2,k}^* \mid E_2\mathbf{c}_{2,k} + \mathbf{n}_k^{[i]}(\mathbf{s})\}\right. \\ &\qquad \times \prod_{k=1}^{r}\prod_{i=1}^{n} \mathrm{Beta}\{p_{2,ik}^* \mid a_{2,1i} + T_{ik}^1(\mathbf{s}^{[i]}), a_{2,2i} + T_{ik}^0(\mathbf{s}^{[i]})\} \\ &\qquad \left.\times \prod_{k=1}^{r}\prod_{i=1}^{n} \mathrm{Ga}\{\beta_{2,ik}^* \mid f_{2,1i} + T_{ik}^1(\mathbf{s}^{[i]}), f_{2,2i} + T_{ik}^1(\mathbf{s}^{[i]})\bar{w}_{ik}(\mathbf{s}^{[i]})\}\right]\end{aligned}$$

for the models additionally incorporating amounts. The high density point was taken as the MCMC iterate corresponding to the maximum value of the observed data likelihood obtained during MCMC sampling. Note that it was necessary to add a random permutation sampling step at the end of the MCMC algorithm in order to correctly compute Chib's estimator; see Section 3.5.1.3 for further details.

To estimate the marginal likelihood using power posteriors, we used data augmentation, appending the hidden states to the set of unknowns. For the hidden Markov models considered in this chapter we chose not to marginalise over the hidden states because the resulting Metropolis Hastings algorithm would then require careful blocking of $\boldsymbol{\theta}_r$ and time consuming tuning, particularly in cases when the number of sites and hence dimension of $\boldsymbol{\theta}_r$ is large. At temperature t, the power posterior is defined as

$$\begin{aligned}&\pi_t(\theta_2, \mathbf{s} \mid \mathbf{d}, r=2) \\ &\quad \propto p(\mathbf{d} \mid \theta_2, \mathbf{s}, r=2)^t p(\theta_2, \mathbf{s} \mid r=2) \\ &\quad = p(\mathbf{d} \mid \theta_{2,\text{obs}}, \mathbf{s}, r=2)^t p(\mathbf{s} \mid \theta_{2,\text{hid}}, r=2)\pi(\theta_{2,\text{hid}} \mid r=2)\pi(\theta_{2,\text{obs}} \mid r=2)\end{aligned} \tag{4.32}$$

for the rainfall occurrence models and

$$\begin{aligned}&\pi_t(\theta_2, \mathbf{s} \mid \mathbf{w}, \mathbf{d}, r=2) \\ &\quad \propto p(\mathbf{w}, \mathbf{d} \mid \theta_{2,\text{obs}}, \mathbf{s}, r=2)^t p(\mathbf{s} \mid \theta_{2,\text{hid}}, r=2)\pi(\theta_{2,\text{hid}} \mid r=2)\pi(\theta_{2,\text{obs}} \mid r=2)\end{aligned} \tag{4.33}$$

for the models which also include amounts. As the temperature variable t is fixed, it is immediately clear that $\theta_{2,\text{hid}}$ (here simply equal to Λ_2) is conditionally independent of t given $\mathbf{s}$, so its full conditional distribution remains as it would be in an ordinary posterior analysis. Since the Bernoulli and exponential distributions are both members of the exponential family we can easily derive the full conditional distributions of the elements of $\theta_{2,\text{obs}}$ as

$$p_{2,ik} \mid \cdots \sim \text{Beta}\{a_{2,1i} + tT^1_{ik}(\mathbf{s}), a_{2,2i} + tT^0_{ik}(\mathbf{s})\} \quad (i,k) \in \{1,2,\ldots,n\} \times \mathcal{S}_2,$$

and when additionally modelling amounts

$$\beta_{2,ik} \mid \cdots \sim \text{Ga}\left\{f_{2,1i} + tT^1_{ik}(\mathbf{s}), f_{2,2i} + tT^1_{ik}(\mathbf{s})\bar{w}_{ik}(\mathbf{s})\right\} \quad (i,k) \in \{1,2,\ldots,n\} \times \mathcal{S}_2.$$

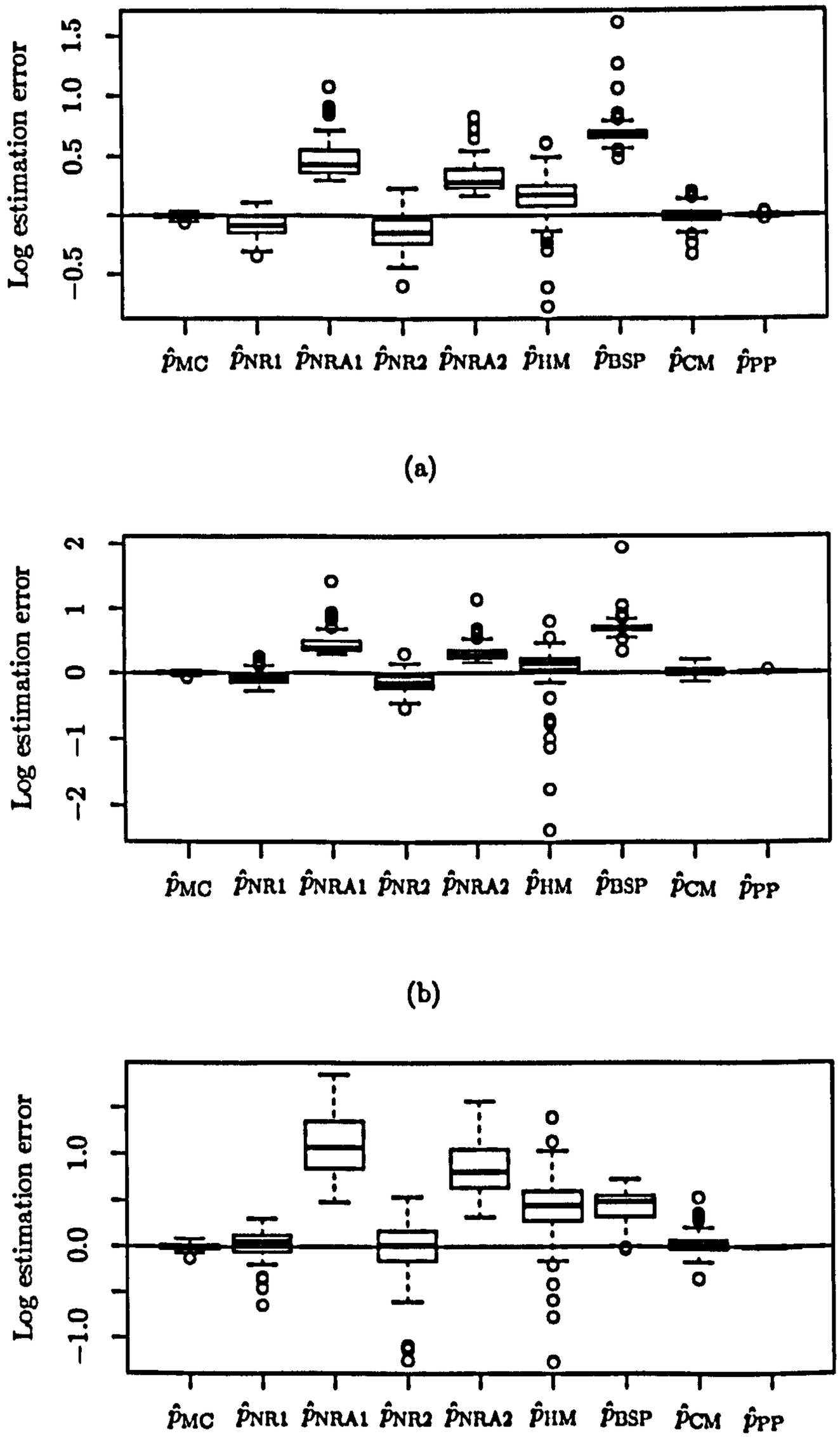
To sample the weather states from their full conditional distribution we simply use a modified version of the forward backward algorithm (Algorithm 3.3.3) in which the density, $p(\mathbf{d}_u \mid S_u = s_u, \theta_{2,\text{obs}}, r = 2)$ or $p(\mathbf{w}_u, \mathbf{d}_u \mid S_u = s_u, \theta_{2,\text{obs}}, r = 2)$, in the filtered probability at time u is raised to the power t.

Estimation of the expected half deviances proceeded according to Algorithm 3.5.1, and the overall estimate of the log marginal likelihood was obtained using the trapezoidal rule, (3.38). For each set of simulated data, we used the "vanilla" power posterior algorithm advocated by Friel & Pettitt (2008), with a geometric spacing of the temperatures, $t_i = (i/n)^c$, for $i = 0,\ldots,n$ where $n = 40$ and $c = 4$. At each temperature 10,000 samples were generated from the power posterior omitting the first 4,000 as burn–in.

4.6.5 Results

For each of the four scenarios, the performance of the estimators was assessed by approximating over the 100 replications the distribution of the estimation error, $\{\log\hat{p}(\mathbf{d} \mid r = 2) - \log p(\mathbf{d} \mid r = 2)\}$ or $\{\log\hat{p}(\mathbf{w},\mathbf{d} \mid r = 2) - \log p(\mathbf{w},\mathbf{d} \mid r = 2)\}$, as appropriate. Figures 4.3–4.6 provide graphical representations of these distributions, in the form of box–and–whisker plots. In each case, subfigures (a)–(c) correspond to parameter sets 1–3, respectively.

In Figures 4.3–4.6 there is generally little difference between the plots (a)–(c). This is surprising since the data were simulated within each scenario with the intention that the resulting posterior distributions would have different shapes depending on the parameter set that was used in generating the data. Specifically, the data were simulated so that the difference between the hidden states was smaller when using parameter set one than parameter set two, with no difference at all when using parameter set three. When the parameters in the two states are very different we would anticipate there to be little posterior uncertainty in the assignment of days to states. As such we might expect the posterior to be fairly concentrated about each of its $2! = 2$ well separated and symmetric (major) modes, meaning that the diffuse unimodal prior would be unlikely to provide a good approximation to the posterior. Consequently we expected the estimators obtained by basing an importance sampling density on the prior to perform more poorly in experiments when data were simulated from hidden Markov models with very distinct states. We expected this to be true especially of the harmonic mean estimator because it only has bounded variance if the prior has thinner tails than the unnormalised posterior, and this

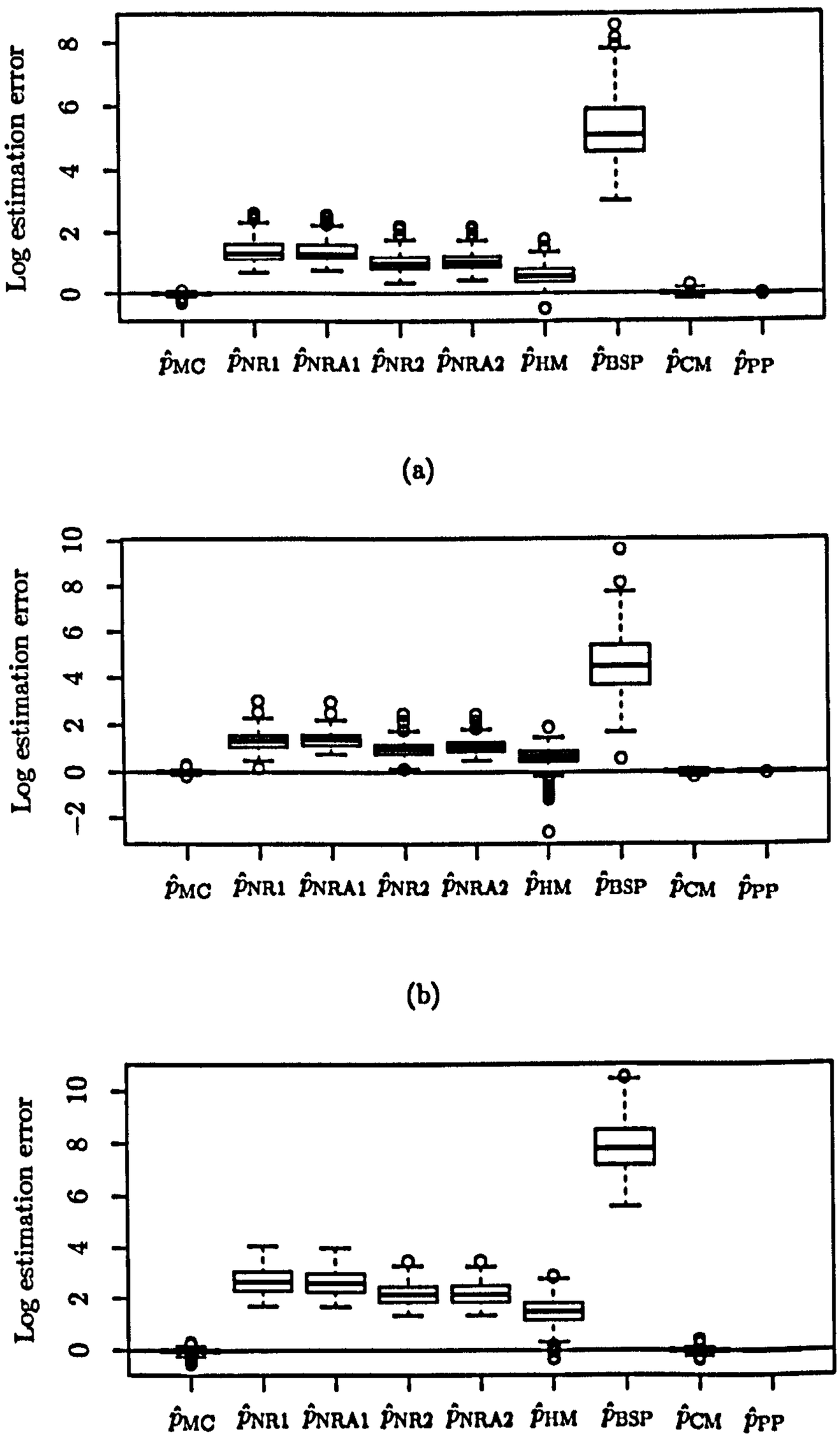


(a)

(b)

(c)

Figure 4.3: Distributions of the estimation error $\log \hat{p}(\mathbf{d} \mid r = 2) - \log p(\mathbf{d} \mid r = 2)$ for different estimators based on data simulated from a $n = 1$ site, $r = 2$ weather state hidden Markov model for rainfall occurrence using the parameter sets (a) 1; (b) 2; and (c) 3. The estimators considered are the Monte Carlo estimator $\hat{p}_{MC}$; the Newton Raftery hybrid estimator based on $\delta = 0.05$, $\hat{p}_{NR1}$, and on $\delta = 0.01$, $\hat{p}_{NR2}$, and the approximate versions $\hat{p}_{NRA1}$ and $\hat{p}_{NRA2}$; the harmonic mean estimator $\hat{p}_{HM}$; a bridge sampling estimator $\hat{p}_{BSP}$; Chib's estimator $\hat{p}_{CM}$; and the power posterior estimator $\hat{p}_{PP}$.



(a)

(b)

(c)

Figure 4.4: Distributions of the estimation error $\log \hat{p}(\mathbf{w}, \mathbf{d} \mid r = 2) - \log p(\mathbf{w}, \mathbf{d} \mid r = 2)$ for different estimators based on data simulated from a $n = 1$ site, $r = 2$ weather state hidden Markov model for rainfall occurrence and amount using the parameter sets (a) 1; (b) 2; and (c) 3. The estimators considered are the Monte Carlo estimator $\hat{p}_{MC}$; the Newton Raftery hybrid estimator based on $\delta = 0.05$, $\hat{p}_{NR1}$, and on $\delta = 0.01$, $\hat{p}_{NR2}$, and the approximate versions $\hat{p}_{NRA1}$ and $\hat{p}_{NRA2}$; the harmonic mean estimator $\hat{p}_{HM}$; a bridge sampling estimator $\hat{p}_{BSP}$; Chib's estimator $\hat{p}_{CM}$; and the power posterior estimator $\hat{p}_{PP}$.

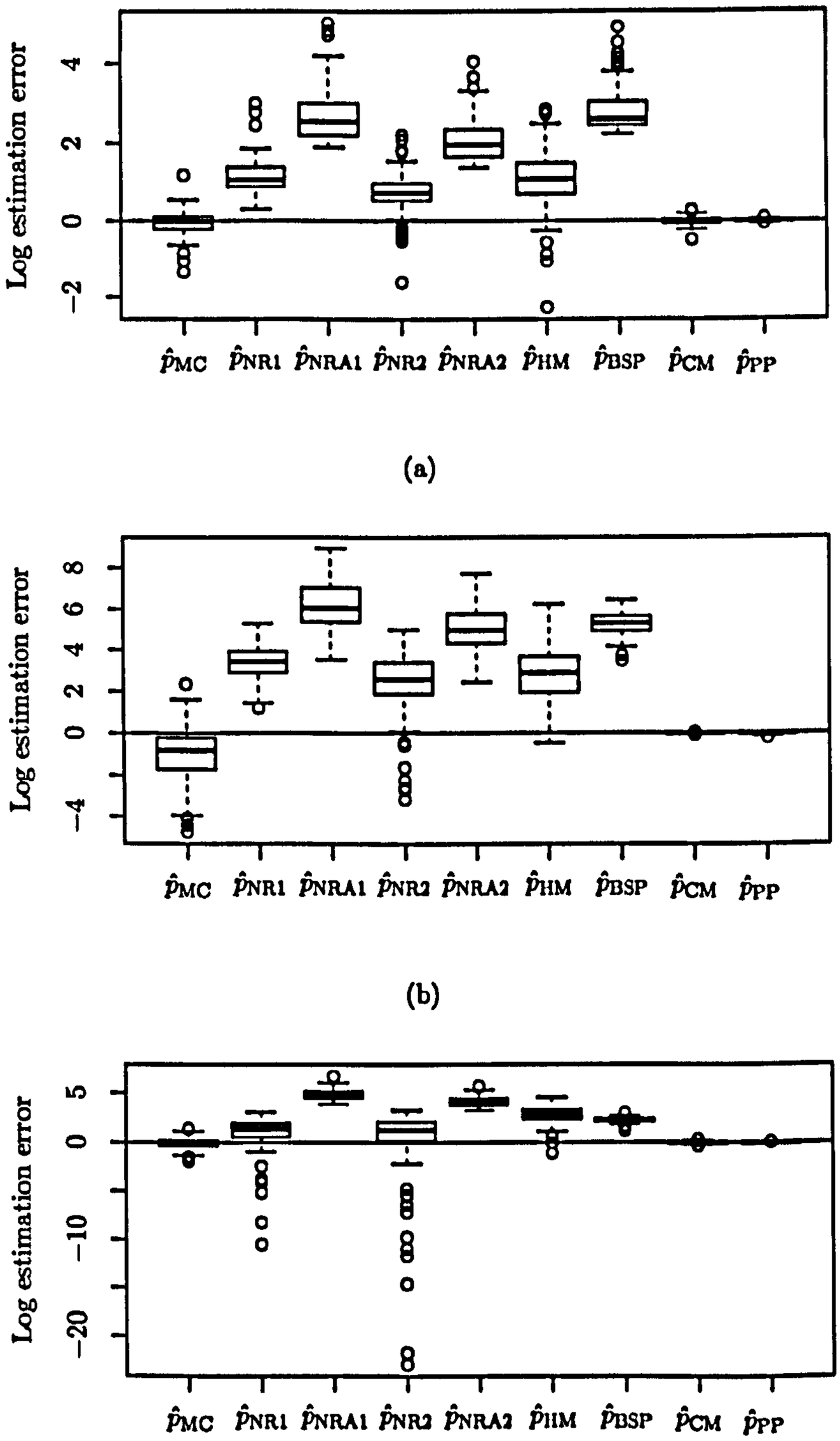


Figure 4.5: Distributions of the estimation error $\log \hat{p}(\mathbf{d} \mid r = 2) - \log p(\mathbf{d} \mid r = 2)$ for different estimators based on data simulated from a $n = 4$ site, $r = 2$ weather state hidden Markov model for rainfall occurrence using the parameter sets (a) 1; (b) 2; and (c) 3. The estimators considered are the Monte Carlo estimator $\hat{p}_{MC}$; the Newton Raftery hybrid estimator based on $\delta = 0.05$, $\hat{p}_{NR1}$, and on $\delta = 0.01$, $\hat{p}_{NR2}$, and the approximate versions $\hat{p}_{NRA1}$ and $\hat{p}_{NRA2}$; the harmonic mean estimator $\hat{p}_{HM}$; a bridge sampling estimator $\hat{p}_{BSP}$; Chib's estimator $\hat{p}_{CM}$; and the power posterior estimator $\hat{p}_{PP}$.

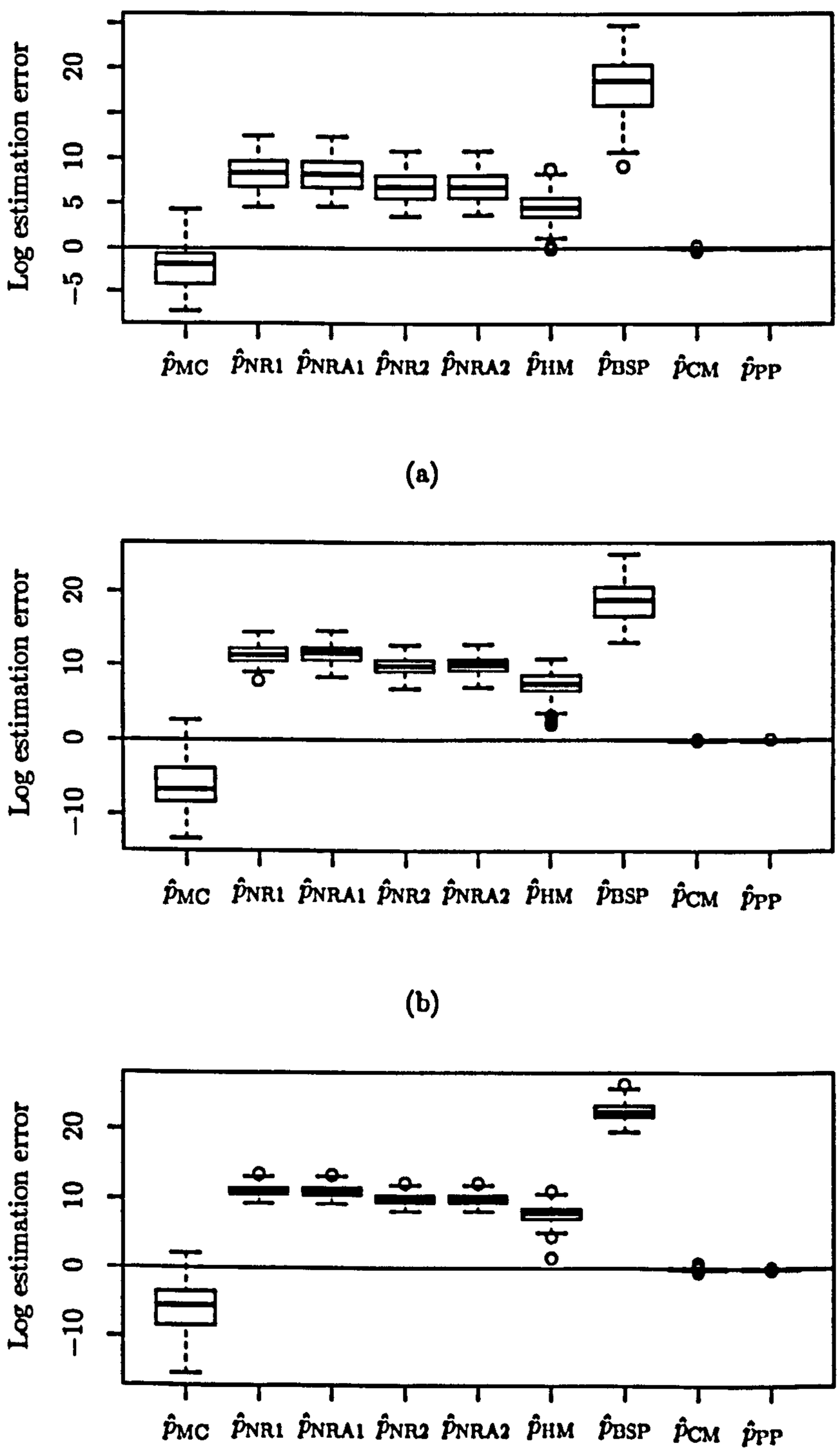


Figure 4.6: Distributions of the estimation error $\log \hat{p}(\mathbf{w}, \mathbf{d} \mid r = 2) - \log p(\mathbf{w}, \mathbf{d} \mid r = 2)$ for different estimators based on data simulated from a $n = 4$ site, $r = 2$ weather state hidden Markov model for rainfall occurrence and amount using the parameter sets (a) 1; (b) 2; and (c) 3. The estimators considered are the Monte Carlo estimator $\hat{p}_{MC}$; the Newton Raftery hybrid estimator based on $\delta = 0.05$, $\hat{p}_{NR1}$, and on $\delta = 0.01$, $\hat{p}_{NR2}$, and the approximate versions $\hat{p}_{NRA1}$ and $\hat{p}_{NRA2}$; the harmonic mean estimator $\hat{p}_{HM}$; a bridge sampling estimator $\hat{p}_{BSP}$; Chib's estimator $\hat{p}_{CM}$; and the power posterior estimator $\hat{p}_{PP}$.

is unlikely to be the case if the posterior is very concentrated. However, these effects were not particularly borne out through the simulation experiments.

In general the Monte Carlo estimator performs rather well except within the simulation experiments involving hidden Markov models for both occurrence and amount at $n = 4$ sites, in which it seems the Monte Carlo estimator underestimates the marginal likelihood and has a very large variance. One possible explanation for this is that the prior for $(\mathcal{B}_2 \mid r = 2)$ assumes independence between sites, whilst the parameters used to simulate the data are chosen to be similar at all sites within a weather state. Therefore drawing from the prior will fail to produce many draws which respect the dependence that is likely to be exhibited in the posterior. With few draws in regions of high likelihood, the estimator will be dominated by a few large likelihood values explaining the large variance and slow convergence to the exact marginal likelihood.

For all scenarios and all parameter sets, the harmonic mean estimator overestimates the marginal likelihood in the majority of the 100 simulated datasets. As we might expect, the Newton Raftery hybrid estimator tends to be more stable than the harmonic mean estimator (with fewer extreme estimation errors), especially when more weight is assigned to the prior, i.e. $\delta = 0.05$ compared to $\delta = 0.01$. The full hybrid estimators generally introduced less bias than their approximate versions. The bridge sampling estimator is the worst amongst those considered, especially in hidden Markov models for rainfall occurrence and amounts when the estimation error is generally very large.

The estimators based on Chib's method and the power posterior approach perform very well in all scenarios, for all parameter sets, with the estimation error being very close to zero. Compared with the other methods, the power posterior approach required slightly more work to program and computing times were slightly longer. Also, in contrast to the other methods, it could not be incorporated directly into standard MCMC sampling from the posterior. However, modifying existing MCMC code which sampled from the joint posterior of the model parameters and hidden states was trivial and led to an estimator which performed better than the others considered, including Chib's estimator. Chib's estimator is generally considered to provide a good approximation to the marginal likelihood in hidden Markov models. Computationally, however, its extended version was deemed too expensive for our model and so Chib's estimator was really only included as a benchmark, emphasising the strength of the power posterior approach.

4.6.6 Concluding Remarks

On the basis of these simulation experiments, it seems that the power posterior approach provides a very good approximation to the marginal likelihood of hidden Markov models and, unlike many of the other estimators, it continues to be usable when some of the full conditional distributions have unknown normalising constants. In this and subsequent chapters, therefore, we will attempt to compute the posterior distribution for r using Bayes Theorem, after estimating the marginal likelihoods via the power posterior approach.

4.7 Application to Yorkshire winter rainfall data

The model and inferential techniques described in the previous sections will now be illustrated by analysing the Yorkshire winter dataset described in Chapter 2. The time series comprises the winter (December–February) rainfall data at $n = 6$ sites over the years 1961/62 to 1990/91, so divides naturally into $Y = 30$ sub–series; one for each winter season. As mentioned in Section 4.4, there are long (nine month) time periods separating consecutive sub–series so it seems reasonable to model them, together with the associated weather states, as independent realisations of a hidden Markov model. The overall length of the data is $T = 2707$ and each sub–series has length $T_y = 90$ (or $T_y = 91$ if the y–th February lies in a leap year).

This section begins by explaining our choice of prior for r and the hyperparameters in our priors for $(\theta_r \mid r)$, $r = 1, \ldots, r_{\max}$. We then use the power posterior approach to estimate the log marginal likelihood for each model, before combining this information with the prior for r to deduce its posterior distribution. For all those values of r with non–negligible posterior support, we use the MCMC scheme outlined in Section 4.5.3 to obtain samples from the posterior distribution $\pi(\theta_r, \mathbf{s} \mid \mathbf{w}, \mathbf{d}, r)$ and analyse the output from the model corresponding to the posterior modal value of r. Finally we assess the fit of the model to the data by comparing the observed values of various test quantities to their posterior predictive distributions, averaged over the posterior for r.

4.7.1 Prior specification

In attempting to model the Yorkshire rainfall data we have two main objectives. The first is prediction, and the fulfilment of this objective will be assessed by comparing observed statistics, summarising important features of the rainfall data, to their posterior predictive distributions; see Section 4.7.4. Our second aim is to provide a parsimonious, interpretable model which could assist in understanding the underlying physical mechanism which generated the rainfall data.

In this section we begin by specifying our prior for r. This involves choosing a value for $r_{\max}$, then a suitable mass function over the support $\{1, \ldots, r_{\max}\}$. Both choices are made to reflect the goals of the analysis, above. Next, we specify our conditional prior for the model parameters, given r. In making this specification we need to consider two issues. First, given the sensitivity of the marginal likelihood to the prior, we choose our conditional priors, $\pi(\theta_1 \mid r = 1), \ldots, \pi(\theta_{r_{\max}} \mid r = r_{\max})$, to give the same predictions for simple marginal quantities like the probability of rain. This is to allow the marginal likelihood to discriminate between models using properties of the *joint* rainfall distribution rather than the marginal distribution of rain on any particular day. The second issue is how to elicit our prior knowledge probabilistically. Section 4.3 provided details of our prior distribution and guidelines regarding the choice of hyperparameters. Our prior specification is concluded with an illustration of this choice.

4.7.1.1 Prior for r

The weather states are just abstract constructs of the hidden Markov model and are not real in any physical sense. However, it is likely that we will be able to provide well defined, distinct

r	1	2	3	4	5	6
Computing Time	1.00	1.96	2.92	4.31	6.67	16.36

Table 4.2: Computing time required to produce 10,000 MCMC draws from the posterior distribution in the analysis of the Yorkshire dataset, conditional on various values of r. The time required when $r = 1$ is taken as a single unit of time, in real terms, around 2 minutes.

physical interpretations of each if the number of weather states is not too large. The hidden Markov model described in this chapter assumes conditional spatial and temporal independence in the precipitation process, given the weather state. This leaves the weather state as the model's only device for capturing the spatio–temporal structure in the data and represents a very simple within weather state model for the observed process. More complex models are available which, for example, allow dependence between the rainfall occurrence indicators, $D_t^1, \ldots, D_t^n$, given the weather state. The model chosen for $(\mathbf{D}_t \mid S_t = k, \theta_r, r)$ and $(\mathbf{W}_t \mid \mathbf{D}_t, S_t = k, \theta_r, r)$ in each state, k, governs the overall *shape* of the contours in the bivariate cross–sections of the joint densities $\Pr(\mathbf{D}_t \mid S_t = k, \theta_r, r)$ and $p(\mathbf{w}_t \mid \mathbf{D}_t = \mathbf{d}_t, S_t = k, \theta_r, r)$. This concept is perhaps easiest to think about in the related problem of constructing models using finite mixtures of multivariate normal distributions. In the bivariate case, consider a situation in which the data appear to arise in elliptical clusters lying on rotations from the principal axes. If a diagonal variance matrix is assumed in the within component model, i.e. the components are constrained to follow the principal axes, then the marginal likelihood might suggest that a larger number of mixture components are required than if a rotation was allowed. Therefore there is a trade–off between the number of weather states (and hence model interpretability) and the complexity of the within–state model. Thus, to help in achieving our second objective of interpretability, we choose the upper bound for r to be reasonably small at $r_{\max} = 5$ and express a preference for a number of states towards the centre of the support, $\{1, \ldots, 5\}$, by choosing a truncated Poisson Po(3) prior, with mean at 2.823. In Section 4.7.2.2 we provide further comments on sensitivity to this prior specification. Note that ideally we would formalise our prior preferences for well defined, well differentiated weather states using the prior for $(\theta_r \mid r)$ (as discussed in Section 3.4.3) but we leave this as an area for future research.

One might argue that limiting the support of r to the small set $\{1, 2, \ldots, 5\}$ is overly restrictive, and indeed might be detrimental to the predictive ability of the model. However, if the model is still unable to provide a good description of the data with $r = 5$ states, then we regard this as symptomatic of an overly simple within weather state model for the precipitation process. The particular choice $r_{\max} = 5$ is motivated by practical as well as theoretical considerations. For the Yorkshire data, conditioning on each of $r = 1, \ldots, 6$, Table 4.2 contains the computing time required to pass through 10,000 sweeps of the MCMC algorithm (with online relabelling), relative to the time required when conditioning on $r = 1$. It is the online relabelling algorithm which is largely responsible for slowing down the execution of the MCMC scheme because, for every draw from the posterior, it is necessary to search through all $r!$ permutations of the MCMC output in order to find that which is the most consistent with the marginal posterior mode estimate for the weather state sequence, $\hat{\mathbf{s}}$. It appears that the computing time sharply increases between $r = 5$ and $r = 6$ states.

4.7.1.2 Prior for $(\theta_r \mid r)$

For the purposes of the following analysis, two important features of our prior distribution for $(\theta_r \mid r)$, $r = 1, \ldots, r_{\max}$, which was outlined in Section 4.3, are that, conditional on r,

(a) we assume *a priori* independence between the parameters of the observed and hidden processes, $\theta_{r,\text{obs}}$ and $\theta_{r,\text{hid}}$;

(b) we adopt a prior for $\theta_{r,\text{obs}}$ in which the parameters are independent and exchangeable across weather states.

We aim to choose the hyperparameters in the priors for $(\theta_r \mid r)$, $r = 1, \ldots, r_{\max}$, so that the first and second order moments in the prior predictive distribution for a single observation $(\mathbf{W}_t, \mathbf{D}_t)$ do not vary with r. Specifically, we are interested in matching each of the prior predictive moments

$$\mathrm{E}(D_t^i \mid r), \qquad \mathrm{E}(W_t^i \mid D_t^i = 1, r),$$
$$\mathrm{E}(D_t^i D_t^j \mid r), \qquad \mathrm{E}\{(W_t^i)^2 \mid D_t^i = 1, r\}, \qquad \mathrm{E}(W_t^i W_t^j \mid D_t^i = 1, D_t^j = 1, r).$$

Note that $\mathrm{E}\{(D_t^i)^2 \mid r\}$ does not appear in this list because $\mathrm{E}(D_t^i \mid r) = \mathrm{E}\{(D_t^i)^2 \mid r\}$ as D_t^i is binary.

Consider a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ from some joint distribution which is parameterised by θ. Suppose that θ is assigned a prior with density $\pi(\theta)$. By the Law of Total Expectation, for any i, j and any $q_i, q_j \in \mathbb{Z}$,

$$\begin{aligned} \mathrm{E}_{(X_i,X_j)}(X_i^{q_i} X_j^{q_j}) &= \mathrm{E}_\theta\{\mathrm{E}_{(X_i,X_j)|\theta}(X_i^{q_i} X_j^{q_j})\} \\ &= \int \mathrm{E}_{(X_i,X_j)|\theta}(X_i^{q_i} X_j^{q_j}) \pi(\theta) d\theta, \end{aligned} \tag{4.34}$$

which clearly reduces to

$$\mathrm{E}_{X_i}(X_i^{q_i}) = \int \mathrm{E}_{X_i|\theta}(X_i^{q_i}) \pi(\theta) d\theta \tag{4.35}$$

if $q_j = 0$.

In the prior predictive distribution for rainfall amount we have

$$\begin{aligned} \mathrm{E}(W_t^i \mid r) &= \sum_{d=0}^{1} \mathrm{E}(W_t^i \mid D_t^i = d, r) \Pr(D_t^i = d \mid r) \\ &= \mathrm{E}(W_t^i \mid D_t^i = 1, r) \Pr(D_t^i = 1 \mid r) && \text{since } \mathrm{E}(W_t^i \mid D_t^i = 0, r) = 0 \\ &= \mathrm{E}(W_t^i \mid D_t^i = 1, r) \mathrm{E}(D_t^i \mid r) && \text{since } D_t^i \text{ is binary} \end{aligned}$$

and so conditional on occurrence

$$\mathrm{E}(W_t^i \mid D_t^i = 1, r) = \frac{\mathrm{E}(W_t^i \mid r)}{\mathrm{E}(D_t^i \mid r)}.$$

It can similarly be shown that

$$\mathrm{E}\{(W_t^i)^2 \mid D_t^i = 1, r\} = \frac{\mathrm{E}\{(W_t^i)^2 \mid r\}}{\mathrm{E}(D_t^i \mid r)}$$

and

$$\mathrm{E}(W_t^i W_t^j \mid D_t^i = 1, D_t^j = 1, r) = \frac{\mathrm{E}(W_t^i W_t^j \mid r)}{\mathrm{E}(D_t^i D_t^j \mid r)}.$$

Therefore having computed the prior predictive moments for the occurrence process, $\mathrm{E}(D_t^i \mid r)$ etc., in order to compute the moments in the prior predictive distribution for rainfall amounts given occurrence, we simply need to calculate $\mathrm{E}(W_t^i \mid r)$, $\mathrm{E}\{(W_t^i)^2 \mid r\}$ and $\mathrm{E}(W_t^i W_t^j \mid r)$. Each of the expectations of interest can be obtained from equation (4.34) or (4.35) by averaging the corresponding conditional expectation (given the model parameters and r) over the prior distribution $\pi(\theta_r \mid r)$. For example the prior predictive means $\mathrm{E}(D_t^i \mid r)$ and $\mathrm{E}(W_t^i \mid r)$ are given by

$$\mathrm{E}(D_t^i \mid r) = \int \mathrm{E}(D_t^i \mid \theta_r, r)\pi(\theta_r \mid r)d\theta_r, \quad \mathrm{E}(W_t^i \mid r) = \int \mathrm{E}(W_t^i \mid \theta_r, r)\pi(\theta_r \mid r)d\theta_r.$$

For rainfall occurrences,

$$\mathrm{E}(D_t^i \mid \theta_r, r) = \sum_{k=1}^{r} \Pr(S_t = k \mid \theta_r, r)\mathrm{E}(D_t^i \mid S_t = k, \theta_r, r) = \sum_{k=1}^{r} g_k(t; \theta_{r,\mathrm{hid}})p_{r,ik}$$

where $g_k(t; \theta_{r,\mathrm{hid}}) = \Pr(S_t = k \mid \theta_r, r)$ and the notation indicates that $g_k(t; \theta_{r,\mathrm{hid}})$ will depend on t and be a function of the parameters of the hidden process, $\theta_{r,\mathrm{hid}} = (\Lambda_r, \nu_r)$. Note that if the chain was in its stationary distribution we would have $g_k(t; \theta_{r,\mathrm{hid}}) = \delta_{r,k}$ at all times, t, where $\delta_r = (\delta_{r,1}, \ldots, \delta_{r,r})$ is the solution to the matrix equation $\delta_r = \delta_r \Lambda_r$. Similarly,

$$\mathrm{E}(D_t^i D_t^j \mid \theta_r, r) = \sum_{k=1}^{r} g_k(t; \theta_{r,\mathrm{hid}})p_{r,ik}p_{r,jk}.$$

For rainfall amounts we have

$$\begin{aligned}
\mathrm{E}(W_t^i \mid \theta_r, r) &= \sum_{d=0}^{1}\sum_{k=1}^{r} \Pr(D_t = d, S_t = k \mid \theta_r, r)\mathrm{E}(W_t^i \mid D_t^i = d, S_t = k, \theta_r, r) \\
&= \sum_{k=1}^{r} \Pr(D_t = 1, S_t = k, \theta_r, r)\mathrm{E}(W_t^i \mid D_t^i = 1, S_t = k, \theta_r, r) \\
&= \sum_{k=1}^{r} \{\Pr(S_t = k \mid \theta_r, r)\Pr(D_t = 1 \mid S_t = k, \theta_r, r)\mathrm{E}(W_t^i \mid D_t^i = 1, S_t = k, \theta_r, r)\} \\
&= \sum_{k=1}^{r} g_k(t; \theta_{r,\mathrm{hid}})p_{r,ik}m_{r,ik}.
\end{aligned}$$

In the same way it can easily be shown that

$$\mathrm{E}\{(W_t^i)^2 \mid \theta_r, r\} = \sum_{k=1}^{r} g_k(t; \theta_{r,\mathrm{hid}}) p_{r,ik} (m_{r,ik})^2 \{(v_{r,ik})^2 + 1\}$$

and

$$\mathrm{E}(W_t^i W_t^j \mid \theta_r, r) = \sum_{k=1}^{r} g_k(t; \theta_{r,\mathrm{hid}}) p_{r,ik} p_{r,jk} m_{r,ik} m_{r,jk}.$$

For any i, j each of these expectations has the form

$$\mathrm{E}(\cdot \mid \theta_r, r) = \sum_{k=1}^{r} g_k(t; \theta_{r,\mathrm{hid}}) h(\theta_{r,\mathrm{obs},k})$$

where $\theta_{r,\mathrm{obs},k}$ is the collection of parameters in $\theta_{r,\mathrm{obs}}$ associated with the weather state k and where $h(\cdot)$ is a polynomial expression. Taking expectations with respect to the prior density yields

$$\begin{aligned}
\mathrm{E}(\cdot \mid r) &= \int_\theta \mathrm{E}(\cdot \mid \theta_r, r) \pi(\theta_r \mid r) d\theta \\
&= \int_\theta \sum_{k=1}^{r} g_k(t; \theta_{r,\mathrm{hid}}) h(\theta_{r,\mathrm{obs},k}) \pi(\theta_r \mid r) d\theta \\
&= \int_{\theta_{r,\mathrm{obs}}} \int_{\theta_{r,\mathrm{hid}}} \sum_{k=1}^{r} g_k(t; \theta_{r,\mathrm{hid}}) h(\theta_{r,\mathrm{obs},k}) \pi(\theta_{r,\mathrm{obs}} \mid r) \pi(\theta_{r,\mathrm{hid}} \mid r) d\theta_{r,\mathrm{obs}} d\theta_{r,\mathrm{hid}} \\
&= \sum_{k=1}^{r} \int_{\theta_{r,\mathrm{hid}}} g_k(t; \theta_{r,\mathrm{hid}}) \pi(\theta_{r,\mathrm{hid}} \mid r) \int_{\theta_{r,\mathrm{obs},k}} h(\theta_{r,\mathrm{obs},k}) \pi(\theta_{r,\mathrm{obs},k} \mid r) d\theta_{r,\mathrm{obs},k} d\theta_{r,\mathrm{hid}} \\
&= \sum_{k=1}^{r} \int_{\theta_{r,\mathrm{hid}}} g_k(t; \theta_{r,\mathrm{hid}}) \pi(\theta_{r,\mathrm{hid}} \mid r) \mathrm{E}_{\theta_{r,\mathrm{obs},k} \mid r}\{h(\theta_{r,\mathrm{obs},k})\} d\theta_{r,\mathrm{hid}},
\end{aligned}$$

where the third and fourth lines follows from the assumptions of *a priori* independence outlined in (a) and (b), respectively. Our assumption of *a priori* exchangeability between weather states means that

$$\mathrm{E}_{\theta_{r,\mathrm{obs},1} \mid r}\{h(\theta_{r,\mathrm{obs},1})\} = \cdots = \mathrm{E}_{\theta_{r,\mathrm{obs},r} \mid r}\{h(\theta_{r,\mathrm{obs},r})\} = F_h, \text{ say,}$$

where F_h is independent of k, so we have

$$\begin{aligned}
\mathrm{E}(\cdot \mid r) &= F_h \sum_{k=1}^{r} \int_{\theta_{r,\mathrm{hid}}} g_k(t; \theta_{r,\mathrm{hid}}) \pi(\theta_{r,\mathrm{hid}} \mid r) d\theta_{r,\mathrm{hid}} \\
&= F_h \left[\sum_{k=1}^{r-1} \int_{\theta_{r,\mathrm{hid}}} g_k(t; \theta_{r,\mathrm{hid}}) \pi(\theta_{r,\mathrm{hid}} \mid r) d\theta_{r,\mathrm{hid}} \right. \\
&\qquad \left. + \int_{\theta_{r,\mathrm{hid}}} \left\{ 1 - \sum_{k=1}^{r-1} g_k(t; \theta_{r,\mathrm{hid}}) \right\} \pi(\theta_{r,\mathrm{hid}} \mid r) d\theta_{r,\mathrm{hid}} \right] \\
&= F_h.
\end{aligned}$$

Therefore, the first and second order moments in the prior predictive distributions for rainfall occurrence, and rainfall amount, given occurrence, on a single day do not depend on the prior chosen for the parameters of the hidden process, $\theta_{r,\text{hid}}$. This seems to be a reasonable result; intuitively a prior chosen for the parameters of the hidden process would be expected to influence only the probability of that day being assigned to a particular weather state. If our prior for the parameters of the observed process makes no distinction between weather states, then the prior for $\theta_{r,\text{hid}}$ will have no bearing on these prior predictive moments.

We can therefore match the prior predictive moments across different values of r simply by choosing the same conditional priors $\pi(\theta_{r,\text{obs},k} \mid r)$, for every $k = 1, \ldots, r$, and each $r = 1, \ldots, r_{\max}$.

We do not have prior knowledge that would allow us to distinguish between the rainfall climate at any of the six sites, so we chose the hyperparameters in the joint prior specification for $(\theta_{r,\text{obs}} \mid r)$ so that the parameters were exchangeable across sites as well as weather states. To help in forming prior opinions about these parameters, information is available from another site in the Yorkshire region. For the winters (December–February) of 1961/62 to 1990/91 the overall proportion of wet days at the extra site, Leeming Reservoir, is 0.494 whilst for rainfall amounts on wet days the mean and coefficient of variation are 3.312 mm and 1.229, respectively. We chose the hyperparameters in the prior for $(\mathcal{P}_r \mid r)$ so that the mean was 0.5, approximately matching the probability of rain at Leeming Reservoir. Although this could be achieved by setting $a_{r,i1} = a_{r,i2} = a$, $i = 1, 2, \ldots, n$, for any $a > 0$, we chose to take $a = 1$ in order to maximise the prior variance subject to the constraint $a \geq 1$. This avoids U–shaped prior distributions which would not be representative of our prior beliefs. For the parameters in $\mathcal{M}_r$ and $\mathcal{V}_r$ we chose the medians of $(m_{r,ik} \mid r)$ and $(v_{r,ik} \mid r)$ to match the mean and coefficient of variation statistics, respectively, for the Leeming Reservoir data. We chose the 95–th percentile of each $(m_{r,ik} \mid r)$ to be 10 mm and the 5–th percentile of each $(v_{r,ik} \mid r)$ to be 0.5 mm. Solving the appropriate systems of equations numerically using R gave hyperparameters $b_{r,1i} = 3.340$, $b_{r,2i} = 10.000$, $c_{r,1i} = 4.671$ and $c_{r,2i} = 3.533$ for each site, $i = 1, 2, \ldots, n$, in each weather state.

Based on our intuitive notion of the length of time it takes a weather front to traverse a region, we chose the expected mean sojourn time in any particular weather state to be 2.5 days and equivalent to $45/r$ weather state transitions per row of Λ_r. This gives an overall prior specification equivalent to 45 weather state transitions which does not represent a particularly strong degree of belief in our prior estimate, especially when we consider that the time series of observed data contains information on $2707 - 30 = 2677$ transitions, where subtraction of 30 from $T = 2707$ accounts for there being $T_y - 1$ transitions in the y–th winter, $y = 1, \ldots, 30$. Moreover, this prior specification guarantees that for all $r = 1, \ldots, r_{\max}$, both parameters in the marginal (Beta) prior for any transition probability, λ_{rjk}, are greater than one. Again, this avoids U–shaped marginal priors which would not have been in keeping with our prior beliefs.

Finally in the prior for ν_r we chose the information content parameter G_r to be equal to r, giving a $\mathcal{D}_r(1, \ldots, 1)$ prior for each model. This choice maximises the prior variance of any component of $\nu_{r,k}$ subject to the constraints imposed by exchangeability ($\mathrm{E}(\nu_{r,k} \mid r) = 1/r$) and avoidance of U–shaped marginals ($G_r/r \geq 1$).

4.7.2 Posterior inference for r

In the following section we provide details regarding the estimation of the log marginal likelihood for each model via power posteriors, including a method for estimating the Monte Carlo standard errors. We then describe the results when the method is applied to the Yorkshire data.

4.7.2.1 Implementation

The power posterior at temperature t is defined as

$$\pi_t(\theta_r, \mathbf{s} \mid \mathbf{w}, \mathbf{d}, r) \propto p(\mathbf{w}, \mathbf{d} \mid \theta_r, \mathbf{s}, r)^t p(\mathbf{s} \mid \theta_{r,\text{hid}}, r)\pi(\theta_{r,\text{hid}} \mid r)\pi(\theta_{r,\text{obs}} \mid r).$$

For each value of $r = 1, \ldots, r_{\max}$, estimation of the expected half deviances proceeds according to Algorithm 3.5.1 in which the weather states $\mathbf{s}$ are treated as parameters in the model. The estimates of the expected half deviances are combined to produce the overall estimate of the log marginal likelihood using the trapezoidal rule, (3.38). At each temperature, t, step 3 of the algorithm involves generating a sample from the power posterior with stationary distribution $\pi(\theta_r, \mathbf{s} \mid \mathbf{w}, \mathbf{d}, r, t)$. Implementation of this step was described in Section 4.6.4 for a simpler model in which rainfall amounts were modelled using the exponential instead of the gamma distribution. Therefore the only modification needed here is to replace draws from the full conditional distributions of the inverse scale parameters, $\mathcal{B}_r = (\beta_{r,ik})$, in the exponential distributions with draws from the full conditional distributions of $\mathcal{M}_r = (m_{r,ik})$ and $\mathcal{V}_r = (v_{r,ik})$ which parameterise the gamma distributions,

$$m_{ik} \mid \cdots \sim \text{IG}\left(b_{1i} + \frac{tT^1_{ik}(\mathbf{s})}{v^2_{ik}}, b_{2i} + \frac{tT^1_{ik}(\mathbf{s})\bar{w}_{ik}(\mathbf{s})}{v^2_{ik}}\right), \qquad (i,k) \in \{1, \ldots, n\} \times \mathcal{S}_r, \tag{4.36}$$

and

$$\pi(v_{ik} \mid \cdots) \propto \Gamma\left(\frac{1}{v^2_{ik}}\right)^{-tT^1_{ik}(\mathbf{s})} v_{ik}^{-\{2tT^1_{ik}(\mathbf{s})/v^2_{ik} - c_{1i} + 1\}} m_{ik}^{-tT^1_{ik}(\mathbf{s})/v^2_{ik}} \bar{w}_{g,ik}(\mathbf{s})^{tT^1_{ik}(\mathbf{s})/v^2_{ik}}$$
$$\times \exp\left\{-\left(c_{2i}v_{ik} + \frac{tT^1_{ik}(\mathbf{s})\bar{w}_{ik}(\mathbf{s})}{v^2_{ik}m_{ik}}\right)\right\} \tag{4.37}$$

for $(i,k) \in \{1, 2, \ldots, n\} \times \mathcal{S}_r$. For each (i,k), the latter non-standard distribution is sampled in a Metropolis Hastings step, generating proposals using a random walk on the gamma scale; see Section 4.5.1. The tuning parameters in this step, ω^i_v, $i = 1, \ldots, n = 6$, were chosen to be $(60, 45, 70, 35, 60, 25)$.

Following the encouraging results of the simulation experiment, we again chose a temperature schedule $t_i = (i/n)^c$ for $i = 0, \ldots, n$ where $n = 40$ and $c = 4$. Within each temperature we collected 10,000 samples from the power posterior of which the first 4,000 were discarded as burn–in. Sensitivity to the spacing of the temperatures in $[0, 1]$ will be discussed further in the following section.

In addition to computing estimates of the log marginal likelihood, we also calculated estimates of the associated Monte Carlo standard error. The Monte Carlo standard error measures the

variability in an estimate of an integral when estimation is via Monte Carlo simulation. Given the trapezoidal rule used to numerically integrate over t, Friel & Pettitt (2008) estimate the overall Monte Carlo standard error by piecing together the individual Monte Carlo standard errors, MCSE_i, for each estimated expectation, $\text{E}_{\theta_r,\mathbf{s}|\mathbf{w},\mathbf{d},r,t_i}\{\log p(\mathbf{w},\mathbf{d} \mid \theta_r,\mathbf{s},r)\}$. The overall Monte Carlo standard error is estimated by

$$\sqrt{\left\{\frac{(t_1-t_0)^2}{4}\text{MCSE}_0^2+\sum_{i=1}^{n-1}\frac{(t_{i+1}-t_{i-1})^2}{4}\text{MCSE}_i^2+\frac{(t_n-t_{n-1})^2}{4}\text{MCSE}_n^2\right\}}.$$

There is a vast literature on methods for estimating the Monte Carlo standard error, for example see Geyer (1992) or Jones *et al.* (2006). Here we choose to estimate each MCSE_i using the simple batch means method with 50 batches; see Roberts (1996) for further details.

4.7.2.2 Results

The expected half deviance, $\text{E}_{\theta_r,\mathbf{s}|\mathbf{w},\mathbf{d},r,t}\{\log p(\mathbf{w},\mathbf{d} \mid \mathbf{s},\theta_r,r)\}$, is plotted against temperature, t, in Figure 4.7 for all the hidden Markov models, $r = 1,\ldots,r_{\max}$. The plot for $r = 1$ has a similar shape to the plots presented by Friel & Pettitt (2008) in which the expected half deviance increases sharply near zero before starting to level off. However in Figure 4.7, the plots for $r > 1$ do not increase smoothly. Following the initial rapid increase and subsequent flattening in shape, the gradient rises sharply again near $t = 0.2$, and once more near $t = 0.4$ for $r > 3$. The plots for the simulated data described in Section 4.6 (not shown) displayed similar patterns when the data were generated using parameter set two, corresponding to hidden Markov models in which there was a large difference between the parameters in the two states. One possible explanation for this behaviour derives from the exchangeability of the prior distribution, which means there is no distinction between any of the hidden states *a priori*. It is conceivable that at lower temperatures, the likelihood contribution is downweighted to such an extent that, when combined with an exchangeable prior, it is not possible to distinguish between any of the states in the power posterior. However, as the temperature is increased, and the likelihood is allowed to impart more influence, a point is reached at which more than one state can be identified. This provides a better explanation of the data and so the expected half deviance, $\text{E}_{\theta_r,\mathbf{s}|\mathbf{w},\mathbf{d},r,t}\{\log p(\mathbf{w},\mathbf{d} \mid \mathbf{s},\theta_r,r)\}$, increases sharply. The additional "jump" when $r > 3$ may arise when it becomes possible to identify yet more states in the power posterior. When the simulation experiments in Section 4.6 were repeated using a non-exchangeable prior that asserted *a priori* belief in one "wet" and one "dry" weather state, plots showed the expected half deviance increasingly smoothly with temperature. Although not intended to provide a definitive justification, this does give some evidence in support of our conjecture.

The temperature schedule, $t_i = (i/n)^c$ with $c > 1$, is designed to ensure many of the chosen temperatures are close to zero, so will be most appropriate in situations where the behaviour of the expected half deviance under each power posterior is "typical", with a single sharp increase near zero. Therefore, for $r > 1$ it is possible that placing more temperatures around the other sharp increases could improve the efficiency of the log marginal likelihood estimates. However, for the models with $r = 2,\ldots,r_{\max}$ states, inserting an extra ten equally spaced temperatures

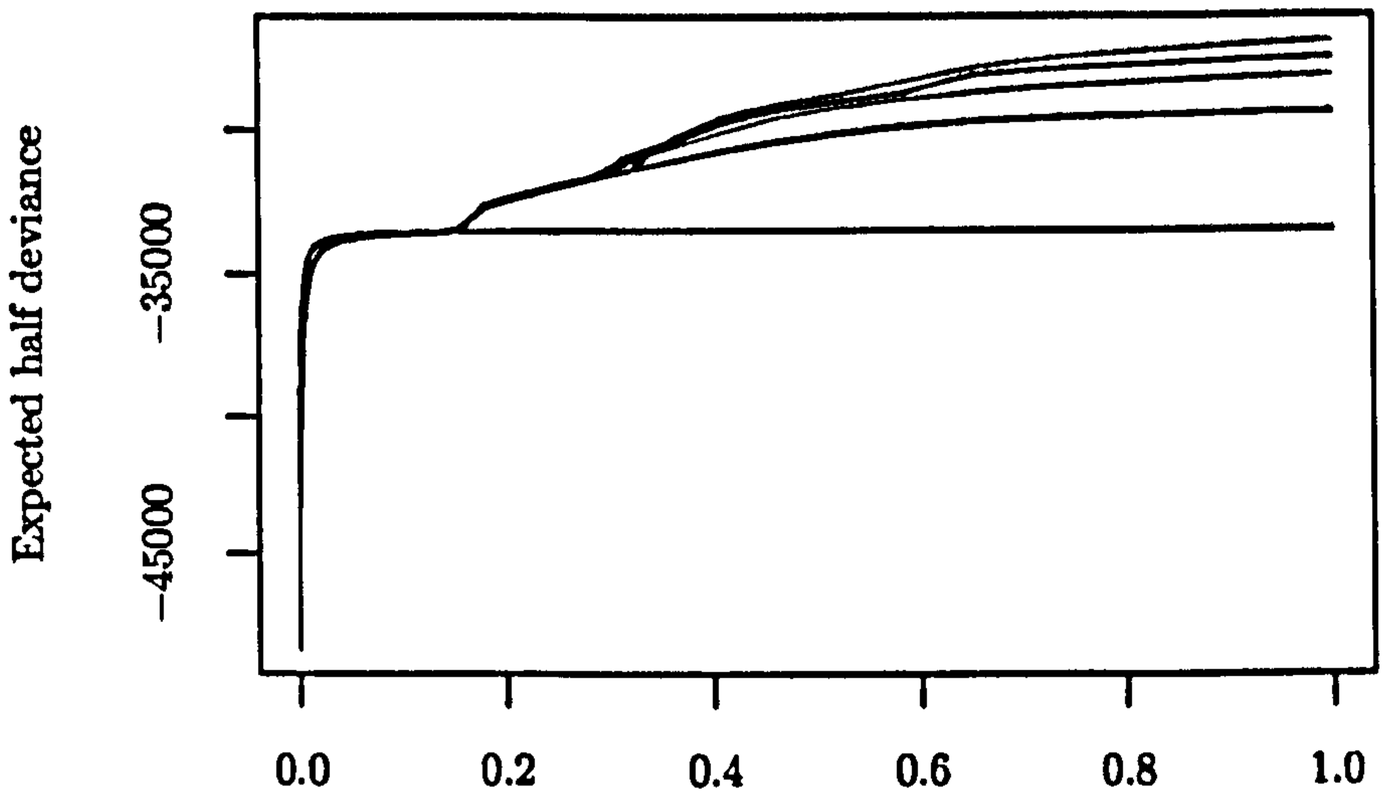


Figure 4.7: From analyses of the Yorkshire data, expected half deviance against temperature for the hidden Markov model with $r = 1$ (———), $r = 2$ (———), $r = 3$ (———), $r = 4$ (———) and $r = 5$ (———) states.

r	1	2	3	4	5
Log marginal likelihood	-33534.08	-30930.88	-30242.23	-29927.35	-29709.68
Monte Carlo std. error	0.10	0.68	0.99	0.68	0.74
Posterior probability	0.00	0.00	1.16×10^{-231}	4.87×10^{-95}	1.00

Table 4.3: Estimates of the log marginal likelihood, the associated Monte Carlo standard error and the posterior distribution for r for the Yorkshire data. The estimates of the log marginal likelihoods were computed via power posteriors.

in the vicinity of the "jumps" and recomputing the estimates led to negligible changes in their values.

Estimates of the log marginal likelihoods, the associated Monte Carlo standard errors and the posterior distribution for r are presented in Table 4.3. Even taking account of the Monte Carlo sampling variability, the magnitude of the differences between the marginal likelihood estimates on the log scale is such that the posterior evidence in favour of $r = 5$ is overwhelming. By the same argument, the posterior for r is not sensitive to the choice of prior $\pi_r(r)$. For example, even if a monotonically decreasing probability mass function, such as a truncated Poisson Po(1) distribution was chosen, the posterior mass would still be stacked up at $r = 5$.

Figure 4.8 displays a plot of the log marginal likelihood estimates against r and suggests that the rate of increase decays with increasing r. However, it appears that extending the support of the prior for r (allowing values greater than $r_{max} = 5$) might result in an increase in the value which maximises the log marginal likelihood. The reason for this, and indeed the reason why the posterior distribution for r offers virtually no support to $r < r_{max}$, is likely to be some

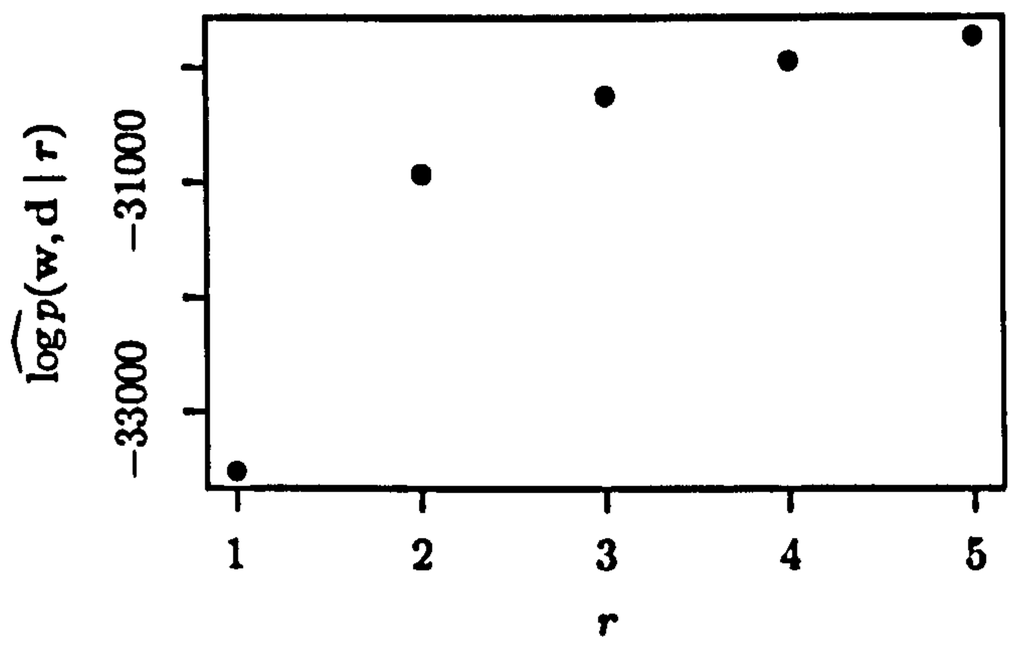


Figure 4.8: Estimates of the log marginal likelihood for the Yorkshire data calculated using the power posterior approach.

combination of the choice of priors and model misspecification. This can be explained as follows. As discussed in Section 3.4, the maximum of the likelihood for an overfitting hidden Markov model with r states will be the same as that for a model with $r-1$ states and so if there were no other unknown parameters over which to marginalise, the likelihood would be non-decreasing as r increased. In reality, however, the models *do* contain other unknown parameters and so what happens to the marginal likelihood as r gets larger depends on the priors assigned to these parameters. Although we did attempt to balance the information in the priors for models with $r = 1, \ldots, 5$ states, it is very difficult to assess the effect of the prior on the marginal likelihood and so it is possible that we inadvertently assigned priors which favoured large values of r. In terms of model misspecification, as explained in Section 4.7.1.1, there is likely to be a trade–off between the number of weather states and the flexibility of the within–state model. In other words, if the within–state distributions provide a poor description of the precipitation patterns in the data, then it is likely that adding extra states will allow refinements in the shape of the resulting mixture distribution. This in turn may increase the likelihood. Since our within–state model is very simple and leaves the weather state as the model's only device for capturing spatio–temporal dependence, this effect is likely to be very strong. Consequently, model misspecification is likely to be the dominant factor causing the posterior distribution for r to offer so little support to $r < r_{\max}$. In Chapter 5 we will consider more complex models for the within weather state rainfall occurrence process and anticipate that this will allow the accumulation of greater posterior mass at smaller values of r.

4.7.3 Posterior inference for $(\theta_r, \mathbf{s} \mid r)$ using MCMC samples

In the following section we focus on the hidden Markov model with $r = 5$ states since the posterior probability for this model is essentially equal to one. First we give details of the implementation of the MCMC scheme. We then illustrate how the model might be useful in improving understanding of the spatio–temporal properties of rainfall in the region, by presenting and discussing the posterior distributions for the model parameters, θ_r, and for the weather states, $\mathbf{s}$.

4.7.3.1 Implementation, convergence and mixing

Fixing $r = 5$, the MCMC algorithm was run from a variety of starting points each of which produced essentially the same results. Each run comprised 1,000,000 iterations. The first $500,000$ were discarded as burn–in and for the remaining $500,000$ iterations, only every 50–th iterate was recorded in order to reduce computing overheads. Our posterior inferences are based on one such run of $N = 10,000$ sampled values. We want to be able to make inferences about the model parameters and the weather states and so need to identify one distinct labelling of the states. Therefore we addressed the problem of label switching by implementing the online relabelling algorithm (Algorithm 3.3.4). The graphical diagnostic checks discussed in Section 4.5.3 gave no evidence of lack of convergence, and the posterior distributions showed no signs of multimodality. Based on the autocorrelation plots, thinning to every 50–th iterate appeared to produce an approximately uncorrelated posterior sample. For example, Figure 4.9 displays trace, autocorrelation and density plots for a representative parameter, $m_{5,53}$. The tuning parameters for the six sites, ω_v^i, were chosen to be $(60, 45, 70, 35, 60, 25)$ leading to acceptance rates for the coefficient of variation parameters which ranged from 0.170 to 0.587.

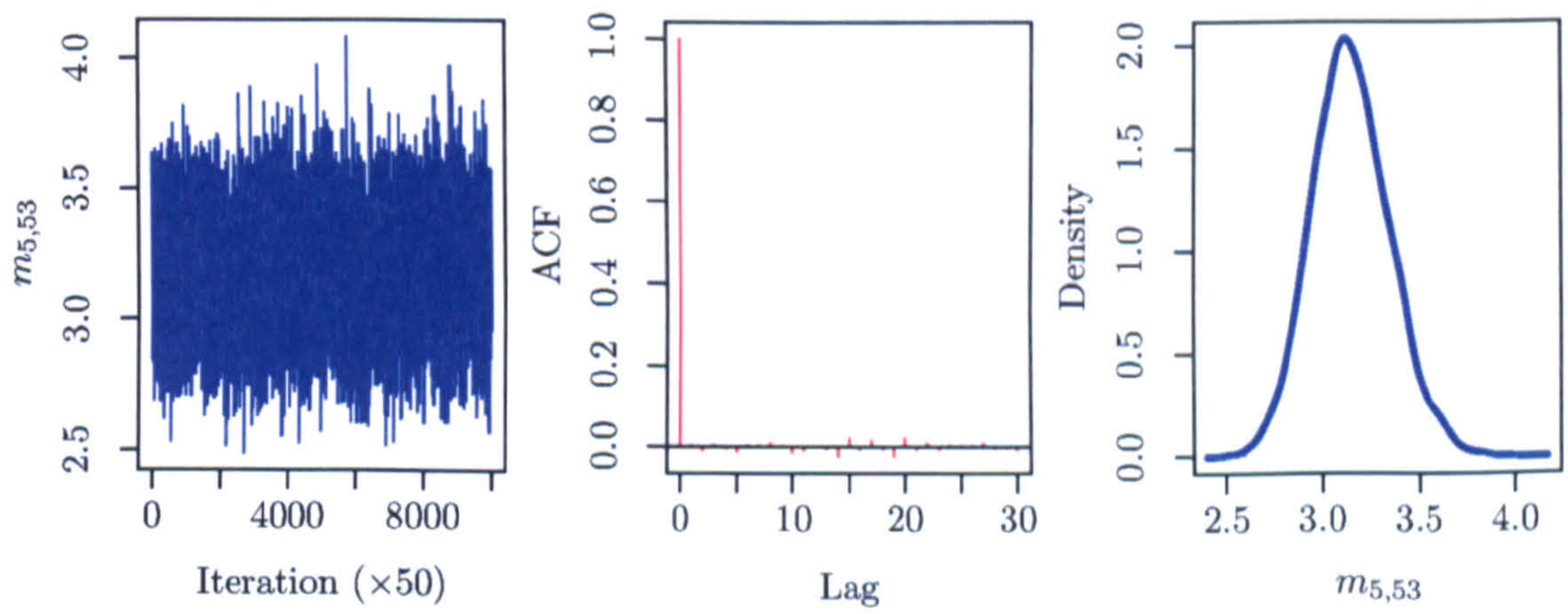


Figure 4.9: Graphical convergence checks for the parameter $m_{5,53}$ in a fixed dimensional analysis of the Yorkshire data with $r = 5$ weather states.

Although not presented here we also performed fixed dimensional analyses of the models with $r = 1, \ldots, 4$ states. The results were similar in terms of the convergence and mixing of the sampler, although the posterior densities for some of the parameters in the model with $r = 4$ states displayed evidence of bimodality. Multimodality in the posterior distributions of parameters in latent variable models is not uncommon, for example see Richardson & Green (1997) or Celeux *et al.* (2000). It generally arises due to the existence of two or more competing descriptions of the data which are comparable in terms of their posterior support. Exploring the two posterior modes in our case we found one was associated with a description of the data which had more support from the prior, and the other was associated with a description which had more support from the likelihood.

4.7.3.2 Posterior for $(\theta_5 \mid r = 5)$

Figure 4.10(a) displays the marginal posterior distributions for the rainfall occurrence probabilities, $\mathcal{P}_5$, through the estimated posterior means together with 95% equi-tailed Bayesian credible regions. Figures 4.10(b) and 4.10(c) display the corresponding plots for the mean and coefficient of variation parameters, $\mathcal{M}_5$ and $\mathcal{V}_5$, in the gamma distributions for non-zero rainfall amounts. From these plots it appears that the weather state labelled 1 is characterised by wet conditions at all sites with high probabilities of rain and large rainfall amounts on wet days. The weather state labelled 2 is also associated with wet conditions at all sites, but typically less rain falls on wet days than in weather state 1.

Weather state 4 can be characterised as "dry", representing opposite conditions to those associated with state 1. In weather state 5, it is typically dry at all but the two Pennine sites, Moorland Cottage (site 3) and Great Walden Edge (site 5), where the posterior means for the rainfall occurrence parameters are both in excess of 0.7. Compared to the other weather states, weather state 3 seems to be associated with "average" conditions at most sites which are neither particularly wet nor particularly dry. The exception is the most Northerly site, Lockwood Reservoir (site 1), at which state 3 is the weather state with the highest probability of rain.

In Chapter 2 we found that one of the Pennine sites, Moorland Cottage (site 3), typically experiences much larger daily rainfall totals on wet days than the other sites. This observation is borne out through Figure 4.10(b) which shows that all but one weather state has an associated posterior mean for $m_{5,3k}$ which is at least equal to 10.838 mm. In fact, at this site there is little difference between the marginal posteriors for $m_{5,3k}$, $k \in \mathcal{S}_5 \setminus 4$ or for $p_{5,3k}$, $k \in \mathcal{S}_5 \setminus 4$.

From the plots in Figures 4.10(b) and 4.10(c), there is still considerable posterior uncertainty surrounding the parameters in the rainfall amounts distribution for site 6 in weather state 4. This can almost certainly be explained by the posterior for $p_{5,64}$ which is concentrated about a mean of 0.026. In other words, it rarely rains at site 6 in weather state 4 and, therefore, there is a paucity of information in the data for updating our prior beliefs about the parameters $m_{5,64}$ and $v_{5,64}$. Coupled with knowledge that weather states will represent similar conditions at all sites, this provides an argument in favour of priors which facilitate *borrowing of strength* between sites within weather states. For example, a prior in which $(m_{r,ik} \mid r)$ were positively correlated between sites within each weather state, k, would allow knowledge that $m_{r,1k}, \ldots, m_{r,i-1\,k}, m_{r,i+1\,k}, \ldots, m_{r,nk}$ were, say smaller than their mean, to update our beliefs about $m_{r,ik}$ so that we would also expect $m_{r,ik}$ to be smaller than its mean.

The posterior distributions for the parameters in $\mathcal{P}_5$ and $\mathcal{M}_5$ are, in general, well separated between weather states. However for every site, there is considerable overlap across weather states in the marginal posterior densities for the coefficient of variation parameters. This indicates that, relative to the mean, the within weather state variability in the gamma distribution for non-zero rainfall amounts is similar in all states.

Table 4.4 displays summaries of the marginal posterior distributions for the parameters of the hidden process, Λ_5 and ν_5. The "dry" weather state, state 4, is clearly very persistent with a posterior mean for the probability of self-transition equal to 0.704. In contrast, the posterior for the probability of self-transition in the wettest weather state, state 1, is concentrated about a

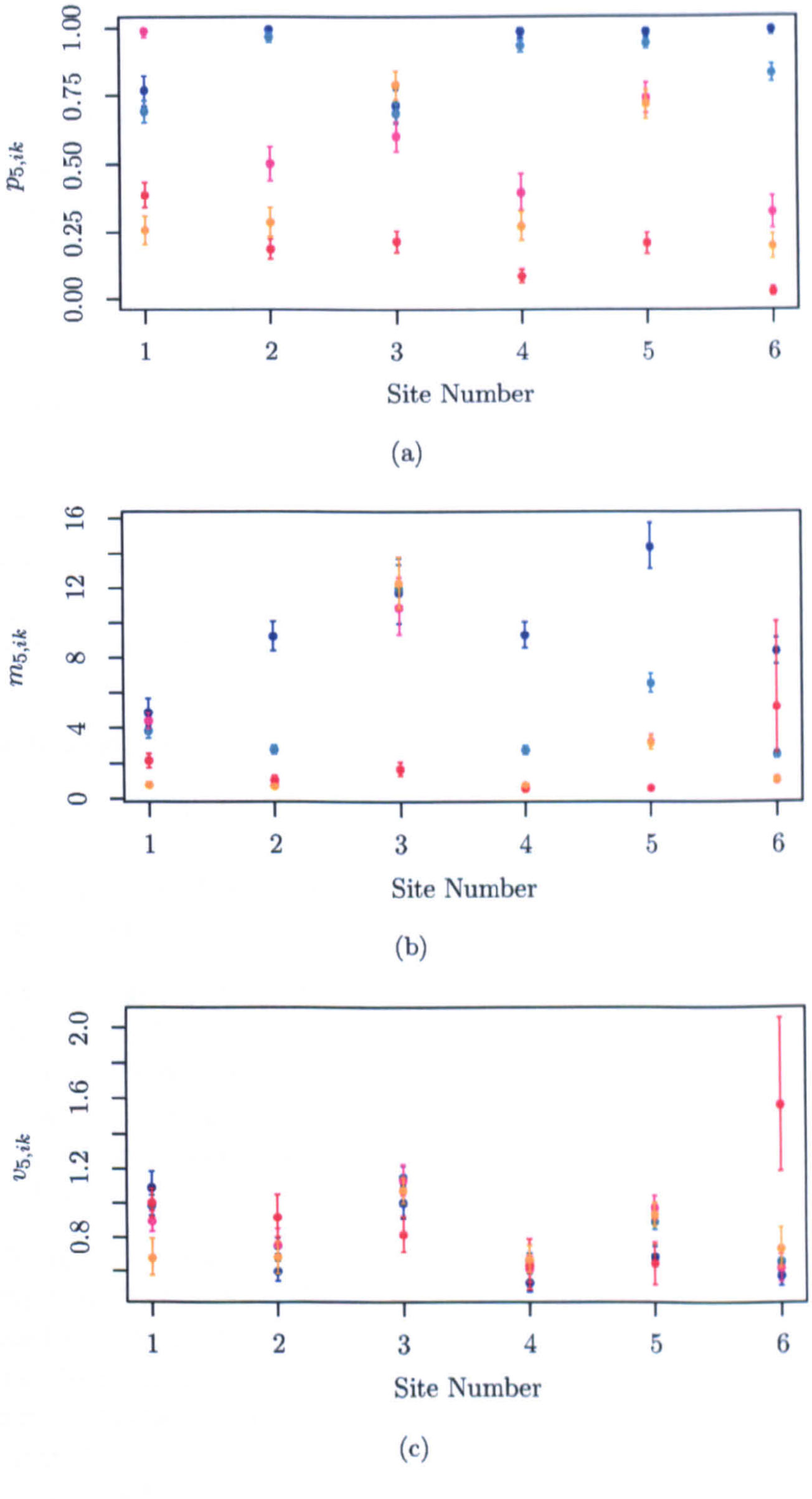


Figure 4.10: Conditional on $r = 5$, posterior means with 95% equi–tailed Bayesian credible intervals for the parameters in (a) $\mathcal{P}_5$; (b) $\mathcal{M}_5$; and (c) $\mathcal{V}_5$ in weather states 1 (———), 2 (———), 3 (———), 4 (———) and 5 (———).

much smaller mean of 0.292, so it appears that runs in this weather state are quickly terminated. These observations are quantified by evaluating $1/(1 - \lambda_{5,jj})$, $j \in \mathcal{S}_5$, for each draw from the posterior. This gives a sample from the posterior of the mean sojourn time in each of the weather

Parameter	Means	Standard Deviations
$\boldsymbol{\Lambda}_5$	$\begin{pmatrix} 0.292 & 0.361 & 0.331 & 0.008 & 0.007 \\ 0.133 & 0.378 & 0.414 & 0.056 & 0.019 \\ 0.082 & 0.269 & 0.174 & 0.224 & 0.252 \\ 0.014 & 0.076 & 0.010 & 0.704 & 0.196 \\ 0.102 & 0.298 & 0.015 & 0.097 & 0.488 \end{pmatrix}$	$\begin{pmatrix} 0.034 & 0.040 & 0.036 & 0.007 & 0.006 \\ 0.015 & 0.024 & 0.025 & 0.016 & 0.010 \\ 0.016 & 0.026 & 0.029 & 0.027 & 0.030 \\ 0.006 & 0.014 & 0.007 & 0.022 & 0.022 \\ 0.017 & 0.026 & 0.011 & 0.019 & 0.029 \end{pmatrix}$
$\boldsymbol{\nu}_5$	$\begin{pmatrix} 0.032 & 0.301 & 0.183 & 0.299 & 0.185 \end{pmatrix}$	$\begin{pmatrix} 0.030 & 0.088 & 0.079 & 0.084 & 0.074 \end{pmatrix}$
$\boldsymbol{\delta}_5$	$\begin{pmatrix} 0.104 & 0.264 & 0.180 & 0.255 & 0.198 \end{pmatrix}$	$\begin{pmatrix} 0.010 & 0.013 & 0.011 & 0.019 & 0.014 \end{pmatrix}$

Table 4.4: Conditional on $r = 5$, posterior means and standard deviations for the transition matrix, $\boldsymbol{\Lambda}_5$, the initial distribution, $\boldsymbol{\nu}_5$ and the solution to the matrix equation, $\boldsymbol{\delta}_5 = \boldsymbol{\delta}_5\boldsymbol{\Lambda}_5$.

states. The means (and standard deviations) are

1.416 (0.068), 1.610 (0.062), 1.212 (0.042), 3.402 (0.251), 1.959 (0.111)

for states 1–5. The means of these posterior distributions are reasonably consistent with our prior expectation of 2.5 days.

As we would intuitively expect, transitions from the two wettest weather states (states 1 and 2) to the two driest (states 4 and 5) are rare. Similarly transitions from the driest two weather states (states 4 and 5) to the wettest (state 1) are rare. The posteriors for transition probabilities into state 3 from states 1 and 2 and from state 3 into states 4 and 5 are concentrated about reasonably large values, so it seems that weather state 3 provides a route from the wet to the more dry weather states.

The solution to the matrix equation $\boldsymbol{\delta}_5 = \boldsymbol{\delta}_5\boldsymbol{\Lambda}_5$ will be the unique stationary distribution of a Markov model with transition matrix $\boldsymbol{\Lambda}_5$ if the Markov chain is irreducible and aperiodic. We can obtain a sample from the posterior distribution of $\boldsymbol{\delta}_5$ by evaluating the solution to this matrix equation at all draws from the posterior for $\boldsymbol{\Lambda}_5$. The posterior means and standard deviations for $\boldsymbol{\nu}_5$ and $\boldsymbol{\delta}_5$ are shown in Table 4.4. Although $\boldsymbol{\nu}_5$ is not formally constructed as an approximation to the stationary distribution of the chain, it is reassuring that there is considerable overlap in the posterior distributions for $\boldsymbol{\nu}_5$ and $\boldsymbol{\delta}_5$. The posterior distributions for the components of $\boldsymbol{\nu}_5$ have larger variances than those for the components of $\boldsymbol{\delta}_5$, but this is to be expected since only the 30 weather states on December 1st 1961–1990 contribute directly to the component of the complete data likelihood involving $\boldsymbol{\nu}_5$, whereas 2677 weather state transitions contribute to the component involving $\boldsymbol{\Lambda}_5$.

4.7.3.3 Posterior for $(\mathbf{s} \mid r = 5)$

A useful summary of the posterior distribution for the weather state sequence is the marginal posterior mode (MPM) estimate, $\hat{\mathbf{s}}$, which is available as a by-product of the relabelling algorithm (Algorithm 3.3.4). Given the results presented in Section 4.7.3.2, it is no surprise that the MPM estimate, displayed in Figure 4.11(a), shows that the sojourn times in the "dry" weather state, labelled 4, are, in general longer than those in any other weather state, and that spells in the wetter weather states (1 and 2) are usually separated from spells in the drier states (4 and 5) by days in weather state 3. It also appears that, in each year, long spells in the dry weather state become more prevalent towards the end of February, which is likely to be due to the advance of spring. This might be regarded as evidence against homogeneity in the weather state sequence. In the next chapter we will consider non-homogeneous hidden Markov models in which atmospheric information influences the transition probabilities between the weather states. It is possible that these atmospheric variables will be able to explain such seasonal effects.

Posterior uncertainty about the weather states can be summarised by the marginal posterior probabilities, $\Pr(S_t = j \mid \mathbf{w}, \mathbf{d}, r = 5)$, $j \in \mathcal{S}_5$, at each time point. Since the weather states are sampled during MCMC, a simple estimate of these probabilities is given by

$$\widehat{\Pr}(S_t = j \mid \mathbf{w}, \mathbf{d}, r) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(S_t^{[i]} = j), \quad j \in \mathcal{S}_r \tag{4.38}$$

for $t = 1, \ldots, T$. These estimates can be computed online (because the sampler relabels online) to avoid having to store the posterior draws of $\mathbf{s}$. The Rao–Blackwellised equivalent,

$$\widehat{\Pr}_{RB}(S_t = j \mid \mathbf{w}, \mathbf{d}, r) = \frac{1}{N} \sum_{i=1}^{N} \Pr(S_t = j \mid \mathbf{w}, \mathbf{d}, \theta_r^{[i]}, r),$$

provides a more precise estimate. However, for each posterior draw, $\theta_r^{[i]}$, computation of the *full sample smoothed probabilities*, $\Pr(S_t = j \mid \mathbf{w}, \mathbf{d}, \theta_r^{[i]}, r)$ for $t = 1, \ldots, T$ requires a recursive algorithm; see Frühwirth-Schnatter (2006) for full details. Clearly this estimator is more computationally expensive than the simpler estimator in (4.38). Therefore we choose the simple estimator to summarise our posterior uncertainty about the weather state allocation.

Figures 4.11(b) and 4.11(c) show the estimates of $\Pr(S_t = j \mid \mathbf{w}, \mathbf{d}, r = 5)$ on every day in the first and last winters in the dataset, respectively. Similar patterns were observed in other years. It appears that there is, in general, little posterior uncertainty regarding the weather state on most days, with the probability of one particular weather state on many days being at or near 1. If there is uncertainty in the posterior distribution for the weather state on any particular day, it tends to be uncertainty between weather states associated with parameters in $\theta_{r,\text{obs}}$ whose posteriors display the greatest overlap. For example between the two wetter weather states, states 1 and 2.

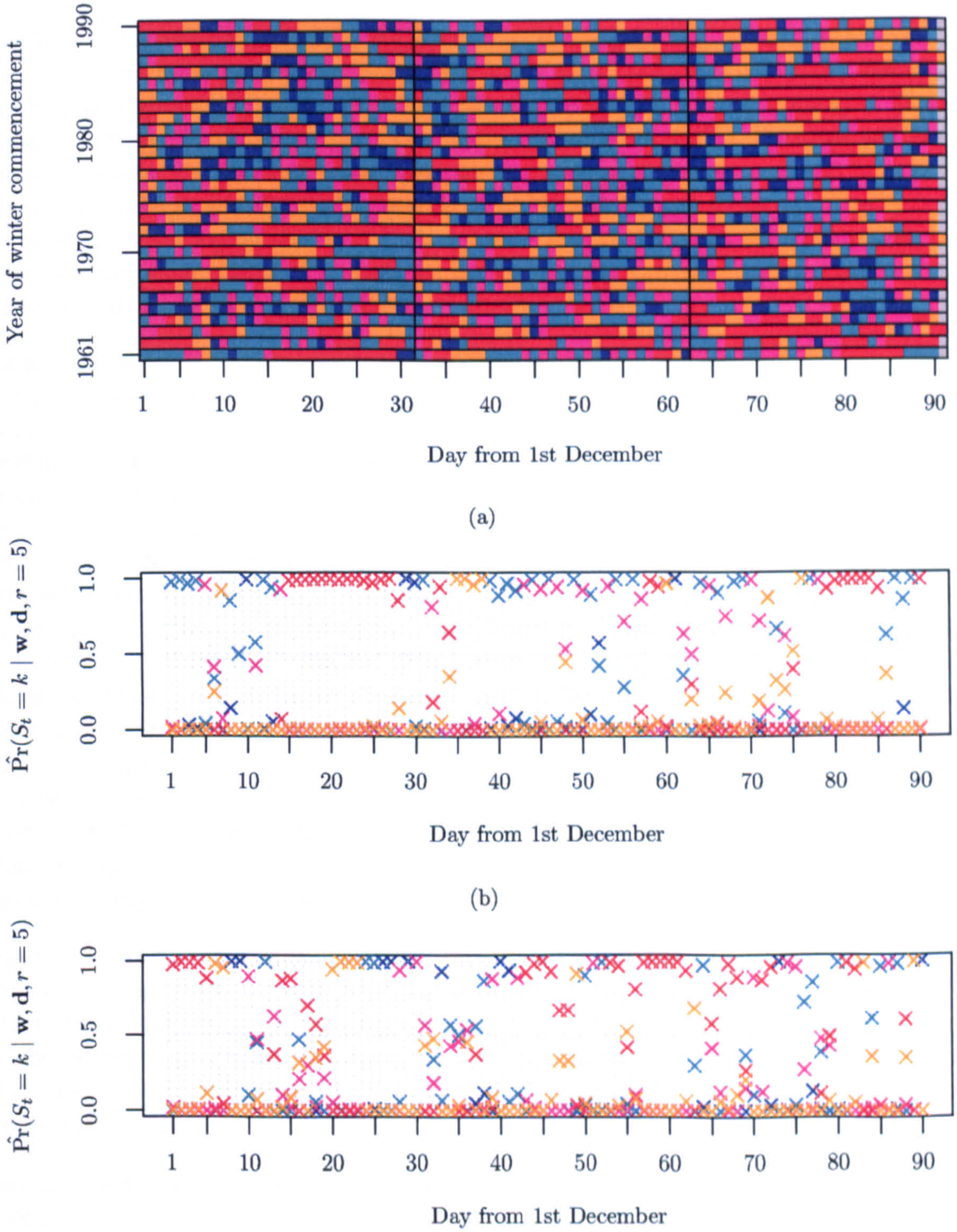


Figure 4.11: Conditional on $r = 5$, (a) marginal posterior mode (MPM) estimate of $\mathbf{s}$; posterior weather state probabilities $\hat{\Pr}(S_t = k \mid \mathbf{w}, \mathbf{d}, r = 5)$ for $k = 1$ (———), $k = 2$ (———), $k = 3$ (———), $k = 4$ (———) and $k = 5$ (———) in the winter (b) 1961/62 and (c) 1990/91.

4.7.4 Model checking

The methods of Bayesian model checking advocated in Chapter 6 of Gelman *et al.* (1995) involve comparing data to hypothetical replicates that could have been observed under the model. The role of the posterior predictive distribution in these procedures was introduced in Section 3.3.1 (see equation (3.2)). The central idea is that if the model fits well then the observed data should look plausible under the posterior predictive distribution, or at least this should be true for those aspects of the model in which we have interest. To this end, denote a *test quantity* by $T\{(\mathbf{w}, \mathbf{d})\}$. This is a scalar summary of the data that we want the model to capture adequately. Gelman *et al.* suggest assessing the fit of a model by comparing the posterior predictive distributions of a number of test quantities to their observed values.

In this section we will use the posterior predictive distribution for various test quantities to assess the model's ability to capture some of the important properties of the rainfall data in this small, dense network of sites. In particular, for each site, separately, we will examine the proportion of wet days and the quantiles in the distribution of rainfall amounts on wet days. The ability of the model to capture the spatial dependence in the occurrence process will be assessed by comparing the observed log odds ratios, defined in Chapter 2, to their posterior predictive distributions. Similarly the ability of the model to capture spatial dependence in the process of rainfall amount (given occurrence) will be assessed by comparing the Spearman's rank correlation coefficients to their posterior predictive distributions. Assessment of the ability of the model to capture the temporal dependence in the occurrence process will be based on comparisons between the observed empirical survivor function of wet and dry spells at each of the sites and their posterior predictive distributions. The empirical survivor function of wet (dry) spells is simply defined as the proportion of runs of consecutive wet (dry) days that persist for at least k days, $k = 1, 2, \ldots$. Finally, for each site, the model's ability to capture temporal dependence in the process of rainfall amount (given occurrence) will be assessed by comparing the observed Spearman's rank correlation coefficients between rainfall amounts at various lags (within uninterrupted wet spells) to the corresponding posterior predictive distribution.

The posterior predictive distributions for these test quantities, $T\{(\mathbf{w}, \mathbf{d})\}$, are not available analytically and so they are simulated by generating replicated data, $(\mathbf{w}^{\text{rep}}, \mathbf{d}^{\text{rep}})^{[i]}$, for each draw from the posterior and computing the test quantities, $T\{(\mathbf{w}^{\text{rep}}, \mathbf{d}^{\text{rep}})^{[i]}\}$, for this replicated data. The posterior predictive distributions are then plotted together with the observed values. Clearly, lack of fit is indicated by the observed value lying far into the tails of the posterior predictive distribution.

As discussed in Chapter 3, in the present context, Bayesian model averaging refers to the technique of making predictions using a weighted average of the posterior predictive distributions, conditional on each value of r, the weights being given by the posterior probabilities, $\pi_r(r \mid \mathbf{w}, \mathbf{d})$. In this way, uncertainty in the value of r is accounted for. However, for the Yorkshire data, the posterior support for $r = r_{\max} = 5$ relative to $r < r_{\max}$ is so overwhelming that Bayesian model averaging is essentially equivalent to basing predictions on the posterior predictive distribution conditional on $r = r_{\max}$. The results in this section are therefore based on the posterior predictive distribution for a model with $r = 5$ weather states.

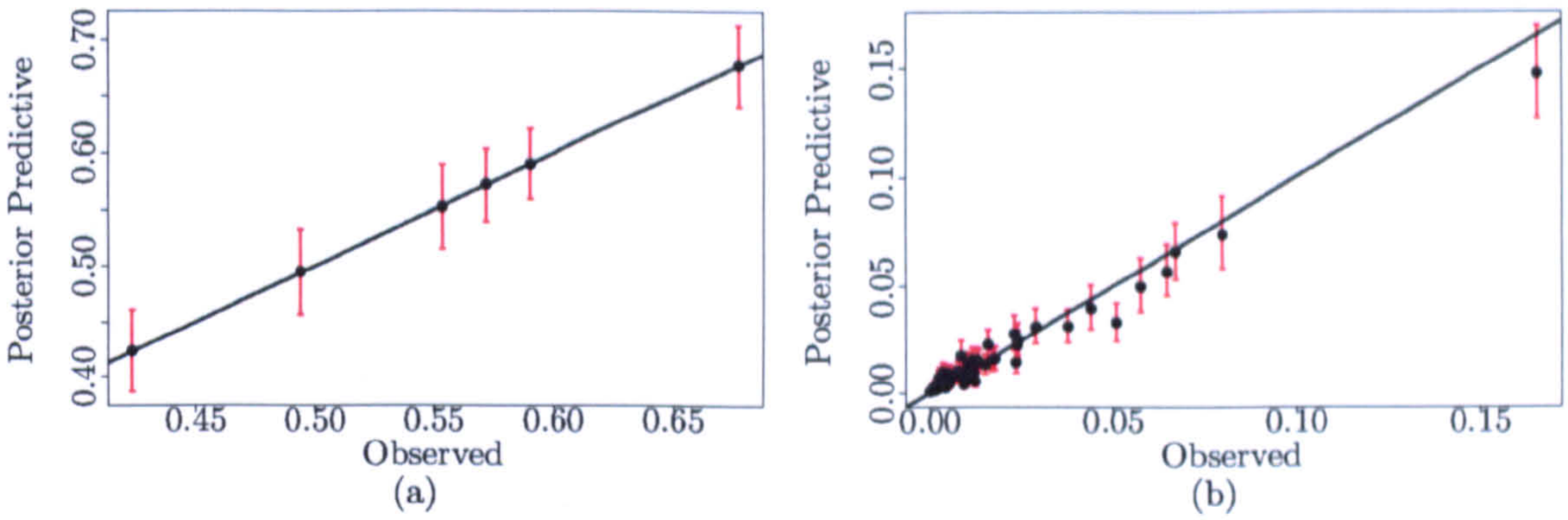


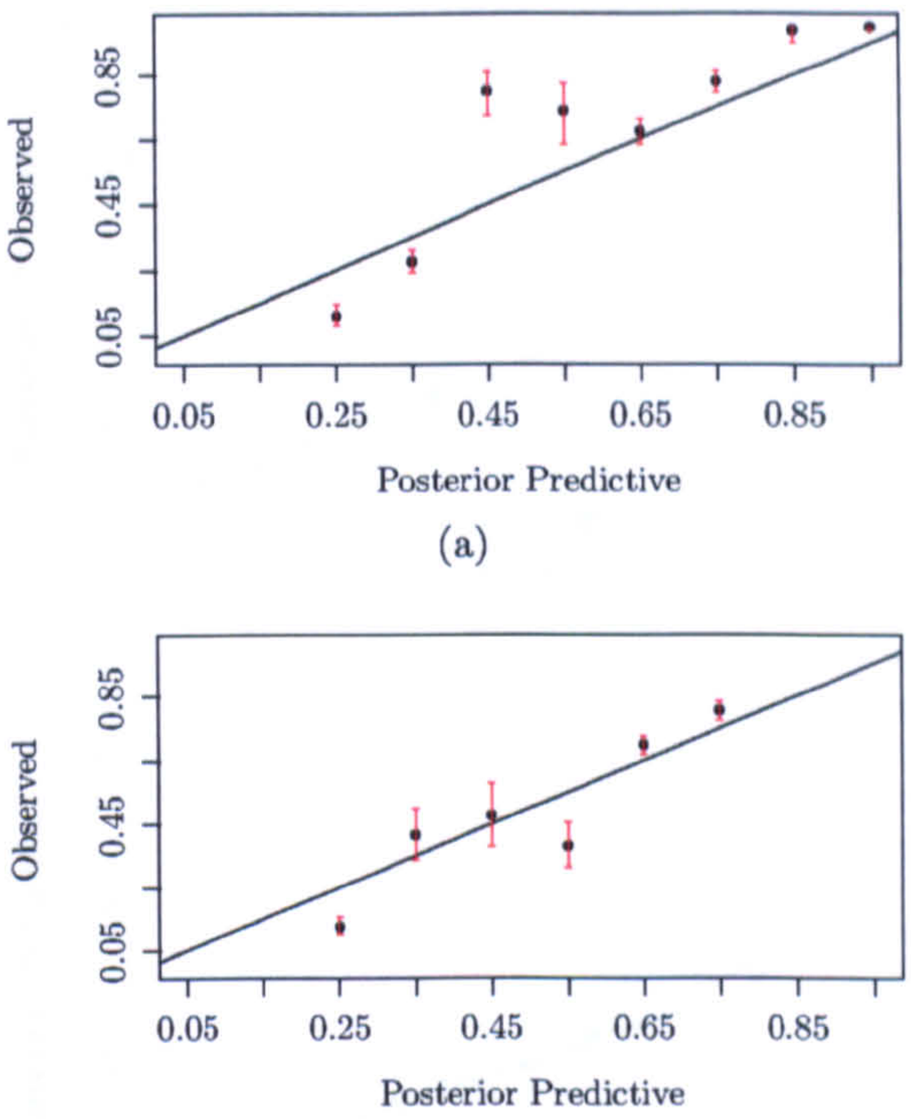
Figure 4.12: Observed values versus posterior predictive means for (a) precipitation occurrence relative frequencies at each Yorkshire site; and (b) relative frequencies of each precipitation occurrence vector for the Yorkshire network. (———) indicate the posterior predictive 95% Bayesian credible regions.

4.7.4.1 Simple marginal properties

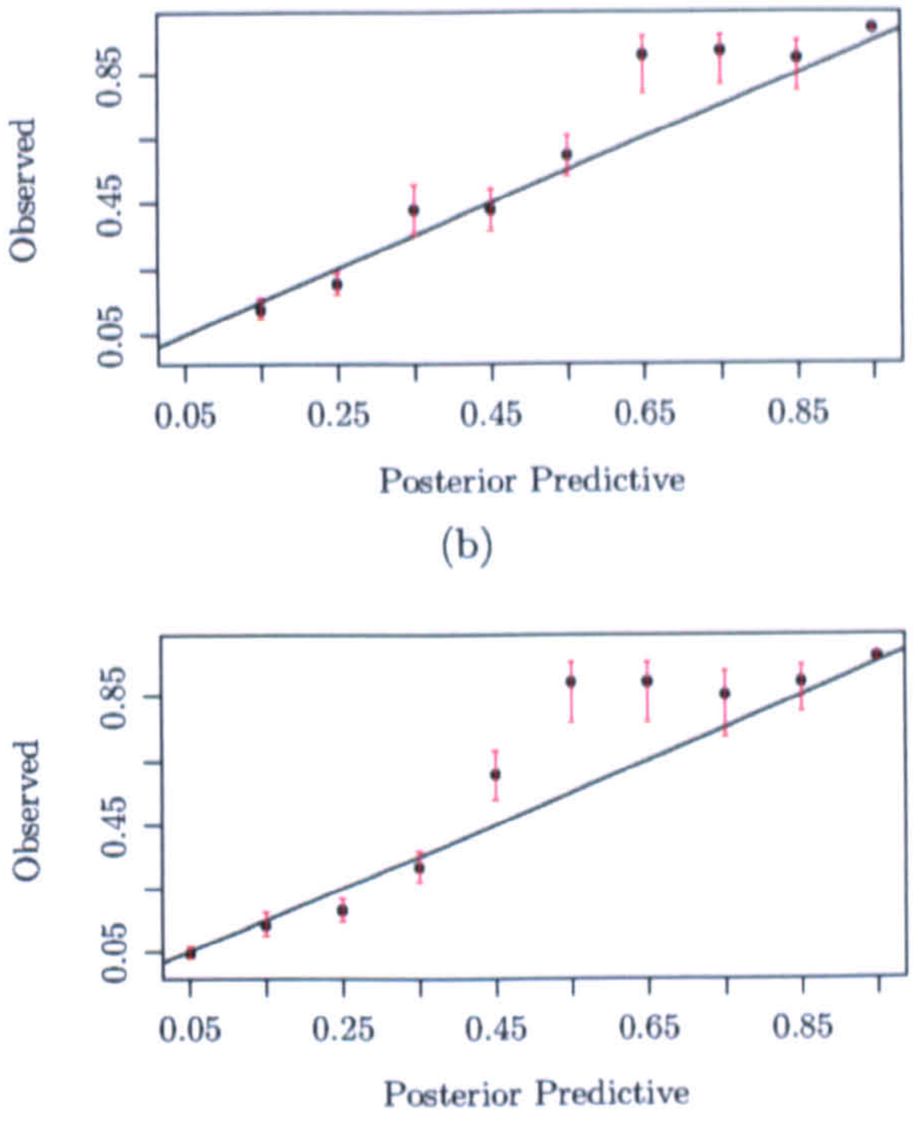
Figure 4.12(a) comprises a plot of the observed and posterior predictive means for the relative frequencies of rainfall occurrence at each of the Yorkshire sites. All of the points lie on the unit diagonal which indicates very good agreement between the two sets of quantities. Defining a rainfall occurrence vector as a vector of wet/dry indicators at the six sites, Figure 4.12(b) displays the observed and posterior predictive means for the relative frequencies of all possible rainfall occurrence vectors. The uncertainty in each posterior predictive distribution is indicated by plotting 95% equi–tailed Bayesian credible regions. If the unit diagonal intersects the plotted credible region, this indicates that the observed statistic lies within the central 95% of its posterior predictive distribution. The observed statistics for most rainfall occurrence vectors lie within the 95% credible regions, but there is some evidence that the model underestimates the most commonly occurring rainfall occurrence vector (rain at all sites) and that which occurs on around 5% of days (dry at all sites except Moorland Cottage). This suggests that the model is not quite capturing the joint distribution of rainfall occurrence at all sites. This will be explored further in Section 4.7.4.2.

Calibration refers to the statistical consistency between distributional forecasts and the observations that materialize. In the context of probabilistic forecasting, it is discussed by, for example, Gneiting *et al.* (2007) who propose tools for checking calibration and *sharpness*, that is, the concentration of the predictive distributions. Figure 4.13 shows *calibration curves* for the posterior predictive probability of rain at each site. Calibration curves are more usually encountered in the elicitation literature (see, for example, Smith, 1988) as a means of judging a person's probability assessments. However we can apply the same ideas here to assess more finely the accuracy of the posterior predictive probability of rain at each site.

For every day, t, at each site, i, we can approximate the posterior predictive probability of rain



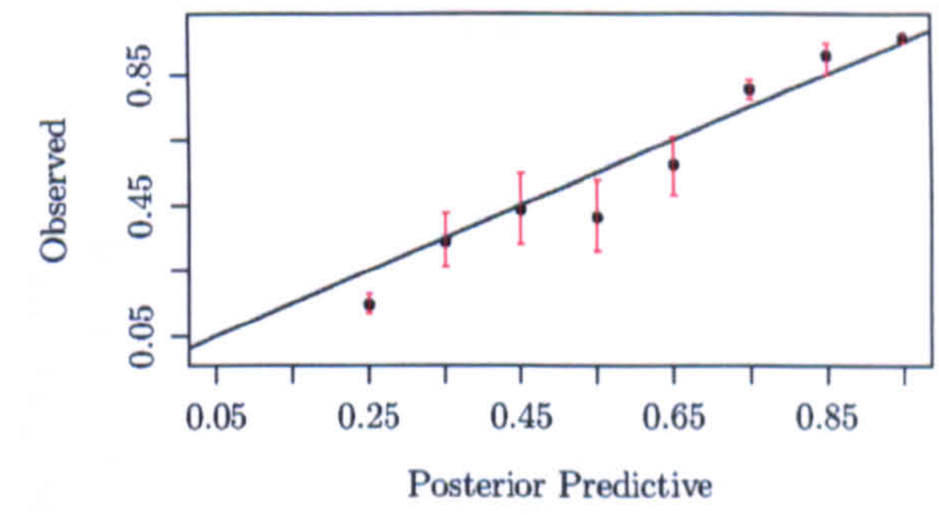
(a)



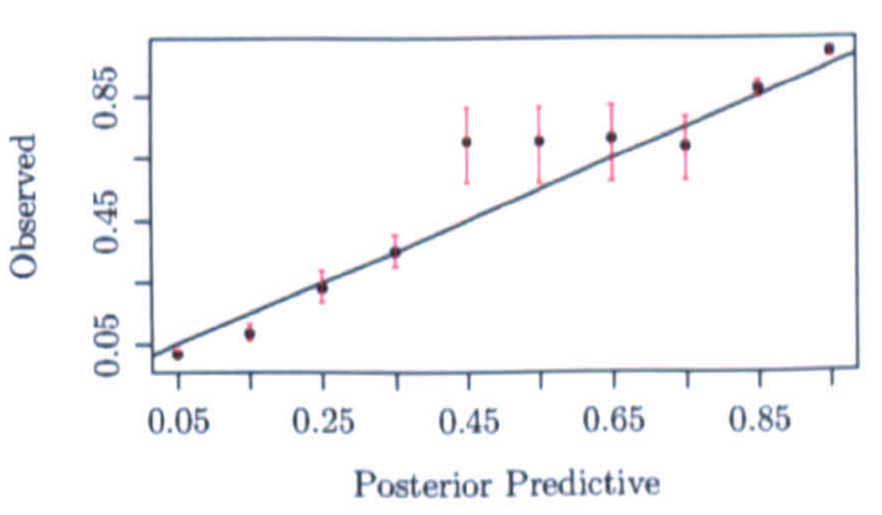
(b)

(c)

(d)



(e)



(f)

Figure 4.13: Calibration curves for the posterior predictive probability of rain at (a) Lockwood Reservoir; (b) Hull, Pearson Park; (c) Moorland Cottage; (d) the Retreat, York; (e) Great Walden Edge; (f) Kirk Bramwith. (———) is a posterior 95% Bayesian interval for the "true" probability based on the observed sample (assumed binomial) and a uniform prior on the "true" probability.

for a hypothetical replication, say D_t^{i*}, of D_t^i by the Monte Carlo estimate

$$\Pr(D_t^{i*} = 1 \mid \mathbf{w}, \mathbf{d}, r) \simeq \frac{1}{N} \sum_{j=1}^{N} \Pr(D_t^i = 1 \mid S_t = s_t^{[j]}, \boldsymbol{\theta}_r^{[j]}, r) = \frac{1}{N} \sum_{j=1}^{N} p_{r,is_t^{[j]}}^{[j]}$$

where $s_t^{[j]}$ is the j–th MCMC draw of the weather state on day t, and $p_{r,ik}^{[j]}$ is the j–th MCMC draw of $p_{r,ik}$. Consider intervals $[0, 0.1)$, $[0.1, 0.2), \ldots, [0.9, 1.0]$ and denote the j–th interval by I_j, $j = 1, \ldots, 10$. To construct the calibration curve for site i, the proportion $\hat{p}_{I_j}$ of days on which rain was observed when the posterior predictive probability lies in the interval I_j is plotted against the midpoint of I_j, $j = 1, \ldots, 10$. Some intervals might be associated with very few observations, so we add 95% error bars to the observed proportions $\hat{p}_{I_j}$ in order to

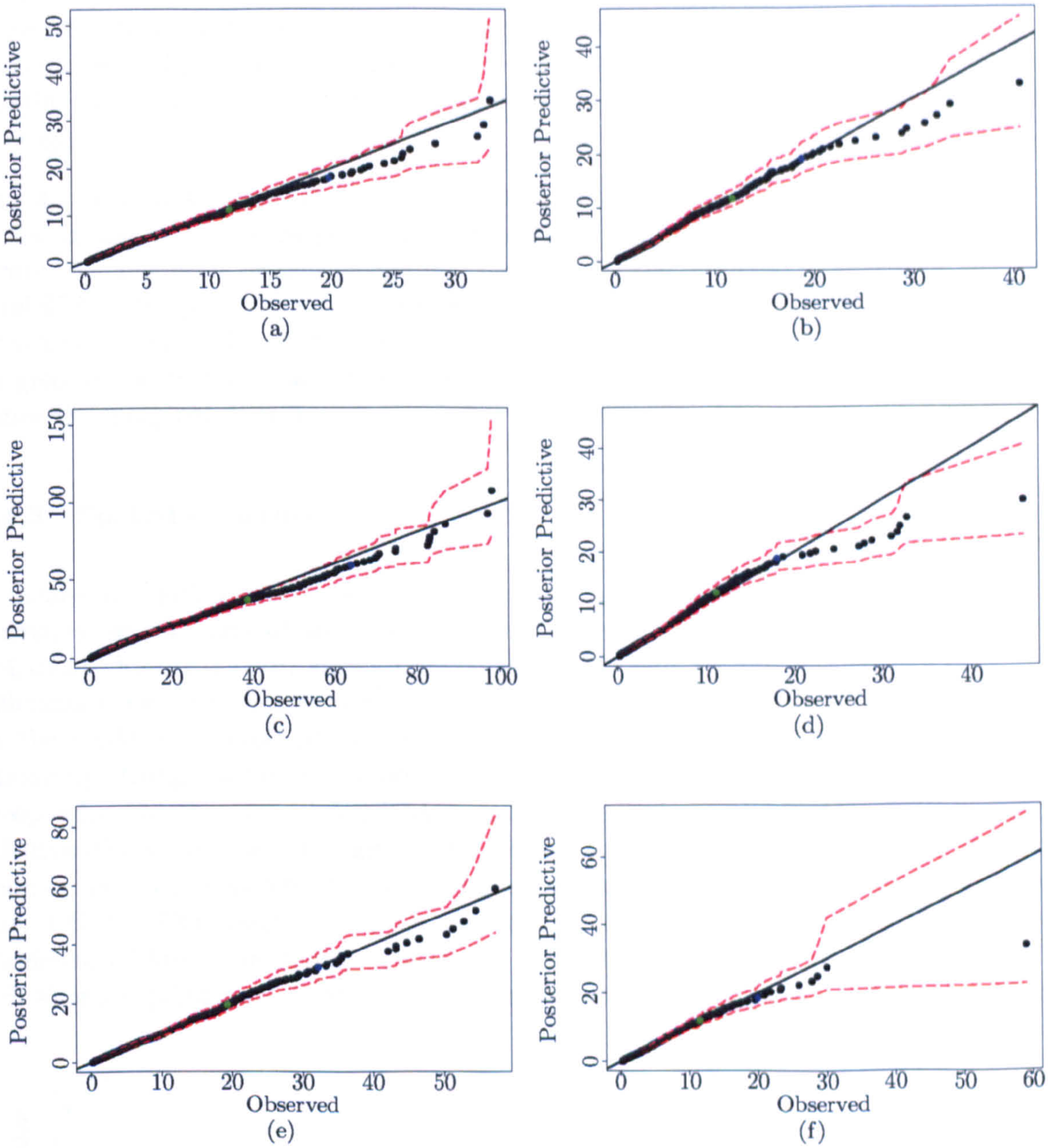


Figure 4.14: Quantile–quantile plots for the observed versus posterior predictive mean rainfall amounts (in mm) at (a) Lockwood Reservoir; (b) Hull, Pearson Park; (c) Moorland Cottage; (d) the Retreat, York; (e) Great Walden Edge; (f) Kirk Bramwith. (------) indicate the posterior predictive 95% Bayesian credible regions. For reference, (•) and (•) indicate roughly the 95–th and 99–th percentiles.

convey this information. These are calculated by regarding the observed number of wet days which gave rise to $\hat{p}_{I_j}$ as a binomial observation, and by assigning a uniform prior to the "true" probability. The 2.5% and 97.5% points in the resulting (beta) posterior distribution for the "true" probability then provide the end points of the error bar for the j–th observed proportion.

Clearly if the posterior predictive distribution can accurately "forecast" the probability of rain, the observed proportions will lie roughly on the unit diagonal. Referring to Figure 4.13, for all

sites, most points lie close to this line with some deviations occurring, most noticeably at sites 1, 4 and 6, when the posterior predictive probability of rain is between 0.4 and 0.6. At these sites the observed proportions remain fairly constant over a few of the more central intervals, indicating that the posterior predictive distribution is less informative about probabilities over this range.

Figure 4.14 contains plots of the sample quantiles of the distribution of non–zero rainfall amounts and the means of the corresponding posterior predictive distributions. The red dotted lines indicate 95% posterior predictive credible regions. All the observed quantiles lie well within the central 95% of the posterior predictive distributions, and slight departure from the unit diagonal only starts to appear for some sites at around the 99–th percentile (indicated by a blue dot). This gives no particular reason to doubt the ability of the (mixture of) gamma distributions to capture the marginal distribution of rainfall amount on wet days at any of the sites.

4.7.4.2 Spatial structure

The means and 95% posterior predictive credible regions for the log odds ratios between rainfall occurrences at all pairs of sites, together with the observed statistics, are displayed in Figure 4.15(a). Figure 4.15(b) shows the corresponding plot for the Spearman's rank correlation coefficients between non–zero rainfall amounts all pairs of sites. From Figure 4.15(a) it appears that the model is capable of capturing the spatial dependence when that dependence is not particularly strong. However, the observed statistics in the upper right hand corner of the plot, corresponding to the highest log odds ratios, lie well above the 97.5% point in the posterior predictive distribution which suggests the model cannot reproduce strong positive spatial association in the occurrence process. We can draw similar conclusions for the amounts process from Figure 4.15(b). This suggests that when we restrict the maximum number of states to $r_{\max} = 5$, unmodelled within–state spatial dependence remains. In Chapter 5 we will investigate a model which allows explicit modelling of within–state dependence in the rainfall occurrence process.

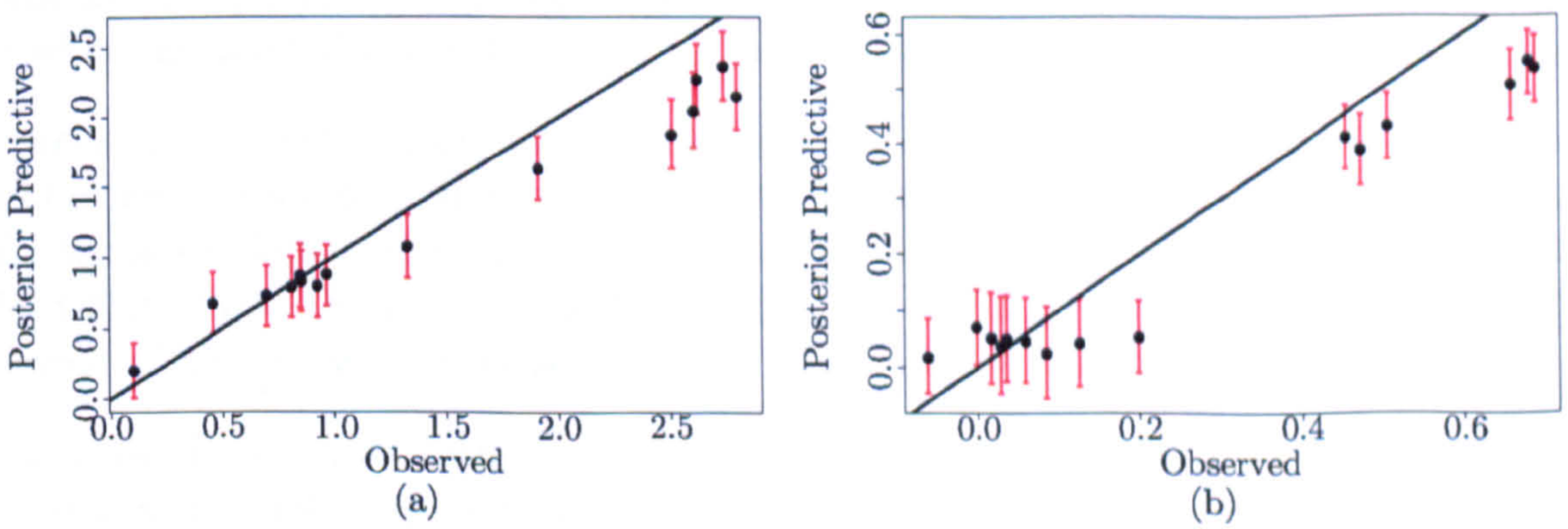


Figure 4.15: Observed values versus posterior predictive means for (a) log odds ratios between rainfall occurrences; and (b) Spearman's rank correlation coefficients between non–zero rainfall amounts at each pair of sites in the Yorkshire network. (———) indicate the posterior predictive 95% Bayesian credible regions.

4.7.4.3 Temporal structure

Figure 4.16 shows the empirical survivor function for wet spells at each site together with the mean and the upper and lower 2.5% points from the corresponding posterior predictive distribution, all plotted on the log scale. For sites 2, 4 and 5, the observed distribution of wet spell duration lies largely within the 95% posterior predictive credible region, and agreement with the posterior predictive mean is particularly good at sites 2 and 4. However, for sites 1 and 3, the observed distribution lies well above the 97.5%-point in the posterior predictive distribution even for some of the shorter wet spell durations. The difference is most evident at site 3 where there appears to be very strong positive association in the observed distribution. Indeed, in Chapter 2, we observed that when the occurrence data at each site are modelled by a q–th order Markov chain, the marginal likelihood for the site 3 data is maximised by a chain of order $q = 3$, compared with $q = 1$ or $q = 2$ for the other sites.

One possible explanation as to why the posterior predictive distribution is able to reproduce the temporal dependence in the occurrence process at some sites and not others is as follows. In Section 4.7.3.2 we observed that the posterior distribution for Λ_5 has considerable support for transitions between the two "wettest" weather states, states 1 and 2. Referring to Figure 4.10(a), at sites 2, 4 and 5 the posterior distributions for $p_{5,i1}$ and $p_{5,i2}$, $i = 2, 4, 5$, are concentrated about means very close to 1 so that wet spells are likely to persist as long as the weather state remains in states 1 or 2, or switches from one to the other. However at site 3, for example, the weather state associated with the largest rainfall probability parameters (state 5) is such that the posterior mean for $p_{5,35}$ is only equal to 0.787. Therefore regardless of the persistence of the "wetter" weather states in this model, there is always a reasonably large probability of a dry day so that wet spells are easily interrupted. Similarly at sites 1 and 6 there is only one, rather than two, states where the posterior for the probability of rain parameter is concentrated close to one.

The corresponding plots for dry spells are displayed in Figure 4.17. Although, on average, the dry spell durations seem to be shorter than the wet spell durations, the plots show comparable patterns and therefore similar conclusions can be drawn. The tendency for predictions to underestimate the persistence in dry spells at sites 1 and 3 can be linked to absence of a weather state whose associated rainfall probability parameters have much posterior density near zero.

The prediction of wet and dry spells which are, on average, shorter than those in the observed data at some sites suggests that the persistence of the weather state is not enough to capture the strong temporal dependence in the rainfall occurrence process. In Chapter 5 we will consider a model which allows the conditional distribution of rainfall occurrence, given the weather state, to depend additionally on whether or not it rained on the previous day.

For each site, Figure 4.18 shows observed Spearman's rank correlation coefficients between rainfall amounts at various lags (within uninterrupted wet spells) and the means and equi-tailed 95% credible regions from the corresponding posterior predictive distributions. The lengths of the 95% credible regions increase as the lag increases because relatively few long runs of consecutive wet days are observed, so there is less information in the data about the correlations at larger lags. At sites 2, 4 and 6 the observed correlations lie within the 95% posterior predictive credible regions for most lags. However, at sites 1, 3 and 5 the observed statistics at the earlier lags, especially lag 1, lie well above the 97.5–th percentile in the posterior predictive distribution. In

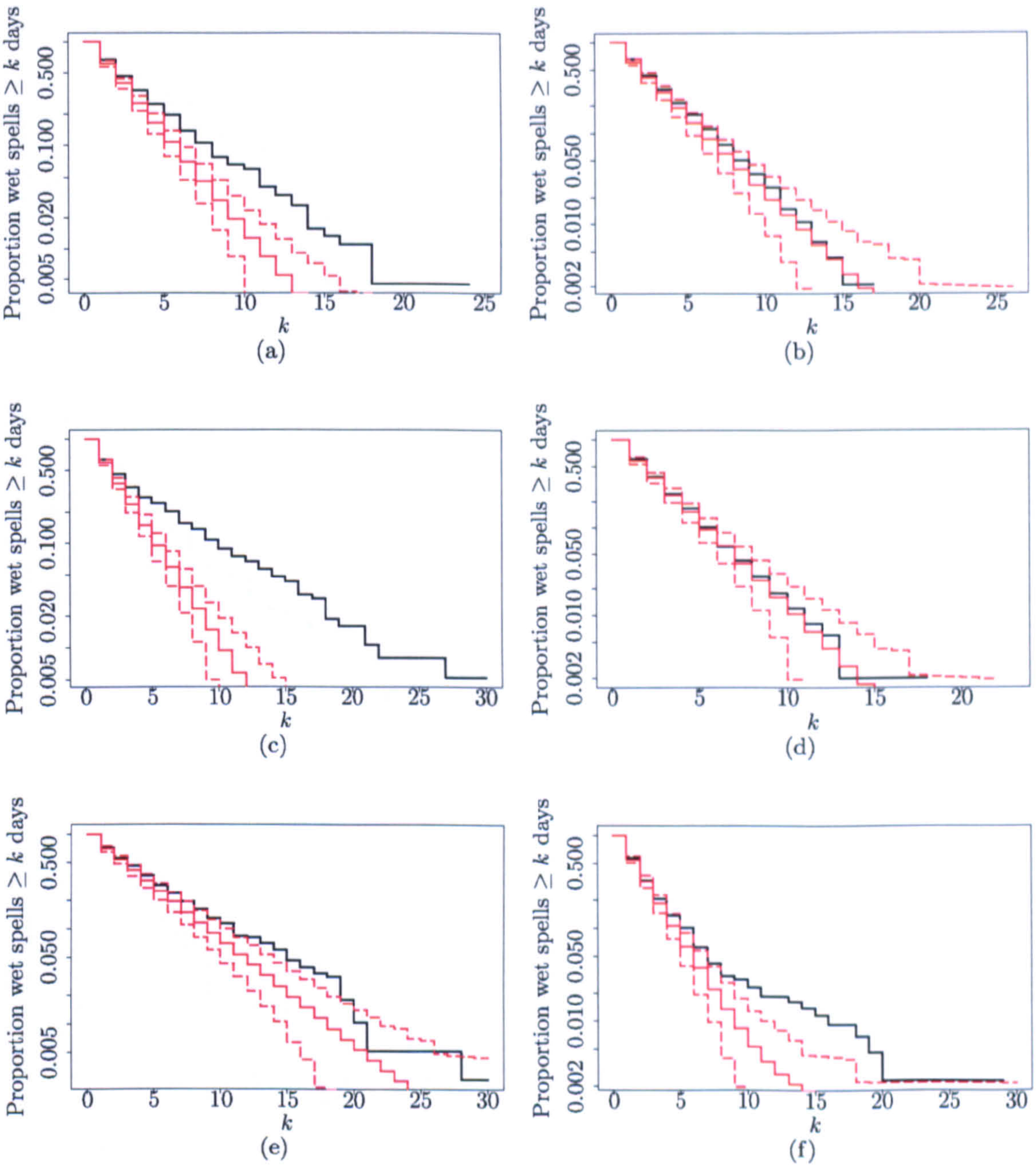


Figure 4.16: Observed (———), posterior predictive mean (———) and posterior predictive 95% Bayesian credible regions (------) for the survival distributions of wet spells at (a) Lockwood Reservoir; (b) Hull, Pearson Park; (c) Moorland Cottage; (d) the Retreat, York; (e) Great Walden Edge; (f) Kirk Bramwith.

fact, the observed lag 1 correlations are highest at sites 3 and 5, both being in excess of 0.26, whilst the means of the corresponding posterior predictive distributions are amongst the lowest, being equal to 0.028 and 0.099 respectively. Figure 4.10(c) suggests that at sites 2, 4 and 6, small values for the coefficient of variation parameters in the "wetter" states (states 1 and 2) are likely *a posteriori*. Therefore as long as these wetter states persist, there is little variability in the gamma distribution modelling rainfall amounts on wet days. This may explain why the model is able to capture the autocorrelation between rainfall amounts within wet spells at these

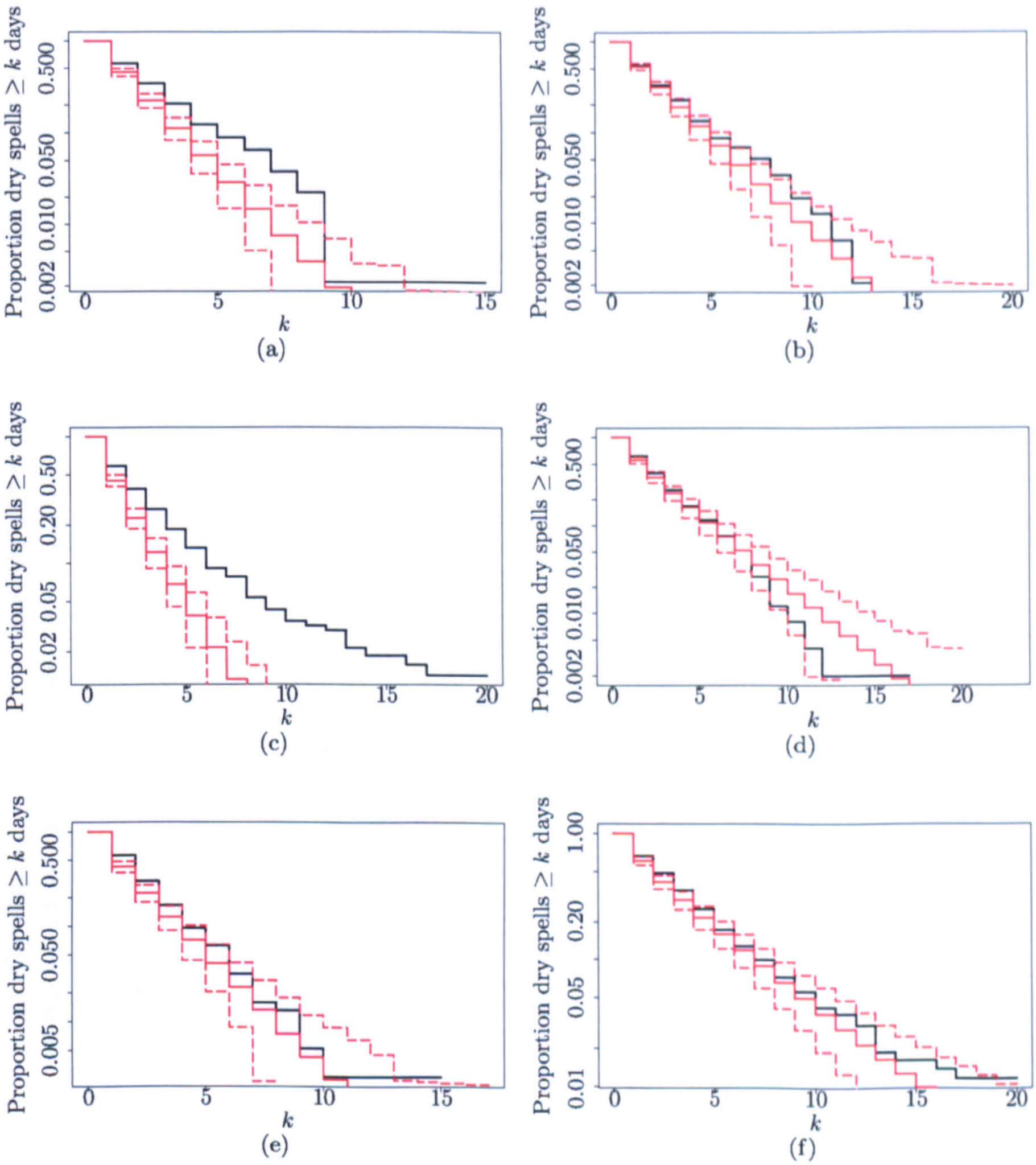


Figure 4.17: Observed (———), posterior predictive mean (———) and posterior predictive 95% Bayesian credible regions (------) for the empirical survival distributions of dry spells at (a) Lockwood Reservoir; (b) Hull, Pearson Park; (c) Moorland Cottage; (d) the Retreat, York; (e) Great Walden Edge; (f) Kirk Bramwith.

sites. However the same is not true at sites 1, 3 and 5 which may account for the inability of the posterior predictive distribution to reproduce the temporal persistence at these sites. Relaxing the assumption of temporal independence in the process of rainfall amount, given occurrence and the weather state, will be addressed in Chapter 6.

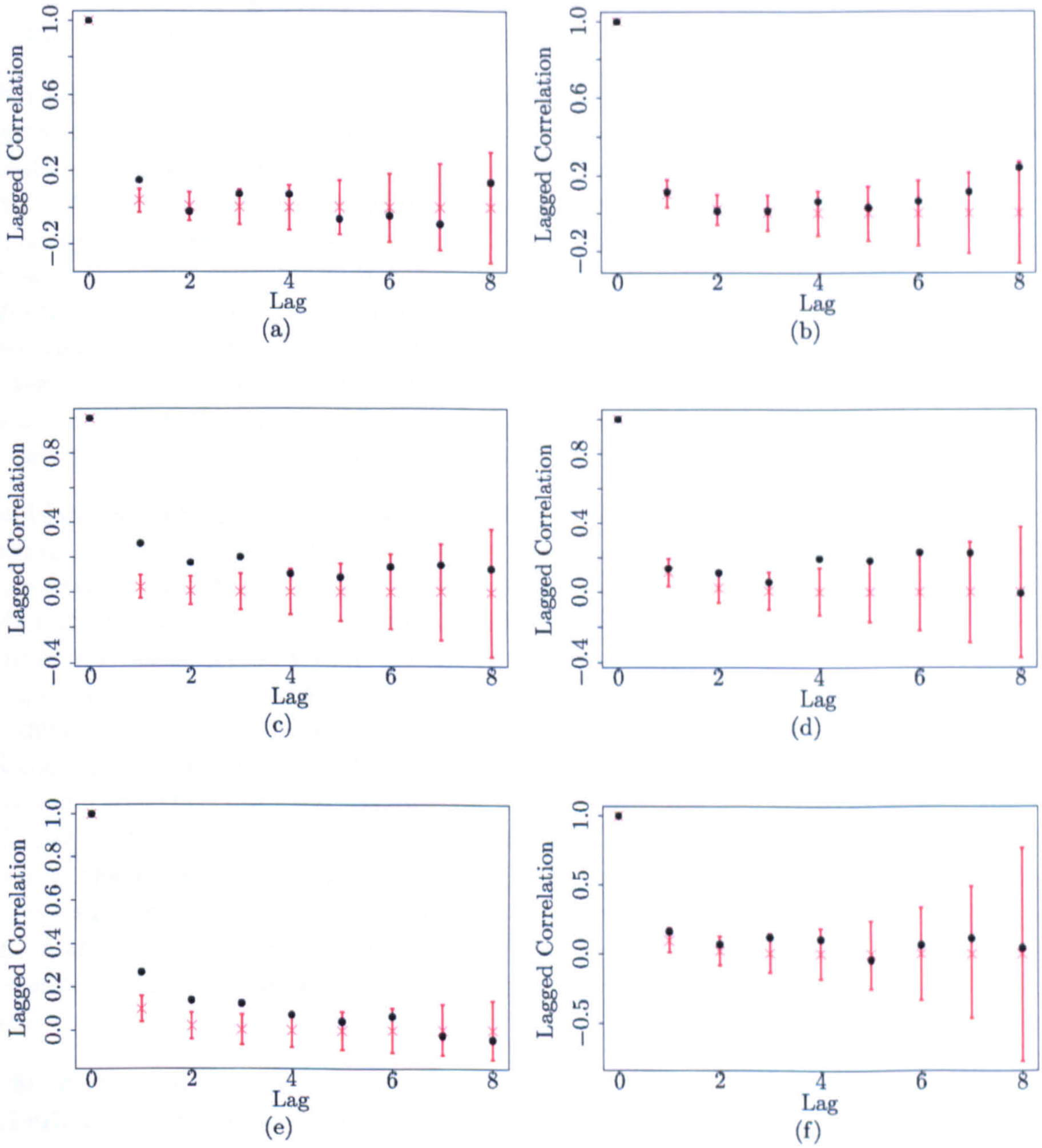


Figure 4.18: Observed (•), posterior predictive mean (×) and posterior predictive 95% Bayesian credible region (———) for the Spearman's rank correlation coefficient between wet days (within runs of consecutive wet days) at various lags at (a) Lockwood Reservoir; (b) Hull, Pearson Park; (c) Moorland Cottage; (d) the Retreat, York; (e) Great Walden Edge; (f) Kirk Bramwith.

4.8 Summary

In this chapter we have described a reasonably simple hidden Markov model for multi–site precipitation and have illustrated some of the Bayesian inferential procedures, discussed in Chapter 3, by applying the model to the Yorkshire dataset. The chapter had two main focuses: finding a satisfactory method for estimating the posterior distribution for the number of weather

states, r; and developing a fully Bayesian approach to inference in hidden Markov models for precipitation, as an alternative to the frequentist versions presented elsewhere in the literature.

With regards to the former focus, most methods from the literature for approximating the marginal likelihood require that the complete data posterior $\pi(\theta \mid \mathbf{s}, \mathbf{w}, \mathbf{d})$, or at least all of the full conditional distributions, have known normalising constants. However, neither of these conditions is met for the hidden Markov model presented in this chapter. Therefore, amongst the more sophisticated methods from the literature, the recently developed power posterior approach (Friel & Pettitt, 2008) is one of the few available that we can apply. In Section 4.6, we conducted a simulation experiment and showed that for models with both univariate ($n = 1$) and multivariate ($n > 1$) within-state distributions, the estimator calculated using power posteriors had negligible bias and small variance, when based on reasonably small MCMC samples. We demonstrated how the approach can easily be implemented for hidden Markov models and used it to estimate the log marginal likelihood in an application to the Yorkshire dataset.

In problems involving model choice or averaging, one of the most appealing features of a fully Bayesian implementation is that post data uncertainty can be completely summarised by the joint posterior distribution of the model parameters and the model indicator, in this case the number of hidden states. This offers a single coherent framework for inference, unlike the frequentist equivalent which treats the model parameters and model indicators separately. By taking a Bayesian approach to inference, therefore, we were able to compare the models directly in terms of their posterior probabilities. We believe this to offer an improvement over frequentist implementations of hidden Markov models for precipitation, in which models have been compared using so-called objective information criterion like the BIC (Hughes *et al.*, 1999), the AIC (Ailliot *et al.*, 2009) or the method of cross validated likelihood (Robertson *et al.*, 2004). Fundamental in the Bayesian paradigm is the specification of a prior distribution. We demonstrated how we can incorporate prior information about meaningful quantities such as the probability of rain and the mean rainfall amount on wet days, and suggested more sophisticated prior structures which allow borrowing of strength between parameters that are correlated in our prior beliefs.

The fit of the model to the Yorkshire data was assessed by comparing the posterior predictive distributions for various test quantities to their observed values. Although the model was able to capture simple marginal characteristics of the data, such as the proportion of wet days at each site and the quantiles in the distribution of non-zero rainfall amounts, the model could not provide an explanation of some of the (higher order) spatio-temporal properties. The posterior predictive distribution underestimated the strongest spatial autocorrelations markedly in both the rainfall occurrence and amounts processes. At some of the sites, where there was strong temporal dependence between rainfall occurrences and between rainfall amounts, the posterior predictive distribution underestimated the durations of wet and dry spells and the lag one autocorrelation between rainfall amounts within wet spells. The model presented here can also be criticised on the grounds that it cannot be used to model non-stationary shifts in the rainfall process, which arise due to changes in atmospheric conditions, so could not be used for statistical downscaling; see Chapter 1 for more details. However, each of these shortcomings presents its own remedy. The model studied in this chapter has a relatively simple spatio-temporal dependence structure, namely conditional independence between sites and in time given the weather state. Modifications to the parameterisation of the precipitation process

which allow the conditional probability of rain at site i to depend on the rainfall status at site j, given the weather state, might provide a more realistic spatial model. Similarly, relaxing assumption A2 to allow the conditional probability of rain at any particular site (given the weather state) to depend additionally on whether or not it rained the previous day may account for the unmodelled temporal autocorrelation that remains at certain sites. These modifications both relate to the occurrence process, but similar suggestions apply to the distribution of rainfall amounts, given occurrence and the weather state. Finally, relaxing assumption A1 to allow atmospheric explanatory variables to influence the transition probabilities in the weather state process would allow the model to respond to changes in the underlying atmospheric conditions which drive precipitation. All of these modifications will be considered further in subsequent chapters.

Chapter 5

A non–homogeneous hidden Markov model for rainfall data

5.1 Introduction

In the hidden Markov model introduced in Chapter 4, the weather state was the only device for capturing spatio–temporal dependence. When applying the model to the Yorkshire dataset, it was found to be unable to predict, *a posteriori*, spatial and temporal dependencies as high as those actually observed amongst rainfall amounts and, in particular, rainfall occurrences. Additionally, in Chapter 1 it was noted that there is currently particular interest in models which can be used in statistical downscaling. To fulfil this function, models need to be able to describe relationships between atmospheric variables and precipitation. In this chapter we generalise the model developed in Chapter 4 by relaxing the "standard" assumptions that; (A1) the hidden states evolve as a homogeneous first order Markov chain and (A2) the observable random quantities are conditionally independent in time, given the hidden states. In allowing a more complex spatio–temporal dependence structure, focus is placed on rainfall occurrences, describing their conditional distribution, given the weather state, as a Markov chain of autologistic models. Dependence on atmospheric information is also introduced in the weather state process by allowing time dependent atmospheric variables to influence the probabilities of transition between the weather states, thus making the model non–homogeneous.

The remainder of this chapter is organised as follows. Section 5.2 provides an introduction to the autologistic model which will be incorporated in the within–state model for rainfall. Section 5.3 describes the new set of conditional independence assumptions on which the non–homogeneous hidden Markov model (NHMM) is based, and specifies the parameterisations chosen for the weather state and precipitation processes. Section 5.4 gives details of the prior distribution, several of whose components are specified in order to encourage borrowing of strength between related parameters. This can be regarded as a compromise between *a priori* independence and a naive approach of fixing various parameters to be equal. This section also provides guidelines for thinking about prior elicitation in the highly parameterised NHMM. The complete data likelihood is derived in Section 5.5, and the full conditional distributions of all model parameters

and the weather states are given in Section 5.6. This section also outlines the MCMC scheme for generating posterior samples for a model with a fixed number of states, r. Finally, in Section 5.7, we apply the model and inferential procedures in an analysis of the Yorkshire dataset. This includes details on estimating the posterior distribution for r via power posteriors, and a detailed exploration of some of the problems in implementing the MCMC scheme. These largely arise due to parameters being only weakly identifiable in the likelihood. We also use the posterior predictive distribution to compare the fit of the NHMM to that obtained using the simple hidden Markov model from Chapter 4.

5.2 The autologistic model

In Section 5.3.3 we model rainfall occurrences as a Markov chain of autologistic models, conditional on the weather state. This section introduces the autologistic model as a means of modelling correlated binary data and describes some of the methods from the literature for handling its analytically intractable normalising constant.

5.2.1 Background

The autologistic model of Besag (1972, 1974) is a popular model for multivariate binary data when a spatial component is incorporated. As a special case it includes the Ising model, originally developed by physicists to model electron spin at each site in a magnetic field. Elsewhere the autologistic model has been used in several ecological applications, for example, to describe spatial patterns of disease presence/absence in agricultural fields (Gumpertz *et al.*, 1997).

Consider a random vector $\mathbf{D} = (D^1, \ldots, D^n)^T$ where D^i is a binary variable corresponding to the i–th spatial location. These spatial locations, called sites hereafter, might be single points that are indexed by sets of coordinates, or areal units into which a geographical region has been divided. The autologistic model belongs to a more general class of auto–models which are formulated in terms of conditional, rather than joint distributions. These models are based on the definitions of sets of neighbours for each site, in which a site, say i, is defined as a neighbour of site $j \neq i$ if and only if the functional form of its full conditional distribution depends on site j. Often the autologistic model is defined on a regular lattice, in which case various neighbourhood structures might be assumed, for example, the first order scheme in which the full conditional distribution for site i depends on its four nearest neighbours. When the sites are irregularly distributed, criteria based on the distances between them are often used to decide the sets of neighbours. Once the sets of neighbours have been defined, assuming pairwise only dependence between sites (that is, ignoring interactions involving more than two sites), the Hammersley–Clifford Theorem (Cressie, 1993) is used to generate the conditional distributions which give rise to a valid joint distribution. In the autologistic model, each full conditional distribution is of logistic form and expresses the log odds of "success" at a particular site as a linear combination of the "successes" or "failures" at sites in the set of neighbours. Letting

$\mathbf{D}^{-i} = (D^1, \ldots, D^{i-1}, D^{i+1}, \ldots, D^n)$, the full conditional and joint distributions are given by

$$\Pr(D^i = d^i \mid \mathbf{D}^{-i} = \mathbf{d}^{-i}, \{\alpha_i\}, \{\beta_{ij}\}) = \frac{\exp\left(\alpha_i d^i + \sum_{j \neq i} \beta_{ij} d^i d^j\right)}{1 + \exp\left(\alpha_i + \sum_{j \neq i} \beta_{ij} d^j\right)}$$

and

$$\Pr(\mathbf{D} = \mathbf{d} \mid \{\alpha_i\}, \{\beta_{ij}\}) = \frac{\exp\left(\sum_{i=1}^{n} \alpha_i d^i + \sum_{i=2}^{n} \sum_{j=1}^{i-1} \beta_{ij} d^i d^j\right)}{\sum_{\tilde{\mathbf{d}}} \exp\left(\sum_{i=1}^{n} \alpha_i \tilde{d}^i + \sum_{i=2}^{n} \sum_{j=1}^{i-1} \beta_{ij} \tilde{d}^i \tilde{d}^j\right)}, \tag{5.1}$$

respectively, where $\beta_{ij} = \beta_{ji}$ and $\beta_{ij} = 0$ unless sites i and j are neighbours. The normalising constant in the denominator of the joint mass function is a sum over all 2^n possible binary vectors.

Besag's original autologistic model has been extended to incorporate both covariates (see, for example, Wu & Huffer, 1997; Gumpertz *et al.*, 1997) and a time component. Some authors, for example Zhu *et al.* (2005) or Zheng & Zhu (2008), have treated time in terms of an extra spatial dimension, redefining the set of neighbours of each site to include response variables at past and future time points. Other authors, for example Zhu *et al.* (2008), model the multivariate binary response variables, $\{\mathbf{D}_t = (D_t^1, \ldots, D_t^n)^T : t = 0, \ldots, T\}$, as a q-th order Markov chain in which $\mathbf{D}_t$ is modelled using an autologistic regression model, conditionally on $(\mathbf{D}_{t-q}, \ldots, \mathbf{D}_{t-1})$. This model belongs to a more general class called MCMF (Markov chain of Markov field) models that were developed by Guyon & Hardouin (2002).

Both the ease with which explanatory variables can be incorporated, and the intuitive appeal of a model in which the probability of success at a particular site depends transparently on the successes/failures at neighbouring sites, have contributed to the popularity of autologistic (regression) models. However, as a serious drawback, there is no closed form expression for the normalising constant, except in special cases. When the number of sites, n, is large this presents a substantial computational challenge. Other drawbacks with autologistic models are the possibility of multicollinearity, when the dependence between different pairs of sites are highly correlated (Gumpertz *et al.*, 1997), and the irregular behaviour of the likelihood when the spatial dependence parameters are large (Møller *et al.*, 2006). These problems can lead to inferential difficulties.

5.2.2 Handling the normalising constant

Reeves & Pettitt (2004) derive an efficient algebraic recursion for computing the normalising constant of the autologistic model on the lattice. The saving in computational time is realised by taking advantage of conditional independence assumptions made in defining the neighbourhood structure. The recursion can handle lattices with up to 20 rows but is not helpful when, for example, conditional independence is not assumed between any pairs of sites.

A common means of addressing the problem of an intractable normalising constant is to approximate its value, for instance, using path sampling (Gelman & Meng, 1998) or the method of Monte Carlo maximum likelihood (Geyer & Thompson, 1992). This is the approach adopted by Hughes *et al.* (1999) who use the autologistic model in an NHMM for rainfall occurrence. In a

Bayesian setting, a recent variation of this approach was provided by Zheng & Zhu (2008) who, at every iteration of their MCMC algorithm, approximated the ratio of normalising constants in the Metropolis Hastings acceptance ratio according to

$$\frac{1/C(\theta^*)}{1/C(\theta)} = \frac{C(\theta)/C(\psi)}{C(\theta^*)/C(\psi)},$$

where $C(\cdot)$ is the normalising constant, θ denotes the parameters of the autologistic model and ψ is a fixed parameter which should be close to the posterior mode for θ. As with the technique of Monte Carlo maximum likelihood, the term $C(\theta)/C(\psi)$ can then be expressed as an expectation of the ratio of unnormalised likelihoods, and this can be approximated via importance sampling. The problem with embedding such approximations within an MCMC scheme is that the acceptance ratio in steps involving parameters of the autologistic model is only approximately equal to the correct acceptance ratio. This has the obvious consequence that the equilibrium distribution of the algorithm can only estimate the true posterior distribution.

To avoid approximating the normalising constant, Møller *et al.* (2006) suggest an MCMC method based on the introduction of an auxiliary variable. This allows the proposal distribution to be constructed in such a way that the normalising constant in the likelihood cancels from the Metropolis Hastings acceptance ratio. For the algorithm to have good mixing and convergence properties, the density of the auxiliary variable should approximate the likelihood. To this end, the authors suggest taking the auxiliary density to be equal to the likelihood evaluated at a fixed estimate of the parameters of the autologistic model, for example, the pseudo–maximum likelihood estimate (Besag, 1975). This is the parameter value maximising the product of the full conditional distributions.

5.3 Description of the NHMM

In this section we introduce the pair of conditional independence assumptions on which the NHMM studied in this chapter is based. We then describe the parameterisations chosen for the weather state process and the precipitation process.

We denote by $\mathbf{X}_t$ the atmospheric data at time t, $t = 1, \ldots, T$, for example, $\mathbf{X}_t$ might include some measure of the sea level pressure over the region in which the sites are located. Like the weather state, S_t, the atmospheric data on day t is common to all sites in the network.

5.3.1 Assumptions of the NHMM

Following Hughes & Guttorp (1994*a*), we relax the "standard" assumption A1 from Chapter 4 to allow the first order hidden Markov chain to be non-homogeneous, with transition probabilities dependent upon time varying atmospheric covariates. For the dataset analysed in Chapter 4, model checks based on the posterior predictive distribution indicated that the temporal dependence induced by the Markovian evolution of the weather state was insufficient to capture the observed persistence in rainfall occurrences at certain sites. Therefore we also relax assumption A2 from Chapter 4 to allow the conditional distribution of rainfall occurrence on day t, $\mathbf{D}_t$,

given all the weather states and the whole history of rainfall occurrences up to and including time $t-1$ to depend on $\mathbf{D}_{t-1}$ as well as S_t. Based on these revised assumptions, the model might more correctly be termed a non-homogeneous Markov switching model (see Section 3.2 for a definition). To our knowledge, the literature does not contain any studies where this type of model has been applied to rainfall data.

Denote the parameters of the NHMM by $\theta = (\theta_{\text{hid}}, \theta_{\text{obs}})$ where θ_{hid} parameterises the hidden process and θ_{obs} parameterises the observed process. The temporal structure in the NHMM considered in this chapter is described by the following assumptions:

A3. $\Pr(S_t \mid S_{0:t-1}, \mathbf{X}_{1:T}, \theta) = \Pr(S_t \mid S_{t-1}, \mathbf{X}_t, \theta_{\text{hid}})$
for $t = 1, \ldots, T$ with $\Pr(S_0 \mid \mathbf{X}_{1:T}, \theta) = \Pr(S_0 = k \mid \theta_{\text{hid}}) = \nu_k$.

A4. $p(\mathbf{w}_t, \mathbf{d}_t \mid \mathbf{w}_{1:t-1}, \mathbf{d}_{0:t-1}, S_{0:T}, \mathbf{X}_{1:T}, \theta) = p(\mathbf{w}_t, \mathbf{d}_t \mid \mathbf{d}_{t-1}, S_t = k, \theta_{\text{obs}})$
for $t = 1, \ldots, T$ with $\Pr(\mathbf{D}_0 \mid S_{0:T}, \mathbf{X}_{1:T}, \theta) = \Pr(\mathbf{D}_0 \mid S_0, \theta_{\text{obs}})$.

Assumption A3 asserts that, given the weather state at the previous time point and the current atmospheric information, the current weather state is conditionally independent of any earlier weather states and of any past or future values of the atmospheric data. The term "non-homogeneous" therefore refers to the way in which the explanatory variables, $\mathbf{X}_t$, adjust the transition probabilities in light of the current atmospheric information. The weather states typically represent particular precipitation patterns which, in turn, are likely to be associated with particular atmospheric patterns. In some sense, therefore, incorporating atmospheric information allows the weather state to provide a classification of both atmospheric and precipitation patterns.

Our second assumption, A4, stipulates that the joint distribution of rainfall occurrence and amount on day t, given the whole hidden process, all atmospheric variables and the history of the observed process up to and including time $t-1$, depends on $\mathbf{D}_{t-1}$ as well as S_t. This allows a refinement in the temporal dependence structure so that on any day, conditional on the weather state, rainfall occurrence on the previous day still affects the joint distribution of rainfall occurrence and amount. In this chapter we do not consider an explicit autoregression on lagged rainfall *amounts* and defer discussion of this more complex structure until Chapter 6.

Note that we have introduced an extra weather state, S_0, so that we can incorporate the atmospheric information on day $t = 1$ in a consistent fashion to that for days $t = 2, \ldots, T$. In the observed process, we also go back an extra time step and introduce a latent occurrence vector, $\mathbf{D}_0$, to reduce the effect of the marginal specification at the beginning of the time series. The ensuing buffering effect could be enhanced by going back further in time, but we choose to avoid this added complication. This is because the dataset studied in Section 5.7 divides naturally into multiple sub–series which we model as independent realisations of the same NHMM, given the atmospheric data. Therefore, we would need to introduce the extra latent variables at the beginning of each sub–series, which would add greatly to the dimension of the parameter space and could impair the performance of the MCMC sampler.

Assumptions A3 and A4 generalise A1 and A2 from Chapter 4, respectively, but we need not consider both generalisations simultaneously. We could adopt assumptions A1 and A4 and consider a (homogeneous) Markov switching model. Similarly, we could adopt assumptions A3

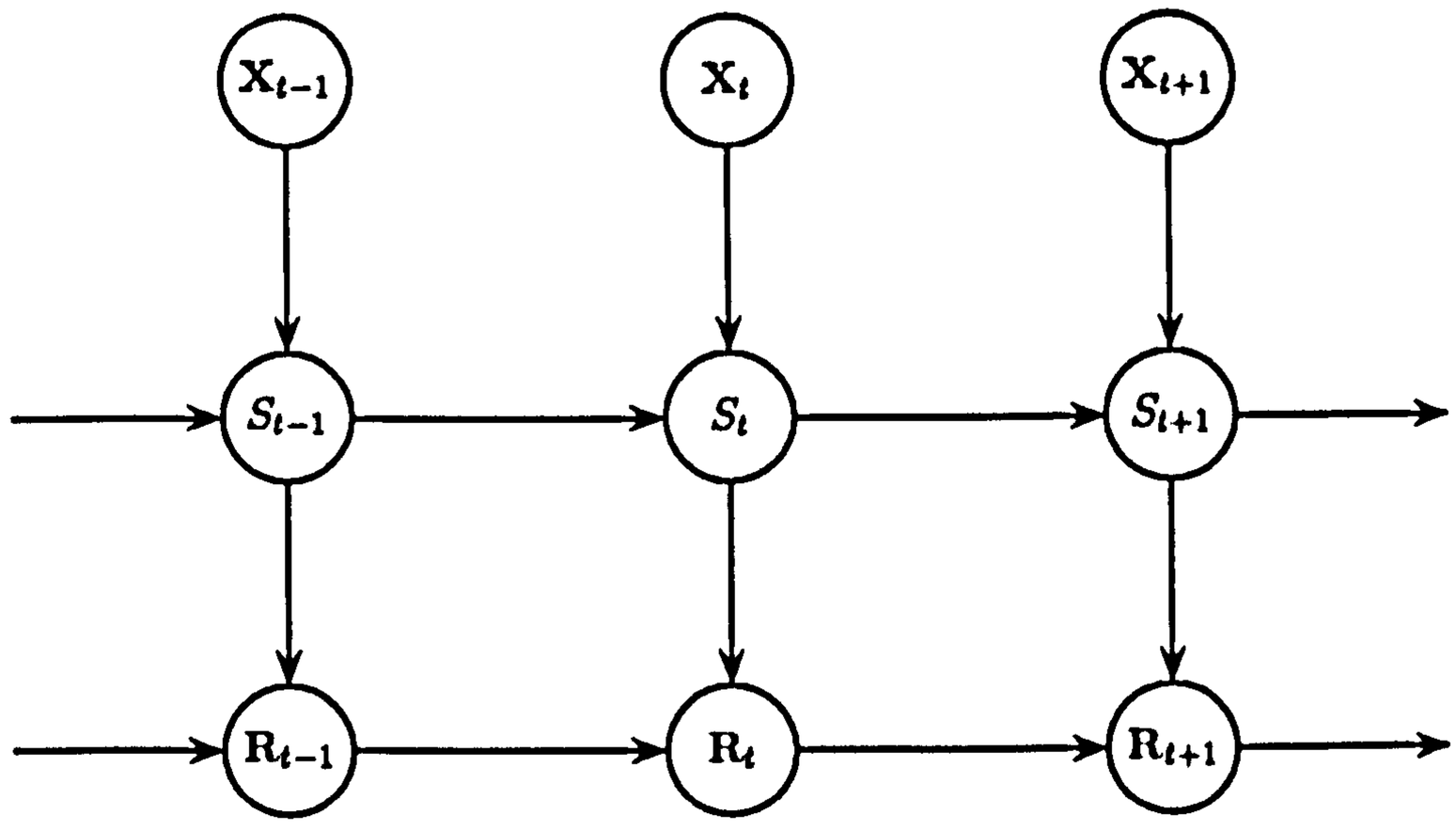


Figure 5.1: A DAG showing the dependence structure in the NHMM described by assumptions A3 and A4, with $\mathbf{R}_t^T = (\mathbf{W}_t^T, \mathbf{D}_t^T)$. Note that $\mathbf{R}_t$ only depends on $\mathbf{R}_{t-1}$ through $\mathbf{D}_{t-1}$.

and A2 and consider a simple non-homogeneous hidden Markov model, for example, a non-homogeneous version of the hidden Markov model studied in Chapter 4. Further comments on this model will be provided in Section 5.3.3.1.

Writing $\mathbf{R}_t^T = (\mathbf{W}_t^T, \mathbf{D}_t^T)$, the temporal dependence structure of the class of models characterised by assumptions A3 and A4 can be represented by the DAG in Figure 5.1. A particular model within this class is defined through the parameterisations chosen for the weather state process $\Pr(S_t \mid S_{t-1}, \mathbf{X}_t, \theta_{\text{hid}})$, and the precipitation process $p(\mathbf{w}_t, \mathbf{d}_t \mid \mathbf{d}_{t-1}, S_t, \theta_{\text{obs}})$.

5.3.2 Parameterisation for the weather state process

This section begins with a brief review of the parameterisations chosen for the weather state process in other NHMMs for rainfall from the literature. The atmospheric data available to us take the form of categorical time varying covariates, so the remainder of the section then discusses possible models for incorporating explanatory variables of this type.

MacDonald & Zucchini (1997) consider two state NHMMs for daily rainfall occurrence at a single site and model seasonality by expressing the logit transformation of each self-transition probability in terms of its truncated Fourier representation. Extending this model when there are more than two states would therefore be non-trivial. In the literature, other NHMMs for precipitation allow the hidden chain to be non-homogeneous by conditioning on time varying atmospheric covariates, as expressed by assumption A3. It is important to appreciate that such NHMMs are specified conditionally on the atmospheric data, which therefore serve only as exogenous or explanatory variables. In other words, a joint conditional distribution for $\{(\mathbf{D}_t, \mathbf{W}_t, S_t)\}$ given $\{\mathbf{X}_t\}$ is specified, rather than a joint distribution for $\{(\mathbf{D}_t, \mathbf{W}_t, S_t, \mathbf{X}_t)\}$

and it is this which makes the model non-stationary. A stationary model could be obtained by adopting a (stationary) distribution for $\{\mathbf{X}_t\}$. Hughes & Guttorp (1994*a*), Hughes *et al.* (1999), Charles *et al.* (1999) and Bellone *et al.* (2000) all use continuous atmospheric data in NHMMs for rainfall, typically comprising linear combinations of high dimensional atmospheric fields which span the region of interest. For example, Hughes & Guttorp (1994*a*) use the first five principal components of the sea level pressure measurements from the multiple node grid covering the study area. In each of these studies, the weather state process was parameterised so that it could be regarded as the product of a baseline transition matrix and a Gaussian kernel of covariates

$$\Pr(S_t = k \mid S_{t-1} = j, \mathbf{X}_t, \theta_{\text{hid}}) \propto \lambda_{jk} \exp\left\{-\frac{1}{2}(\mathbf{X}_t - \mu_{jk})^T \mathbf{V}^{-1}(\mathbf{X}_t - \mu_{jk})\right\}, \qquad (5.2)$$

with the (arbitrary) variance matrix $\mathbf{V}$ set equal to the raw variance matrix of $\mathbf{X}_t$, and the constraints $\sum_{k=1}^{r} \lambda_{jk} = 1$ and $\sum_{k=1}^{r} \mu_{jk} = \mathbf{0}$ imposed to ensure identifiability. Similarly, Robertson *et al.* (2004) incorporated continuous atmospheric data in an NHMM for rainfall occurrence using a multinomial logistic regression model, shown to be equivalent to (5.2) under certain conditions.

The available atmospheric data for the Yorkshire network are Lamb weather types. These were introduced in Chapter 2 and their relationship with daily precipitation was explored. This type of atmospheric data is similar to that used by other authors in that it comprises summaries of atmospheric information, but it differs in that the Lamb weather types are categorical, rather than continuous. As such, incorporating the Lamb weather type data presents a new challenge.

We denote the Lamb weather type on day t by X_t, where $X_t \in \mathcal{Q} = \{1, \ldots, 27\}$. Details of the labelling were presented in Table 2.2. The most natural way of incorporating these data might appear to be a multinomial logistic model (see, for example, Gelman *et al.*, 1995; Congdon, 2005) such as

$$\Pr(S_t = k \mid S_{t-1} = j, X_t = x, \theta_{\text{hid}}) = \frac{\exp(\eta_{jxk})}{\sum_{\ell=1}^{r} \exp(\eta_{jx\ell})}, \qquad (5.3)$$

with

$$\eta_{jxk} = \alpha_k + \beta_{jk} + \gamma_{xk} + \delta_{jxk}.$$

For transitions *into* weather state k, α_k is an overall mean, β_{jk} is the effect of the preceding day's weather state (for weather state j), γ_{xk} is the effect of the current day's Lamb weather type (for type x) and δ_{jxk} is an interaction effect. To ensure that the parameters are identifiable in the likelihood, constraints would be required, for example, fixing α_1 and each β_{j1}, β_{1k}, γ_{x1}, γ_{1k}, δ_{jx1}, δ_{1xk} and δ_{j1k} to be zero. However, since there is no obvious "baseline" weather state, it might be preferable to impose a more symmetric set of constraints such as $\sum_{k=1}^{r} \alpha_k = 0$ and

$$\sum_{j=1}^{r} \beta_{jk} = \sum_{k=1}^{r} \beta_{jk} = \sum_{x=1}^{27} \gamma_{xk} = \sum_{k=1}^{r} \gamma_{xk} = \sum_{j=1}^{r} \delta_{jxk} = \sum_{x=1}^{27} \delta_{jxk} = \sum_{k=1}^{r} \delta_{jxk} = 0$$

for each j, x, k or pair thereof. We could assign a multivariate normal prior to all effects except, say, α_r and each β_{rk}, β_{jr}, $\gamma_{27,k}$, γ_{xr}, δ_{rxk}, $\delta_{j,27,k}$ and δ_{jxr} which are defined, by subtraction, to satisfy the constraints above. Within this prior it might be sensible to give the interaction effects, δ_{jxk}, zero means and small variances to penalise more complicated models. It would also be

reasonably straightforward to build a prior dependence structure which encouraged borrowing of strength between Lamb weather types. For example, we might specify high *a priori* correlations between the parameters $\gamma_{1k}, \gamma_{2k}, \ldots, \gamma_{27,k}$ for each $k \in \mathcal{S}_r$.

Although the parameterisation (5.3) is appealing for its potential to build a sophisticated prior dependence structure, the entanglement of the very large number of parameters in the likelihood, and the absence of a conjugate prior for any of them, make it unattractive in terms of performing inference via MCMC. A parameterisation which might lend itself more naturally to this kind of analysis would be

$$\Pr(S_t = k \mid S_{t-1} = j, X_t = x, \theta_{\text{hid}}) = A^x_{jk} \tag{5.4}$$

for $t = 1, \ldots, T$, where we ensure identifiability by imposing the constraints $\sum_{k=1}^{r} A^x_{jk} = 1$ for each pair $(j, x) \in \mathcal{S}_r \times \mathcal{Q}$. In other words, for every combination of the previous day's weather state and the current day's Lamb weather type, a different stochastic vector $\mathbf{A}^x_j = (A^x_{j1}, \ldots, A^x_{jr}) \in \mathcal{S}_r$ governs the probabilities of transition to the current day's weather state. Using this parameterisation, a conjugate prior in the form of the Dirichlet distribution is now available for each stochastic vector $\mathbf{A}^x_j$. For notational convenience we denote the collection of stochastic vectors by

$$\mathbf{A} = (\mathbf{A}^1_1, \ldots, \mathbf{A}^{27}_1, \ldots, \mathbf{A}^1_r, \ldots, \mathbf{A}^{27}_r) \in \mathcal{S}_r^{27r}.$$

By adopting a hierarchical Dirichlet prior for $\mathbf{A}$, we can benefit from (semi)-conjugacy whilst encouraging borrowing of strength between the elements of $\mathbf{A}$. This will facilitate (indirect) learning about some of the more rare (j, x) combinations. Further details will be given in Section 5.4.

In the context of homogeneous hidden Markov models, various possibilities for specifying the initial distribution, in this case the distribution of S_0, were discussed in Chapter 3. One of these was the stationary distribution of the Markov chain. For the NHMM, however, the absence of a model for the atmospheric data, $\{X_t\}$, makes the Markov chain non-stationary, and so it cannot be initialised at its stationary distribution. Denote $\Pr(S_0 = k \mid \theta_{\text{hid}}) = \nu_k$ with $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_r) \in \mathcal{S}_r$. As with the real data applications in Chapter 4, we choose to make S_0 random, with a distribution, $\boldsymbol{\nu}$, that does not depend on any of the transition probabilities in $\mathbf{A}$.

5.3.3 Parameterisation for the precipitation process

In this section, we begin by outlining a simple non-homogeneous extension to the hidden Markov model from Chapter 4. We then describe the more complex within-state model for precipitation that will be studied in the remainder of this chapter.

5.3.3.1 A simple within-state model

In building an NHMM for rainfall, a more natural development of the hidden Markov model described in Chapter 4 might have been to alter only the assumptions of the hidden process. This would lead to a model in which assumption A3 from Section 5.3.1 characterised the temporal dependence structure of the hidden process, whilst assumption A2 from Chapter 4 characterised

that of the observed process. The weather state process could be parameterised according to equation (5.4), whilst the precipitation process could be modelled according to Section 4.2.2, leading to a non-homogeneous model which assumes conditional spatial and temporal independence in the joint distribution of rainfall occurrence and amount, given the weather state. We fitted this model to the Yorkshire dataset, but its predictive capabilities did not noticeably differ from those of the model from Chapter 4. Therefore what follows is only a brief summary of the results of this analysis.

In the simple NHMM described above, the priors chosen for the hidden process parameters were the same as those which will be outlined later in this chapter (see Section 5.4), whilst the priors for the observed process parameters and the number of states, r, were the same as those chosen in Chapter 4 (see Section 4.3). Within these priors, we elicited the same values for the fixed hyperparameters as those chosen in Sections 5.7.1 and 4.7.1.2 for the hidden and observed process parameters, respectively. For the Yorkshire dataset, we computed the log marginal likelihood for the NHMMs with $r = 1, \ldots, 5$ states. Comparison with the corresponding values for the hidden Markov model revealed that the NHMM had a moderately larger log marginal likelihood for each $r = 1, \ldots, 5$. This suggested that under the chosen prior specifications, the simple NHMM offers a more likely explanation of the data than the corresponding homogeneous hidden Markov model. In spite of these differences, the plots used in model checking for the homogeneous hidden Markov model (see Section 4.7.4) and this NHMM (not shown) were virtually indistinguishable. In particular, the NHMM was still unable to reproduce the temporal and spatial dependence when that dependence was high. In the remainder of this section, we describe the parameterisation of the precipitation process that will be studied further in this chapter.

5.3.3.2 Allowing spatio-temporal dependence in the within-state model

From assumption A4, the joint density of the variables in the NHMM is factorised in such a way that the conditional distribution of $(\mathbf{W}_t, \mathbf{D}_t)$ depends on $\mathbf{D}_{t-1}$ as well as S_t. We factorise the joint mixed density and mass function of $(\mathbf{W}_t, \mathbf{D}_t)$, given $(\mathbf{D}_{t-1}, S_t)$, as

$$\begin{aligned} &p(\mathbf{w}_t, \mathbf{d}_t \mid \mathbf{d}_{t-1}, S_t = k, \theta_{\text{obs}}) \\ &\qquad = \Pr(\mathbf{D}_t = \mathbf{d}_t \mid \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, S_t = k, \theta_{\text{obs,k}}) p(\mathbf{w}_t \mid \mathbf{D}_t = \mathbf{d}_t, S_t = k, \theta_{\text{obs,k}}), \end{aligned} \tag{5.5}$$

then the overall model can be represented graphically by the DAG in Figure 5.2. In this chapter we focus on adding spatio-temporal structure to the occurrence process and assume that $\mathbf{W}_t$ is conditionally independent of $\mathbf{D}_{t-1}$, given $\mathbf{D}_t$ and the weather state, S_t. Further, we assume conditional spatial independence of rainfall amounts, given occurrences and the weather state, so that

$$p(\mathbf{w}_t \mid \mathbf{D}_t = \mathbf{d}_t, S_t = k, \theta_{\text{obs,k}}) = \prod_{i=1}^{n} p(w_t^i \mid D_t^i = d_t^i, S_t = k, \theta_{\text{obs,k}}), \tag{5.6}$$

where

$$\Pr(W_t^i = 0 \mid D_t^i = 0) = 1, \quad (W_t^i \mid D_t^i = 1, S_t = k, \theta_{\text{obs,k}}) \sim \text{Ga}\left(\frac{1}{v_{ik}^2}, \frac{1}{v_{ik}^2 m_{ik}}\right), \tag{5.7}$$

for sites $i = 1, \ldots, n$, which is identical to the rainfall amounts model adopted in Chapter 4. Dependence on $\mathbf{D}_{t-1}$ could be incorporated by adopting a different set of parameters in each

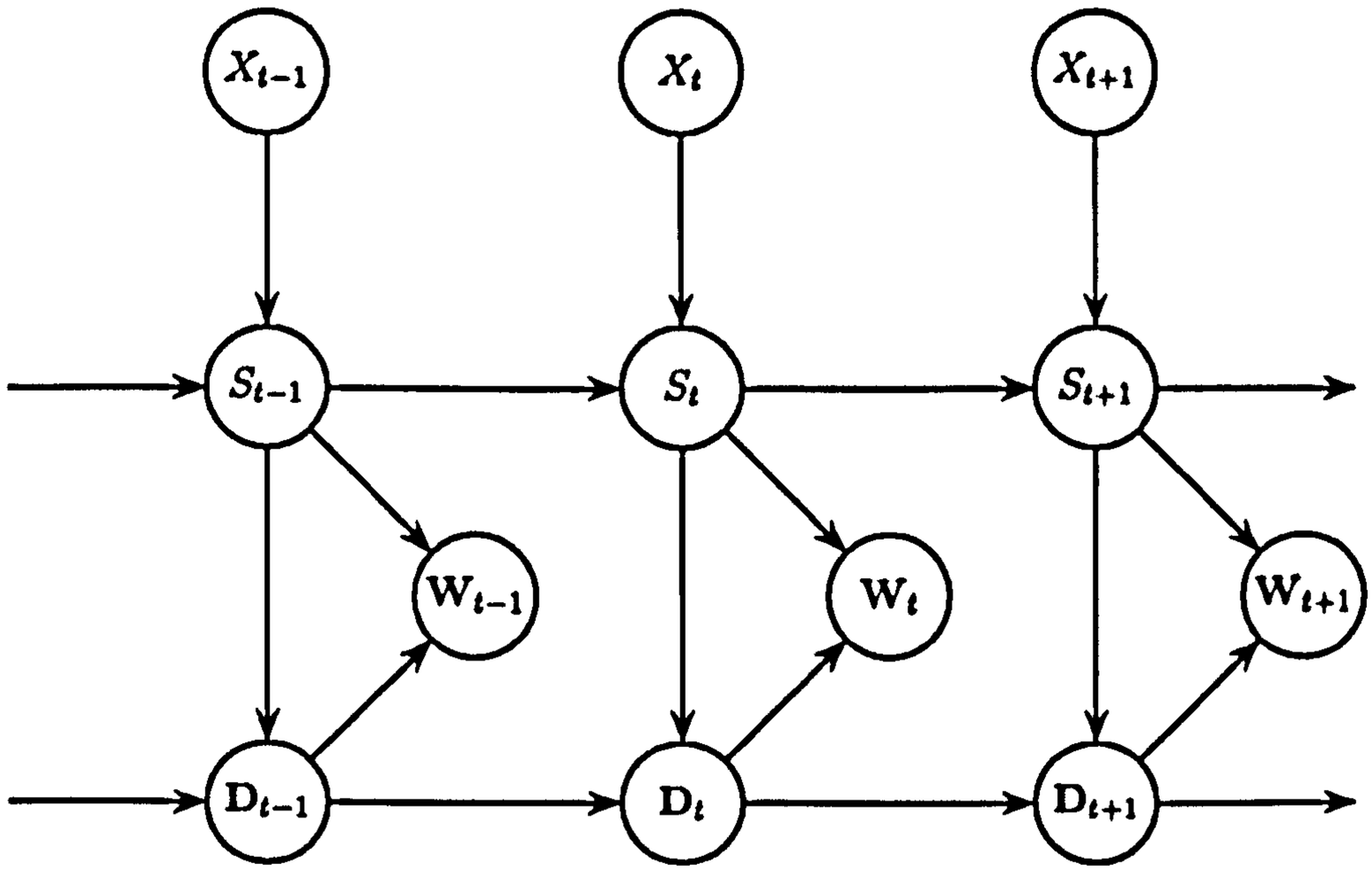


Figure 5.2: A DAG showing the (temporal) dependence structure in the class of NHMMs described by assumptions A3 and A4 and the factorisation of the joint mixed density and mass function $p(\mathbf{w}_t, \mathbf{d}_t \mid \mathbf{d}_{t-1}, S_t, \boldsymbol{\theta}_{\text{obs}})$ given in equation (5.5).

gamma distribution, depending on whether or not it rained the previous day. However, other authors have not found such extensions to be useful; see, for example, Stern & Coe (1984) or Woolhiser & Roldán (1982). Adding further spatio-temporal structure to the amounts process, such as autoregressions on the values at neighbouring sites, or on lagged values at the site in question, is difficult because of the presence of zeros. We provide further comments in Section 5.8 when considering extensions to the model, but postpone the introduction of more sophisticated models for non-zero rainfall amounts until Chapter 6.

In the initial rainfall occurrence distribution, $\Pr(\mathbf{D}_0 = \mathbf{d}_0 \mid S_0 = k, \boldsymbol{\theta}_{\text{obs},k})$, we assume that $D_0^1, \ldots, D_0^n$ are independent of S_0 and of each other with

$$D_0^i \sim \text{Bern}(p_0^i), \qquad \text{independently for } i = 1, \ldots, n$$

where $p_0^i \in [0,1]$ is fixed. A more sophisticated approach would have been to allow $p_0^1, \ldots, p_0^n$ to be variable and to give them a prior which encouraged borrowing of strength between sites; see Section 4.3 for an example of this kind of prior.

To incorporate spatio-temporal structure in the within-state model for rainfall occurrences, we adopt a Markov chain of autologistic models, as introduced in Section 5.2. Our model for rainfall

occurrence is given by

$$\Pr(\mathbf{D}_t = \mathbf{d}_t \,|\mathbf{D}_{t-1} = \mathbf{d}_{t-1}, S_t = k, \theta_{\text{obs}}) = \frac{\exp\left(\sum_{i=1}^{n} \alpha_{ik} d_t^i + \sum_{i=2}^{n}\sum_{j=1}^{i-1} \beta_{ijk} d_t^i d_t^j + \sum_{i=1}^{n} \gamma_{ik} d_t^i d_{t-1}^i\right)}{\sum_{\mathbf{d}} \exp\left(\sum_{i=1}^{n} \alpha_{ik} d^i + \sum_{i=2}^{n}\sum_{j=1}^{i-1} \beta_{ijk} d^i d^j + \sum_{i=1}^{n} \gamma_{ik} d^i d_{t-1}^i\right)} \tag{5.8}$$

$$\propto \exp\left(\sum_{i=1}^{n} \alpha_{ik} d_t^i + \sum_{i=2}^{n}\sum_{j=1}^{i-1} \beta_{ijk} d_t^i d_t^j + \sum_{i=1}^{n} \gamma_{ik} d_t^i d_{t-1}^i\right), \tag{5.9}$$

for $t = 1, \ldots, T$. The sum in (5.8) is over all 2^n possible rainfall occurrence vectors, and $\beta_{ijk} = 0$ unless sites i and j are neighbours. The "normalising constant" at time t depends on S_t and $\mathbf{D}_{t-1}$ and henceforth will be denoted by

$$C_{s_t}\{\mathcal{I}(\mathbf{d}_{t-1}) \mid \theta_{\text{obs},s_t}\} = \sum_{\mathbf{d}} \exp\left(\sum_{i=1}^{n} \alpha_{is_t} d^i + \sum_{i=2}^{n}\sum_{j=1}^{i-1} \beta_{ijs_t} d^i d^j + \sum_{i=1}^{n} \gamma_{is_t} d^i d_{t-1}^i\right), \tag{5.10}$$

where, for notational convenience, we use $\mathcal{I}(\mathbf{d}_{t-1}) = \sum_{i=1}^{n} d_{t-1}^i 2^{n-i} \in \{0, 1, \ldots, 2^n - 1\}$ to denote the numerical labelling of $\mathbf{d}_{t-1}$.

The parameters α_{ik} and β_{ijk} can be interpreted as spatial trend and spatial dependence parameters, respectively, whilst the γ_{ik} can be viewed as temporal dependence parameters. In the special case when $\gamma_{ik} = 0$ for all i, k we recover Besag's standard autologistic model, as used by Hughes & Guttorp (1994*b*) and Hughes *et al.* (1999) in their NHMMs for rainfall occurrence. When $\beta_{ijk} = 0$ for all i, j, k we obtain a model which assumes conditional spatial, but not temporal, independence amongst rainfall occurrences, given the weather state. Temporally, the occurrence process at each site would be described as a first order Markov chain, conditional on the weather state, with transition matrix

$$\begin{pmatrix} 1 - p_{ik}^0 & p_{ik}^0 \\ 1 - p_{ik}^1 & p_{ik}^1 \end{pmatrix} = \begin{pmatrix} \dfrac{1}{1 + \exp(\alpha_{ik})} & \dfrac{\exp(\alpha_{ik})}{1 + \exp(\alpha_{ik})} \\ \dfrac{1}{1 + \exp(\alpha_{ik} + \gamma_{ik})} & \dfrac{\exp(\alpha_{ik} + \gamma_{ik})}{1 + \exp(\alpha_{ik} + \gamma_{ik})} \end{pmatrix}$$

for site i and weather state k. In the above notation, p_{ik}^0 and p_{ik}^1 would be the probabilities of rain at site i in weather state k, given that the previous day at site i was dry and wet, respectively. Finally, when $\beta_{ijk} = \gamma_{ik} = 0$ for all i, j, k we recover the conditional spatial independence model for occurrences studied in Chapter 4, with $p_{ik} = e^{\alpha_{ik}}/(1 + e^{\alpha_{ik}})$.

The absolute magnitude of β_{ijk} indicates the strength of the pairwise dependence, with larger absolute values indicating stronger associations, whilst the sign indicates whether the spatial dependence is positive or negative. We follow Hughes *et al.* (1999) and assume that every site, $j \neq i$, belongs to the set of neighbours of site i, for $i = 1, \ldots, n$. In other words, we do not make any assumptions of conditional independence between sites and so all the β_{ijk} $(i \neq j)$ are assumed non-zero. A simplification suggested by Hughes & Guttorp (1994*b*) but not considered here, would be to define a distance, say $h > 0$, such that only sites separated by a distance less than h are regarded as neighbours. Note that without making any assumptions of conditional

independence between sites, we cannot simplify computation of the normalising constant using the forward recursion discussed in Section 5.2.2.

Conditional on the weather state, the absolute magnitude of γ_{ik} indicates the strength of the lag one persistence in wet spells, with larger absolute values indicating stronger persistence, and the sign of γ_{ik} indicating whether the (conditional) temporal dependence is positive or negative. Since

$$\Pr(D_t^i = 1 \mid \mathbf{D}_t^{-i} = \mathbf{d}_t^{-i}, \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, S_t = k, \theta_{\text{obs},k}) = \frac{\exp\left(\alpha_{ik} + \sum_{j \neq i} \beta_{ijk} d_t^j + \gamma_{ik} d_{t-1}^i\right)}{1 + \exp(\alpha_{ik} + \sum_{j \neq i} \beta_{ijk} d_t^j + \gamma_{ik} d_{t-1}^i)}, \tag{5.11}$$

where $\beta_{ijk} = \beta_{jik}$, the model is defined so that the only lagged rainfall occurrence to affect the conditional probability $\Pr(D_t^i = 1 \mid \mathbf{D}_t^{-i} = \mathbf{d}_t^{-i}, \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, S_t = k, \theta_{\text{obs},k})$ is that at site i. A more sophisticated temporal model might include temporal interactions across sites, that is, terms of the form $d_t^i d_{t-1}^j$ for $i \neq j$. Such models are considered in Guyon & Hardouin (2002).

5.4 Prior distribution

This section provides a description and justification of the prior distribution chosen for the model parameters from equations (5.4), (5.7) and (5.8). Several of the prior components are specified with hierarchical structure in order to encourage borrowing of strength between similar parameters, meaning that correlations between them will have to be considered, as well as means and variances. When, for example, parameters have the same function at different sites, such correlated priors seem intuitively reasonable. Moreover, from a pragmatic perspective, they are often necessary for the convergence of the MCMC sampler. The section is concluded with a series of guidelines for choosing hyperparameters in order to incorporate prior knowledge. For example, we suggest an elicitation strategy for choosing the hyperparameters in a hierarchical Dirichlet prior, in such a way that marginal correlations do not have to be specified directly.

For convenience in notation, we collect the set of spatial trend parameters in the occurrence process into a $n \times r$ matrix $\mathcal{A}$ with (i,k)–th entry α_{ik}. Similarly, we collect the spatial and temporal dependence parameters into $n(n-1)/2 \times r$ and $n \times r$ matrices $\mathcal{B} = (\beta_{ijk})$ and $\mathcal{G} = (\gamma_{ik})$, respectively. For the parameters in the distributions of non–zero rainfall amounts we adopt the notation from Chapter 4 and collect the mean and coefficient of variation parameters into $n \times r$ matrices $\mathcal{M} = (m_{ik})$ and $\mathcal{V} = (v_{ik})$, respectively. The set of all model parameters is therefore denoted by $\theta = (\theta_{\text{hid}}, \theta_{\text{obs}})$ where

$$\theta_{\text{hid}} = (\mathbf{A}, \nu) \in \mathscr{S}_r^{27r} \times \mathscr{S}_r \quad \text{and} \quad \theta_{\text{obs}} = (\mathcal{A}, \mathcal{B}, \mathcal{G}, \mathcal{M}, \mathcal{V}) \in \mathbb{R}^{nr} \times \mathbb{R}^{n(n-1)r/2} \times \mathbb{R}^{nr} \times \mathbb{R}_+^{nr} \times \mathbb{R}_+^{nr}.$$

Uncertainty about the unknown model parameters, *a priori*, is expressed through a prior distribution of the form

$$\pi(\theta) = \pi(\theta_{\text{hid}})\pi(\theta_{\text{obs}}) = \pi(\mathbf{A})\pi(\nu)\pi(\mathcal{A})\pi(\mathcal{B})\pi(\mathcal{G})\pi(\mathcal{M})\pi(\mathcal{V}), \tag{5.12}$$

which we will assume to be exchangeable across weather states. This prior distribution expresses an assumption of *a priori* independence between θ_{hid} and θ_{obs}, and between each of the parameter blocks within these two components of θ. Since the quantities about which we hold

prior beliefs are likely to include the marginal probability of rain at each site, this assumption might be considered untenable for the parameters of the occurrence process $(\mathcal{A}, \mathcal{B}, \mathcal{G})$ as changing prior beliefs about one of these parameter blocks would necessitate a change in belief about the others. For example, α_{ik} and β_{ijk} are actually likely to be negatively correlated in our prior beliefs because increasing one of them without reducing the other would increase the marginal probability of rain at site i in weather state k. However, especially when the number of sites, n, is large, assessing the effect of learning the value of one parameter on our beliefs about others is very difficult. In the absence of credible knowledge about the dependence structure, therefore, we argue that it is less harmful to assume *a priori* independence between $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{G}$ than to adopt a more sophisticated prior dependence structure whose correlation parameters would be difficult to solicit reliably. From a Bayesian perspective, this lack of transparency might be regarded as a criticism of the autologistic model.

Beginning with the parameters which also appeared in the model from Chapter 4, we continue to adopt the conjugate Dirichlet prior,

$$\boldsymbol{\nu} \sim \mathscr{D}_r(G\mathbf{g}), \tag{5.13}$$

for the initial distribution, $\boldsymbol{\nu}$. Note that the information content parameter, $G \in \mathbb{R}^+$, and mean, $\mathbf{g} \in \mathscr{S}_r$, are fixed hyperparameters. For the parameters of the rainfall amounts process, $\mathcal{M}$ and $\mathcal{V}$, we adopt the prior distributions described in Section 4.3 for the simple hidden Markov model.

Consider the parameters of the occurrence process, $(\mathcal{A}, \mathcal{B}, \mathcal{G})$. In the priors for the spatial trend parameters in $\mathcal{A}$, we assume independence between weather states, and similarly for the spatial and temporal dependence parameters in $\mathcal{B}$ and $\mathcal{G}$, respectively. This seems reasonable since the weather states are likely to be associated with broadly unrelated patterns of rainfall. In Section 5.2.1 we remarked that multicollinearity is a commonly reported problem concerning autologistic models, arising when the dependencies between site i and sites $j \neq i$ are highly correlated. Although this will be discussed further in Section 5.7.3.1, briefly, (partial) multicollinearity leads to flat ridges on the likelihood surface in the directions of affected parameters, making these parameters difficult to identify. It is therefore likely that we will need priors which encourage borrowing of strength. If there are some parameters about which the data *are* informative, such correlated priors allow this information to update belief about related parameters, for which the data are less informative, through the influence of the prior in their posterior distributions.

We therefore adopt hierarchical priors such that, for each $k \in S_r$,

$$\alpha_{ik} \mid \alpha_k, \sigma^2_{\alpha,k} \sim \mathrm{N}(\alpha_k, \sigma^2_{\alpha,k}) \quad \text{independently for } i \in \{1, \ldots, n\}, \tag{5.14}$$

$$\beta_{ijk} \mid \beta_k, \sigma^2_{\beta,k} \sim \mathrm{N}(\beta_k, \sigma^2_{\beta,k}) \quad \text{independently for } (i,j) \in \{(i,j) : i = 2, \ldots, n, \; j = 1, \ldots, i-1\}, \tag{5.15}$$

$$\gamma_{ik} \mid \gamma_k, \sigma^2_{\gamma,k} \sim \mathrm{N}(\gamma_k, \sigma^2_{\gamma,k}) \quad \text{independently for } i \in \{1, \ldots, n\}, \tag{5.16}$$

where

$$\alpha_k \sim \mathrm{N}(a_{0,\alpha,k}, a^2_{1,\alpha,k}), \qquad \sigma^2_{\alpha,k} \sim \mathrm{IG}(h_{0,\alpha,k}, h_{1,\alpha,k}), \tag{5.17}$$

$$\beta_k \sim \mathrm{N}(a_{0,\beta,k}, a^2_{1,\beta,k}), \qquad \sigma^2_{\beta,k} \sim \mathrm{IG}(h_{0,\beta,k}, h_{1,\beta,k}), \tag{5.18}$$

$$\gamma_k \sim \mathrm{N}(a_{0,\gamma,k}, a^2_{1,\gamma,k}), \qquad \sigma^2_{\gamma,k} \sim \mathrm{IG}(h_{0,\gamma,k}, h_{1,\gamma,k}). \tag{5.19}$$

Note that we could, more simply, construct a hierarchical prior in which the variances $\sigma^2_{\alpha,k}$, $\sigma^2_{\beta,k}$ and $\sigma^2_{\gamma,k}$ were fixed. Under this prior, the amount of information that, say, α_{ik} could convey about α_{jk} (through the prior) would be fixed, even if the data actually suggested that α_{ik} and α_{jk} were not particularly alike. The effect of allowing $\sigma^2_{\alpha,k}$, $\sigma^2_{\beta,k}$ and $\sigma^2_{\gamma,k}$ to be random variables is to allow the amount of information provided by α_{ik} about α_{jk} to depend on how similar the data suggest α_{ik} and α_{jk} are. In other words the amount of borrowing of strength can be influenced by the data, rather than being fixed *a priori*.

Using the law of total expectation it can be demonstrated that, marginally,

$$\mathrm{E}(\alpha_{ik}) = a_{0,\alpha,k}, \qquad \mathrm{Var}(\alpha_{ik}) = \frac{h_{1,\alpha,k}}{h_{0,\alpha,k}-1} + a^2_{1,\alpha,k}, \qquad i = 1,\ldots,n$$

$$\mathrm{Corr}(\alpha_{ik},\alpha_{jk}) = \frac{a^2_{1,\alpha,k}}{h_{1,\alpha,k}/(h_{0,\alpha,k}-1) + a^2_{1,\alpha,k}} \qquad i,j = 1,\ldots,n,\ i \neq j$$

for each $k \in \mathcal{S}_r$, with analogous expressions for the β_{ijk} and γ_{ik}. The term $a^2_{1,\alpha,k}$ represents *shared variance*, whilst the mean of the distribution for $\sigma^2_{\alpha,k}$, namely $h_{1,\alpha,k}/(h_{0,\alpha,k}-1)$, represents *specific variance*. We can increase the marginal correlations by choosing values for the hyperparameters which make the shared variance large compared to the specific variance.

The choice of hyperparameters $\{a_{0,\alpha,k}, a^2_{1,\alpha,k}, h_{0,\alpha,k}, h_{1,\alpha,k} : k \in \mathcal{S}_r\}$ for the α_{ik}, and equivalently for the β_{ijk} and γ_{ik}, will be discussed further in Section 5.4.1. For notational convenience, we denote $\mathcal{A}^0 = (\alpha_1,\ldots,\alpha_r,\sigma^2_{\alpha,1},\ldots,\sigma^2_{\alpha,r})$ with $\mathcal{B}^0$ and $\mathcal{G}^0$ defined analogously. We append $(\mathcal{A}^0,\mathcal{B}^0,\mathcal{G}^0)$ to the (observed process) model parameters, θ_{obs}, and replace the product $\pi(\mathcal{A})\pi(\mathcal{B})\pi(\mathcal{G})$ in the prior distribution (5.12) with

$$\pi(\mathcal{A} \mid \mathcal{A}^0)\pi(\mathcal{B} \mid \mathcal{B}^0)\pi(\mathcal{G} \mid \mathcal{G}^0)\pi(\mathcal{A}^0)\pi(\mathcal{B}^0)\pi(\mathcal{G}^0)$$

which then factorises as

$$\prod_{k=1}^{r}\Bigg\{\pi(\alpha_k)\pi(\sigma^2_{\alpha,k})\prod_{i=1}^{n}\pi(\alpha_{ik} \mid \alpha_k,\sigma^2_{\alpha,k}) \times \pi(\beta_k)\pi(\sigma^2_{\beta,k})\prod_{i=2}^{n}\prod_{j=1}^{i-1}\pi(\beta_{ijk} \mid \beta_k,\sigma^2_{\beta,k}) \times \pi(\gamma_k)\pi(\sigma^2_{\gamma,k})\prod_{i=1}^{n}\pi(\gamma_{ik} \mid \gamma_k,\sigma^2_{\gamma,k})\Bigg\}.$$

Figure 2.2 indicated that there are some Lamb weather types that occur very infrequently. This means that the data are unlikely to be very informative about some of the stochastic vectors in A. A natural way of thinking about how the atmospheric data might influence the transition probabilities is to suppose there is some underlying baseline transition matrix. We might then imagine that particular Lamb weather types could lead to deviations from this baseline. The parameterisation (5.2) exemplifies this way of thinking, with the atmospheric data adjusting the baseline transition matrix in a multiplicative way. We can encourage a similar structure by adopting a set of two–stage or hierarchical Dirichlet priors,

$$\mathbf{A}^x_j \mid \boldsymbol{\xi}_j \overset{\text{iid}}{\sim} \mathscr{D}_r(\Xi_j\boldsymbol{\xi}_j), \quad \boldsymbol{\xi}_j \sim \mathscr{D}_r(E_j\mathbf{e}_j), \tag{5.20}$$

independently for each $j \in \mathcal{S}_r$, where $\Xi_j \in \mathbb{R}^+$, $E_j \in \mathbb{R}^+$ and $\mathbf{e}_j \in \mathscr{S}_r$ are fixed hyperparameters, whilst $\boldsymbol{\xi}_j$ is the (variable) mean in the conditional prior for $(\mathbf{A}_j^x \mid \boldsymbol{\xi}_j)$, $x \in Q$. Here $\mathbf{A}_j^x = (A_{j1}^x, \ldots, A_{jr}^x)$, where $A_{jk}^x = \Pr(S_t = k \mid S_{t-1} = j, X_t = x, \boldsymbol{\theta}_{\text{hid}})$. This is analogous to a hierarchical normal prior; if $\boldsymbol{\xi}_j$ and the $(\mathbf{A}_j^x \mid \boldsymbol{\xi}_j)$ were multivariate normal, then $\boldsymbol{\xi}_j$ would represent a baseline vector from which $\mathbf{A}_j^x$, for each Lamb weather type x, deviated in an additive manner. We can therefore think of the $(r \times r)$ matrix with j-th row equal to $\boldsymbol{\xi}_j$ as being like a baseline transition matrix, although the posterior will not actually take this hierarchical form. More formally, the prior (5.20) is defined so that the blocks of stochastic vectors $(\mathbf{A}_j^1, \ldots, \mathbf{A}_j^{27})$ and $(\mathbf{A}_k^1, \ldots, \mathbf{A}_k^{27})$ are independent for each distinct pair of weather states $j, k \in \mathcal{S}_r$. However, within each block the stochastic vectors $\mathbf{A}_j^1, \ldots, \mathbf{A}_j^{27}$ are positively correlated, expressing the belief that if, for example, A_{jk}^x was found to be larger (smaller) than its mean, this would lead to an upward (downward) revision of our beliefs about the mean of A_{jk}^y for $y \neq x$, $x, y \in Q$.

The distribution (5.20) belongs to a family which has been termed a (continuous) mixture of Dirichlets (see, for example, Albert & Gupta, 1982). One of its main benefits is that it is (semi)-conjugate to the multinomial form of the complete data likelihood. Note that we could additionally allow Ξ_j to be variable and assign to it a distribution on $\mathbb{R}^+$. As explained previously, the effect of assigning a (non-degenerate) distribution to the second-stage variance-like parameter in a hierarchical prior is to allow the amount of borrowing of strength to be influenced by the data.

Using the law of total expectation it can be shown that the marginal means, variances and covariances across Lamb weather types are

$$\mathrm{E}(\mathbf{A}_j^x) = \mathrm{E}\{\mathrm{E}(\mathbf{A}_j^x \mid \boldsymbol{\xi}_j)\} = \mathbf{e}_j,$$

$$\mathrm{Var}(A_{jk}^x) = \mathrm{E}\{\mathrm{Var}(A_{jk}^x \mid \boldsymbol{\xi}_j)\} + \mathrm{Var}\{\mathrm{E}(A_{jk}^x \mid \boldsymbol{\xi}_j)\} = \frac{e_{jk}(1-e_{jk})(\Xi_j + E_j + 1)}{(E_j + 1)(\Xi_j + 1)},$$

$$\mathrm{Cov}(A_{jk}^x, A_{jk}^y) = \mathrm{E}\{\mathrm{Cov}(A_{jk}^x, A_{jk}^y \mid \boldsymbol{\xi}_j)\} + \mathrm{Cov}\{\mathrm{E}(A_{jk}^x \mid \boldsymbol{\xi}_j), \mathrm{E}(A_{jk}^y \mid \boldsymbol{\xi}_j)\} = \frac{e_{jk}(1-e_{jk})}{E_j + 1},$$

so that, marginally, the correlation between A_{jk}^x and A_{jk}^y for $x \neq y$ is given by

$$\mathrm{Corr}(A_{jk}^x, A_{jk}^y) = \frac{\mathrm{Cov}(A_{jk}^x, A_{jk}^y)}{\sqrt{\mathrm{Var}(A_{jk}^x)\mathrm{Var}(A_{jk}^y)}} = \frac{\Xi_j + 1}{E_j + \Xi_j + 1}. \tag{5.21}$$

Note that the parameter E_j reflects uncertainty shared by $\mathbf{A}_j^1, \mathbf{A}_j^2, \ldots, \mathbf{A}_j^{27}$ whilst Ξ_j reflects uncertainty specific to a particular $\mathbf{A}_j^x$. For any $j \in \mathcal{S}_r$, in the limit as $\Xi_j \to \infty$, then $\mathrm{Var}(A_{jk}^x \mid \boldsymbol{\xi}_j) \to 0$, and we obtain $\mathbf{A}_j^x = \boldsymbol{\xi}_j \sim \mathscr{D}_r(E_j \mathbf{e}_j)$ for each $x \in Q$. In other words, $\mathbf{A}_j^1, \mathbf{A}_j^2, \ldots, \mathbf{A}_j^{27}$ become perfectly positively correlated and we recover a (homogeneous) hidden Markov model. This fact is borne out by equation (5.21) which shows that the marginal correlation between A_{jk}^x and A_{jk}^y, $x \neq y$, approaches one as Ξ_j approaches infinity. Likewise in the limit as the shared uncertainty approaches zero ($E_j \to \infty$), then $\boldsymbol{\xi}_j = \mathbf{e}_j$, and we obtain a prior which assumes independence between all stochastic vectors. Again this is substantiated by equation (5.21) which shows that the marginal correlation between A_{jk}^x and A_{jk}^y, $x \neq y$, approaches zero as E_j approaches infinity.

We can therefore express prior belief in strong correlations amongst $\mathbf{A}_j^1, \mathbf{A}_j^2, \ldots, \mathbf{A}_j^{27}$ by making the shared uncertainty large relative to the specific uncertainty, that is, by choosing E_j to be

small relative to Ξ_j. In practice it will be difficult to elicit values for E_j and Ξ_j based on these vague notions of shared and specific uncertainty or in terms of the above moments. The objective of Section 5.4.2 is therefore to outline an alternative approach. In this and subsequent sections, it will be convenient to introduce the notation $\mathcal{E} = (\xi_1, \ldots, \xi_r) \in \mathcal{S}_r^r$ and append $\mathcal{E}$ to the set of (hidden process) model parameters θ_{hid}. The term $\pi(\mathbf{A})$ in the prior distribution (5.12) should then be replaced with $\pi(\mathbf{A} \mid \mathcal{E})\pi(\mathcal{E})$ which factorises as

$$\prod_{j=1}^{r} \pi(\xi_j) \prod_{x=1}^{27} \pi(\mathbf{A}_j^x \mid \xi_j).$$

5.4.1 Prior beliefs about the parameters of the rainfall occurrence process

To give a prior for the parameters of the rainfall occurrence process which is exchangeable across weather states, we need to choose the hyperparameters to be the same for each weather state, for example, $a_{0,\alpha,k} = a_{0,\alpha}$, $a_{1,\alpha,k}^2 = a_{1,\alpha}^2$, $h_{0,\alpha,k} = h_{0,\alpha}$ and $h_{1,\alpha,k} = h_{1,\alpha}$ for all $k \in S_r$ in the prior for $\mathcal{A}^0$. Eliciting values for these hyperparameters is not easy because the expectations of quantities that we can, in principle, observe correspond to very complicated combinations of the parameters in $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{G}$. For precisely this reason it is very difficult to think about the priors for each of $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{G}$ separately due to their correlation in our prior beliefs. However, as we have assumed *a priori* independence between them, we are forced to consider each of $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{G}$ independently. Given these difficulties, in this section we attempt only to provide a series of guidelines about how to make a sensible prior specification.

For fixed values of $\alpha_{ik} = \alpha$, $\beta_{ijk} = \beta$ and $\gamma_{ik} = \gamma$, we can easily simulate data from the Markov chain of autologistic models. These data can then be summarised in terms of statistics, such as the overall proportion of wet days, about which we might hold prior beliefs. By experimenting with different values, we can find a set that leads to simulated samples with sensible properties, then choose the expectations $a_{0,\alpha}$, $a_{0,\beta}$ and $a_{0,\gamma}$ in the hierarchical priors to be equal to this set of values. However, we still need to elicit values for three other hyperparameters in each of $\pi(\mathcal{A}^0)$, $\pi(\mathcal{B}^0)$ and $\pi(\mathcal{G}^0)$. In each case we can do this by choosing a value for the marginal variance and correlation, and the variance of the distribution for σ_α^2, σ_β^2 or σ_γ^2. The mean of this distribution is the specific variance, whilst its variance controls the extent to which the amount of borrowing of strength can be influenced by the data. Having specified these three values, we can solve the resulting set of simultaneous equations to fix the values of the three remaining hyperparameters, for example, $a_{1,\alpha}$, $h_{0,\alpha}$ and $h_{1,\alpha}$ in the case of $\pi(\mathcal{A}^0)$.

Choosing values for the marginal variances is difficult for precisely the reasons previously discussed. However, we can gain insight into the scale on which values might be considered "large" by thinking about a simple logistic model

$$\Pr(D = 1 \mid \alpha) = \frac{e^\alpha}{1 + e^\alpha} \qquad \text{with prior} \qquad \alpha \sim \mathrm{N}(m, v^2).$$

In order to illustrate how m and v^2 might be chosen in this simpler problem, suppose our prior estimate of $g(\alpha) = \Pr(D = 1 \mid \alpha) = e^\alpha/(1 + e^\alpha)$ is 0.5. By symmetry, we can make the prior mean of $g(\alpha)$ equal to 0.5 by setting $m = 0$. In order to choose the variance, v^2, a sensible

strategy might be to pick the largest value before the prior for $g(\alpha)$ becomes bimodal. This will be the largest value of v^2 for which

$$\frac{\partial^2}{\partial g(\alpha)^2}\pi_{g(\alpha)}\{g(\alpha)\}\Big|_{g(\alpha)=1/2} \leq 0.$$

It is straightforward to show that $v^2 = 2$, indicating that marginal prior variances of around 2 might be a sensible choice.

When choosing the variances in the priors for σ^2_α, σ^2_β and σ^2_γ and the marginal correlations (e.g. $\text{Corr}(\alpha_{ik}, \alpha_{jk})$), our experience has shown that, unless the variances and marginal correlations are chosen to be (reasonably) small and large, respectively, the MCMC sampler suffers from convergence difficulties due to problems of likelihood non-identifiability for certain parameters.

5.4.2 Prior beliefs about the weather state transition probabilities

To satisfy exchangeability across weather states in the hierarchical prior specification, the marginal prior mean $\mathbf{e}_j$ must have the form

$$\left(\frac{1-a}{r-1}, \ldots, \frac{1-a}{r-1}, a, \frac{1-a}{r-1}, \ldots, \frac{1-a}{r-1}\right)$$

where the j–th element is a, and r is the number of states. Further, the information content parameters at each stage in the hierarchical prior, Ξ_j and E_j, must be the same for all weather states, say, $\Xi_j = \Xi$ and $E_j = E$ for all $j \in S_r$.

Correlation is a measure of linear association and, as such, is not a particularly natural quantity to think about for random variables with constraints on their support, especially when the constraints are interlinking, such as those on the simplex. This makes it difficult to elicit values for the marginal correlations in the prior for the weather state transition probabilities, $\mathbf{A}$. The objective of this section is to suggest an alternative, more intuitive, way of thinking about the dependence structure in the prior for $(\mathbf{A}^1_j, \ldots, \mathbf{A}^{27}_j)$, $j \in S_r$.

Suppose that we have specified values for the marginal means and variances of the self-transition probability parameters, say m and v, so that

$$\text{E}^2(A^x_{jj}) = e_{jj} = m \qquad \text{and} \qquad \text{Var}^2(A^x_{jj}) = \frac{e_{jj}(1-e_{jj})(\Xi+E+1)}{(E+1)(\Xi+1)} = v,$$

where the superscript 2 denotes expectation/variance with respect to the two–stage (hierarchical) prior (5.20). We can rewrite the expression for the marginal variance as

$$\text{Var}^2(A^x_{jj}) = \frac{m(1-m)(\Xi+E+1)}{(E+1)(\Xi+1)} = \frac{m(1-m)}{\zeta+1} = v \tag{5.22}$$

where $\zeta = \Xi E/(\Xi+E+1)$. Then by rearranging (5.22) we can also express ζ, whose value is fixed, in terms of the chosen marginal means and variances

$$\zeta = \frac{m(1-m)}{v} - 1.$$

Our task is therefore to decide on values of $E \in \mathbb{R}^+$ and $\Xi \in \mathbb{R}^+$, subject to the constraint $\Xi E/(\Xi + E + 1) = \zeta$. To this end, consider the following hypothetical scenario.

Instead of the two–stage Dirichlet prior, suppose we adopted a simple one–stage Dirichlet prior

$$\mathbf{A}_j^x \sim \mathscr{D}_r(\zeta \mathbf{e}_j), \qquad \text{independently for all } (j,x) \in S_r \times Q \tag{5.23}$$

in which the means and variances of the self-transition probabilities are the same as those in the hierarchical specification, (5.20). That is, $\mathrm{E}^1(A_{jj}^x) = e_{jj} = m$ and $\mathrm{Var}^1(A_{jj}^x) = e_{jj}(1 - e_{jj})/(\zeta + 1) = m(1-m)/(\zeta+1) = v$, where the superscript 1 denotes expectation/variance with respect to the one–stage prior (5.23). Now suppose that we could learn everything about $\mathbf{A}_j^1, \ldots, \mathbf{A}_j^{x-1}, \mathbf{A}_j^{x+1}, \ldots, \mathbf{A}_j^{27}$, or (roughly) equivalently, suppose we could learn that $\boldsymbol{\xi}_j$ was exactly equal to its mean, $\mathbf{e}_j$. If we had adopted the simple prior above, we might judge that this information is worth observing N transitions from weather state j, terminating on days where the Lamb weather type is x. Under (5.23) we can compute the prior expectation of the posterior variance of A_{jj}^x if we really did observe N such transitions. Let n_{jk}^x denote the number of these transitions *to* weather state k and write $\mathbf{n}_j^x = (n_{j1}^x, n_{j2}^x, \ldots, n_{jr}^x)$. Then $\sum_{k=1}^r n_{jk}^x = N$ and we have

$$\mathbf{A}_j^x \mid \text{data} \sim \mathscr{D}_r(\zeta \mathbf{e}_j + \mathbf{n}_j^x),$$

from which we can compute

$$\begin{aligned}
\mathrm{E}_{\text{data}}\{\mathrm{Var}^1(A_{jj}^x \mid \text{data})\} &= \mathrm{E}_{\text{data}}\left\{\frac{(\zeta e_{jj} + n_{jj}^x)(\zeta + N - \zeta e_{jj} - n_{jj}^x)}{(\zeta+N)^2(\zeta+N+1)}\right\} \\
&= \mathrm{E}^1_{\mathbf{A}_j^x}\left[\mathrm{E}_{\text{data}|\mathbf{A}_j^x}\left\{\frac{(\zeta e_{jj} + n_{jj}^x)(\zeta + N - \zeta e_{jj} - n_{jj}^x)}{(\zeta+N)^2(\zeta+N+1)}\right\}\right] \\
&= \frac{\zeta e_{jj}(1-e_{jj})}{(\zeta+N)(\zeta+1)} \\
&= \frac{\zeta m(1-m)}{(\zeta+N)(\zeta+1)}.
\end{aligned} \tag{5.24}$$

Here the (outer) expectation on the second line is taken with respect to the simple one–stage prior (5.23). According to our equivalence assessment, above, this value of $\mathrm{E}\{\mathrm{Var}^1(A_{jj}^x \mid \text{data})\}$ should be equal to the prior variance in the hierarchical model, (5.20), if we learnt that $\boldsymbol{\xi}_j$ was exactly equal to $\mathbf{e}_j$, in which case (5.20) reduces to $\mathbf{A}_j^x \mid \boldsymbol{\xi}_j = \mathbf{e}_j \sim \mathscr{D}_r(\Xi \mathbf{e}_j)$. That is, (5.24) should be equal to

$$\mathrm{Var}^2(A_{jj}^x \mid \boldsymbol{\xi}_j = \mathbf{e}_j) = \frac{e_{jj}(1-e_{jj})}{(\Xi+1)} = \frac{m(1-m)}{(\Xi+1)}, \tag{5.25}$$

where, here, the variance is taken with respect to the two–stage prior (5.20) when $\boldsymbol{\xi}_j = \mathbf{e}_j$. Equating (5.24) and (5.25) and solving for Ξ gives

$$\Xi = \frac{\zeta^2 + \zeta N + N}{\zeta},$$

then E is obtained from the expression $\zeta = \Xi E/(\Xi + E + 1)$ as

$$E = \frac{\zeta(\Xi+1)}{\Xi - \zeta}.$$

Note that if this calculation had used the variance expressions for A^x_{jk}, rather than A^x_{jj}, then we may have arrived at different formulae for Ξ and E. Nevertheless, we prefer to work in terms of the self-transition, or persistence probabilities, as we would formulate our beliefs in terms of the expected sojourn times, $1/(1 - A^x_{jj})$, in the case of independent priors.

This elicitation strategy relies on the provision of values for the marginal mean and variance of A^x_{jj}. The transition matrix, $\mathbf{\Lambda} = (\lambda_{jk})$, of the homogeneous hidden Markov model from Chapter 4 was given a prior comprising an independent Dirichlet distribution for each row. In Section 4.3.3 we chose the hyperparameters in this prior by thinking about the moments or quantiles in the distribution of the expected sojourn time, $1/(1 - \lambda_{jj})$, in any state, $j \in \mathcal{S}_r$. Calculations of these quantities were based on a property of the Dirichlet distribution that the univariate marginals are beta distributions. Marginalising over $\{(\mathbf{A}, \mathcal{E}) \setminus A^x_{jj}\}$ in the joint prior density for $(\mathbf{A}, \mathcal{E})$ does not lead to a beta distribution for A^x_{jj}. Nevertheless, for each $j \in \mathcal{S}_r$, we choose the marginal mean and variance for A^x_{jj}, $x = 1, \ldots, 27$, to match the values $\mathrm{E}(\lambda_{jj})$ and $\mathrm{Var}(\lambda_{jj})$ that we would hypothetically choose if we were analysing a homogeneous hidden Markov model. Consequently, in the limit as the specific uncertainty approaches zero (that is, $\Xi \to \infty$), $\mathbf{A}^1_j, \ldots, \mathbf{A}^{27}_j$ become perfectly positively correlated and we would obtain a homogeneous hidden Markov model with the same prior as that specified in Chapter 4.

5.5 Likelihood

The Yorkshire data are analysed in Section 5.7. As explained in Section 4.4, this dataset divides naturally into Y subsets, one for each winter season, and so, conditionally on the atmospheric data $\mathbf{x}$, the sub-series $\{(\mathbf{w}_{Tv+1}, \mathbf{d}_{Tv+1}, s_{Tv+1}), \ldots, (\mathbf{w}_{Tv+1}, \mathbf{d}_{Tv+1}, s_{Tv+1})\}$ for $y = 1, \ldots, Y$ are modelled as independent realisations of the same NHMM. Let $s_{0,y}$ and $\mathbf{d}_{0,y} = (d^1_{0,y}, \ldots, d^m_{0,y})^T$ denote the initial weather state and occurrence vector for the y-th sub-series then write $\mathbf{s}_0 = (s_{0,1}, \ldots, s_{0,Y})$ and $\mathbf{d}_0 = (\mathbf{d}_{0,1}, \ldots, \mathbf{d}_{0,Y})$. Posterior inference will be via MCMC with data augmentation (see Section 5.6), meaning the derivation of the full conditional distributions will require the complete data likelihood

$$p(\mathbf{w}, \mathbf{d}, \mathbf{d}_0, \mathbf{s}, \mathbf{s}_0 \mid \boldsymbol{\theta}, \mathbf{x}) = p(\mathbf{w}, \mathbf{d}, \mathbf{d}_0 \mid \mathbf{s}, \mathbf{s}_0, \boldsymbol{\theta}_{\text{obs}}) p(\mathbf{s}, \mathbf{s}_0 \mid \boldsymbol{\theta}_{\text{hid}}, \mathbf{x}) \tag{5.26}$$

where,

$$\begin{aligned} p(\mathbf{w}, \mathbf{d}, \mathbf{d}_0 \mid \mathbf{s}, \mathbf{s}_0, \boldsymbol{\theta}_{\text{obs}}) &= \prod_{y=1}^{Y} p(\mathbf{w}_{Tv+1:Tv+1}, \mathbf{d}_{Tv+1:Tv+1}, \mathbf{d}_{0,y} \mid \mathbf{s}_{Tv+1:Tv+1}, s_{0,y}, \boldsymbol{\theta}_{\text{obs}}) \\ &= p(\mathbf{d}_0) \prod_{k=1}^{r} \Bigg\{ \prod_{\{y : s_{Tv+1} = k\}} p(\mathbf{w}_{Tv+1}, \mathbf{d}_{Tv+1} \mid s_{Tv+1} = k, \mathbf{d}_{0,y}, \boldsymbol{\theta}_{\text{obs},k}) \\ &\quad \times \prod_{\substack{\{t : s_t = k, \\ t \neq T^v + 1\}}} p(\mathbf{w}_t, \mathbf{d}_t \mid S_t = k, \mathbf{d}_{t-1}, \boldsymbol{\theta}_{\text{obs},k}) \Bigg\}, \end{aligned}$$

that is,

$$p(\mathbf{w}, \mathbf{d}, \mathbf{d}_0 \mid \mathbf{s}, \mathbf{s}_0, \theta_{\text{obs}}) = p(\mathbf{d}_0) \times \prod_{k=1}^{r} \prod_{i=1}^{n} \prod_{t=1}^{T} \text{Ga}\left(w_t^i \middle| \frac{1}{v_{ik}^2}, \frac{1}{v_{ik}^2 m_{ik}}\right)^{\mathbb{I}(s_t=k, d_t^i=1)}$$
$$\times \prod_{k=1}^{r} \left[\prod_{\ell=0}^{2^n-1} C_k(\ell \mid \theta_{\text{obs},k})^{-L_{k\ell}(\mathbf{s},\mathbf{d}_0)} \right.$$
$$\left. \times \exp\left\{ \sum_{i=1}^{n} T_{ik}^1(\mathbf{s})\alpha_{ik} + \sum_{i=2}^{n} \sum_{j=1}^{i-1} T_{ijk}^{11}(\mathbf{s})\beta_{ijk} + \sum_{i=1}^{n} {}^1T_{ik}^1(\mathbf{s})\gamma_{ik} \right\} \right] \quad (5.27)$$

and

$$p(\mathbf{s}, \mathbf{s}_0 \mid \theta_{\text{hid}}, \mathbf{x}) = \prod_{y=1}^{Y} \pi(\mathbf{s}_{T^{v}+1:T^{v+1}}, s_{0,y} \mid \theta_{\text{hid}}, \mathbf{x}_{T^{v}+1:T^{v+1}})$$
$$= \prod_{y=1}^{Y} \left(\nu_{s_{0,y}} A_{s_{0,y}, s_{T^{v}+1}}^{x_{T^{v}+1}} \prod_{t=T^{v}+2}^{T^{v+1}} A_{s_{t-1}s_t}^{x_t} \right)$$
$$= \prod_{j=1}^{r} \nu_j^{m_j(\mathbf{s}_0)} \times \prod_{j=1}^{r} \prod_{x=1}^{27} \prod_{k=1}^{r} (A_{jk}^x)^{n_{jk}^x(\mathbf{s},\mathbf{s}_0)}, \quad (5.28)$$

in which

$$T_{ik}^1(\mathbf{s}) = \sum_{t=1}^{T} \mathbb{I}(d_t^i = 1, s_t = k), \qquad T_{ijk}^{11}(\mathbf{s}) = \sum_{t=1}^{T} \mathbb{I}(d_t^i = 1, d_t^j = 1, s_t = k),$$
$${}^1T_{ik}^1(\mathbf{s}) = \sum_{y=1}^{Y} \left\{ \mathbb{I}(d_{T^{v}+1}^i = 1, d_{0,y}^i = 1, s_{T^{v}+1} = k) + \sum_{t=T^{v}+2}^{T^{v+1}} \mathbb{I}(d_t^i = 1, d_{t-1}^i = 1, s_t = k) \right\},$$
$$m_j(\mathbf{s}_0) = \sum_{y=1}^{Y} \mathbb{I}(s_{0,y} = j), \quad (5.29)$$
$$n_{jk}^x(\mathbf{s}, \mathbf{s}_0) = \sum_{y=1}^{Y} \left\{ \mathbb{I}(s_{0,y} = j, s_{T^{v}+1} = k, x_{T^{v}+1} = x) + \sum_{t=T^{v}+2}^{T^{v+1}} \mathbb{I}(s_{t-1} = j, s_t = k, x_t = x) \right\},$$
$$L_{k\ell}(\mathbf{s}, \mathbf{d}_0) = \sum_{y=1}^{Y} \left[\mathbb{I}\{\mathcal{I}(\mathbf{d}_{0,y}) = \ell, s_{T^{v}+1} = k\} + \sum_{t=T^{v}+2}^{T^{v+1}} \mathbb{I}\{\mathcal{I}(\mathbf{d}_{t-1}) = \ell, s_t = k)\} \right]$$

denote the relevant counts and $\mathbb{I}(x)$ is the indicator function.

5.6 Posterior inference via MCMC

In Section 4.5, the reasons for choosing to sample from the posterior via MCMC with data augmentation were outlined. These arguments apply equally to the NHMM studied in this chapter,

and so we apply the generic Algorithm 3.3.2 to sample from the joint posterior distribution of the model parameters, θ, and the weather states, $(\mathbf{s}, \mathbf{s}_0)$. However, we need to add a third step in which we sample the initial rainfall occurrence vectors, $\mathbf{d}_0$, from their full conditional distributions:

- Step 1(a) involves sampling $\theta_{\text{hid}} = (\nu, \mathbf{A}, \mathcal{E})$ from $\pi(\theta_{\text{hid}} \mid \mathbf{s}, \mathbf{s}_0, \mathbf{x})$ and will be described in Section 5.6.1.
- Similarly, step 1(b) involves sampling $\theta_{\text{obs}} = (\mathcal{A}, \mathcal{B}, \mathcal{G}, \mathcal{M}, \mathcal{V}, \mathcal{A}^0, \mathcal{B}^0, \mathcal{G}^0)$ from $\pi(\theta_{\text{obs}} \mid \mathbf{w}, \mathbf{d}, \mathbf{d}_0, \mathbf{s})$ and more details are provided in Section 5.6.2.
- Step 2 is the data augmentation step in which $(\mathbf{s}, \mathbf{s}_0)$ is simulated from $\pi(\mathbf{s}, \mathbf{s}_0 \mid \theta, \mathbf{w}, \mathbf{d}, \mathbf{d}_0, \mathbf{x})$. Further details are outlined in the paragraph below.
- Finally, step 3 involves sampling the initial rainfall occurrence vectors $\mathbf{d}_0$ from $\pi(\mathbf{d}_0 \mid \mathbf{d}, \mathbf{s}, \theta_{\text{obs}})$ and is outlined in Section 5.6.3.

The generic forward backward algorithm (Algorithm 3.3.3) for sampling the weather states from their full conditional distribution was derived under very general conditions which are satisfied by an NHMM based on assumptions A3 and A4. Conditionally on the atmospheric data $\mathbf{x}$, each sub–series is modelled as an independent realisation of the same NHMM and so we apply the forward backward algorithm separately to each sub–series. However, we need to modify both the (forward) filtering and backward recursions so that the first time point is time $t = 0$ and not $t = 1$. Moreover, because $\mathbf{D}_0$ is assumed to be independent of S_0 in our model, we can simplify the initialisation of the filtering algorithm so that

$$\Pr(S_0 = \ell \mid \mathbf{d}_0, \theta) = \Pr(S_0 = \ell \mid \theta) = \nu_\ell.$$

In the current notation, within the one step–ahead predictive probabilities and the filtered probabilities in equations (3.14)–(3.16) we have

$$\Pr(S_t = \ell \mid S_{t-1} = k, \mathbf{w}_{1:t-1}, \mathbf{d}_{0:t-1}, \theta, x_t) = \Pr(S_t = \ell \mid S_{t-1} = k, \theta, x_t) = A^{x_t}_{k\ell} \tag{5.30}$$

and

$$\begin{aligned} p(\mathbf{w}_t, \mathbf{d}_t \mid S_t = \ell, \mathbf{w}_{1:t-1}, \mathbf{d}_{0:t-1}, \theta) \\ &= p(\mathbf{w}_t, \mathbf{d}_t \mid \mathbf{d}_{t-1}, S_t = \ell, \theta) \\ &= \Pr(\mathbf{D}_t = \mathbf{d}_t \mid \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, S_t = \ell, \theta) p(\mathbf{w}_t \mid \mathbf{D}_t = \mathbf{d}_t, S_t = \ell, \theta) \end{aligned} \tag{5.31}$$

with $\Pr(\mathbf{D}_t = \mathbf{d}_t \mid \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, S_t = \ell, \theta)$ and $p(\mathbf{w}_t \mid \mathbf{D}_t = \mathbf{d}_t, S_t = \ell, \theta)$ given in equations (5.8) and (5.6)–(5.7), respectively. Equation (5.30) also holds in (3.18) within the backward sweep. Note that compared with the equations in the generic algorithm, we have to condition on the explanatory variable, x_t, in (5.30) to account for the non–homogeneity of the transition probabilities in the hidden Markov chain.

Evaluation of the normalising constant in the expression for $\Pr(\mathbf{D}_t = \mathbf{d}_t \mid \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, S_t = \ell, \theta)$ requires summation over 2^n terms and can become computationally infeasible for large n. Although this chapter does not consider any applications in which we cannot compute the normalising constants exactly, we could, if necessary, approximate their values using one of the methods outlined in Section 5.2.2.

5.6.1 Sampling from the complete data posterior $\pi(\theta_{\text{hid}} \mid \mathbf{s}, \mathbf{s}_0, \mathbf{x})$

From (5.13) and (5.28) we can immediately deduce the posterior for the initial distribution, ν, as

$$\pi(\nu \mid \mathbf{s}_0) \propto \pi(\nu) p(\mathbf{s}_0 \mid \nu) \propto \prod_{j=1}^{r} \nu_j^{m_j(\mathbf{s}_0)} \times \prod_{j=1}^{r} \nu_j^{Gg_j - 1} = \prod_{j=1}^{r} \nu_j^{Gg_j + m_j(\mathbf{s}_0) - 1},$$

whence

$$\nu \mid \mathbf{s}_0 \sim \mathscr{D}_r\Big(G\mathbf{g} + \mathbf{m}(\mathbf{s}_0)\Big)$$

where $\mathbf{m}(\mathbf{s}_0) = \Big(m_1(\mathbf{s}_0), \ldots, m_r(\mathbf{s}_0)\Big)$. Therefore we can sample from the full conditional distribution of ν directly.

Combining the appropriate part of the complete data likelihood with the prior for $(\mathbf{A}, \mathcal{E})$ yields

$$\begin{aligned}
&\pi(\mathbf{A}, \mathcal{E} \mid \mathbf{s}, \mathbf{s}_0, \mathbf{x}) \\
&\quad \propto \pi(\mathbf{A} \mid \mathcal{E}) \pi(\mathcal{E}) p(\mathbf{s} \mid \mathbf{A}, \mathbf{s}_0, \mathbf{x}) \\
&\quad \propto \prod_{j=1}^{r} \left[\left\{ \prod_{x=1}^{27} \prod_{k=1}^{r} \Gamma(\Xi_j \xi_{jk})^{-1} (A_{jk}^x)^{\Xi_j \xi_{jk} - 1} \right\} \left(\prod_{k=1}^{r} \xi_{jk}^{E_j e_{jk} - 1} \right) \times \left\{ \prod_{x=1}^{27} \prod_{k=1}^{r} (A_{jk}^x)^{n_{jk}^x(\mathbf{s}, \mathbf{s}_0)} \right\} \right] \\
&\quad = \prod_{j=1}^{r} \left[\left\{ \prod_{k=1}^{r} \xi_{jk}^{E_j e_{jk} - 1} \Gamma(\Xi_j \xi_{jk})^{-27} \right\} \times \left\{ \prod_{x=1}^{27} \prod_{k=1}^{r} (A_{jk}^x)^{\Xi_j \xi_{jk} + n_{jk}^x(\mathbf{s}, \mathbf{s}_0) - 1} \right\} \right], \qquad (5.32)
\end{aligned}$$

from which we can immediately deduce that the sets of stochastic vectors $(\xi_j, \mathbf{A}_j^1, \mathbf{A}_j^2, \ldots, \mathbf{A}_j^{27})$ and $(\xi_k, \mathbf{A}_k^1, \mathbf{A}_k^2, \ldots, \mathbf{A}_k^{27})$ are conditionally independent for all $j \neq k$, $j, k \in \mathcal{S}_r$. It is also clear that

$$\pi(\mathbf{A} \mid \mathcal{E}, \mathbf{s}, \mathbf{s}_0, \mathbf{x}) \propto \prod_{x=1}^{27} \prod_{k=1}^{r} (A_{jk}^x)^{\Xi_j \xi_{jk} + n_{jk}^x(\mathbf{s}, \mathbf{s}_0) - 1}$$

from which we recognise that the stochastic vectors $\mathbf{A}_j^1, \mathbf{A}_j^2, \ldots, \mathbf{A}_j^{27}$ are conditionally independent, given $(\xi_j, \mathbf{s}, \mathbf{s}_0, \mathbf{x})$, with

$$\mathbf{A}_j^x \mid \xi_j, \mathbf{s}, \mathbf{s}_0, \mathbf{x} \sim \mathscr{D}_r\Big(\Xi_j \xi_j + \mathbf{n}_j^x(\mathbf{s}, \mathbf{s}_0)\Big) \qquad (5.33)$$

where $\mathbf{n}_j^x(\mathbf{s}, \mathbf{s}_0) = \Big(n_{j1}^x(\mathbf{s}, \mathbf{s}_0), \ldots, n_{jr}^x(\mathbf{s}, \mathbf{s}_0)\Big)$. Using (5.32) we can readily obtain the (non-standard) full conditional distribution of ξ_j for each $j \in \mathcal{S}_r$. However, compared to a two block MCMC sampler which draws from the full conditional distribution of $\mathcal{E}$ and then the full conditional distribution of $\mathbf{A}$, mixing was found to improve by implementing a one block sampler which draws from the (non-standard) joint full conditional distribution of $(\mathbf{A}, \mathcal{E})$. We use a Metropolis Hastings step to update $(\xi_j, \mathbf{A}_j^1, \mathbf{A}_j^2, \ldots, \mathbf{A}_j^{27})$ for each $j \in \mathcal{S}_r$ as follows. Denoting $\mathbf{A}_j = (\mathbf{A}_j^1, \mathbf{A}_j^2, \ldots, \mathbf{A}_j^{27})$, we first propose ξ_j^* with density $q_1(\xi^* \mid \xi_j, \mathbf{A}_j)$ and next $\mathbf{A}_j^*$ with density $q_2(\mathbf{A}_j^* \mid \xi_j^*, \xi_j, \mathbf{A}_j)$. In practice, we take the proposal densities to be such that $q_1(\xi^* \mid \xi_j, \mathbf{A}_j) = q_1(\xi^* \mid \xi_j)$ and $q_2(\mathbf{A}_j^* \mid \xi_j^*, \xi_j, \mathbf{A}_j) = q_2(\mathbf{A}_j^* \mid \xi_j^*)$.

For $q_1(\xi_j^* \mid \xi_j)$ we use a Dirichlet distribution

$$\xi_j^* \mid \xi_j \sim q_1(\xi_j^* \mid \xi_j) \equiv \mathscr{D}_r(\omega_d \xi_j + \epsilon \mathbf{1}_r),$$

where $\mathbf{1}_r$ is an r–vector of 1's, $\omega_d \in \mathbb{R}^+$ is a tuning parameter which can be adjusted to control the acceptance rate and $\epsilon \in \mathbb{R}^+$ is an additional parameter which should be set equal to some small value and can improve the mixing of the chain. In practice we used the values $\omega_d = 120$ and $\epsilon = 0.005$. The proposal distribution has mean

$$\frac{\omega_d}{\omega_d + r\epsilon}\xi_j + \frac{\epsilon}{\omega_d + r\epsilon}\mathbf{1}_r \to \xi_j \quad \text{as} \quad \epsilon \to 0, \tag{5.34}$$

and so, for $\epsilon = 0$, could be regarded as a random walk on the Dirichlet scale. The variance of the k-th component of the proposal is

$$\frac{(\omega_d\xi_{jk} + \epsilon)\{\omega_d(1-\xi_{jk}) + (r-1)\epsilon\}}{(\omega_d + r\epsilon)^2(\omega_d + r\epsilon + 1)} \to \frac{\xi_{jk}(1-\xi_{jk})}{\omega_d + 1} \quad \text{as} \quad \epsilon \to 0, \quad k \in \mathcal{S}_r. \tag{5.35}$$

Before taking the limit, this expression is quadratic in ω_d in the numerator and cubic in ω_d in the denominator. Therefore increasing ω_d reduces the variance and encourages more moves to be accepted. To demonstrate the purpose of the additional parameter ϵ, consider generating and accepting a proposal at the ℓ-th iteration such that the resulting $\xi_{jk}^{[\ell]}$ is very close to zero or one for some $k \in \mathcal{S}_r$. With reference to equations (5.34) and (5.35), if $\epsilon = 0$ the mean of the proposal distribution at the $(\ell+1)$-th iteration would be approximately zero or one, respectively, and the proposal variance would be approximately zero. This means the sampler can effectively get stuck at zero or one in this component. However, taking ϵ to be small but strictly positive ensures that in such extreme cases the mean of the proposal distribution is drawn slightly away from zero or one and the proposal variance is greater than zero.

The proposal distribution $q_2(\mathbf{A}_j^* \mid \boldsymbol{\xi}_j^*)$ is taken to be the same as the full conditional distribution of $\mathbf{A}_j$, (5.33), conditioned on the value $\boldsymbol{\xi}_j^*$ proposed from $q_1(\boldsymbol{\xi}_j^* \mid \boldsymbol{\xi}_j)$. It is clear from (5.32) that the joint full conditional distribution for $(\mathbf{A}_j, \boldsymbol{\xi}_j)$ factorises as $\pi(\mathbf{A}_j, \boldsymbol{\xi}_j \mid \mathbf{s}, \mathbf{s}_0, \mathbf{x}) = \pi(\boldsymbol{\xi}_j \mid \mathbf{s}, \mathbf{s}_0, \mathbf{x})\pi(\mathbf{A}_j \mid \boldsymbol{\xi}_j, \mathbf{s}, \mathbf{s}_0, \mathbf{x})$ and so the terms involving $\mathbf{A}_j$ will cancel from the acceptance ratio,

$$\begin{aligned}\alpha\{(\mathbf{A}_j, \boldsymbol{\xi}_j), (\mathbf{A}_j^*, \boldsymbol{\xi}_j^*)\} &= \min\left\{1, \frac{\pi(\mathbf{A}_j^*, \boldsymbol{\xi}_j^* \mid \mathbf{s}, \mathbf{s}_0, \mathbf{x})q_1(\boldsymbol{\xi}_j \mid \boldsymbol{\xi}_j^*)q_2(\mathbf{A}_j \mid \boldsymbol{\xi}_j)}{\pi(\mathbf{A}_j, \boldsymbol{\xi}_j \mid \mathbf{s}, \mathbf{s}_0, \mathbf{x})q_1(\boldsymbol{\xi}_j^* \mid \boldsymbol{\xi}_j)q_2(\mathbf{A}_j^* \mid \boldsymbol{\xi}_j^*)}\right\} \\ &= \min\{1, A\}\end{aligned} \tag{5.36}$$

where

$$A = \frac{\prod_{k=1}^{r}\left\{(\xi_{jk}^*)^{E_j e_{jk} - \omega_d\xi_{jk} - \epsilon}\Gamma(\omega_d\xi_{jk} + \epsilon)\Gamma(\Xi_j\xi_{jk})^{27}\prod_{x=1}^{27}\Gamma(\Xi_j\xi_{jk}^* + n_{jk}^x)\right\}}{\prod_{k=1}^{r}\left\{\xi_{jk}^{E_j e_{jk} - \omega_d\xi_{jk}^* - \epsilon}\Gamma(\omega_d\xi_{jk}^* + \epsilon)\Gamma(\Xi_j\xi_{jk}^*)^{27}\prod_{x=1}^{27}\Gamma(\Xi_j\xi_{jk} + n_{jk}^x)\right\}}. \tag{5.37}$$

With probability $\alpha\{(\mathbf{A}_j, \boldsymbol{\xi}_j), (\mathbf{A}_j^*, \boldsymbol{\xi}_j^*)\}$ we accept the proposal $(\mathbf{A}_j^*, \boldsymbol{\xi}_j^*)$ as the next iteration, otherwise we retain $(\mathbf{A}_j, \boldsymbol{\xi}_j)$.

5.6.2 Sampling from the complete data posterior $\pi(\boldsymbol{\theta}_{\text{obs}} \mid \mathbf{w}, \mathbf{d}, \mathbf{d}_0, \mathbf{s})$

We can write $p(\mathbf{w}, \mathbf{d}, \mathbf{d}_0 \mid \mathbf{s}, \mathbf{s}_0, \boldsymbol{\theta}_{\text{obs}}) = p(\mathbf{w}, \mathbf{d} \mid \mathbf{d}_0, \mathbf{s}, \boldsymbol{\theta}_{\text{obs}})p(\mathbf{d}_0)$ and so the complete data posterior $\pi(\boldsymbol{\theta}_{\text{obs}} \mid \mathbf{w}, \mathbf{d}, \mathbf{d}_0, \mathbf{s})$ is given by

$$\pi(\boldsymbol{\theta}_{\text{obs}} \mid \mathbf{w}, \mathbf{d}, \mathbf{d}_0, \mathbf{s}) \propto \pi(\boldsymbol{\theta}_{\text{obs}})p(\mathbf{w}, \mathbf{d} \mid \mathbf{d}_0, \mathbf{s}, \boldsymbol{\theta}_{\text{obs}}),$$

from which we can deduce the full conditional distributions of each component of $(\mathcal{A}, \mathcal{B}, \mathcal{G}, \mathcal{M}, \mathcal{V})$ as follows.

For the mean rainfall amount parameters in $\mathcal{M}$ and the coefficient of variation parameters in $\mathcal{V}$, derivation of the full conditional distributions proceeds as described in Section 4.5.1, with the full conditional distributions being given by (4.22) and (4.24).

Using the notation "$| \cdots$" to denote conditioning on all variables, the full conditional density for α_{ik} is given, up to proportionality, as

$$\begin{aligned}\pi(\alpha_{ik} \mid \cdots) &\propto \pi(\alpha_{ik} \mid \alpha_k) \times \left\{ \prod_{\ell=0}^{2^n-1} C_k(\ell \mid \theta_{\text{obs},k})^{-L_{k\ell}(\mathbf{s},\mathbf{d}_0)} \right\} \exp\{T^1_{ik}(\mathbf{s})\alpha_{ik}\} \\ &\propto \left\{ \prod_{\ell=0}^{2^n-1} C_k(\ell \mid \theta_{\text{obs},k})^{-L_{k\ell}(\mathbf{s},\mathbf{d}_0)} \right\} \times \exp\left\{ T^1_{ik}(\mathbf{s})\alpha_{ik} - \frac{(\alpha_{ik}-\alpha_k)^2}{2\sigma^2_{\alpha,k}} \right\}, \qquad (5.38)\end{aligned}$$

for $(i,k) \in \{1,\ldots,n\} \times \mathcal{S}_r$, where $C_k(\ell \mid \theta_{\text{obs},k})$ is the normalising constant in the autologistic model when conditioning on $\mathcal{I}(\mathbf{d}_{t-1}) = \ell$ and $S_t = k$; see equation (5.10). It can similarly be shown that the full conditional distributions for β_{ijk} and γ_{ik} are, respectively,

$$\pi(\beta_{ijk} \mid \cdots) \propto \left\{ \prod_{\ell=0}^{2^n-1} C_k(\ell \mid \theta_{\text{obs},k})^{-L_{k\ell}(\mathbf{s},\mathbf{d}_0)} \right\} \times \exp\left\{ T^{11}_{ijk}(\mathbf{s})\beta_{ijk} - \frac{(\beta_{ijk}-\beta_k)^2}{2\sigma^2_{\beta,k}} \right\}, \qquad (5.39)$$

for $(i,j,k) \in \{(i,j,k) : i = 2,\ldots,n,\ j = 1,\ldots,i-1,\ k = 1,\ldots,r\}$ and

$$\pi(\gamma_{ik} \mid \cdots) \propto \left\{ \prod_{\ell=0}^{2^n-1} C_k(\ell \mid \theta_{\text{obs},k})^{-L_{k\ell}(\mathbf{s},\mathbf{d}_0)} \right\} \times \exp\left\{ {}^1T^1_{ik}(\mathbf{s})\gamma_{ik} - \frac{(\gamma_{ik}-\gamma_k)^2}{2\sigma^2_{\gamma,k}} \right\}, \qquad (5.40)$$

for $(i,k) \in \{1,\ldots,n\} \times \mathcal{S}_r$. Across weather states, the parameters of the occurrence process are independent *a posteriori* because both the complete data likelihood and the prior factorise into a product of r functions, one for each weather state. However, the presence of the product of normalising constants, $C_k(\ell \mid \theta_{\text{obs},k})$, in each of the full conditional distributions, (5.38)–(5.40), means that the parameters $(\alpha_{1k},\ldots,\alpha_{nk},\beta_{21k},\ldots,\beta_{n,n-1,k},\gamma_{1k},\ldots,\gamma_{nk})$ for any particular state, $k \in \mathcal{S}_r$, are correlated in the posterior.

For the parameters in $\mathcal{A}^0 = (\alpha_1,\ldots,\alpha_r,\sigma^2_{\alpha,1},\ldots,\sigma^2_{\alpha,r})$, combining information from the prior, $\pi(\mathcal{A}^0)$, with information from the "likelihood", $\pi(\mathcal{A} \mid \mathcal{A}^0)$, is straightforward due to the semi-conjugacy of the normal and inverse gamma distributions to the Gaussian likelihood function. Focusing first on the mean parameters $(\alpha_1,\ldots,\alpha_r)$, their full conditional distribution is such that $\alpha_1,\ldots,\alpha_r$ are conditionally independent with

$$\alpha_k \mid \alpha_{1k},\ldots,\alpha_{nk},\sigma^2_{\alpha,k} \sim \text{N}\left(\frac{a^2_{1,\alpha}\sum_{i=1}^n \alpha_{ik} + a_{0,\alpha}\sigma^2_{\alpha,k}}{na^2_{1,\alpha}+\sigma^2_{\alpha,k}}, \frac{a^2_{1,\alpha}\sigma^2_{\alpha,k}}{na^2_{1,\alpha}+\sigma^2_{\alpha,k}} \right), \qquad \text{for } k \in \mathcal{S}_r.$$

The full conditional distribution for the variance parameters $(\sigma^2_{\alpha,1},\ldots,\sigma^2_{\alpha,r})$ is such that $\sigma^2_{\alpha,1},\ldots,\sigma^2_{\alpha,r}$ are conditionally independent with

$$\sigma^2_{\alpha,1} \mid \alpha_{1k},\ldots,\alpha_{nk},\alpha_k \sim \text{IG}\left(\frac{1}{2}n + h_{0,\alpha}, \frac{1}{2}\sum_{i=1}^n (\alpha_{ik}-\alpha_k)^2 + h_{1,\alpha} \right), \qquad \text{for } k \in \mathcal{S}_r.$$

In an analogous fashion we can obtain the full conditional distributions for the parameters in $\mathcal{B}^0$ and $\mathcal{G}^0$ as

$$\beta_k \mid \beta_{21k}, \ldots, \beta_{n,n-1,k}, \sigma^2_{\beta,k} \sim \mathrm{N}\left(\frac{a^2_{1,\beta}\sum_{i=2}^{n}\sum_{j=1}^{i-1}\beta_{ijk} + a_{0,\beta}\sigma^2_{\beta,k}}{ma^2_{1,\beta} + \sigma^2_{\beta,k}}, \frac{a^2_{1,\beta}\sigma^2_{\beta,k}}{ma^2_{1,\beta} + \sigma^2_{\beta,k}}\right),$$

$$\sigma^2_{\beta,1} \mid \beta_{21k}, \ldots, \beta_{n,n-1,k}, \beta_k \sim \mathrm{IG}\left(\frac{1}{2}m + h_{0,\beta}, \frac{1}{2}\sum_{i=2}^{n}\sum_{j=1}^{i-1}(\beta_{ijk} - \beta_k)^2 + h_{1,\beta}\right),$$

$$\gamma_k \mid \gamma_{1k}, \ldots, \gamma_{nk}, \sigma^2_{\gamma,k} \sim \mathrm{N}\left(\frac{a^2_{1,\gamma}\sum_{i=1}^{n}\gamma_{ik} + a_{0,\gamma}\sigma^2_{\gamma,k}}{na^2_{1,\gamma} + \sigma^2_{\gamma,k}}, \frac{a^2_{1,\gamma}\sigma^2_{\gamma,k}}{na^2_{1,\gamma} + \sigma^2_{\gamma,k}}\right),$$

$$\sigma^2_{\gamma,1} \mid \gamma_{1k}, \ldots, \gamma_{nk}, \gamma_k \sim \mathrm{IG}\left(\frac{1}{2}n + h_{0,\gamma}, \frac{1}{2}\sum_{i=1}^{n}(\gamma_{ik} - \gamma_k)^2 + h_{1,\gamma}\right),$$

for $k \in \mathcal{S}_r$, where here $m = n(n-1)/2$.

The full conditional distributions for the parameters in $\mathcal{M}$, $\mathcal{A}^0$, $\mathcal{B}^0$ and $\mathcal{G}^0$ are standard and can be sampled directly. However, the full conditional distributions for the parameters in $\mathcal{V}$, $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{G}$ have unknown normalising constants and samples are drawn using Metropolis Hastings steps. The Metropolis Hastings scheme for the parameters in $\mathcal{V}$ was described in Section 4.5.1.

For the parameters of the occurrence process in $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{G}$, the presence of the normalising constant presents a challenge. For any $k \in \mathcal{S}_r$, each normalising constant, $C_k(\ell \mid \theta_{\mathrm{obs},k})$, involves a summation over 2^n terms, and computing the product $\prod_{\ell=0}^{2^n-1} C_k(\ell \mid \theta_{\mathrm{obs},k})^{-L_{k\ell}(\mathbf{s},\mathbf{d}_0)}$ requires calculation of (up to) 2^n of these normalising constants, one for each value, ℓ, for which the count, $L_{k\ell}(\mathbf{s}, \mathbf{d}_0)$, is non-zero. Calculating the value of this product of normalising constants is therefore computationally demanding and becomes infeasible as n increases. Techniques for dealing with the normalising constant were discussed in Section 5.2.2. Amongst these we prefer to avoid methods which approximate the normalising constant because of the error that this introduces. Although the auxiliary variable Metropolis Hastings approach of Møller *et al.* (2006) obviates the need to make such approximations, it is not well suited for our purpose. This is because its success relies on the choice of a density function for the auxiliary variable which provides a good approximation to the likelihood. The "complete data" $(\mathbf{w}, \mathbf{d}, \mathbf{d}_0, \mathbf{s}, \mathbf{s}_0)$ changes at every MCMC iteration, meaning a new approximation to the density $p(\mathbf{d} \mid \mathbf{d}_0, \mathbf{s}, \theta_{\mathrm{obs}})$ would need to be produced at every MCMC iteration. It is likely that this would be computationally prohibitive and so we did not pursue this approach further.

In the remainder of this chapter it will be assumed that the normalising constants can be computed exactly. For the Yorkshire dataset, there are only $n = 6$ sites and calculating a normalising constant as a sum over $2^6 = 64$ terms is not computationally unreasonable. For datasets with a larger number of sites, exact computation of the normalising constants could be made feasible by simplifying the model. For example, if the sites were divided into a few non-overlapping groups of neighbours then, effectively, we would have a conditionally independent autologistic model for each group, given the weather state. Computation of each normalising constant may then be analytically tractable because the problem would reduce to that of finding the product of the (simpler) normalising constants for each group. We provide more comments about modelling larger networks of sites with this simplified model in Chapter 7.

A simple way of updating the parameters of the autologistic model is to implement a sequence of Metropolis Hastings steps in which the parameters are updated one-at-a-time using symmetric Gaussian random walks. For example, in the Metropolis Hastings step in which we update α_{ik}, we would propose

$$\alpha_{ik}^{*} \mid \alpha_{ik} \sim q(\alpha_{ik}, \alpha_{ik}^{*}) \equiv \mathrm{N}(\alpha_{ik}, \omega_{\alpha}^{i}).$$

The term $\omega_{\alpha}^{i} \in \mathbb{R}^{+}$ is a tuning parameter for site i which can be adjusted to control the acceptance rate. The symmetry of the proposal density, $q(\alpha_{ik}^{*}, \alpha_{ik}) = q(\alpha_{ik}, \alpha_{ik}^{*})$, means that the acceptance probability of the proposed move is simply given by

$$\begin{aligned} \alpha(\alpha_{ik}, \alpha_{ik}^{*}) &= \min\left\{1, \frac{\pi(\alpha_{ik}^{*} \mid \cdots)}{\pi(\alpha_{ik} \mid \cdots)}\right\} \\ &= \min\{1, A\} \end{aligned} \tag{5.41}$$

where

$$A = \frac{\left\{\prod_{\ell=0}^{2^n-1} C_k(\ell \mid \theta_{\mathrm{obs},k}^{*})^{-L_{k\ell}(\mathbf{s},\mathbf{d}_0)}\right\} \times \exp\left\{T_{ik}^{1}(\mathbf{s})\alpha_{ik}^{*} - \frac{(\alpha_{ik}^{*}-\alpha_k)^2}{2\sigma_{\alpha,k}^{2}}\right\}}{\left\{\prod_{\ell=0}^{2^n-1} C_k(\ell \mid \theta_{\mathrm{obs},k})^{-L_{k\ell}(\mathbf{s},\mathbf{d}_0)}\right\} \times \exp\left\{T_{ik}^{1}(\mathbf{s})\alpha_{ik} - \frac{(\alpha_{ik}-\alpha_k)^2}{2\sigma_{\alpha,k}^{2}}\right\}} \tag{5.42}$$

and here $\theta_{\mathrm{obs},k}^{*} = (\alpha_{1k}, \ldots, \alpha_{i-1,k}, \alpha_{ik}^{*}, \alpha_{i+1,k}, \ldots, \alpha_{nk}, \beta_{21k}, \ldots, \beta_{n,n-1,k}, \gamma_{1k}, \ldots, \gamma_{nk})$.

One-at-a-time Metropolis within Gibbs sampling can lead to slow mixing of the Markov chain. This might be improved by implementing a block updating scheme, in which changes in several parameters are considered simultaneously; see, for example, Gamerman & Lopes (2006) for more details. To be successful in realising this objective, a block updating scheme should take account of the dependence structure within the posterior. One way of achieving this is to use a tailored independence chain (Chib & Greenberg, 1998) which is a Metropolis Hastings scheme which tailors the proposal density to the unnormalised target density. For a general parameter θ with unnormalised (joint) full conditional density $g(\theta \mid \cdots)$, a proposal is taken to be $\theta^{*} = \hat{\theta} + \omega$ where ω is an increment random vector and $\hat{\theta}$ is the approximate mode of $\log g(\theta \mid \cdots)$. Chib & Greenberg (1998) suggest using the multivariate-t distribution to generate the increment random vectors, but a simpler alternative takes ω to be multivariate normal with mean zero and variance matrix $\tau^2\mathbf{C}$, where τ^2 is a tuning parameter and $\mathbf{C}$ is the inverse of the negative Hessian matrix of $\log g(\theta \mid \cdots)$, evaluated at the approximate mode, $\hat{\theta}$. At every draw from the posterior, the approximate mode $\hat{\theta}$ is obtained using two or three steps of the Newton-Raphson algorithm, initialised at the mode from the previous draw. Note that the normal approximation to the posterior tends to have thinner tails than the actual posterior density, and so satisfactory exploration of the tails of this distribution requires $\tau^2 > 1$. Although it has the potential to improve mixing, finding the normal approximation to the posterior at every iteration of the MCMC algorithm can be computationally demanding. Further comments on this subject will be provided in Section 5.7.3.1.

5.6.3 Missing data and initial occurrence vectors

The Yorkshire dataset that we analyse in Section 5.7 contains missing values. As discussed in Section 4.5.2, we assume that the missing data mechanism is ignorable. In Chapter 4,

our assumptions of conditional independence in time and space, given the weather state, allowed analytic marginalisation over the missing data. In this chapter, we retain the conditional independence assumption for rainfall amounts, given occurrences and the weather state, and so can proceed according to Section 3.3.6 in order to marginalise analytically over the missing rainfall amounts. However, for rainfall occurrences, we cannot handle the missing data in this way because the same conditional independence assumptions are not made. As the state space of $\mathbf{D}_t$ is discrete and finite, if we omitted the dependence on temporally lagged variables, $\mathbf{D}_{t-1}$, we could easily marginalise over the missing data on day t by simply summing $\Pr(\mathbf{D}_t = \mathbf{d}_t \mid S_t = s_t, \theta_{\text{obs},s_t})$ over all possible values for the missing occurrence(s). However, since we do not assume this simple temporal structure, there is no straightforward way of computing the probabilities $\Pr(\mathbf{D}_t = \mathbf{d}_t \mid \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, S_t = s_t, \theta_{\text{obs},s_t})$ if $\mathbf{d}_{t-1}$ is missing completely or only partially observed. To simulate from the posterior of the weather states given the model parameters, either by the forward backward algorithm or the naive one-at-a-time Gibbs updating scheme described in Section 3.3.4.2, a closed form expression for this probability must be available, which is not the case here. Therefore we append the missing occurrence data to the set of unknown quantities and sample values from their full conditional distributions.

If (d_t^i, w_t^i) is missing, the full conditional distribution of D_t^i depends on the value of t which dictates whether the vector $\mathbf{D}_t$ is connected on both sides, with $\mathbf{D}_{t-1}$ and $\mathbf{D}_{t+1}$, or just on one side. We now consider each case separately. Denote $\mathbf{D}_t^{-i} = (D_t^1, \ldots, D_t^{i-1}, D_t^{i+1}, \ldots, D_t^n)$ and $\mathbf{d}^{-t} = (\mathbf{d}_{1:t-1}, \mathbf{d}_{t+1:T})$. Then, if $T^y + 1 \leq t \leq T^{y+1} - 1$ for $y = 1, \ldots, Y$, the full conditional distribution of D_t^i is given by

$$
\begin{aligned}
&\Pr(D_t^i = d \mid \mathbf{d}_t^{-i}, \mathbf{d}^{-t}, \mathbf{d}_0, \mathbf{s}, \theta_{\text{obs}}) \\
&\quad \propto \Pr(D_t^i = d, \mathbf{D}_t^{-i} = \mathbf{d}_t^{-i} \mid \mathbf{D}_{t-1}, S_t = s_t, \theta_{\text{obs},s_t}) \\
&\qquad \times \Pr(\mathbf{D}_{t+1} \mid D_t^i = d, \mathbf{D}_t^{-i} = \mathbf{d}_t^{-i}, S_{t+1} = s_{t+1}, \theta_{\text{obs},s_{t+1}}) \\
&\quad \propto \exp\left\{ \alpha_{is_t} d + \sum_{\ell \neq i} \left(\alpha_{\ell s_t} d_t^\ell + \beta_{i\ell s_t} d d_t^\ell + \sum_{\substack{\{j: j<\ell, \\ j \neq i\}}} \beta_{\ell j s_t} d_t^\ell d_t^j + \gamma_{\ell s_t} d_{t-1}^\ell d_t^\ell \right) + \gamma_{i s_t} d d_{t-1}^i \right\} \\
&\qquad \times [C_{s_{t+1}}\{\mathcal{I}(d_t^i = d, \mathbf{d}_t^{-i}) \mid \theta_{\text{obs},s_{t+1}}\}]^{-1} \exp(\gamma_{i s_{t+1}} d_{t+1}^i d)
\end{aligned}
$$

and so

$$
\begin{aligned}
&\Pr(D_t^i = 1 \mid \mathbf{d}_t^{-i}, \mathbf{d}^{-t}, \mathbf{d}_0, \mathbf{s}, \theta_{\text{obs}}) \\
&= C_{s_{t+1}}\{\mathcal{I}(d_t^i = 0, \mathbf{d}_t^{-i}) \mid \theta_{\text{obs},s_{t+1}}\} \exp\left(\alpha_{is_t} + \sum_{\ell \neq i} \beta_{i\ell s_t} d_t^\ell + \gamma_{is_t} d_{t-1}^i + \gamma_{is_{t+1}} d_{t+1}^i \right) \\
&\quad \times \Bigg[C_{s_{t+1}}\{\mathcal{I}(d_t^i = 1, \mathbf{d}_t^{-i}) \mid \theta_{\text{obs},s_{t+1}}\} \\
&\quad + C_{s_{t+1}}\{\mathcal{I}(d_t^i = 0, \mathbf{d}_t^{-i}) \mid \theta_{\text{obs},s_{t+1}}\} \exp\left(\alpha_{is_t} + \sum_{\ell \neq i} \beta_{i\ell s_t} d_t^\ell + \gamma_{is_t} d_{t-1}^i + \gamma_{is_{t+1}} d_{t+1}^i \right) \Bigg]^{-1}
\end{aligned}
\tag{5.43}
$$

where $\beta_{ijk} = \beta_{jik}$ and it is understood that $\mathbf{d}_{t-1} = \mathbf{d}_{0,y}$ if $t = T^y + 1$, some $y = 1, \ldots, Y$. If $t = T^{y+1}$ so that $\mathbf{D}_t$ has no child at the next time point, the full conditional distribution of D_t^i has logistic form

$$\begin{aligned}
&\Pr(D_t^i = d \mid \mathbf{d}_t^{-i}, \mathbf{d}^{-t}, \mathbf{d}_0, \mathbf{s}, \theta_{\text{obs}}) \\
&\quad \propto \Pr(D_t^i = d, \mathbf{D}_t^{-i} = \mathbf{d}_t^{-i} \mid \mathbf{D}_{t-1}, S_t = s_t, \theta_{\text{obs},s_t}) \\
&\quad \propto \exp\left\{ \alpha_{is_t} d + \sum_{\ell \neq i} \left(\alpha_{\ell s_t} d_t^\ell + \beta_{i\ell s_t} d d_t^\ell + \sum_{\substack{\{j: j<\ell, \\ j \neq i\}}} \beta_{\ell j s_t} d_t^\ell d_t^j + \gamma_{\ell s_t} d_{t-1}^\ell d_t^\ell \right) + \gamma_{is_t} d d_{t-1}^i \right\}
\end{aligned}$$

and so

$$\Pr(D_t^i = 1 \mid \mathbf{d}_t^{-i}, \mathbf{d}^{-t}, \mathbf{d}_0, \mathbf{s}, \theta_{\text{obs}}) = \frac{\exp\left(\alpha_{is_t} + \sum_{\ell \neq i} \beta_{i\ell s_t} d_t^\ell + \gamma_{is_t} d_{t-1}^i\right)}{1 + \exp\left(\alpha_{is_t} + \sum_{\ell \neq i} \beta_{i\ell s_t} d_t^\ell + \gamma_{is_t} d_{t-1}^i\right)} \tag{5.44}$$

where $\beta_{ijk} = \beta_{jik}$.

Therefore if the observation on day t at site i is missing, we draw a value for D_t^i from the Bernoulli distribution with success probability $\Pr(D_t^i = 1 \mid \mathbf{d}_t^{-i}, \mathbf{d}^{-t}, \mathbf{d}_0, \mathbf{s}, \theta_{\text{obs}})$ given by (5.43) or (5.44), as appropriate.

The initial rainfall occurrence vectors are latent and it is convenient to draw them from their full conditional distributions. Denoting the vector for the y–th sub–series by $\mathbf{D}_{0,y} = (D_{0,y}^1, \ldots, D_{0,y}^n)$, the i–th component has full conditional distribution

$$\begin{aligned}
&\Pr(D_{0,y}^i = d \mid \mathbf{d}_{0,y}^{-i}, \mathbf{d}, \mathbf{s}, \theta_{\text{obs}}) \\
&\quad \propto \Pr(D_{0,y}^i = d) \Pr(\mathbf{D}_{Tv+1} = \mathbf{d}_{Tv+1} \mid D_{0,y}^i = d, \mathbf{D}_{0,y}^{-i} = \mathbf{d}_{0,y}^{-i}, S_{Tv+1} = s_{Tv+1}, \theta_{\text{obs},s_{Tv+1}}) \\
&\quad \propto (p_0^i)^d (1 - p_0^i)^{1-d} \times [C_{s_{Tv+1}}\{\mathcal{I}(d_{0,y}^i = d, \mathbf{d}_{0,y}^{-i}) \mid \theta_{\text{obs},s_{Tv+1}}\}]^{-1} \exp(\gamma_{is_{Tv+1}} d_{Tv+1}^i d)
\end{aligned}$$

and so

$$\begin{aligned}
&\Pr(D_{0,y}^i = 1 \mid \mathbf{d}_{0,y}^{-i}, \mathbf{d}, \mathbf{s}, \theta_{\text{obs}}) \\
&\quad = \frac{C_{s_{Tv+1}}\{\mathcal{I}(d_{0,y}^i = 0, \mathbf{d}_{0,y}^{-i})\} \exp(\gamma_{is_{Tv+1}} d_{Tv+1}^i) p_0^i}{C_{s_{Tv+1}}\{\mathcal{I}(d_{0,y}^i = 0, \mathbf{d}_{0,y}^{-i})\} \exp(\gamma_{is_{Tv+1}} d_{Tv+1}^i) p_0^i + C_{s_{Tv+1}}\{\mathcal{I}(d_{0,y}^i = 1, \mathbf{d}_{0,y}^{-i})\}(1 - p_0^i)}.
\end{aligned} \tag{5.45}$$

Therefore at every iteration of the MCMC scheme, we draw values for each $D_{0,y}^i$, $(i, y) \in \{1, \ldots, n\} \times \{1, \ldots, Y\}$, from the Bernoulli distribution with success probability $\Pr(D_{0,y}^i = 1 \mid \mathbf{d}_{0,y}^{-i}, \mathbf{d}, \mathbf{s}, \theta_{\text{obs}})$ given in equation (5.45).

5.6.4 MCMC scheme

The general form of the MCMC algorithm was outlined at the beginning of Section 5.6 and full details can be found in Appendix A. Note that in order to obtain posterior samples from an identified NHMM, we address the problem of label switching by using the online relabelling algorithm (Algorithm 3.3.4) described in Chapter 3.

5.7 Application to Yorkshire winter rainfall data

In this section we illustrate application of the model and inferential procedures through an analysis of the Yorkshire winter dataset. This time series comprises observations at $n = 6$ sites, on $T = 2707$ days, over the $Y = 30$ consecutive winters (December–February) from 1961/62 to 1990/91. Conditional on the atmospheric data (Lamb weather types), the winter periods, together with the weather states, are modelled as independent realisations of the same r–state NHMM. Each winter period has length $T_y = 90$ or 91.

The number of states, $r \in \{1, \ldots, r_{\max}\}$, is also an unknown quantity about which we would like to make inference, and so our interest lies in the joint posterior distribution $\pi(r, \theta_r, \mathbf{s}, \mathbf{s}_0, \mathbf{d}_0 \mid \mathbf{w}, \mathbf{d})$. To indicate the model from which the different parameters and hyperparameters arise, we attach r as their first subscript. For example, the spatial trend parameters for the occurrence process in an r–state model are denoted by $\mathcal{A}_r = (\alpha_{r,ik})$, whilst the hyperparameters at the lowest level in their hierarchical prior are written as $a_{r,0,\alpha}$, $a_{r,1,\alpha}$, $h_{r,0,\alpha}$ and $h_{r,1,\alpha}$. This is the style of notation that was adopted in Chapters 3 and 4 when there was uncertainty regarding the number of states.

This section begins by explaining our prior specification for r, and the choice of hyperparameters in our conditional priors for $(\theta_r \mid r)$. We then use the power posterior approach to estimate the posterior distribution for r, and compare the log marginal likelihoods to those obtained for the simple hidden Markov model from Chapter 4. For each value of r, we use the MCMC scheme outlined in Appendix A to obtain samples from the posterior distribution, $\pi(\theta_r, \mathbf{s}, \mathbf{s}_0, \mathbf{d}_0 \mid \mathbf{w}, \mathbf{d}, r)$, and analyse the distribution conditioned on the posterior mode. By comparing the posterior distributions for $\mathbf{A}^x_{r,j}$ for $x = 1, \ldots, 27$, we assess whether there is enough information in the data to discriminate between Lamb weather types. Finally, using the posterior predictive distributions of various test quantities, we assess the fit of the model and compare its performance to that of the simple hidden Markov model.

5.7.1 Prior specification

Proceeding as in Chapter 4, we restricted $r_{\max}$ to be equal to 5, and adopted a truncated Poisson Po(3) prior distribution for r. This prior expresses our preference for small values of r, in the hope that the states in such models will hold meteorological meaning.

We adopted a prior for the model parameters, $(\theta_r \mid r)$, which is exchangeable across weather states and which factorises as $\pi(\theta_r \mid r) = \pi(\theta_{r,\text{obs}} \mid r)\pi(\theta_{r,\text{hid}} \mid r)$. For the model in Chapter 4, we showed that under these conditions, for each $r = 1, \ldots, r_{\max}$, the first and second order moments in the prior predictive distributions for $\mathbf{D}_t$ and $(\mathbf{W}_t \mid \mathbf{D}_t = 1)$ will be the same if identical priors, $\pi(\theta_{r,\text{obs}} \mid r)$, are specified for each r. By analogy, the same is true for the more complicated NHMM considered in this chapter. Therefore, for all $r \in \{1, \ldots, r_{\max}\}$, we chose the same hyperparameters in our priors for the parameters of the observed process, $\theta_{r,\text{obs}}$. As we do not distinguish between sites *a priori*, we also adopt priors which are exchangeable across sites.

For the mean and coefficient of variation parameters in the gamma distributions for non–zero

rainfall amounts, $\mathcal{M}_r$ and $\mathcal{V}_r$, we chose the same hyperparameters as those selected in Section 4.7.1.2 for the homogeneous hidden Markov model. Following the guidelines outlined in Section 5.4.1, we chose the hyperparameters in the hierarchical priors for the occurrence process parameters to be

$$\begin{aligned} a_{r,0,\alpha} &= -1.8, & a^2_{r,1,\alpha} &= 2.0, & h_{r,0,\alpha} &= 2.1, & h_{r,1,\alpha} &= 0.1155, \\ a_{r,0,\beta} &= 0.6, & a^2_{r,1,\beta} &= 2.0, & h_{r,0,\beta} &= 2.1, & h_{r,1,\beta} &= 0.1155, \\ a_{r,0,\gamma} &= 0.6, & a^2_{r,1,\gamma} &= 2.0, & h_{r,0,\gamma} &= 2.1, & h_{r,1,\gamma} &= 0.1155. \end{aligned}$$

This gave a prior which is exchangeable across sites with marginal variances equal to 2.105 and marginal correlations equal to 0.950. The choice of marginal variances was based on the suggestion from Section 5.4.1 that variances of around 2 might be sensible for the parameters of an autologistic model. The high marginal correlations were actually necessary to give an MCMC sampler with good convergence properties; see Section 5.7.3.1. For example, when the marginal correlations were only 0.75, some of the parameters were so weakly identified in the posterior, that the sampler did not converge. The values for the marginal means, $a_{r,0,\alpha}$, $a_{r,0,\beta}$ and $a_{r,0,\gamma}$, were decided by simulating samples from the following Markov chain of autologistic models,

$$\Pr(\mathbf{D}_t \mid \mathbf{D}_{t-1}, \alpha, \beta, \gamma) \propto \exp\left(\sum_{i=1}^{n} \alpha d_t^i + \sum_{i=2}^{n}\sum_{j=1}^{i-1} \beta d_t^i d_t^j + \sum_{i=1}^{n} \gamma d_t^i d_{t-1}^i\right),$$

with α, β and γ fixed at a selection of values. Setting $\alpha = -1.8$, $\beta = 0.6$ and $\gamma = 0.6$ led to a sample with the following properties:

(i) The overall proportion of wet days at each site was close to 0.5

(ii) The proportion of days on which the rainfall status at each pair of sites was the same was close to 0.6

(iii) The proportion of wet days following wet (dry) days at each site was close to 0.6 (0.4).

These properties seemed reasonable, and so the hyperparameters $a_{r,0,\alpha}$, $a_{r,0,\beta}$ and $a_{r,0,\gamma}$, which represented the marginal means in the hierarchical priors, were taken to be equal to these values of α, β and γ. It is noted that the posterior was broadly insensitive to changes in the values of these hyperparameters.

The fixed parameters, p_0^i, in the distribution for the initial rainfall occurrences, $\mathbf{D}_{0,y}$, were chosen to be 0.5, in keeping with our prior beliefs about the probability of rain at any site in the winter period. It follows that $\pi(\mathbf{d}_0)$ is the discrete uniform distribution over $\{0,1\}^{nY}$.

Following the elicitation strategy outlined in Section 5.4.2, for each $j \in \mathcal{S}_r$ and each $r \in \{1,\ldots,5\}$, we chose the marginal prior means and variances for the self-transition probabilities, $A^x_{r,jj}$, $x = 1,\ldots,27$, to match the values $\mathrm{E}(\lambda_{r,jj} \mid r)$ and $\mathrm{Var}(\lambda_{r,jj} \mid r)$ that we would choose if we were analysing a homogeneous hidden Markov model with transition matrix Λ_r. In Section 4.7.1.2 we applied a simple homogeneous hidden Markov model to the Yorkshire dataset, with independent Dirichlet priors for the rows of $(\Lambda_r \mid r)$. The hyperparameters in the prior for each $(\Lambda_r \mid r)$ were chosen to have the information content of a sequence of length 46 days and

to represent belief that the mean sojourn time in any state is around 2.5 days. For example, conditional on $r = 3$, this led to a prior in which the mean and variance for $\lambda_{3,jj}$ were $m = 0.56$ and $v = 0.0154$. Therefore, in the hierarchical prior for $(\mathbf{A}^1_{3,j}, \ldots, \mathbf{A}^{27}_{3,j})$, we set $e_{3,jj} = 0.56$, $e_{3,jk} = (1 - e_{3,jj})/(3 - 1) = 0.22$, $j \neq k$, and fixed $\zeta = m(1 - m)/v - 1 = 15$.

For each $r \in \{1, \ldots, r_{\max}\}$ and every $j \in \mathcal{S}_r$ we then judged that learning that $\xi_{r,j}$ was equal to its mean, $\mathbf{e}_{r,j}$, would be equivalent to observing $N = T/(27r) \times 1.6$ transitions from weather state j, terminating on days with Lamb weather type x, if we had adopted a simpler prior in which the $\mathbf{A}^x_{r,j}$ were all independent. If the time series contained an equal number of each of these $(s_{t-1} = j, x_t = x)$–type transitions, $(j, x) \in \mathcal{S}_r \times \mathcal{Q}$, then there would be $T/(27r)$ of each kind. The figure for N that we selected is 1.6 times this value. Note that we chose a reasonably large value for N because we anticipated that there would be some stochastic vectors, $\mathbf{A}^x_j$, about which we would learn very little. From the specification of the marginal means and variances, and the choice of N, we deduced values for Ξ_r and E_r using the formulae in Section 5.4.2. Continuing the example with $r = 3$, we took $N = 2707/(27 \times 3) \times 1.6 \simeq 53$, finally leading to $\Xi_3 = 71.53$ and $E_3 = 19.25$. This procedure led to prior distributions in which the marginal correlations between $A^x_{r,jj}$ and $A^y_{r,jj}$, $x \neq y$, were just under 0.8 for all $r \in \{1, \ldots, r_{\max}\}$.

Finally, for the initial distribution, we adopted the exchangeable prior specification from Chapter 4, taking the information content parameter, G_r, to be equal to r, to give $(\boldsymbol{\nu}_r \mid r)$ a Dirichlet $\mathcal{D}_r(1, \ldots, 1)$ distribution for $r \in \{1, \ldots, r_{\max}\}$.

5.7.2 Posterior inference for r

In this section we provide further details on estimation of the posterior distribution for r via power posteriors and then present the results when the method is applied to the Yorkshire data.

5.7.2.1 Implementation

Assume first that there are no missing data. Recall that the power posterior method is a within model simulation technique and so we estimate the marginal likelihood for each model (i.e. for each value of r) separately. For a model with r states, the power posterior at temperature t is given by

$$\begin{aligned}
&\pi_t(\boldsymbol{\theta}_r, \mathbf{s}, \mathbf{s}_0, \mathbf{d}_0 \mid \mathbf{w}, \mathbf{d}, \mathbf{x}, r) \\
&\propto p(\mathbf{w}, \mathbf{d} \mid \boldsymbol{\theta}_r, \mathbf{s}, \mathbf{s}_0, \mathbf{d}_0, \mathbf{x}, r)^t \pi(\boldsymbol{\theta}_r, \mathbf{s}, \mathbf{s}_0, \mathbf{d}_0 \mid \mathbf{x}, r) \\
&= p(\mathbf{w}, \mathbf{d} \mid \mathbf{d}_0, \mathbf{s}, \boldsymbol{\theta}_{r,\text{obs}})^t p(\mathbf{s}, \mathbf{s}_0 \mid \boldsymbol{\theta}_{r,\text{hid}}, \mathbf{x}, r) \pi(\boldsymbol{\theta}_{r,\text{hid}} \mid r) \pi(\boldsymbol{\theta}_{r,\text{obs}} \mid r) p(\mathbf{d}_0).
\end{aligned}$$

The expected half deviances, $\mathrm{E}_{\boldsymbol{\theta}_r, \mathbf{s}, \mathbf{s}_0, \mathbf{d}_0 \mid \mathbf{w}, \mathbf{d}, \mathbf{x}, r, t}\{\log p(\mathbf{w}, \mathbf{d} \mid \mathbf{d}_0, \mathbf{s}, \mathbf{s}_0, \boldsymbol{\theta}_r, r)\}$, at all temperatures, t, are calculated according to Algorithm 3.5.1, in which the weather states, $(\mathbf{s}, \mathbf{s}_0)$, and the initial occurrence vectors, $\mathbf{d}_0$, are treated as model parameters. Application of the power posterior algorithm to hidden Markov models was discussed in detail in Section 4.6.4 and so we provide only brief details here. The temperature variable, t, is fixed and so the full conditional distributions of $\boldsymbol{\theta}_{r,\text{hid}}$ and $(\mathcal{A}^0_r, \mathcal{B}^0_r, \mathcal{G}^0_r)$ which do not appear in the expression $p(\mathbf{w}, \mathbf{d} \mid \mathbf{d}_0, \mathbf{s}, \boldsymbol{\theta}_{r,\text{obs}})^t$ remain as they would be in an ordinary posterior analysis. The full conditional distributions for

the parameters in $\mathcal{M}_r$ and $\mathcal{V}_r$ were deduced in Section 4.7.2.1 and are given in equations (4.36) and (4.37), respectively. For the parameters in the autologistic model, derivation of the full conditional distributions is straightforward, for example, for α_{ik} we have

$$\pi(\alpha_{r,ik} \mid \cdots) \propto \left\{ \prod_{\ell=0}^{2^n-1} C_k(\ell \mid \theta_{r,\text{obs},k}, r)^{-L_{k\ell}(\mathbf{s},\mathbf{d}_0)} \right\}^t \times \exp\left\{ tT_{ik}(\mathbf{s})\alpha_{r,ik} + \frac{(\alpha_{r,ik} - \alpha_{r,k})^2}{2\sigma^2_{r,\alpha,k}} \right\},$$

for each $(i,k) \in \{1,\ldots,n\} \times \mathcal{S}_r$, with similar expressions for the elements in $\mathcal{B}_r$ and $\mathcal{G}_r$. These distributions are sampled in Metropolis Hastings steps, generating proposals using symmetric Gaussian random walks. The full conditional distribution for the initial rainfall occurrence, $D^i_{0,y}$, is Bernoulli, with success probability

$$\frac{C_{s_{Tv+1}}\{\mathcal{I}(d^i_{0,y}=0, \mathbf{d}^{-i}_{0,y})\}^t \exp(t\gamma_{is_{Tv+1}} d^i_{Tv+1}) p^i_0}{C_{s_{Tv+1}}\{\mathcal{I}(d^i_{0,y}=0, \mathbf{d}^{-i}_{0,y})\}^t \exp(t\gamma_{is_{Tv+1}} d^i_{Tv+1}) p^i_0 + C_{s_{Tv+1}}\{\mathcal{I}(d^i_{0,y}=1, \mathbf{d}^{-i}_{0,y})\}^t (1-p^i_0)},$$

and can be sampled directly. In order to simulate from the full conditional distribution of the weather states we adopt the block updating scheme outlined in Section 4.6.4, where a sample is generated using an adapted version of the forward backward algorithm in which the distributions $p(\mathbf{w}_u, \mathbf{d}_u \mid \mathbf{D}_{u-1} = \mathbf{d}_{u-1}, S_u = s_u, \theta_{r,\text{obs},s_u}, r)$ in the filtered probabilities are raised to the power t.

Now suppose that some of the data, $(\mathbf{w}_{\text{miss}}, \mathbf{d}_{\text{miss}})$, are missing, with $(\mathbf{w}^{-\text{miss}}, \mathbf{d}^{-\text{miss}})$ denoting the observed data. In an ordinary posterior analysis, we handle the missing occurrence data by appending them to the set of unknowns and simulating values from the full conditional distributions on every iteration of the MCMC scheme; see Section 5.6.3 for details. If we handled the missing data in this way within the power posterior algorithm, the expectation that we would need to compute at every temperature would be

$$\begin{aligned} &\mathrm{E}_{\theta_r,\mathbf{s},\mathbf{s}_0,\mathbf{d}_0,\mathbf{d}_{\text{miss}} \mid \mathbf{w}^{-\text{miss}},\mathbf{d}^{-\text{miss}},\mathbf{x},r,t}\{\log p(\mathbf{w}^{-\text{miss}}, \mathbf{d}^{-\text{miss}} \mid \mathbf{d}_0, \mathbf{d}_{\text{miss}}, \mathbf{s}, \mathbf{s}_0, \theta_r, r)\} \\ &\qquad \simeq \frac{1}{N}\sum_{j=1}^{N} \log p(\mathbf{w}^{-\text{miss}}, \mathbf{d}^{-\text{miss}} \mid \mathbf{d}_0^{[j]}, \mathbf{d}_{\text{miss}}^{[j]}, \mathbf{s}^{[j]}, \mathbf{s}_0^{[j]}, \theta_r^{[j]}, r), \end{aligned}$$

where here $(\mathbf{d}_0^{[j]}, \mathbf{d}_{\text{miss}}^{[j]}, \mathbf{s}^{[j]}, \mathbf{s}_0^{[j]}, \theta_r^{[j]})$ denotes the j-th draw from the power posterior at temperature t. However, in the conditional density,

$$p(\mathbf{w}^{-\text{miss}}, \mathbf{d}^{-\text{miss}} \mid \mathbf{d}_0, \mathbf{d}_{\text{miss}}, \mathbf{s}, \mathbf{s}_0, \theta_r, r) = p(\mathbf{w}^{-\text{miss}} \mid \mathbf{d}^{-\text{miss}}, \mathbf{s}, \theta_r, r) p(\mathbf{d}^{-\text{miss}} \mid \mathbf{d}_0, \mathbf{d}_{\text{miss}}, \mathbf{s}, \theta_r, r),$$

there is no closed form expression for $p(\mathbf{d}^{-\text{miss}} \mid \mathbf{d}_0, \mathbf{d}_{\text{miss}}, \mathbf{s}, \theta_r, r)$. It can be written as

$$\frac{p(\mathbf{d}^{-\text{miss}}, \mathbf{d}_{\text{miss}} \mid \mathbf{d}_0, \mathbf{s}, \theta_r, r)}{p(\mathbf{d}_{\text{miss}} \mid \mathbf{d}_0, \mathbf{s}, \theta_r, r)} \quad \text{where} \quad p(\mathbf{d}_{\text{miss}} \mid \mathbf{d}_0, \mathbf{s}, \theta_r, r) = \sum_{\mathbf{d}^{-\text{miss}}} p(\mathbf{d}^{-\text{miss}}, \mathbf{d}_{\text{miss}} \mid \mathbf{d}_0, \mathbf{s}, \theta_r, r),$$

but calculation of this expression for the denominator is generally not computationally feasible.

The 115 missing values in the Yorkshire dataset are roughly equally split between the winters of 1980/81 and 1990/91, so we chose to omit these two years when computing estimates of the log marginal likelihood. Of course, this means that we cannot deduce the complete joint posterior

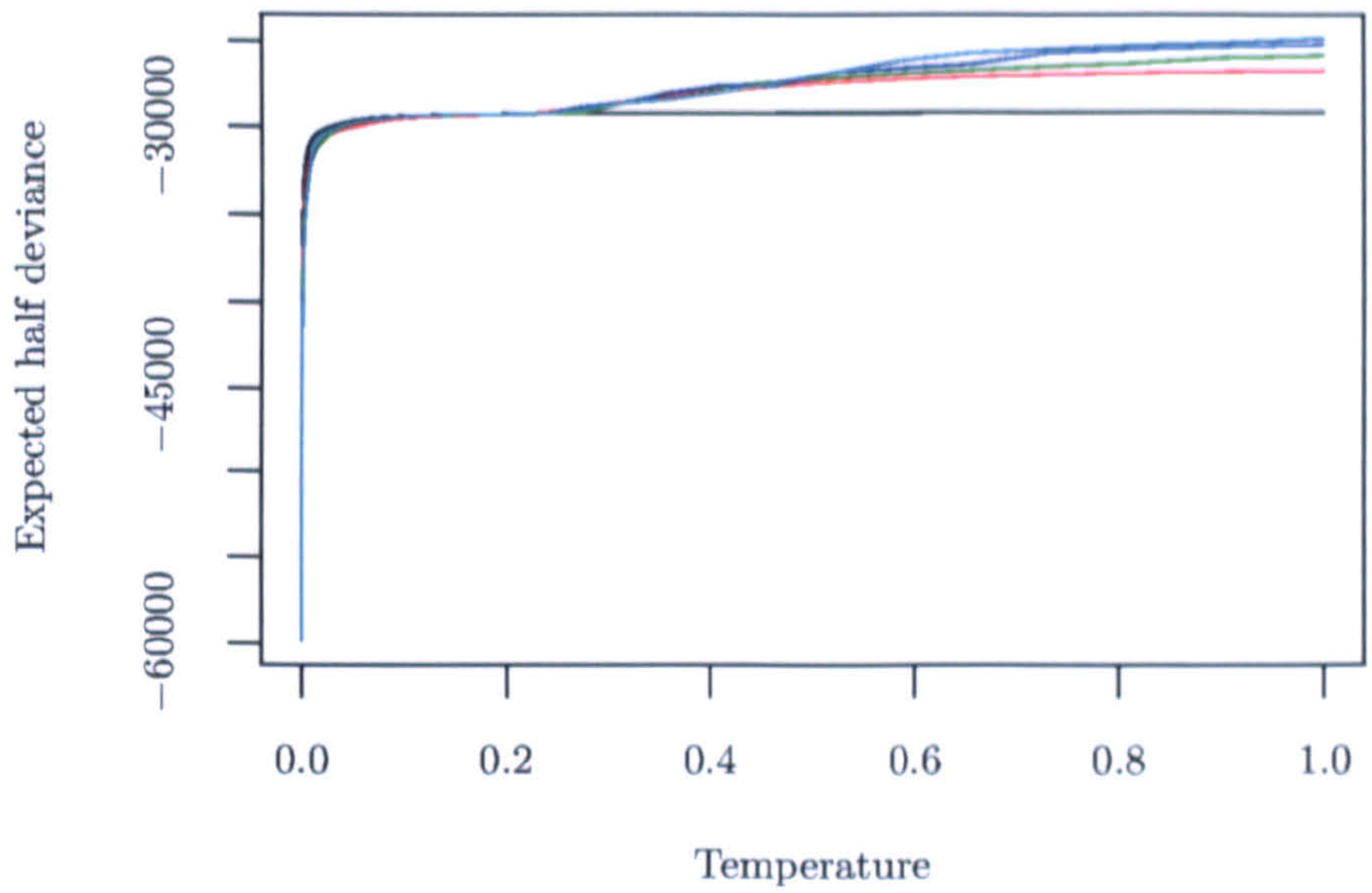


Figure 5.3: From analyses of the Yorkshire data, expected half deviance against temperature for the NHMM with $r = 1$ (———), $r = 2$ (———), $r = 3$ (———), $r = 4$ (———) and $r = 5$ (———) states.

$\pi(r, \boldsymbol{\theta}_r, \mathbf{s}, \mathbf{s}_0, \mathbf{d}_0 \mid \mathbf{w}, \mathbf{d})$, and $\pi_r(r \mid \mathbf{w}^{-\text{miss}}, \mathbf{d}^{-\text{miss}})$ can only be regarded as an approximation to $\pi_r(r \mid \mathbf{w}, \mathbf{d})$.

In the power posterior algorithm, we use the temperature schedule $t_i = (i/n)^c$, $i = 0, \ldots, n$, with $n = 40$ and $c = 4$, and generate 100,000 draws from the power posterior at each temperature, of which the first 40,000 are discarded as burn–in. For every temperature, note that we have increased the number of posterior draws by a factor of 10 compared with the analyses from Chapter 4. This was due to very high autocorrelations in the MCMC output (see Section 5.7.3.1) which meant that every posterior draw was worth considerably less than it would have been, had the draws been less highly correlated. The estimates of the expected half deviance at each temperature were combined using the trapezoidal rule to give the overall estimate of the log marginal likelihood, and the Monte Carlo standard errors were estimated using the procedure outlined in Section 4.7.2.1.

5.7.2.2 Results

For the NHMMs with $r = 1, \ldots, 5$ weather states, Figure 5.3 displays plots of the expected half deviance, $\mathrm{E}_{\boldsymbol{\theta}_r, \mathbf{s}, \mathbf{s}_0, \mathbf{d}_0 \mid \mathbf{w}, \mathbf{d}, \mathbf{x}, r, t}\{\log p(\mathbf{w}, \mathbf{d} \mid \mathbf{d}_0, \mathbf{s}, \mathbf{s}_0, \boldsymbol{\theta}_r, r)\}$, against temperature, t. The shapes are similar to those that we observed in Chapter 4, with occasional sharp increases in the expected half deviance at temperatures when the likelihood is given sufficient weight to allow additional weather states to be identified in the power posterior.

Estimates of the log marginal likelihoods, their Monte Carlo standard errors and the posterior distribution for r are presented in Table 5.1. The posterior for r is such that virtually all the posterior mass lies at $r = 5$ and it seems unlikely that this result would change if we had been able to use all of the data to compute the log marginal likelihood estimates. In Chapter 4

r	1	2	3	4	5
Log marginal likelihood	-29258.4	-27975.13	-27662.16	-27367.62	-27245.64
Monte Carlo std. error	0.12	0.43	0.57	0.53	3.08
Posterior probability	0.00	3.41×10^{-317}	2.86×10^{-181}	1.77×10^{-53}	1.00

Table 5.1: Estimates of the log marginal likelihood, the associated Monte Carlo standard error and the posterior distribution for r for the 28 years in the Yorkshire dataset with no missing values. The estimates of the log marginal likelihoods were computed via power posteriors.

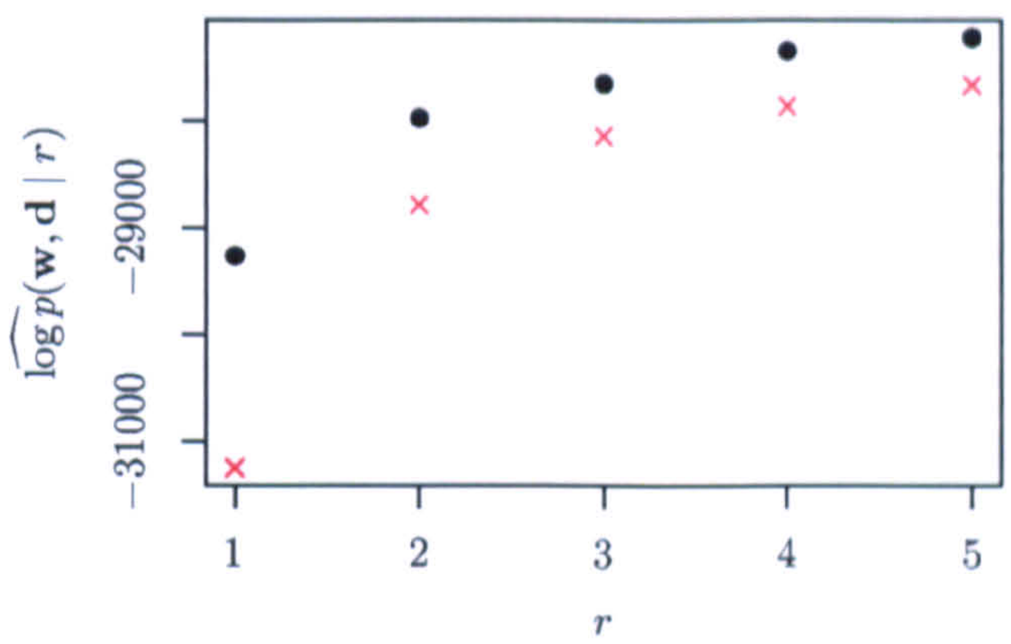


Figure 5.4: Estimates of the log marginal likelihood, for the 28 complete years in the Yorkshire dataset, when modelling the data using the r–state NHMM (•) and simple hidden Markov model (×) from Chapter 4. Estimates were calculated using the power posterior approach.

we obtained a similarly extreme posterior distribution for r. Again, this is likely to be due to some combination of the choice of priors and model misspecification. For example, even though rainfall occurrences are allowed to be positively correlated within weather states, we still assume non–zero rainfall amounts to be conditionally independent. Consequently, a large number of states may be required to compensate for the simplicity of the within–state model for rainfall amounts on wet days.

The Monte Carlo standard error for the estimate when $r = 5$ is noticeably larger than that for any other value of r. Although trace plots of the half mean deviance, $\log\{p(\mathbf{w}, \mathbf{d} \mid \mathbf{d}_0, \mathbf{s}, \mathbf{s}_0\boldsymbol{\theta}_r, r)\}$, displayed quite large variances at some temperatures, this did not appear to be due to any lack of convergence. The large variances may be because some of the weather states in the 5 state model are not well supported by the data when the likelihood is raised to a power less than 1, that is, when the influence of the likelihood is downweighted. It follows that at the lower temperatures there might be a lot of uncertainty surrounding some of the weather states in the 5 state model, and perhaps more variability in the power posterior for the half mean deviance.

For comparative purposes, we estimated the log marginal likelihood for the simple homogeneous hidden Markov model (see Chapter 4) with $r = 1, \ldots, 5$ states, using only the 28 years of data which contained no missing values. The resulting estimates are plotted together with the corresponding values for the NHMM in Figure 5.4. For all values of r, the marginal likelihood suggests that the NHMM provides a better description of the data than the simple hidden Markov

model with the same number of states. Indeed, evidence in favour of the NHMM is sufficient that the estimate of the log marginal likelihood for the NHMM with $r = 2$ states is similar to that for the simple hidden Markov model with $r = 5$ states. It also appears that the log marginal likelihood increases more slowly with r for the NHMM than the simple hidden Markov model. A possible explanation for this is that the within–state model for the NHMM explains some of the spatio–temporal dependence in the occurrence process, whereas the weather state is the only device for capturing this structure in the simple hidden Markov model. Moreover, the large number of parameters in $\theta_{r,\text{hid}}$ for the NHMM means that, in comparison to the hidden Markov model, every increase in r constitutes a greater escalation in model complexity. Therefore, in terms of providing a better fitting, but parsimonious model, less is gained per increase in r for the NHMM.

5.7.3 Posterior inference for $(\theta_r, \mathbf{s} \mid r)$ using MCMC samples

In this section we describe some of the problems encountered during MCMC sampling. We then present summaries of the posterior distributions for the model parameters and the weather states. We focus on the model with $r = 5$ states since the posterior probability for this model is essentially equal to one. However, we also generated posterior samples from models with $r = 1, \ldots, 4$ states, and encountered similar MCMC problems, although to a lesser extent.

5.7.3.1 Implementation, convergence and mixing

Fixing the number of states at $r = 5$, the MCMC algorithm was run from a variety of starting points, all of which produced essentially the same results, up to the labelling of the states. In each run we generated 2,500,000 draws from the posterior, omitting the first 500,000 as burn–in, and thinning the remaining output so that only every 200–th sample was stored. This gave a posterior sample of size $N = 10000$. The posterior distributions presented in this section are based on one such run. Convergence was assessed using the usual graphical diagnostic checks (see Section 3.3.1), and although the posterior distributions showed no signs of non–convergence, the mixing for the parameters in some weather states was poor. In some cases, the ACF plots showed that even thinning to every 200–th iterate did not eliminate the autocorrelation between successive draws. For example, for the most highly autocorrelated parameter, the effective sample size, which gives a measure of the sample size adjusted for autocorrelation (Kass *et al.*, 1998), was only 1,574.

Mixing was worst amongst the $\alpha_{5,ik}$ and some of the $\beta_{5,ijk}$ parameters in weather states 4 and 5, which we categorise as "wet" weather states in the next section. Mixing problems arise when, in certain weather states, there are observations on too few of the 2^n possible rainfall occurrence vectors. For example, conditioning on the marginal posterior mode for $\mathbf{s}$, it appeared that, on average, only six distinct rainfall occurrence vectors occurred in the wettest weather state (state 5). The consequence of observing too few occurrence vectors in some states is that some parameters are, at best, only weakly identifiable in the likelihood. We explore this issue below.

On any day, the rainfall occurrence indicator can take one of 2^n possible values. Therefore, conditional on the weather state and the rainfall occurrence indicator on the previous day, the

likelihood is of multinomial form. To explain the source of the identifiability problems, it will be convenient to begin by thinking about the multinomial distribution. To this end, consider a multinomial distribution with index parameter equal to 1 and probability vector $(p_1, \ldots, p_K)$. This is simply the generalisation of the Bernoulli distribution to $K > 2$ possible outcomes. By definition,

$$\sum_{i=1}^{K} p_i = 1, \tag{5.46}$$

and so the multinomial likelihood will be completely determined by any $(K-1)$ of $p_1, \ldots, p_K$ or, equivalently, by the *relative* sizes of $p_1, \ldots, p_K$. For example, if we regarded p_1 as a baseline, then the likelihood could be determined through the ratios $\phi_2, \ldots, \phi_K$ where $\phi_i = p_i/p_1$. Suppose that we are using a model with $(K-1)$ parameters, for example, $\phi_2, \ldots, \phi_K$, in which case $p_1 = 1/(1 + \phi_2 + \cdots + \phi_K)$ and $p_i = \phi_i/(1 + \phi_2 + \cdots + \phi_K)$ for $i = 2, \ldots, K$. Based on independent realisations from this multinomial distribution, the likelihood is

$$L \propto \prod_{i=1}^{K} p_i^{n_i}$$

where n_i is the total number of times outcome i is observed.

Suppose that for one i, say $i = a$, we have $n_a = 0$. Then p_a does not appear in the likelihood above, but it can still be identified because it follows from (5.46) that $p_a = 1 - \sum_{i \neq a} p_i$. Suppose now that we have zero frequencies for two (or more) outcomes, say $i = a$ and $i = b$. Then we cannot identify p_a and p_b since, even using (5.46), they only appear in the likelihood as $(p_a + p_b)$. Therefore, we cannot identify ϕ_a and ϕ_b.

Now consider modelling multivariate binary data using the (atemporal) autologistic model in equation (5.1). Let $\beta = (\alpha_1, \ldots, \alpha_n, \beta_{21}, \ldots, \beta_{n,n-1})^T$ and denote the probabilities of the 2^n possible vectors, $(d^1, \ldots, d^n)^T$, by $p_1, \ldots, p_K$ where $K = 2^n$. In this model, each $\eta_i = \log \phi_i$ is just a linear combination of the parameters in β, say, $\eta_i = \tilde{\mathbf{x}}_i^T \beta$. The vector $\tilde{\mathbf{x}}_i$ will comprise 1's, 0's and (-1)'s depending on which probability is chosen as the baseline. If we have two or more zero frequencies then some of the $\{\eta_i\}$ are not directly identified, that is, they would not be identified if they were separate parameters. It *may* then be the case that we cannot identify certain elements of β from the remaining $\{\eta_i\}$, typically because they always occur together in the same combination. Suppose we take the baseline to be a probability corresponding to one of the binary vectors with zero observed frequency. If we denote just the identified $\{\eta_i\}$ as η, then we can write $\eta = \tilde{\mathbf{X}}\beta$. Now the parameters in β will only all be identified if the matrix $\tilde{\mathbf{X}}$ has rank equal to the number of parameters in β or, equivalently, if the columns in $\tilde{\mathbf{X}}$ are linearly independent. When this is not the case, the problem is akin to the issue of multicollinearity, which arises in classical linear regression when two or more explanatory variables are highly correlated.

As an example, suppose that we are modelling binary vectors of length $n = 2$ using the (atemporal) autologistic model. Let the probabilities of observing $(0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$ be p_1, p_2, p_3 and p_4, respectively, and suppose we take p_3 to be the baseline. Then

$$\eta_1 = \log \phi_1 = \log(p_1/p_3) = -\alpha_1$$

and, similarly, $\eta_2 = \alpha_2 - \alpha_1$ whilst $\eta_4 = \alpha_2 + \beta_{21}$. Suppose that we only ever observe $(0,0)$'s and $(1,1)$'s. If we observe n_1 of the former and n_4 of the latter then $\eta = (\eta_1, \eta_4)^T$ and

$$\tilde{\mathbf{X}} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

which has rank $2 < 3$. Therefore we cannot identify all the parameters in β. In this example we do not learn about $\eta_2 = \alpha_2 - \alpha_1$. This can also be seen by writing out the likelihood in terms of $(\alpha_1, \alpha_2, \beta_{21})$, which gives

$$\frac{e^{n_4(\alpha_1+\alpha_2+\beta_{21})}}{(1+e^{\alpha_1}+e^{\alpha_2}+e^{\alpha_1+\alpha_2+\beta_{21}})^{n_1+n_4}}.$$

Clearly the likelihood depends on α_1 and α_2 in exactly the same way, and so, we do not learn about $\alpha_2 - \alpha_1$.

In our NHMM for rainfall, the problem is more complicated because, conditionally on the weather state, we have a Markov chain of autologistic models. However, for each weather state, this simply means that there are many multinomial distributions, one for each occurrence vector on the previous day, which are linked by common parameters. In practice, this just means that we have many constraints of the form (5.46).

The identifiability problems discussed above are likely to have been responsible for the poor mixing in the wet weather states, states 4 and 5. As remarked earlier in this section, the wettest weather state (state 5) was only associated with six distinct rainfall occurrence vectors. It was therefore impossible to identify all 27 parameters in the likelihood for rainfall occurrences in this state. Similarly, on most days in state 4, there were some pairs of sites which were either both wet or both dry, again leading to multicollinearity problems. When there are parameter combinations which are, at best, only weakly likelihood-identifiable, this leads to flat ridges in the likelihood in some directions of the parameter space. This in turn leads to mixing problems during MCMC sampling.

When the parameters are not identifiable in the likelihood, or only weakly identifiable, the same need not be true in the posterior. If the combinations of parameters that give rise to the same value of the likelihood lead to different values of the prior, then the parameters *will* be identifiable in the posterior distribution. Clearly, the greater the distinction between these sets of parameters in the prior, the easier it will be to distinguish between them in the posterior. The idea of borrowing of strength can be used to construct a prior which is informative for directions in the parameter space about which the data provide little information. In the hierarchical priors (5.14)–(5.19) we had to introduce high *a priori* correlations so that the prior was sufficiently informative to overcome identifiability problems in the likelihood.

Often mixing can be improved by updating parameters simultaneously in blocks. Since mixing seemed to be worst for the parameters in $\mathcal{A}_5$, for each $k \in \mathcal{S}_r$ we considered updating $(\alpha_{5,1k}, \ldots, \alpha_{5,nk})$ in a single block using a tailored independence chain, as outlined in Section 5.6.2. For the parameters in some of the drier weather states this led to faster decay in the ACF function, that is, mixing improved. However, for the parameters in the wetter weather states, where the ACF function decayed the most slowly, block updating made very little difference to the ACF plots. In approximating the mode of the log of the unnormalised target

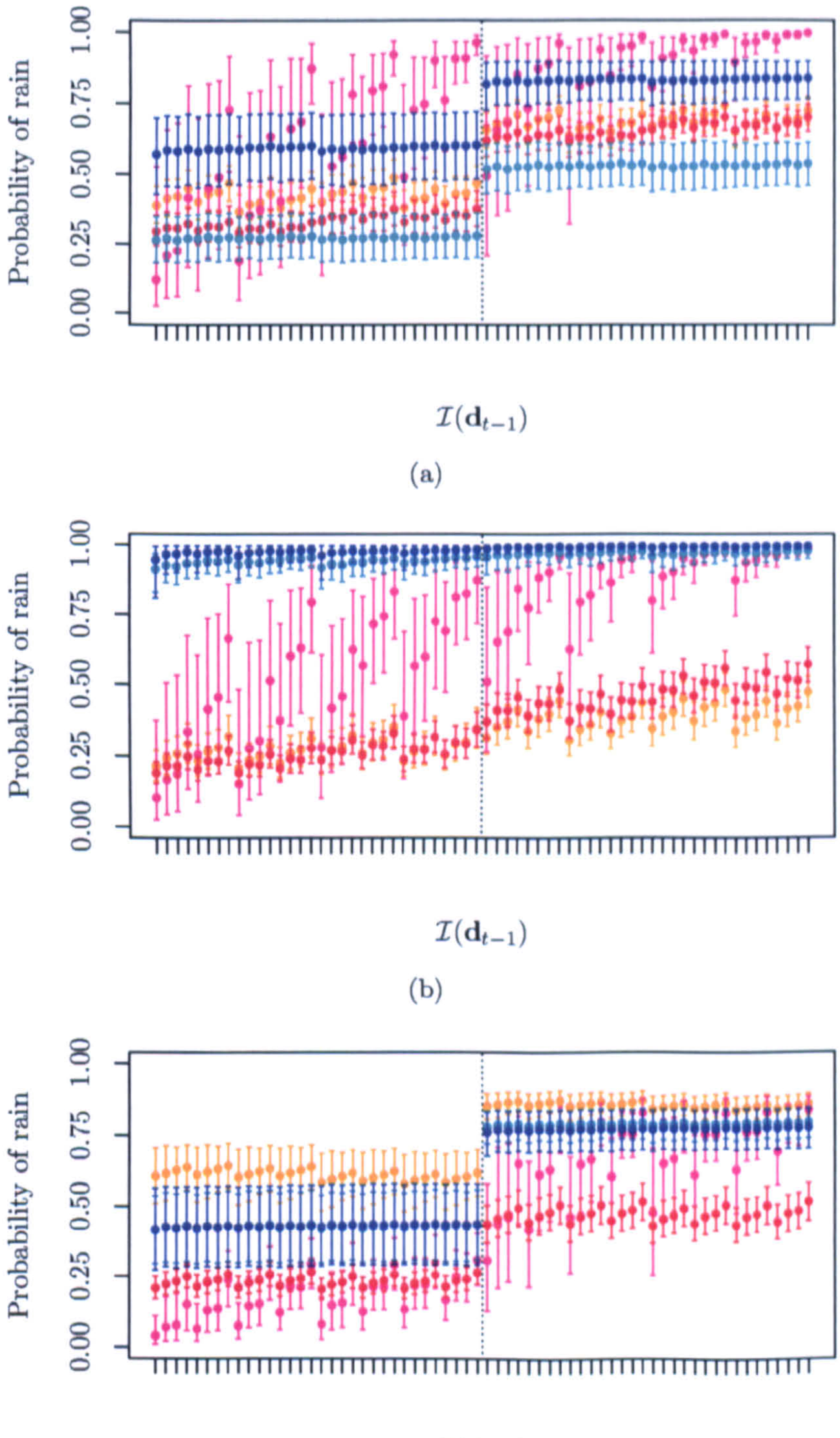
density, computation of its first and second derivatives at every step of the Newton Raphson scheme involved multiple summations over 2^n terms. Therefore, computationally, the scheme was considerably slower than one–at–a–time updating based on symmetric Gaussian random walks. Further, in the wetter weather states, the Newton Raphson scheme was rather unstable and often failed to converge to the mode of the log target density. This is likely to have been because the posterior was still rather flat in certain directions of the parameter space, making the second derivatives small, even in the vicinity of the mode. Within the small number of recursions considered, this might have caused the Newton Raphson algorithm to stray from the mode. We therefore abandoned such block–updating schemes and the results presented in the following sections are based on one–at–a–time updating of the α_{ik}, β_{ijk} and γ_{ik}.

5.7.3.2 Posterior for $(\theta_5 \mid r = 5)$

Conditional on the posterior mode, $r = 5$, Figures 5.5 and 5.6 display the posterior distributions for the conditional probabilities of rainfall at each site, given the weather state and the previous day's rainfall occurrence vector, $\mathbf{D}_{t-1}$. That is, the posteriors for $\Pr(D_t^i = 1 \mid \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, S_t = k, \theta_{5,\text{obs}}, r = 5)$ for each possible value of $\mathbf{d}_{t-1}$ and each $k \in \mathcal{S}_5$. At each site, the numerical labels of the 2^n possible values for $\mathbf{D}_{t-1}$, $\mathcal{I}(\mathbf{d}_{t-1})$, are ordered such that the first 2^{n-1}, on the left hand side of the dotted line, correspond to no rain at the site in question on day $t-1$, and conversely for the last 2^{n-1} values. Each posterior distribution is visualised through its mean and 95% equi–tailed Bayesian credible regions. Figure 5.7 displays plots of the marginal posterior distributions for the mean parameters, $\mathcal{M}_5$, in the gamma distributions for rainfall amounts on wet days.

From these plots it seems that weather state 5 is wet at all sites, with large rainfall amounts on wet days and high probabilities of rain, irrespective of the rainfall occurrence vector at lag one. The converse is true for the "dry" weather state, state 3, but in this case, the probability of rainfall is sensitive to rainfall occurrence at lag one, with lower probabilities of rain if the preceding day at the site in question was dry. Weather state 4 is generally wet, with high probabilities of rainfall at all sites (except Lockwood Reservoir, site 1), and reasonably large non–zero rainfall amounts. At the high altitude site, Lockwood Reservoir, weather state 2 was the wettest state, although elsewhere this state represented conditions that were neither particularly wet nor dry. Similarly, whilst the wettest weather state for Moorland Cottage (site 3) was weather state 1, at other sites this state was associated with dry conditions. Since Moorland Cottage is a high altitude site in the Pennines, it is likely that orographic effects (i.e. the influence of rising altitudes) are responsible for the occurrence of days characterised by this kind of behaviour. Although the labelling of the states was different, the weather states identified by the simple hidden Markov model in Chapter 4 had similar characteristics to those identified here.

The effect of rain on the preceding day is the most pronounced at sites 1 and 3 in all weather states. This may be because the behaviour at these two sites often differs from the behaviour elsewhere. Therefore it could be that the weather state and autoregression on neighbouring sites is less able to explain the probability of rainfall occurrence at these sites, so that autoregression on the previous day is still an important factor. It also seems that in the wetter weather states, states 5 and 4, the conditional probability of rainfall depends only on whether or not it



(a)

(b)

(c)

Figure 5.5: Conditional on $r = 5$, posterior means with 95% equi–tailed Bayesian credible intervals for the probabilities $\Pr(D_t^i = 1 \mid \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, S_t = k, \boldsymbol{\theta}_{5,\text{obs},k}, r = 5)$, $\mathcal{I}(\mathbf{d}_{t-1}) = 0, \ldots, 2^n - 1$, at sites (a) 1 (b) 2 and (c) 3, in weather states 1 (———), 2 (———), 3 (———), 4 (———) and 5 (———). The values $\mathcal{I}(\mathbf{d}_{t-1})$ are ordered so that the first 2^{n-1} correspond to $d_{t-1}^i = 0$, in ascending order, and the last 2^{n-1} correspond to $d_{t-1}^i = 1$, in ascending order.

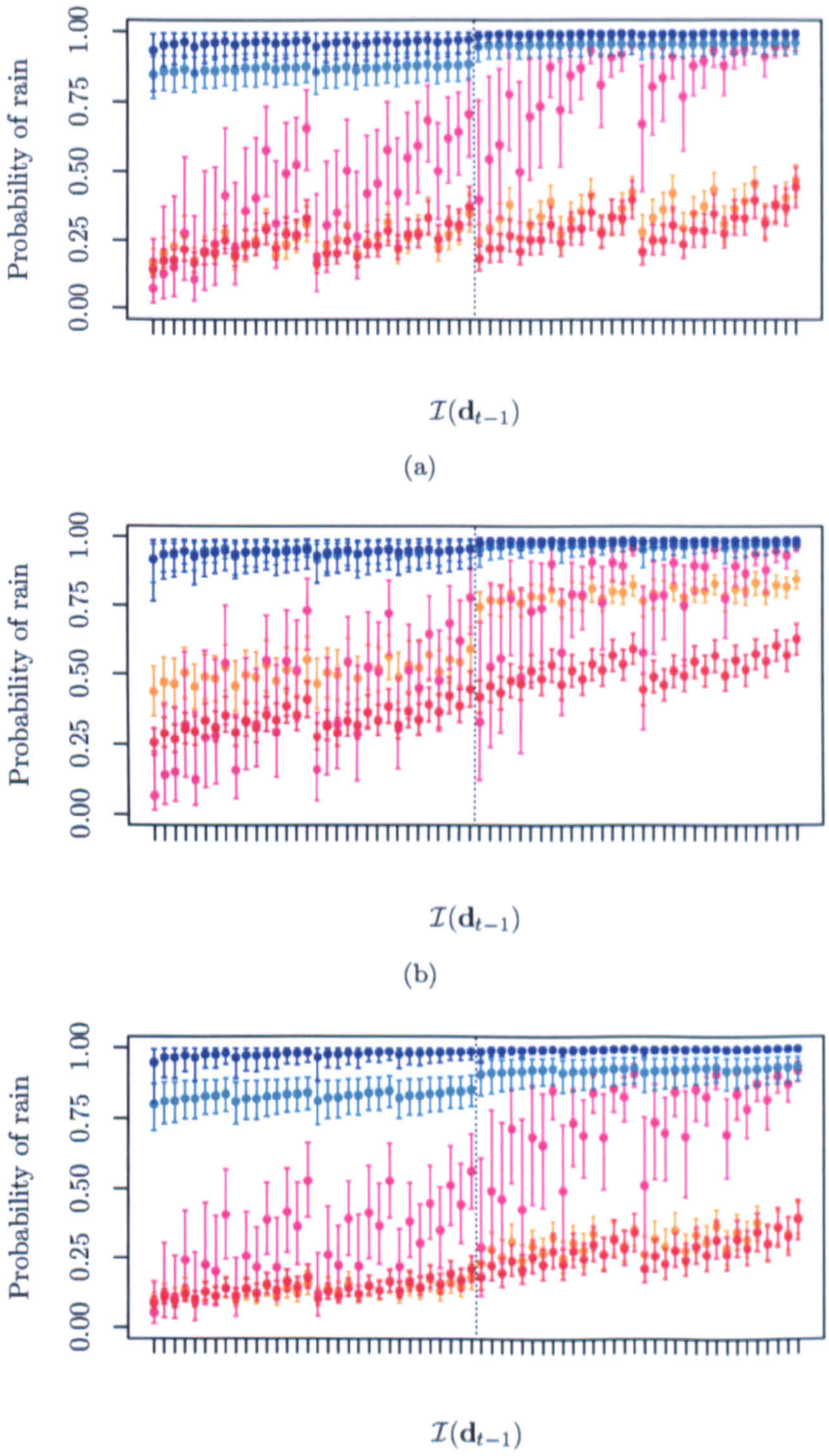


Figure 5.6: Conditional on $r = 5$, posterior means with 95% equi–tailed Bayesian credible intervals for the probabilities $\Pr(D_t^i = 1 \mid \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, S_t = k, \boldsymbol{\theta}_{5,\text{obs},k}, r = 5)$, $\mathcal{I}(\mathbf{d}_{t-1}) = 0, \ldots, 2^n - 1$, at sites (a) 4 (b) 5 and (c) 6, in weather states 1 (———), 2 (———), 3 (———), 4 (———) and 5 (———). The values $\mathcal{I}(\mathbf{d}_{t-1})$ are ordered so that the first 2^{n-1} correspond to $d_{t-1}^i = 0$, in ascending order, and the last 2^{n-1} correspond to $d_{t-1}^i = 1$, in ascending order.

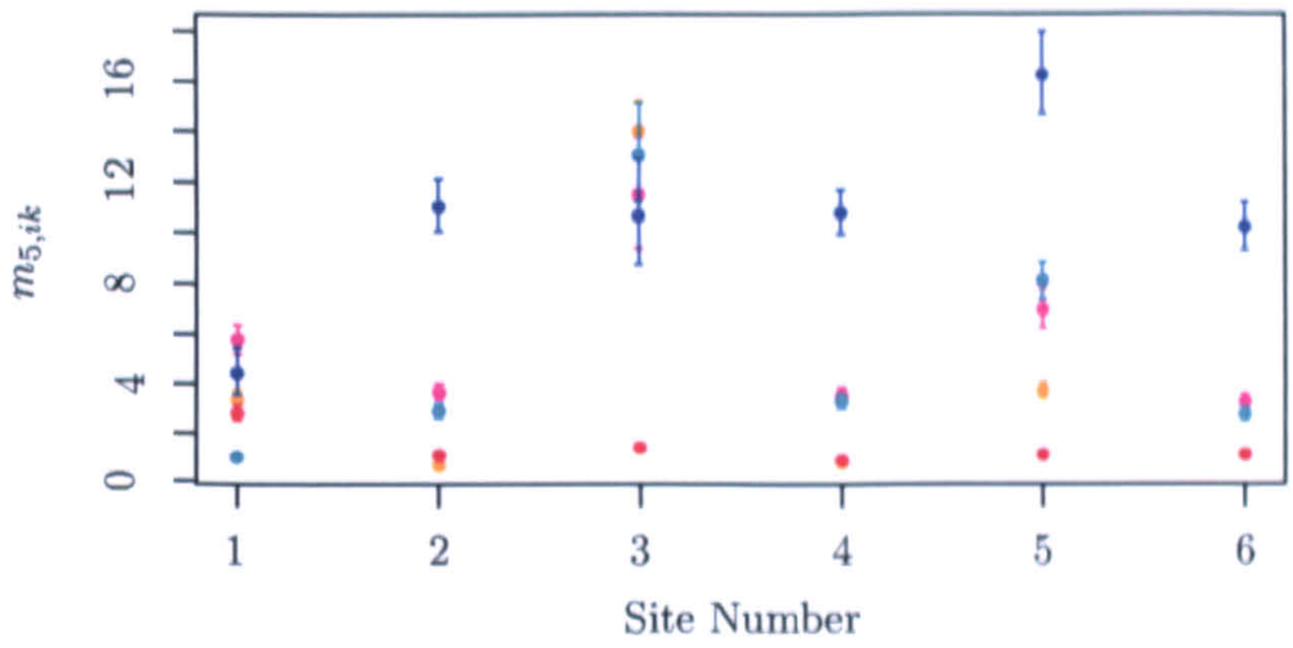


Figure 5.7: Conditional on $r = 5$, posterior means with 95% equi–tailed Bayesian credible intervals for the parameters in $\mathcal{M}_5$ in weather states 1 (———), 2 (———), 3 (———), 4 (———) and 5 (———).

rained at the site in question the previous day, and not on the lag one rainfall occurrence at other sites. For each site, this is apparent from the similarity in the posterior distributions for $\Pr(D_t^i \mid \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, S_t = k, \boldsymbol{\theta}_{5,\text{obs},k}, r = 5)$, where $k = 4$ or $k = 5$, amongst all 2^{n-1} values of $\mathbf{d}_{t-1}$ for which $d_t^i = 0$, and amongst all 2^{n-1} values for which $d_t^i = 1$. In the other weather states, the lag one rainfall occurrences at sites $j \neq i$ also seem to affect the conditional probability of rain at site i. For example, for each site, the first and last values of $\mathcal{I}(\mathbf{d}_{t-1})$ correspond to $\mathbf{d}_{t-1} = (0,0,0,0,0,0)^T$ and $\mathbf{d}_{t-1} = (1,1,1,1,1,1)^T$, respectively, and represent the lowest and highest probabilities of rain in state k, where $k = 1$, 2 or 3. The disparities between the posterior distributions given the different lag one rainfall occurrence vectors is evidence that they can be distinguished *a posteriori*. This supports the inclusion of an autoregression on lagged rainfall occurrences.

The plots of the marginal posterior distributions for the coefficient of variation parameters in $\mathcal{V}_5$ show the same patterns as the corresponding plots for the hidden Markov model in Chapter 4 (see Figure 4.10(c)) and so are not shown. They again seem to suggest more variation between sites than between weather states.

Figure 5.8 displays the marginal posterior distributions for some of the weather state transition probabilities, $A_{5,jk}^x = \Pr(S_t = k \mid S_{t-1} = j, X_t = x, \boldsymbol{\theta}_{5,\text{hid}}, r = 5)$. The Lamb weather types are labelled so that 1 (and 8–9) are anticyclonic (hybrids), 10–17 are pure directional types, 18 (and 19–26) are cyclonic (hybrids) and 27 is unclassified; see Table 2.2 for more details. Each plot also shows the marginal posterior distribution for the corresponding $\xi_{5,jk}$'s and the marginal prior distribution for the transition probabilities, which is the same for $A_{5,jk}^x$ and $A_{5,jk}^y$, where $x \neq y$. Figure 5.8(d) shows the marginal posteriors for $A_{5,53}^x$, $x = 1, \ldots, 27$, and is typical of the posteriors for all probabilities of transition *from* weather states 5 and 4, $A_{5,5k}^x$ and $A_{5,4k}^x$. For these transition probabilities there is considerable overlap in the marginal posteriors across Lamb weather types, indicating that the atmospheric data are not particularly helpful in explaining transitions *from* the wetter weather states. However, the transition probabilities from and, to a lesser extent, into the "dry" weather state are heavily influenced by the Lamb weather type. Figure 5.8(c), for example, displays the marginal posterior distributions for $A_{5,33}^x$, $x = 1, \ldots, 27$. This is the probability of remaining in the dry weather state given that the current Lamb weather

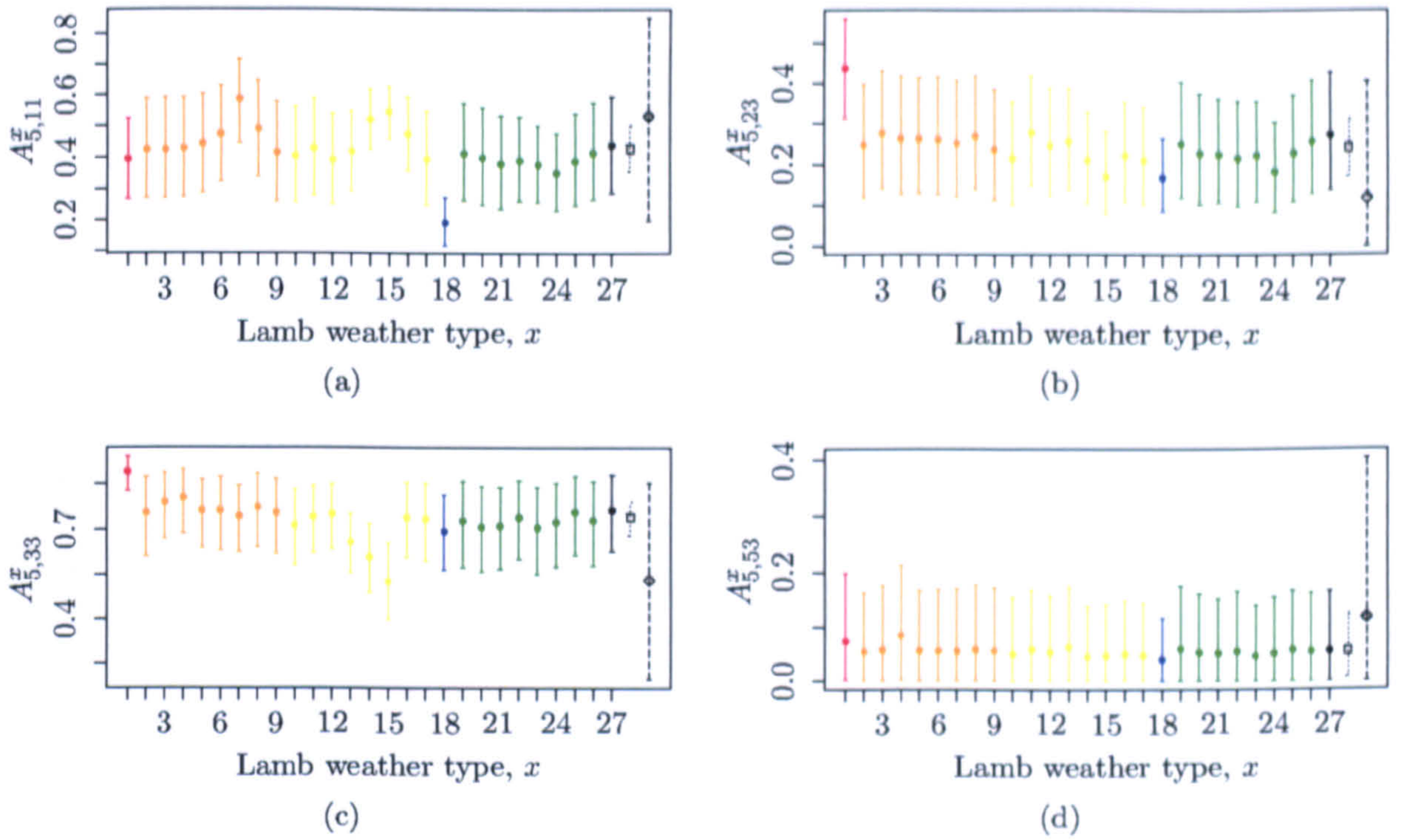


Figure 5.8: Conditional on $r = 5$, posterior means with 95% equi–tailed Bayesian credible intervals for $\mathbf{A}^x_{5,jk}$, $x = 1, \ldots, 27$, (———) and $\boldsymbol{\xi}_{5,jk}$ (·····) when (a) $j = 1$, $k = 1$; (b) $j = 2$, $k = 3$; (c) $j = 3$, $k = 3$; and (d) $j = 5$, $k = 3$. Also shown are the marginal prior means with 95% equi–tailed Bayesian credible intervals (– – –) for the corresponding transition probabilities $\mathbf{A}^x_{5,jk}$, $x = 1, \ldots, 27$.

type is x and, for all x, the transition probabilities are high, indicating that the dry weather state is persistent. Further, if the Lamb weather type is pure anticyclonic ($x = 1$), the posterior for the probability of making this transition, $A^1_{5,33}$, has more density at larger values than the posteriors for $A^x_{5,33}$ for any other Lamb weather type $x \neq 1$. Given that the anticyclonic weather type is generally associated with dry conditions, this is not surprising. Conversely, when the Lamb weather type on the current day is pure south–westerly ($x = 14$) or pure westerly ($x = 15$), the posterior for the probability of remaining in state 3 has more density at lower values than the posteriors for $A^x_{5,33}$ with $x \neq 14$ or 15. As discussed in Chapter 2, this may be because the westerly weather types tend to bring rain to the Pennines and so it is unlikely that the weather state would remain "dry" if the Lamb weather type was pure westerly.

Many of the posterior distributions for transition probabilities $A^x_{5,jk}$, $x = 1, \ldots, 27$, are only noticeably different when the Lamb weather type is pure anticyclonic or pure cyclonic ($x = 18$). For example, the marginal posterior distributions for the transition probabilities $A^x_{5,23}$ and $A^x_{5,11}$, $x = 1, \ldots, 27$, are shown in Figures 5.8(b) and 5.8(a), respectively. Weather state 1 is broadly characterised as dry which provides an explanation as to why the probability of self–transition, $A^x_{5,11}$, is lower when the Lamb weather type is pure cyclonic, since cyclonic Lamb weather types tend to be rain bearing. Similarly, transition from state 2 into the dry weather state, state 3, is more likely, *a posteriori*, if the current Lamb weather type is pure anticyclonic.

For some distinct pairs of Lamb weather types, x and y, and some state j to state k transitions, the central 95% of the posteriors for $A^x_{5,jk}$ and $A^y_{5,jk}$ do not overlap, for example, those for

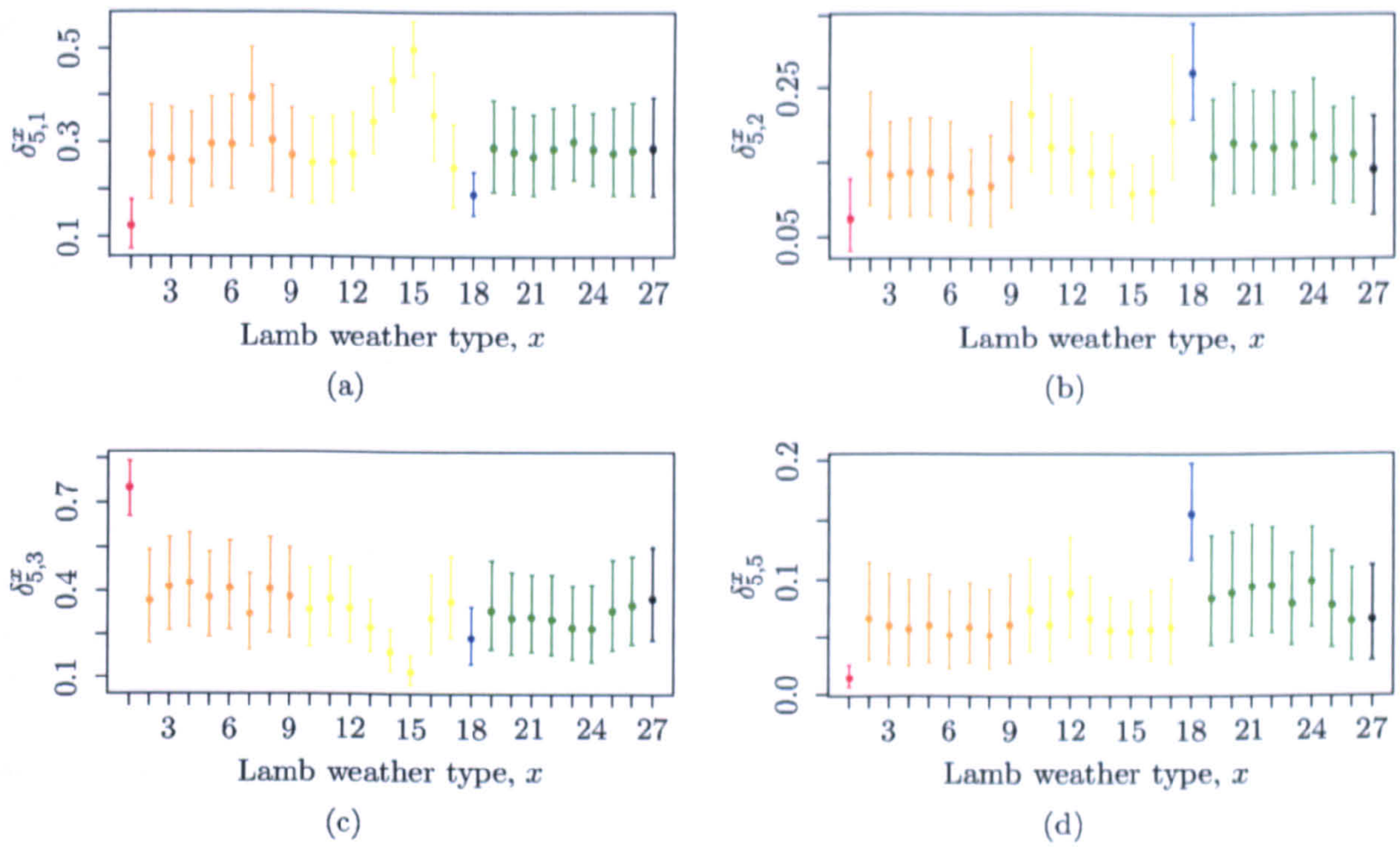


Figure 5.9: Marginal posterior means and 95% equi–tailed Bayesian credible regions for the solution to the matrix equation $\boldsymbol{\delta}^x_5 \boldsymbol{\Lambda}^x_5 = \boldsymbol{\delta}^x_5$, $x = 1, \ldots, 27$. Here $\boldsymbol{\Lambda}^x_5$ is the 5×5 stochastic matrix with j–th row equal to $\mathbf{A}^x_{5,j}$, $\boldsymbol{\delta}^x_5 = (\delta^x_{5,1}, \delta^x_{5,2}, \delta^x_{5,3}, \delta^x_{5,4}, \delta^x_{5,5}) \in \mathscr{S}_5$ and the plots show (a) $\delta^x_{5,1}$, (b) $\delta^x_{5,2}$, (c) $\delta^x_{5,3}$ and (d) $\delta^x_{5,5}$. $\sum_{j=1}^5 \delta^x_{5,j} = 1$ and the plot for $\delta^x_{5,4}$ is not shown.

$A^1_{5,33}$ and $A^{15}_{5,33}$, demonstrating that the Lamb weather types can help to explain some of the transitions between states in the hidden process.

Defining $\boldsymbol{\Lambda}^x_5$ as the 5×5 stochastic matrix with j–th row equal to $\mathbf{A}^x_{5,j}$, we can compute the posterior distribution for the solution to the matrix equation $\boldsymbol{\delta}^x_5 \boldsymbol{\Lambda}^x_5 = \boldsymbol{\delta}^x_5$ for each value of x. The solution, $\boldsymbol{\delta}^x_5$, can then be interpreted as the stationary distribution of the (homogeneous) hidden Markov model that would prevail if the Lamb weather type was always equal to x. Posterior means and 95% equi–tailed credible regions for these hypothetical stationary distributions are displayed in Figure 5.9 for each value of x, and these highlight clear patterns amongst the Lamb weather types. The pure cyclonic type ($x = 18$) offers much more support to weather states 2 and 5 than any other Lamb weather type. Given that state 5 is the wet weather state, and cyclonic Lamb weather types are typically associated with wet conditions, this agrees with our intuition. Most of the cyclonic hybrids, for example, cyclonic south–easterly ($x = 21$) also offer more support to weather state 5 than the anticyclonic or pure directional Lamb weather types. The pure anticyclonic type ($x = 1$) strongly favours weather state 3, as we would expect given that state 3 is dry and anticyclonic Lamb weather types tend to be associated with dry conditions. Many of the anticyclonic hybrids also offer more support to state 3 than the cyclonic or pure directional types. Several of the westerly hybrids ($x = 6$–9 and $x = 23$–25), the pure north–westerly and south–westerly types ($x = 16$ and $x = 14$) and, in particular, the pure westerly type ($x = 15$) offer much more support to weather state 1 than the other Lamb weather types. Similarly, the pure northerly ($x = 17$) and north–easterly ($x = 10$) types offer more support

to weather state 2 than most of the other Lamb weather types. Weather states 1 and 2 are associated with wet conditions at Moorland Cottage, in the Pennines, and Lockwood Reservoir, in East Yorkshire, respectively. The westerly Lamb weather types are the main precipitation bearers for the Pennines, whilst the northerly and easterly types tend to bring rain to Eastern parts of Yorkshire, so these observations are in accordance with the known relationships between Lamb weather types and precipitation.

5.7.3.3 Posterior for $(\mathbf{s} \mid r = 5)$

The marginal posterior mode estimate of the weather state sequence, $\hat{\mathbf{s}}$, conditional on there being $r = 5$ weather states, is shown in Figure 5.10(a). The colour coding for the states is chosen so that at least the clear–cut wet (blue) and dry (red) states are coloured in the same way as their hidden Markov model counterparts from Chapter 4 (see Figure 4.11). As was the case for the hidden Markov model, the dry weather state (here state 3) is, again, the most persistent, and becomes more prevalent towards the end of February. It also appears that the sojourn times in the two wetter weather states, states 4 and 5, are generally shorter than those in states 1–3.

The marginal posterior probabilities, $\Pr(S_t = j \mid \mathbf{w}, \mathbf{d}, \mathbf{x}, r = 5)$, $j \in \mathcal{S}_5$, at each time point were approximated using the simple estimate

$$\hat{\Pr}(S_t = j \mid \mathbf{w}, \mathbf{d}, \mathbf{x}, r = 5) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(s_t^{[i]} = j), \qquad j \in \mathcal{S}_5,$$

where $s_t^{[i]}$ is the i–th (thinned) MCMC draw of s_t. Figures 5.10(b) and 5.10(c) display the estimates for the first and last winter periods, and are representative of patterns seen in other years. Comparison with the corresponding figures for the hidden Markov model in Chapter 4 (Figures 4.11(b) and 4.11(c)) suggests that the NHMM generally leads to more posterior uncertainty in the allocation of days to weather states. One possible reason for this is that the within–state model used in the NHMM can capture a greater range of spatial patterns than that in the hidden Markov model, which assumed conditional independence in space, given the weather state. Therefore, compared with the simpler model, more of the weather states may be able to offer a plausible explanation for the observed patterns of rainfall (occurrence) on any particular day.

5.7.4 Model checking

In this section we compare the posterior predictive distributions for the test quantities introduced in Chapter 4 with the observed statistics, in order to assess the fit of the model and to compare its performance with that of the simple hidden Markov model. The posterior predictive distributions are, again, simulated in the manner described in Section 4.7.4. Since the posterior distribution for r has essentially all of its mass at a single value ($r = 5$), averaging the posterior predictive distribution for a test quantity over the posterior for r is equivalent to using the posterior predictive distribution conditioned on $r = 5$.

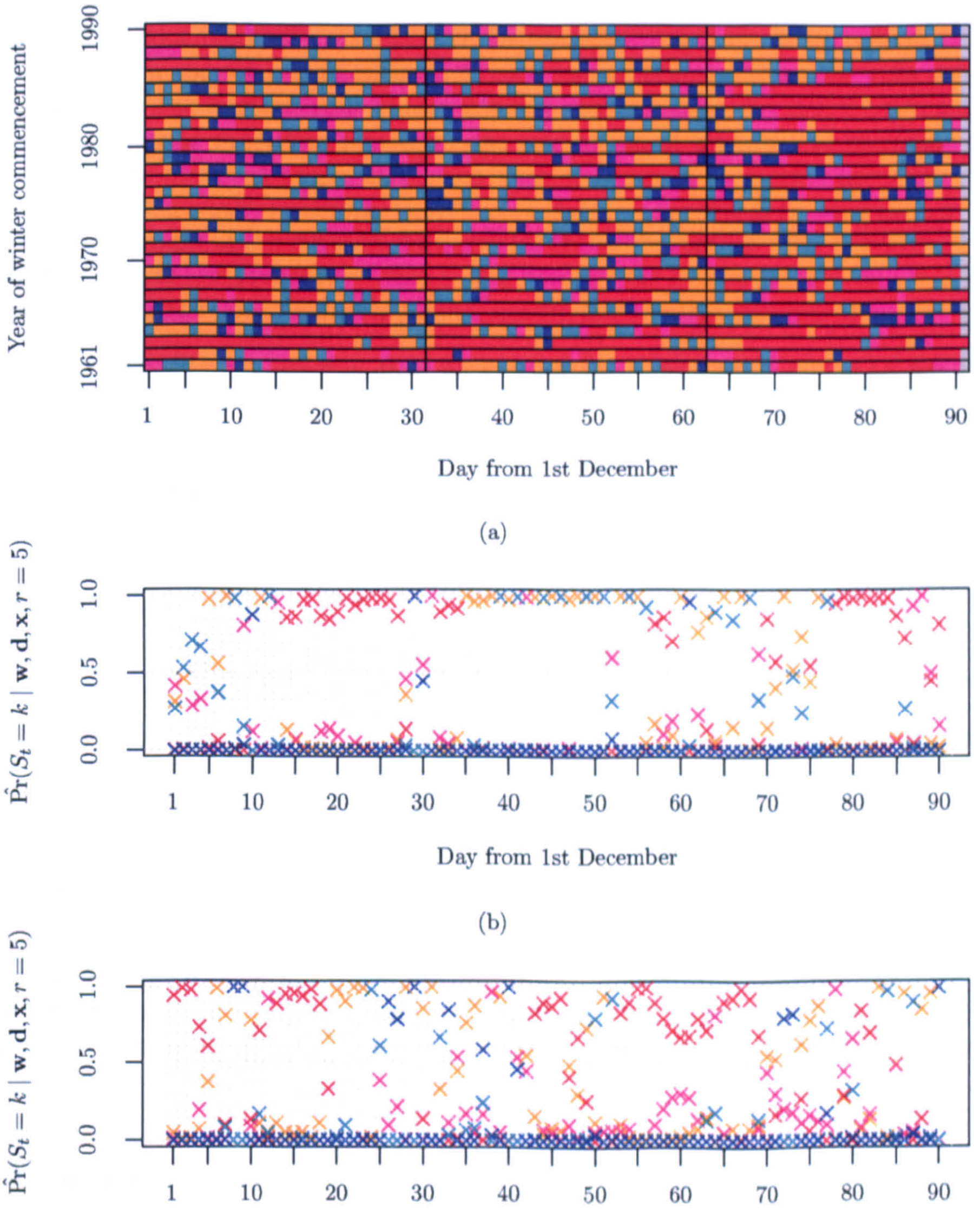


Figure 5.10: (a) Conditional on $r = 5$, marginal posterior mode (MPM) estimate of $\mathbf{s}$; posterior weather state probabilities $\hat{\Pr}(S_t = k \mid \mathbf{w}, \mathbf{d}, \mathbf{x}, r = 5)$ for $k = 1$ (———), $k = 2$ (———), $k = 3$ (———), $k = 4$ (———) and $k = 5$ (———) in the winter (b) 1961/62 and (c) 1990/91.

5.7.4.1 Simple marginal properties

Plots of the posterior predictive distributions for the relative frequencies of rainfall occurrence at each site and for the relative frequencies of each rainfall occurrence vector were very similar to those based on the homogeneous hidden Markov model and so, for brevity, are not shown. In brief, the posterior predictive means for the proportion of wet days at each site matched almost exactly the observed proportions. The same was true of all rainfall occurrence vectors, except the most frequently occurring, $\mathbf{D}_t = (1,1,1,1,1,1)^T$, for which the observed proportion lay slightly to the right of the mean in the posterior predictive distribution. However, this slight underestimation was less pronounced than it had been for the hidden Markov model, suggesting an improvement in the joint model for rainfall occurrence on any particular day.

The probability $\Pr(D_t^{*i} = 1 \mid S_t = s_t, \theta, \mathbf{x})$ for a hypothetical replication D_t^{*i} of $\mathbf{D}_t^i$ has no simple closed form and so the calibration curves in Figure 5.11 are based on the posterior predictive probabilities

$$\begin{aligned}
&\Pr(D_t^{*i} = 1 \mid \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, \mathbf{w}, \mathbf{d}, \mathbf{x}) \\
&\qquad \simeq \frac{1}{N}\sum_{j=1}^{N} \Pr(D_t^{*i} = 1 \mid \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, S_t = s_t^{[j]}, \theta^{[j]}_{\mathrm{obs},s_t^{[j]}}) \\
&\qquad = \frac{1}{N}\sum_{j=1}^{N}\sum_{\mathbf{d}^{-i}} \Pr(D_t^{*i} = 1, \mathbf{D}_t^{*-i} = \mathbf{d}^{-i} \mid \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, S_t = s_t^{[j]}, \theta^{[j]}_{\mathrm{obs},s_t^{[j]}}),
\end{aligned}$$

where $\mathbf{d}_{t-1}$ is the observed rainfall occurrence vector on day $t-1$. Compared with the calibration curves constructed from the hidden Markov model (see Figure 4.13) the plots for the NHMM indicated generally improved fit, although at some sites there is still some discrepancy between the observed and posterior predictive probabilities. This suggests that the posterior predictive distributions are still uninformative about the probabilities over some intervals.

The plots displaying the sample quantiles of the distribution of non-zero rainfall amounts at each site and summaries of the corresponding posterior predictive distributions (not shown) were indistinguishable from those obtained using the hidden Markov model. Again, they gave no reason to question the choice of (a mixture of) gamma distributions to model rainfall amounts on wet days.

5.7.4.2 Spatial structure

For all pairs of sites, the means and 95% equi-tailed posterior predictive distributions for the log odds ratios between rainfall occurrences and the Spearman's rank correlation coefficients between non-zero rainfall amounts are displayed in Figures 5.12(a) and 5.12(b), respectively, together with the observed statistics. One of the deficiencies of the simple hidden Markov model was that it underestimated the larger log odds ratios, but this problem has been largely eradicated by introducting a Markov chain of autologistic models within weather states. From Figure 5.12(a), all of the observed log odds ratios lie within the central 95% of their posterior predictive distributions with the larger observed values positioned only slightly higher than the posterior predictive means.

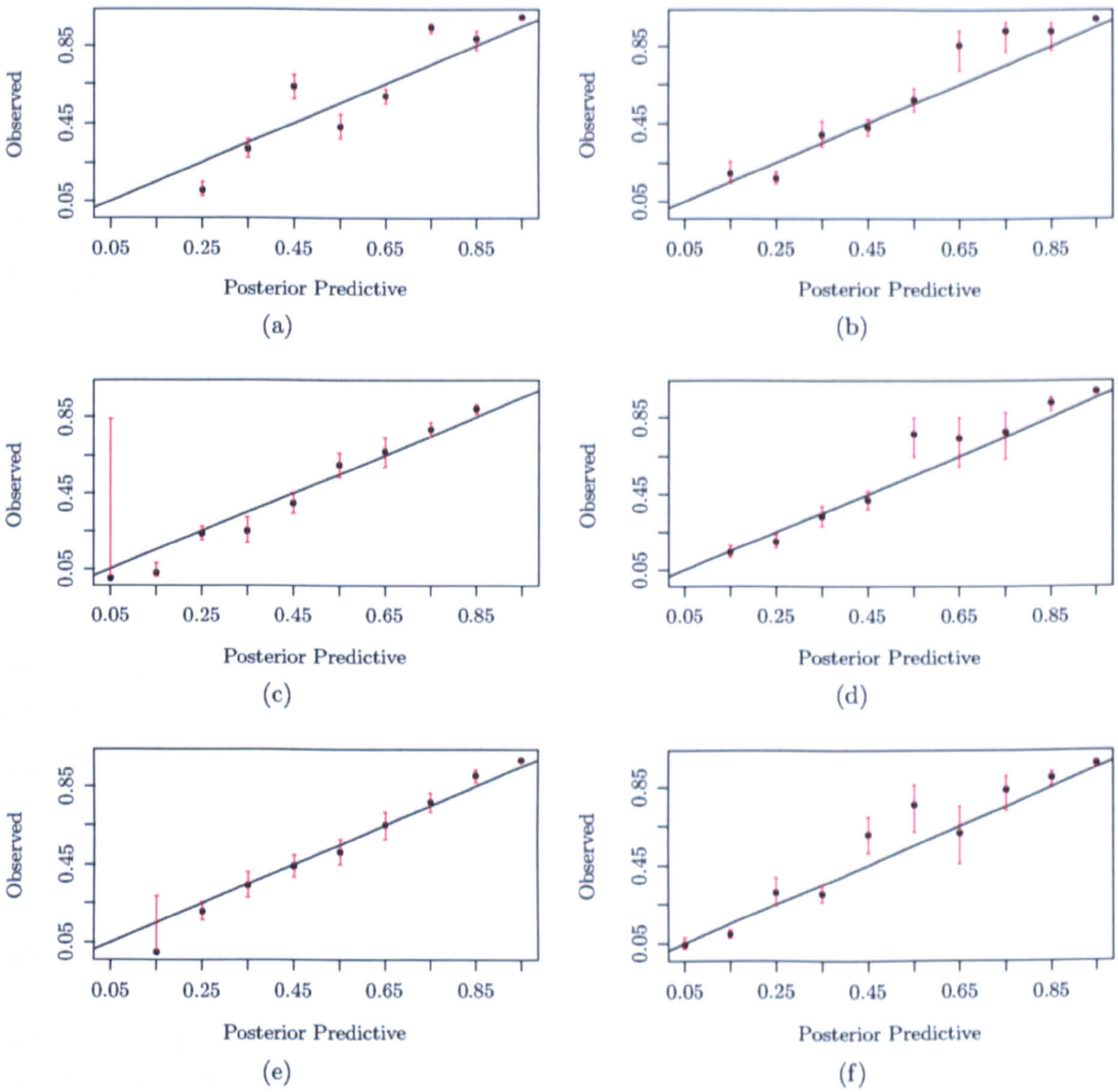


Figure 5.11: Calibration curves for the posterior predictive probability of rain at (a) Lockwood Reservoir; (b) Hull, Pearson Park; (c) Moorland Cottage; (d) the Retreat, York; (e) Great Walden Edge; (f) Kirk Bramwith. (———) is a posterior 95% Bayesian interval for the "true" probability based on the observed sample (assumed binomial) and a uniform prior on the "true" probability.

The hidden Markov model in Chapter 4 assumed conditional independence between rainfall amounts given occurrences and the weather state and was unable to reproduce the highest correlations between pairs of sites. Although the NHMM studied in this chapter continues to make this assumption, it seems to perform better in this regard. The largest observed correlations continue to lie beyond the 97.5% points in their posterior predictive distributions but it appears that the posterior predictive distributions associated with the NHMM assign more density to larger values than those based on the hidden Markov model. This may be due to differences in the properties of the weather states identified by the hidden Markov model and the NHMM. The improvement was not observed when we studied the straightforward non–homogeneous generalisation of the simple hidden Markov model (see Section 5.3.3.1) and so must

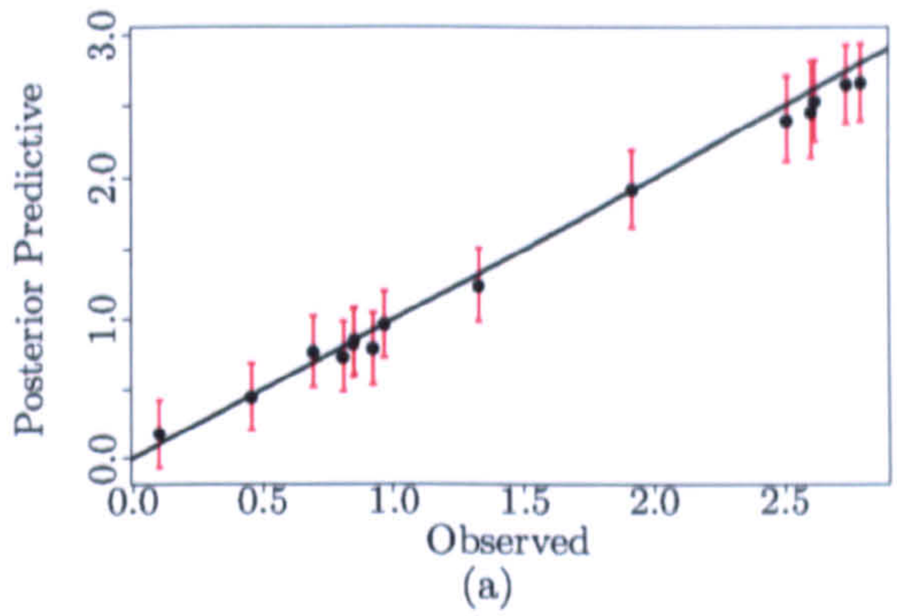


Figure 5.12: Observed values versus posterior predictive means for (a) log odds ratios between rainfall occurrences; and (b) Spearman's rank correlation coefficients between non–zero rainfall amounts at each pair of sites in the Yorkshire network. (———) indicate the posterior predictive 95% Bayesian credible regions.

be due to the improved modelling of rainfall occurrences via the Markov chain of autologistic models, rather than the incorporation of atmospheric information.

In later work (Chapter 6) we will consider models which allow rainfall amounts to be correlated within weather states which may explain the extra correlation left over after conditioning on the weather state.

5.7.4.3 Temporal structure

The empirical survivor functions for wet spells at sites 1, 3 and 6, together with the mean and upper and lower 2.5% points from the posterior predictive distributions are displayed in Figure 5.13. The simple hidden Markov model led to good agreement between the observed survivor functions and the posterior predictive means at sites 2, 4 and 5, and this remains the case for the NHMM. At sites 1, 3 and 6, the simple hidden Markov model underestimated the proportions of longer duration wet spells (see Figure 4.16), with the empirical survivor function lying beyond the 97.5% point in the posterior predictive distribution. From Figure 5.13 it is clear that at site 1, the empirical survivor function now lies within the central 95% of its posterior predictive distribution. At sites 3 and 6, the model still underestimates the proportions of longer duration wet spells, but considerably less so than the simple hidden Markov model.

The hidden Markov model also underestimated the proportions of longer duration dry spells at sites 1 and 3 (see Figure 4.17). However, from Figure 5.14, this problem has largely been rectified by the more complex temporal dependence structure of the NHMM.

Finally, plots (not shown) displaying the observed Spearman's rank correlation coefficients between rainfall amounts at lags 0–8, and the corresponding posterior predictive distributions, exhibited no discernible difference compared to the plots obtained using the hidden Markov model (see Figure 4.18). This means there is still a slight tendency for the NHMM to underestimate the dependence between consecutive rainfall amounts in wet spells, at sites where

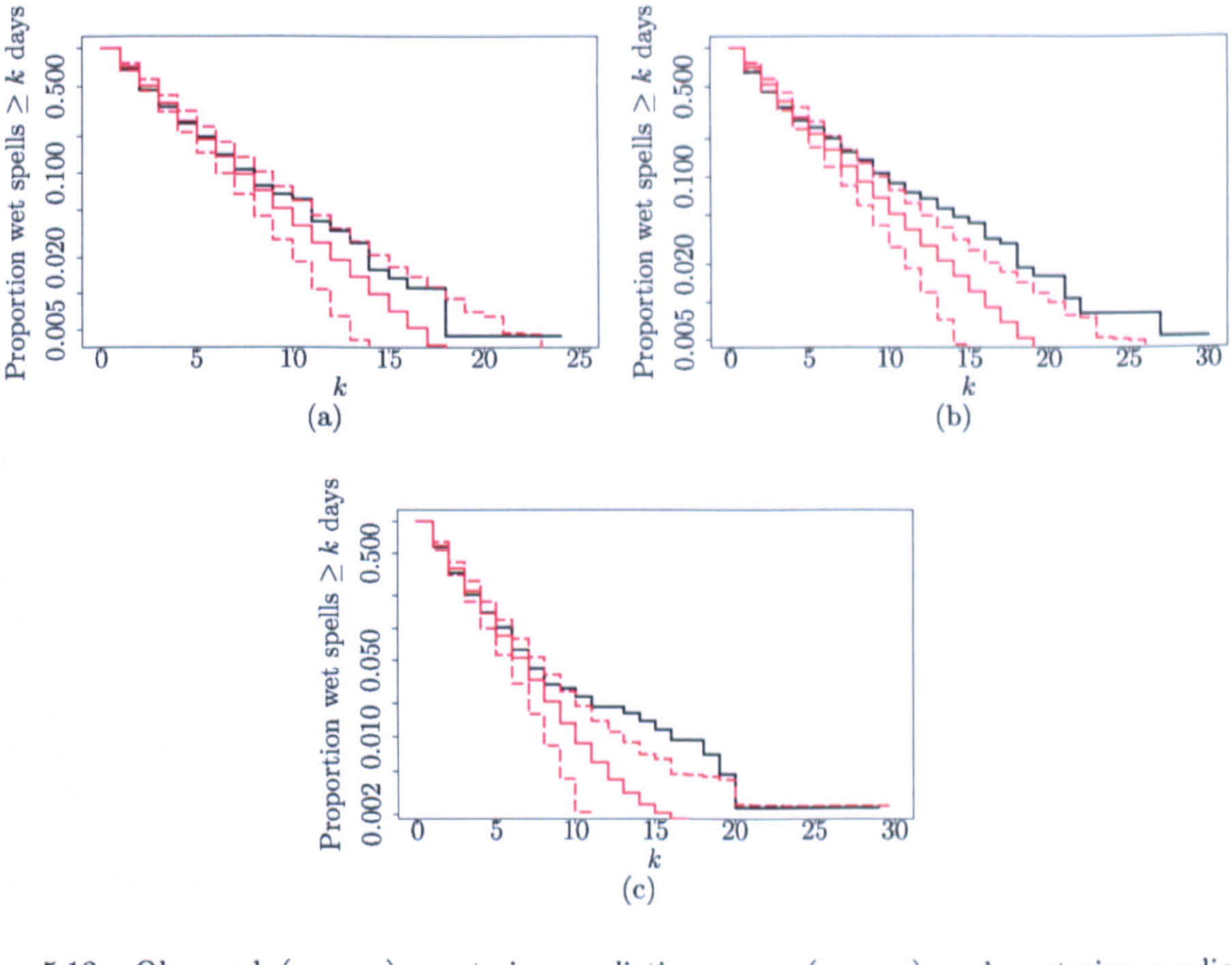


Figure 5.13: Observed (———), posterior predictive mean (———) and posterior predictive 95% Bayesian credible regions (------) for the survival distributions of wet spells at (a) Lockwood Reservoir; (b) Moorland Cottage; (c) Kirk Bramwith.

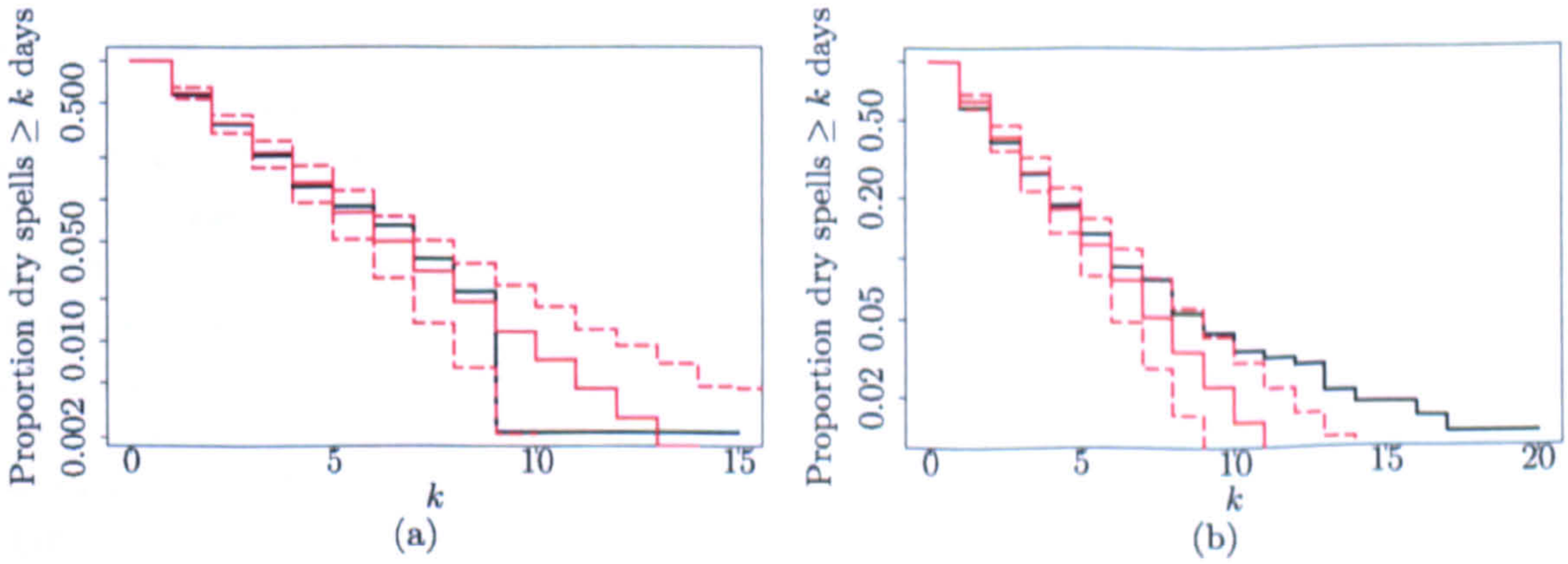


Figure 5.14: Observed (———), posterior predictive mean (———) and posterior predictive 95% Bayesian credible regions (------) for the survival distributions of dry spells at (a) Lockwood Reservoir; (b) Moorland Cottage.

this dependence is strong. The problem of explicitly incorporating dependence between rainfall amounts in wet spells, given the weather state, will be discussed further in Chapter 6.

5.8 Summary

In this chapter we had two main objectives; (i) to develop a model for the weather state process which incorporated categorical atmospheric variables (Lamb weather types) in a manner amenable to posterior inference via MCMC, (ii) to introduce a more sophisticated model for rainfall occurrences, given the weather state. In our fully Bayesian framework, satisfaction of each of these objectives required careful specification of a prior distribution that encouraged borrowing of strength between parameters. This was necessary as compensation for the paucity of information in the data about some of the unknowns in this highly parameterised model.

Regarding the first objective, we suggested describing each vector of (s_{t-1}, x_t)–type transitions as a separate stochastic vector $\mathbf{A}_j^x$ and then introduced *a priori* dependence between $\mathbf{A}_j^1, \ldots, \mathbf{A}_j^{27}$ through a hierarchical Dirichlet prior. We also presented an intuitive strategy for choosing the hyperparameters in these priors which avoided having to think directly about marginal correlations or covariances. In the Yorkshire dataset, we found differences between the stochastic vectors $\mathbf{A}_j^1, \ldots, \mathbf{A}_j^{27}$, *a posteriori*, suggesting that different Lamb weather types can be associated with different precipitation patterns. Indeed, evidence of some of the known relationships between Lamb weather types and Yorkshire precipitation could be gleaned from the analysis.

Considering now the second objective, in this chapter we introduced a Markov chain of autologistic models to describe the joint distribution of rainfall occurrence, given the weather state. Both the marginal likelihood and checks of model fit based on the posterior predictive distribution suggested that the NHMM provided a better description of rainfall in the Yorkshire network than the simple hidden Markov model from Chapter 4.

In spite of the improvements over the simple hidden Markov model, there are several problems with the model proposed in this chapter. Although it was manageable for the small network of Yorkshire sites, computation of the normalising constants in the Markov chain of autologistic models would quickly become analytically intractable if the number of sites was increased. In this case, either the normalising constants would need to be approximated, which would introduce error, or the model would need to be simplified so that its normalising constants could be computed exactly. However, such simplifications may compromise the ability of the model to capture spatial dependence between rainfall occurrences. A second problem is that, *a posteriori*, the model could not predict correlations between non–zero rainfall amounts as strong as some of those observed in the Yorkshire dataset. However, there is no particularly natural way of developing the within weather state model studied in this chapter to incorporate dependence between non–zero rainfall amounts. In Chapter 6 we consider NHMMs in which latent and partially latent multivariate normal random vectors are introduced to jointly model dependence between both rainfall occurrences *and* between rainfall amounts within weather states. The complete data likelihood of this model is also free of potentially intractable normalising constants.

Another major problem with the Markov chain of autologistic models is that there are often weather states in which some parameters are, at best, only weakly likelihood–identifiable. This

leads to ridges in the likelihood in certain directions of the parameter space which, in turn, can cause mixing problems in the MCMC analysis. In order to remove these ridges from the posterior, it is necessary to introduce very strong priors, for example by making parameters with the same function at different sites highly correlated *a priori*. However, another criticism of the autologistic model is that it is difficult to identify the exact nature of the relationships between the various parameters in the likelihood. This in turn makes it very difficult to encapsulate prior information in our prior specification.

The latter criticisms highlight a more general point about hidden Markov models with highly parameterised within-state distributions. Hidden Markov models are designed to partition data into more homogeneous segments. Inevitably, therefore, given enough hidden states, there will be some associated with only a limited range of data patterns. If we adopt a highly parameterised model within each hidden state, then the consequence is likely to be identifiability problems in the likelihood for some states. Therefore, to encourage identifiability in the posterior, it will be helpful to construct priors which allow borrowing of strength between parameters. To this end, it is important to have a framework in which we can think about the relationships between different parameters and are able to construct a prior which conveys this information. Germain *et al.* (2010*b*) consider the problem of constructing a prior for a variance matrix which allows a full specification of all *a priori* beliefs about the variances and covariances, within a distributional framework that guarantees positive definiteness. This prior is used in Chapter 6 for NHMMs which rely on the introduction of latent multivariate normal random vectors.

Chapter 6

Hidden Markov models and latent Gaussian variables in models for rainfall data

6.1 Introduction

Much of the recent work on statistical rainfall modelling has been based on the introduction of latent multivariate normal random variables. For example, Sansó & Guenni (2000), Allcroft & Glasbey (2003) and Ailliot *et al.* (2009) have all taken this approach. Often, the latent variables are related to the observable rainfall amounts through a process of truncation and transformation in which rainfall only occurs if the value of a latent variable exceeds a particular threshold. When this threshold is exceeded, the observed rainfall amount is then equal to some transformation of the latent random quantity. The same unobserved random vector is therefore responsible for capturing the dependence amongst rainfall occurrences as well as non–zero rainfall amounts. Although this provides a parsimonious means of modelling spatial association in the two processes, it prevents independent changes in the probability of rain and the distribution of non–zero rainfall amounts. However, by giving the transformed non–zero rainfall amounts a non–degenerate *distribution*, conditional on the latent Gaussian variables, a much more flexible model can be obtained. In this chapter we use these ideas to build spatio–temporal dependence amongst rainfall occurrences and non–zero rainfall amounts, adding an extra latent layer between the weather states and the observables in a more sophisticated hierarchical NHMM. This resolves two of the difficulties encountered in Chapter 5. Specifically, MCMC can proceed without the need to compute intractable normalising constants, whilst spatial dependence between non–zero rainfall amounts can be incorporated naturally through the dependence structure of the latent multivariate normal random variables.

The remainder of this chapter is organised as follows. In Section 6.2, we discuss how latent Gaussian variables can be used to model spatial dependence between rainfall occurrences, then propose an NHMM in which, effectively, the rainfall occurrences form a Markov chain of multivariate probit models, conditional on the weather state. Section 6.3 extends the ideas from the

previous section, describing a general framework for the spatial modelling of rainfall amounts that is based on two (partially) latent Gaussian variables. We explain how many of the models from the literature fit into this framework, then embed a simplified version into an NHMM for daily rainfall amounts. For this NHMM, Sections 6.4 and 6.5 outline the prior and likelihood, respectively, then Section 6.6 details posterior inference via MCMC for a model with a fixed number of states. In Section 6.7 we explain how the power posterior approach for computing marginal likelihoods needs a correction term when applied to certain models, namely, those for which the set of parameter (and latent variable) values with non-zero posterior density/mass is a proper subset of that comprising values with non-zero prior density/mass. The proposed NHMM belongs to this class and we describe how its marginal likelihood can be approximated. In Section 6.8 we apply the model to the Yorkshire dataset, and compare our results to those obtained in earlier chapters. The posterior predictive performance of the model is assessed through comparisons with data used to fit the model as well as observations from outside of the model-fitting dataset.

6.2 Modelling rainfall occurrence

The hidden Markov model in Chapter 4 assumed that the weather states formed a homogeneous first order Markov chain and that rainfall occurrences were conditionally independent in time and space, given the weather state. The NHMM studied in Chapter 5 allowed atmospheric data to influence the transition probabilities between weather states and modelled rainfall occurrences using a Markov chain of autologistic models, given the weather state. The NHMM was able to reproduce the spatial structure in the observed data much more closely than the simpler model from Chapter 4. In this section we consider models which borrow the dependence structure of the multivariate normal distribution in order to induce dependence between binary variables. We focus on modelling spatial association because our objective is to develop NHMMs for rainfall in which the temporal dynamics will be largely captured through the weather state process. An NHMM for rainfall occurrence data is discussed in Section 6.2.4.

6.2.1 Hierarchical models for spatial binary data

When the autologistic model is used to describe binary data, spatial dependence is introduced at the first stage of the model specification. An alternative approach is to adopt a hierarchical model and introduce spatial association at the second stage using the components of latent multivariate normal random variables as spatially varying random effects. Consider a vector of rainfall occurrence indicators $\mathbf{D} = (D^1, \ldots, D^n)^T$ where D^i is equal to 1 if it rains at site i and is equal to 0 otherwise. A simple hierarchical model might assume that, given model parameters $\boldsymbol{\beta}$ and latent variables Z^i, the rainfall occurrence indicators D^i are conditionally independent Bernoulli random variables,

$$D^i \mid Z^i, \boldsymbol{\beta} \overset{iid}{\sim} \text{Bern}\left(g^{-1}(\eta^i)\right), \qquad i = 1, \ldots, n, \tag{6.1}$$

where $\eta^i = (\mathbf{x}^i)^T\boldsymbol{\beta} + Z^i$ for some link function $g : [0,1] \to \mathbb{R}$ with inverse g^{-1}. The Z^i would then be modelled as spatial random effects coming from a Gaussian process. Therefore the second

stage specification might be

$$\mathbf{Z} = (Z^1, \ldots, Z^n)^T \mid \sigma^2, \phi \sim \mathrm{N}_n\Big(0, \sigma^2 \mathbf{H}(\phi)\Big), \tag{6.2}$$

where $\mathbf{H}(\phi)$ would typically be an isotropic correlation function, for example, the exponential correlation function, where $\mathrm{Corr}(Z^i, Z^j \mid \phi) = \exp(-\phi d_{ij})$ and d_{ij} is the distance between the i–th and j–th sites. At the third stage, priors would be assigned to the parameters β, σ^2 and ϕ. Suitable link functions for binary data are the logit or probit links. This is an example of a generalized linear spatial process model in which, more generally, the data can be assumed to arise from any member of the class of exponential family models. For further details, see Diggle *et al.* (1998) who developed the model framework, or Banerjee *et al.* (2004) who suggested using the Bernoulli version as a model for multi–site rainfall occurrence data on a single day.

Choosing the logit link, Velarde *et al.* (2004) model weekly rainfall occurrence at 25 sites in the central region of Brazil using a modified version of the three stage Bernoulli model, in which the spatial random effects are CAR effects. Conditionally autoregressive (CAR) models, first introduced by Besag (1974), are a class of models for multivariate data which have the property that the joint distribution of the data is uniquely determined by the full conditionals, assuming a particular neighbourhood structure. For example, the autologistic model, studied in Chapter 5, is a CAR model for binary data. The Gaussian version (sometimes called the autonormal model) is often used at the second stage of a hierarchical model to capture spatial association via random effects. Specifically, let $\mathbf{Z} = (Z^1, \ldots, Z^n)^T$ be a vector of spatially varying random effects and let w_{ij} be equal to 1 if location j is a neighbour of location i and equal to 0 otherwise. A very simple CAR model takes

$$Z^i \mid Z^1, \ldots, Z^{i-1}, Z^{i+1}, \ldots, Z^n, \tau^2 \sim \mathrm{N}(\bar{Z}^i, \tau^2/m_i), \tag{6.3}$$

where $m_i = \sum_{j \neq i} w_{ij}$ is the number of neighbours of location i and $\bar{Z}^i = \sum_{j \neq i} w_{ij} Z^j / m_i$ is the average of the Z^j, $j \neq i$, for locations j that are neighbours of i. This leads to a joint distribution of pairwise difference form

$$p(Z^1, \ldots, Z^n \mid \tau^2) \propto \exp\left\{ -\frac{1}{2\tau^2} \sum_{i=1}^{n-1} \sum_{j>i} w_{ij} (Z^i - Z^j)^2 \right\}, \tag{6.4}$$

which is improper and unaffected by the addition of a constant to all of the Z^i. Thus to identify an intercept term in the linear predictor, a constraint must be imposed, for example, $\sum_{i=1}^n Z^i = 0$, although this does not make the density proper. The joint distribution in (6.4) can also be written as

$$p(Z^1, \ldots, Z^n \mid \tau^2) \propto \exp\left\{ -\frac{1}{2\tau^2} \mathbf{Z}^T (\mathbf{D}_w - \mathbf{W}) \mathbf{Z} \right\}, \tag{6.5}$$

where $\mathbf{D}_w = \mathrm{diag}(m_1, \ldots, m_n)$ and the matrix $\mathbf{W}$ contains w_{ij} in the (i,j)–th position for $i \neq j$ and zero's on the diagonal. The impropriety in the joint distribution (6.5) can be remedied by introducing a "propriety parameter" which makes the precision matrix, $(\mathbf{D}_w - \mathbf{W})/\tau^2$, non–singular. For theoretical details and properties of CAR models see, for example, Besag *et al.* (1991) or Banerjee *et al.* (2004).

Rainfall data would usually be classified as point data, meaning the rainfall amounts are measured at locations whose coordinates vary continuously over multidimensional Euclidean space. Typically, models with CAR effects are used for areal data in which measurements are often sums or averages of variables over the areal units into which a geographical region has been divided. For such data, it may be natural to use neighbour–based notions of proximity between spatial locations, for example, two units are neighbours if they share a common boundary. However, with point data, as long as a sensible neighbourhood structure can be defined, CAR effects offer computational advantages over more general Gaussian process random effects in that their conditional specification makes Gibbs sampling particularly convenient. Additionally, the CAR model directly parameterises the precision matrix of the random effects, unlike a Gaussian process model which will typically parameterise the variance matrix using an isotropic covariance function. Therefore, because likelihood evaluations require computation of a quadratic involving the precision matrix, the use of CAR effects can simplify such calculations. In their rainfall model, Velarde *et al.* (2004) define the neighbourhood set of each site using a distance–based criterion. Independent and identically distributed sets of CAR effects of the form (6.3) are then incorporated in the model for rainfall occurrence in each week. For the rainfall occurrence indicator at the i–th site in week t, D_t^i, the fixed effects component of the linear predictor $((\mathbf{x}_t^i)^T\boldsymbol{\beta}_t^i)$ incorporated seasonal effects and temporal association, the latter by including the terms D_{t-q}^i in $\mathbf{x}_t^i$ for $q = 1, 2, 3$. One drawback of this approach is the very limited scope for prior elicitation in the distribution of CAR effects, in this case being confined to the choice of prior for a single parameter τ^2. Also note that although the neighbourhood sets were defined using distances between sites, this will not necessarily lead to any distance based association in the variance matrix of the random effects.

In these hierarchical models for binary data, spatial dependence is introduced using normally distributed random effects at the level of the linear predictor. An alternative way to induce dependence through the multivariate normal distribution is to directly define each binary variable as some transformation of a latent normal variable. This idea is central to the multivariate probit (MVP) model, in which the latent multivariate normal random variables deterministically fix the values of the binary variables by means of a threshold specification.

6.2.2 The multivariate probit model

The multivariate probit model was introduced by Ashford & Sowden (1970) as a generalisation of the standard binary probit model. The flexible dependence structure of the multivariate normal distribution is used to model dependence between binary variables by introducing a latent multivariate normal random vector, whose mean is typically expressed as a linear combination of observed covariates. The binary data are then assumed to arise through a threshold specification on the underlying latent vector. Appealing features of the MVP model therefore include the ease with which covariates can be incorporated and the separation of trend and dependence parameters in the linear predictor and the variance matrix, respectively. From a Bayesian perspective, the ability to think independently about the parameters in the linear predictor and the variance matrix could, in principle, offer an advantage over the autologistic model, discussed in Chapter 5, for which it was difficult to assess separately the spatial trend and spatial association parameters.

In general, evaluation of the observed data likelihood in MVP models is difficult because it involves computing multidimensional normal integrals. This can be avoided through a combination of Gibbs sampling and data augmentation in which the latent variables are appended to the set of unknowns. An MCMC scheme then alternates between generating draws from the conditional posterior distribution of the latent variables, given the model parameters, and the conditional posterior distribution of the model parameters, given the latent variables. Chib & Greenberg (1998) were the first to present this as an MCMC solution to numerical analyses of MVP models.

Denote by $\mathbf{D}_t = (D_t^1, \ldots, D_t^n)^T$ a binary random vector whose i–th entry, D_t^i, corresponds to the t–th observation on the i–th variable for $t = 1, \ldots, T$ and $i = 1, \ldots, n$, and let $\mathbf{Z}_t = (Z_t^1, \ldots, Z_t^n)^T$ be an n–variate normal random vector for $t = 1, \ldots, T$. Let $\mathbf{X}_t = (\mathbf{X}_{t,0}, \ldots, \mathbf{X}_{t,m-1})$ denote a $(n \times mn)$ design matrix for the t–th observation where, for $j = 0, \ldots, m-1$, $\mathbf{X}_{tj} = \text{diag}(\mathbf{x}_{tj}^T)$. Here, $\mathbf{x}_{tj} = (x_{tj}^1, \ldots, x_{tj}^n)^T$ is a vector of covariates whose i–th entry is associated with the t–th observation on the i–th variable. Suppose that $\mathbf{Z}_t$ has distribution

$$\mathbf{Z}_t \mid \boldsymbol{\beta}, \boldsymbol{\Sigma} \overset{iid}{\sim} \mathrm{N}_n(\mathbf{X}_t\boldsymbol{\beta}, \boldsymbol{\Sigma})$$

for $t = 1, \ldots, T$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^T, \ldots, \boldsymbol{\beta}_{m-1}^T)^T$ and $\boldsymbol{\beta}_j = (\beta_j^1, \ldots, \beta_j^n)^T \in \mathbb{R}^n$ for $j = 0, \ldots, m-1$. Therefore, each $\boldsymbol{\beta}_j$ is a vector of unknown regression coefficients whose i–th entry is associated with the i–th variable. Finally, we relate $\mathbf{D}_t$ and $\mathbf{Z}_t$ through the signs of the elements of Z_t^i, specifically

$$D_t^i = \mathbb{I}(Z_t^i > 0), \qquad i = 1, \ldots, n, \quad t = 1, \ldots, T.$$

The probability that $\mathbf{D}_t = \mathbf{d}_t$, conditioned on parameters $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ and on covariates $\mathbf{X}_t$, is given by

$$\Pr(\mathbf{D}_t = \mathbf{d}_t \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \int_{B_t^n} \cdots \int_{B_t^1} \phi_n(\mathbf{Z}_t \mid \mathbf{X}_t\boldsymbol{\beta}, \boldsymbol{\Sigma})\, d\mathbf{Z}_t$$

where $\phi_n(\mathbf{Z}_t \mid \mathbf{X}_t\boldsymbol{\beta}, \boldsymbol{\Sigma})$ is the density of the n–variate normal distribution, $\mathrm{N}_n(\mathbf{X}_t\boldsymbol{\beta}, \boldsymbol{\Sigma})$, evaluated at $\mathbf{Z}_t$, and

$$B_t^i = \begin{cases} (0, \infty) & \text{if } d_t^i = 1, \\ (-\infty, 0] & \text{if } d_t^i = 0. \end{cases}$$

Let $B_t = B_t^1 \times B_t^2 \times \cdots \times B_t^n$ so that B_t depends purely on the observed data, $\mathbf{D}_t$.

As presented above, the parameters $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ are not identifiable in the observed data likelihood. For example, for any positive definite $(n \times n)$ diagonal matrix, $\mathbf{C} = \text{diag}(C_1, \ldots, C_n)$, suppose that $\boldsymbol{\Omega}$ is an $(n \times n)$ matrix defined by $\boldsymbol{\Omega} = \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^T$ and $\boldsymbol{\gamma} = (\gamma_0^1, \ldots, \gamma_0^n, \ldots, \gamma_{m-1}^1, \ldots, \gamma_{m-1}^n)^T$ is an mn–vector whose elements are defined by $\gamma_j^i = C_i\beta_j^i$. It is then easy to show that $\Pr(\mathbf{D}_t = \mathbf{d}_t \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \Pr(\mathbf{D}_t = \mathbf{d}_t \mid \boldsymbol{\gamma}, \boldsymbol{\Omega})$. Therefore, for the parameters $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ to be likelihood identifiable, restrictions need to be placed on $\boldsymbol{\beta}$ or $\boldsymbol{\Sigma}$ that prevent arbitrary rescaling of the linear predictor, $\mathbf{X}_t\boldsymbol{\beta}$, or the variance matrix, $\boldsymbol{\Sigma}$. Section 6.2.3 provides further discussion of ways in which the non–identifiability problem can be handled.

When the probit link is used in the hierarchical Bernoulli model, its equivalence to the MVP model can easily be verified. Here, by equivalence, we simply mean that any joint mass function for the binary variables that is produced by the MVP model can also be generated from

the hierarchical model through an appropriate choice of parameters, and *vice versa*. However, the MVP model offers a computational advantage over hierarchical Bernoulli models. This is because the full conditional distributions for the latent variables in the MVP model are truncated normal distributions (see Section 6.6.2) from which draws can easily be generated. In the hierarchical Bernoulli model, however, the full conditional distributions for the spatial effects are non–standard and therefore more difficult to sample.

6.2.3 Handling the non–identifiability problem in MVP models

In the previous subsection we remarked that the parameters $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ in the MVP model are not likelihood identifiable. Edwards & Allenby (2003) suggest handling this problem by ignoring it and post–processing the posterior draws of $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ using the transformations $\mathbf{R} = \mathbf{S}^{-1}\boldsymbol{\Sigma}\mathbf{S}^{-1}$ and $\alpha_j^i = S_i^{-1}\beta_j^i$ for $j = 0, \ldots, m-1$ and $i = 1, \ldots, n$, where $\mathbf{S} = \text{diag}(S_1, \ldots, S_n)$ is a diagonal matrix of standard deviations and $\mathbf{R}$ is the correlation matrix. The parameters $\boldsymbol{\alpha} = (\alpha_0^1, \ldots, \alpha_0^n, \ldots, \alpha_{m-1}^1, \ldots, \alpha_{m-1}^n)^T$ and $\mathbf{R}$ are then identifiable in the likelihood and will have proper posterior distributions as long as the prior distributions for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are proper. Draws of the identified parameters $(\boldsymbol{\alpha}, \mathbf{R})$ can be used to assess the convergence of the MCMC sampler and for posterior predictive inference. The marginal posterior of $(\boldsymbol{\alpha}, \mathbf{R})$ is theoretically the same as that obtained by working only with the identified parameters and the joint prior for $(\boldsymbol{\alpha}, \mathbf{R})$ which was induced by the prior for $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$. The main advantages of this approach are that it is typically easy to sample from the full conditional distributions of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, especially if conjugate priors are chosen, and the performance of the MCMC sampler might be improved compared to a situation in which identifiability constraints were placed on $\boldsymbol{\beta}$ or $\boldsymbol{\Sigma}$. This effect was observed by McCulloch *et al.* (2000) in an analysis of the related multinomial probit model. From the perspective of prior elicitation, however, a subjective Bayesian should feel uncomfortable specifying priors for non–identified parameters. Ideally, we should relate expressions of prior belief about unknown parameters to expressions of belief concerning observable quantities (see, for example, Garthwaite *et al.*, 2005), but this is not possible if parameters cannot be identified in the likelihood. Although it would be possible to use numerical techniques to deduce the marginal prior for the identified parameters, $(\boldsymbol{\alpha}, \mathbf{R})$, in general, this density will not be available in closed form. This makes it difficult to express prior beliefs about $(\boldsymbol{\alpha}, \mathbf{R})$ through the prior for $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$.

The most common strategy for dealing with the non–identifiability problem in MVP models is to constrain the variance matrix to be a correlation matrix. However, the space of correlation matrices is subject to complex constraints, requiring matrices to be positive definite, with fixed diagonal elements equal to one. These constraints present an obstacle to prior elicitation, making it difficult to specify a prior for the correlation matrix that conveys substantive initial information; see, for example, Germain *et al.* (2010*b*) for further details. They also introduce computational problems in sampling the correlation matrix from its full conditional distribution (FCD). Much of the work in the literature on MVP models has focused on the latter issue.

Noting that it is easy to derive the range of an individual correlation, R_{ij}, over which a correlation matrix, $\mathbf{R}$, remains positive definite, Barnard *et al.* (2000) suggest using a griddy Gibbs sampler to draw the correlations one–at–a–time from their FCDs. However, high correlations between components of the correlation matrix can lead to poor mixing and slow convergence of the

MCMC sampler. This technique is also likely to be computationally expensive due to the $n(n-1)/2$ evaluations of the FCD of $\mathbf{R}$ required per complete draw. Chib & Greenberg (1998) suggest updating the correlation matrix using either a random walk sampler, or a particular kind of independence sampler, in which the shape of the proposal density is tailored to the unnormalised FCD of $\mathbf{R}$. For example, this can be achieved by using a t–distribution with mean equal to the approximate mode of the log FCD, and variance matrix equal to a scaled version of the negative Hessian of the log FCD, evaluated at the mode. A problem with both of these proposal densities is that they cannot be guaranteed to generate valid correlation matrices. Further, the random walk sampler might suffer from poor mixing due to high autocorrelations between successive draws. Although the independence sampler might offer an improvement in this regard, numerical approximation of the mode and Hessian of the log FCD at every iteration can be computationally demanding.

Borrowing ideas from parameter expansion algorithms (Liu *et al.*, 1998), Liu (2001) suggests an MCMC scheme that introduces a reparameterisation of the regression coefficients, $\boldsymbol{\beta}$, the correlation matrix, $\mathbf{R}$, and the latent variables, $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_T)$, when sampling the correlation matrix. In the reparameterised space, the correlation matrix is transformed to a variance matrix and this allows a draw of $\mathbf{R}$ to be generated by drawing the variance matrix from its FCD, then transforming back via $\mathbf{R} = \mathbf{S}^{-1}\boldsymbol{\Sigma}\mathbf{S}^{-1}$. Here, $\mathbf{S}$ is a diagonal matrix of standard deviations which take their scale from the latent variables in the reparameterised space. By choosing the prior for $\mathbf{R}$ to be that which is induced by Jeffrey's prior on a variance matrix, $\pi(\mathbf{R}) \propto |\mathbf{R}|^{-(n+1)/2}$, the FCD for the variance matrix in the reparameterised space is an inverse Wishart distribution. This makes it easy to sample the variance and hence correlation matrix. As a by–product of the sampling procedure, draws of the scale parameters, $\mathbf{S}$, are obtained and can be used to adjust the current draws of the regression coefficients and latent variables. This can improve the mixing and convergence of the MCMC sampler. However, to be implemented as described, Jeffrey's prior must be chosen for the correlation matrix which, by design, cannot reflect substantive prior information. Liu & Daniels (2006) propose the "parameter expanded reparameterisation and Metropolis Hastings algorithm" as an extension of the work by Liu (2001) to allow $\mathbf{R}$ to have any prior. To accommodate this extra flexibility, $\mathbf{R}$ must be sampled via $\boldsymbol{\Sigma}$ in a Metropolis Hastings step, in which the inverse Wishart distribution is only used to generate a *proposal* for $\boldsymbol{\Sigma}$ in the reparameterised space, and hence a proposal for $\mathbf{R}$.

Zhang *et al.* (2006) suggest the "parameter extended Metropolis Hastings algorithm" for sampling the correlation matrix in MVP models. In some sense, this is a cross between the "parameter expanded reparameterisation and Metropolis Hastings algorithm" of Liu & Daniels (2006) and the approach of Edwards & Allenby (2003) involving non–identifiable parameters. In this algorithm, the likelihood is defined only in terms of identifiable parameters, $(\boldsymbol{\beta}, \mathbf{R})$, but a joint prior must be specified for $(\boldsymbol{\beta}, \mathbf{R}, \mathbf{S})$, with $\mathbf{S}$ being a diagonal matrix of (non–identified) standard deviations. Interest then lies in the joint posterior distribution $\pi(\boldsymbol{\beta}, \mathbf{R}, \mathbf{S}, \mathbf{Z} \mid \mathbf{D})$. The pair, $(\mathbf{R}, \mathbf{S})$, are sampled jointly in a Metropolis Hastings step in which a proposal, $(\mathbf{R}^*, \mathbf{S}^*)$, is generated by drawing a variance matrix, $\boldsymbol{\Sigma}^*$, from a Wishart proposal distribution. This distribution is centred at the current value of $\boldsymbol{\Sigma} = \mathbf{SRS}$, and the relevant Jacobian must be incorporated in the proposal density of the acceptance ratio. The advantage of this technique is that it allows the correlations to be updated in a single block and, unlike the methods suggested by Chib & Greenberg (1998), proposals are guaranteed to be valid correlation matrices. However, because this Metropolis Hastings method is essentially a random walk on the Wishart scale, this can

lead to poor mixing due to high autocorrelations between successive draws. In eliciting the prior for $(\mathbf{R}, \mathbf{S})$, the authors suggest choosing a Wishart or inverse Wishart distribution for $\boldsymbol{\Sigma}$ and then deducing the prior this induces for $(\mathbf{R}, \mathbf{S})$. However, unless the (fixed) scale matrix in the original prior for $\boldsymbol{\Sigma}$ is diagonal, it is not possible to analytically obtain the marginal prior for $\mathbf{R}$, making it difficult to assess its properties. This would not, however, be a problem if the modeller assumed *a priori* independence between $\mathbf{R}$ and $\mathbf{S}$, and adopted any proper prior for $\mathbf{S}$, together with a separately assessed prior for $\mathbf{R}$.

The methods proposed by Liu & Daniels (2006), Zhang *et al.* (2006) and Chib & Greenberg (1998) all provide techniques for sampling the correlation matrix in MVP models without restricting the choice of its prior. However, as explained earlier in this subsection, choosing a prior for a correlation matrix which is representative of genuine beliefs can be difficult due to the constraints on the space of correlation matrices. In Germain *et al.* (2010*b*) we develop a prior for the variance matrix in multivariate normal distributions. Further details are provided in Section 6.4 but, briefly, this prior is based on the modified Cholesky decomposition of the precision matrix and can be used to provide a complete pre–data summary of our uncertainty about the value of $\boldsymbol{\Sigma}$. To be able to use this prior directly we need to parameterise the likelihood in terms of the full variance matrix. We therefore make use of an alternative approach to handling the non–identifiability problem, which has been neglected in the literature in favour of constraining the variance matrix. Specifically, we can prevent arbitrary rescaling of the linear predictor $\mathbf{X}_t\boldsymbol{\beta}$ by placing restrictions on the regression coefficients $\boldsymbol{\beta}$. This also has the advantage of removing the difficult problem of sampling a correlation matrix.

One appropriate way of constraining $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^T, \ldots, \boldsymbol{\beta}_{m-1}^T)^T$ is to insist that, say, $\boldsymbol{\beta}_{m-1} \in \{-1, 1\}^n$. This is a particularly natural solution when the explanatory variables $\mathbf{x}_{t,m-1} = (x_{t,m-1}^1, \ldots, x_{t,m-1}^n)^T$ are indicator variables. In this case

$$\Pr(D_t^i = 1 \mid x_{t,m-1}^i = d, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \Phi\left(\frac{\beta_0^i x_{t,0}^i + \cdots + \beta_{m-2}^i x_{t,m-2}^i + \beta_{m-1}^i d}{\sqrt{\Sigma^{ii}}}\right),$$

for $i = 1, \ldots, n$. Therefore, compared to a baseline where $x_{t,m-1}^i = 0$, changing only this indicator to $x_{t,m-1}^i = 1$ can *either* increase ($\beta_{m-1}^i = 1$) *or* decrease ($\beta_{m-1}^i = -1$) the marginal probability that $D_t^i = 1$. When $\mathbf{x}_{t,m-1}$ does not consist of indicator variables, it might be more appropriate to choose $\boldsymbol{\beta}_{m-1} \in \{-\ell, \ell\}^n$ for some other $\ell \in \mathbb{R}$. The value chosen should, of course, be taken into consideration when eliciting priors for the other regression coefficients, $\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_{m-2}$.

Introducing identifiability constraints is a well known cause of convergence and mixing problems in MCMC samplers and the particular way that an identifiability problem is remedied can affect the performance of the sampler. Although it is not a primary objective, the constrain–$\boldsymbol{\beta}$ approach to handling the non–identifiability problem might lead to an MCMC algorithm with better mixing properties than some of those previously discussed, which constrain $\boldsymbol{\Sigma}$. To investigate this further, omitting full details for brevity, we conducted a simulation experiment for an MVP model in which

$$\mathbf{Z}_t \mid \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, \boldsymbol{\beta}, \boldsymbol{\Sigma} \sim \mathrm{N}_n(\mathbf{X}_t\boldsymbol{\beta}, \boldsymbol{\Sigma})$$

for $t = 1, \ldots, T$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^T, \boldsymbol{\beta}_1^T)^T$ and $\mathbf{X}_t = (\mathbf{X}_{t0}, \mathbf{X}_{t1})$, with $\mathbf{X}_{t0} = \mathbf{I}_n$ and $\mathbf{X}_{t1} =$

$\text{diag}(d_{t-1}^1, \ldots, d_{t-1}^n)$. For simplicity, the sequence was initialised with independent discrete uniform distributions, $D_0^i \sim \text{U}\{0,1\}$, $i = 1, \ldots, n$. The binary sequence $\{\mathbf{D}_t : t = 1, \ldots, T\}$ then formed a Markov chain of n–vectors of correlated binary variables. Taking $n = 7$, $T = 1000$ binary vectors were simulated from a model in which $\boldsymbol{\Sigma}$ was taken to be a correlation matrix and each component of $\boldsymbol{\beta}_1 = (\beta_1^1, \ldots, \beta_1^n)^T$ was set equal to 1. We then generated 50,000 posterior samples using four MCMC implementations; (i) constrain–$\boldsymbol{\beta}$ and, for simplicity, restrict $\beta_1^i = 1$ for $i = 1, \ldots, n$, (ii) constrain–$\boldsymbol{\Sigma}$ and update the correlation matrix, $\mathbf{R}$, in four blocks using the simple symmetric random walk proposed by Chib & Greenberg (1998), (iii) constrain–$\boldsymbol{\Sigma}$ and update the correlation matrix using the parameter expanded reparameterisation and Metropolis Hastings algorithm of Liu & Daniels (2006) and (iv) constrain–$\boldsymbol{\Sigma}$ and update the correlation matrix using the parameter extended Metropolis Hastings algorithm of Zhang *et al.* (2006).

Let $\boldsymbol{\Sigma} = \mathbf{SRS}$, where $\mathbf{S}$ is a diagonal matrix of standard deviations. *A priori* independence was assumed between $\mathbf{R}$, $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ in (ii)–(iv) and between $\mathbf{R}$, $\mathbf{S}$ and $\boldsymbol{\beta}_0$ in (i). In all cases, multivariate normal priors were chosen for $\boldsymbol{\beta}_0$, whilst the jointly uniform prior (Barnard *et al.*, 2000) over the set of all correlation matrices, $\pi(\mathbf{R}) \propto 1$, was chosen for the correlation matrix. For (i), independent zero–median lognormal distributions were chosen for the diagonal elements in $\mathbf{S}$ and the resulting joint prior for $(\mathbf{R}, \mathbf{S})$ induced a prior for $\boldsymbol{\Sigma}$. In (ii)–(iv), multivariate normal priors were chosen for $\boldsymbol{\beta}_1$, with unit means. Note that in (iv), where a prior for $(\mathbf{R}, \mathbf{S})$ is required, following Zhang *et al.* (2006), the jointly uniform prior is achieved for $\mathbf{R}$, marginally, by choosing a Wishart prior for the variance matrix, $\boldsymbol{\Sigma}$, with identity scale matrix and $n+1$ degrees of freedom. This is then converted to a joint prior for $(\mathbf{R}, \mathbf{S})$. So that the prior specifications for (i) and for (ii)–(iv) could be regarded as roughly "equivalent", the remaining hyperparameters in the priors for $\boldsymbol{\beta}_1$ and $\mathbf{S}$ were adjusted until they led to prior predictive distributions for $\mathbf{D}_t$ in which the first and second moments, obtained by Monte Carlo integration, approximately matched.

The full conditional distributions for Z_t^i, $t = 1, \ldots, T$, $i = 1, \ldots, n$, and for $\boldsymbol{\beta}$ (or $\boldsymbol{\beta}_0$ in (i)) are truncated normal and multivariate normal distributions, respectively, which can be updated in simple Gibbs steps (see Section 6.6 for more details). In (i), the full conditional distribution for $\boldsymbol{\Sigma}$ is non–standard and was updated in a Metropolis Hastings step, in which the normalised likelihood (an inverse Wishart density) was used to generate the proposals. The methods for updating the correlation matrix in (ii)–(iv) were specific to the algorithm and outlined, briefly, earlier in this subsection.

In terms of recovering the parameters used to generate the data, there was little difference between any of the four implementations. However, the posterior standard deviations for the elements of the variance matrix, $\boldsymbol{\Sigma}$, in (i) were, as expected, larger than those for the more constrained elements of the correlation matrix, $\mathbf{R}$, in (ii)–(iv). The time taken to generate the posterior samples was broadly similar across implementations (i)–(iii), but around 40% slower in implementation (iv). Based on posterior samples that had been thinned to every 20–th iterate, trace and autocorrelation plots revealed better mixing in implementations (i) and (iii) compared with (ii) and (iv). In the latter pair of implementations, the lag–1 autocorrelations in the trace plots for some R_{ij} were still in excess of 0.8. In contrast, virtually no autocorrelation remained in the trace plots for parameters updated using implementation (iii), whilst the small amount of autocorrelation remaining in trace plots for parameters updated using the constrain–$\boldsymbol{\beta}$ implementation, (i), died out quickly. It therefore seems that the approach of constraining $\boldsymbol{\beta}$

can lead to an MCMC algorithm with better mixing properties than some of the constrain-Σ suggestions from the literature. In particular, this includes those which are based on random walk Metropolis Hastings updates of the correlation matrix. It is worth noting that the mixing of the constrain-β implementation might have improved further if Σ had been assigned a semi-conjugate prior, so that the variance matrix could be updated via Gibbs, rather than Metropolis Hastings, steps.

6.2.4 An NHMM for rainfall occurrence

In Section 6.3.3, we present our final NHMM for rainfall occurrence and amounts. Effectively, this model is an extension of an NHMM for rainfall occurrence in which the MVP model describes the binary observables, conditional on the weather state. An outline of this foundational model for rainfall occurrence is provided in the remainder of this section.

There are no closed form expressions for the multi-dimensional normal integrals in the observed data likelihood for the MVP model. For simplicity, we therefore include the latent multivariate normal random vectors, $\mathbf{Z}_t = (Z_t^1, \ldots, Z_t^n)^T$, $t = 1, \ldots, T$, underlying the MVP model in the specification of the NHMM for rainfall occurrence. Continuing to use the notation from Chapters 4 and 5, let $\mathbf{D}_t = (D_t^1, \ldots, D_t^n)^T$, $S_t \in \mathcal{S}_r = \{1, \ldots, r\}$ and $\mathbf{X}_t$ denote the observable random vector of rainfall occurrence indicators, the hidden weather state and the observable atmospheric data, respectively, on day t where $t = 1, \ldots, T$. Also, denote the parameters of the NHMM by $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\text{hid}}, \boldsymbol{\theta}_{\text{obs}})$ where $\boldsymbol{\theta}_{\text{hid}}$ parameterises the weather state process and, although now associated with latent variables, $\boldsymbol{\theta}_{\text{obs}}$ denotes the parameters which do not directly relate to the weather state process.

The temporal structure of this NHMM is based on the following set of conditional independence assumptions:

A3. $\Pr(S_t \mid S_{0:t-1}, \mathbf{X}_{1:T}, \boldsymbol{\theta}) = \Pr(S_t \mid S_{t-1}, \mathbf{X}_t, \boldsymbol{\theta}_{\text{hid}})$
for $t = 1, \ldots, T$ with $\Pr(S_0 \mid \mathbf{X}_{1:T}, \boldsymbol{\theta}) = \Pr(S_0 = k \mid \boldsymbol{\theta}_{\text{hid}}) = \nu_k$.

A5. (a) $D_t^i = \mathbb{I}(Z_t^i > 0)$ for $i = 1, \ldots, n$ and $t = 1, \ldots, T$.

(b) $p(\mathbf{d}_t, \mathbf{z}_t \mid \mathbf{d}_{0:t-1}, \mathbf{z}_{1:t-1}, S_{0:T}, \mathbf{X}_{1:T}, \boldsymbol{\theta}) = p(\mathbf{z}_t \mid \mathbf{d}_{t-1}, S_t = k, \boldsymbol{\theta}_{\text{obs}})$
for $t = 1, \ldots, T$ with an initial model $\Pr(\mathbf{D}_0 \mid S_{0:T}, \mathbf{X}_{1:T}, \boldsymbol{\theta}) = \Pr(\mathbf{D}_0 \mid S_0, \boldsymbol{\theta}_{\text{obs}})$.

Note that assumption A3, concerning the hidden process, has been taken directly from Chapter 5. Within A5(b), the simplification

$$\begin{aligned} p(\mathbf{d}_t, \mathbf{z}_t \mid \mathbf{d}_{0:t-1}, \mathbf{z}_{1:t-1}, S_{0:T}, \mathbf{X}_{1:T}, \boldsymbol{\theta}) &= p(\mathbf{z}_t \mid \mathbf{d}_{t-1}, S_t = k, \boldsymbol{\theta}_{\text{obs}}) p(\mathbf{d}_t \mid \mathbf{z}_t) \\ &= p(\mathbf{z}_t \mid \mathbf{d}_{t-1}, S_t = k, \boldsymbol{\theta}_{\text{obs}}) \end{aligned}$$

arises due to assumption A5(a) which means that $p(\mathbf{d}_t \mid \mathbf{z}_t) = 1$ for any actually observed $\mathbf{d}_t$. The information in A3 and A5 can also be summarised by the DAG in Figure 6.1, where the double circles show that $\mathbf{D}_t$ depends deterministically on $\mathbf{Z}_t$.

The weather state process and initial rainfall occurrence distribution can be parameterised exactly as described in Sections 5.3.2 and 5.3.3, respectively. At times $t = 1, \ldots, T$, we suggest

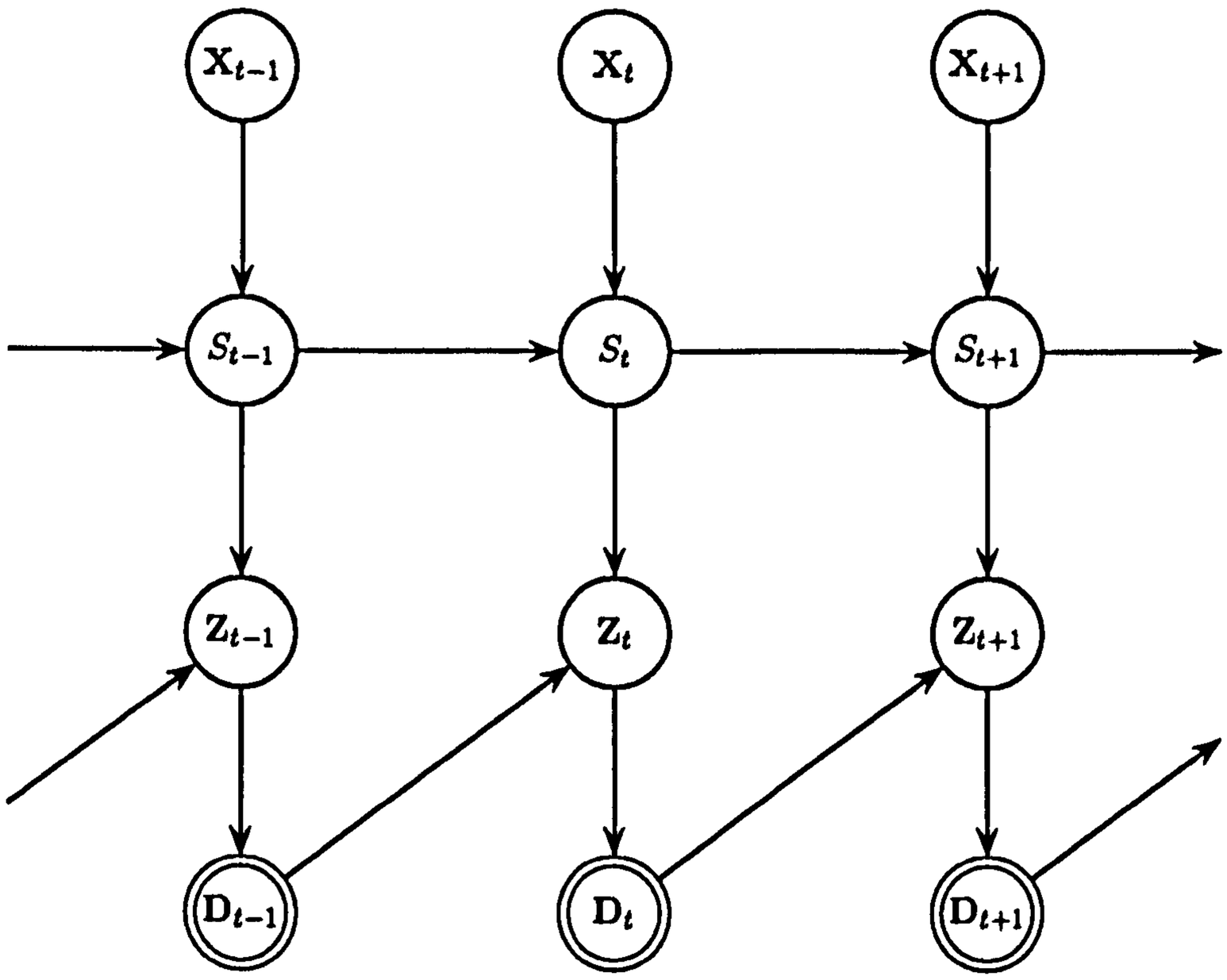


Figure 6.1: A DAG showing the temporal dependence structure in the NHMM described by assumptions A3 and A5.

parameterising the rainfall occurrence process (in A5(b)) using the MVP model with

$$\mathbf{Z}_t \mid \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, S_t = k, \theta_{\text{obs}} \sim \text{N}_n(\mathbf{X}_t \boldsymbol{\beta}_k, \boldsymbol{\Sigma}_k)$$

where $\boldsymbol{\beta}_k = (\boldsymbol{\beta}_{0k}^T, \boldsymbol{\beta}_{1k}^T)^T$ and $\mathbf{X}_t = (\mathbf{X}_{t0}, \mathbf{X}_{t1})$, with $\mathbf{X}_{t0} = \mathbf{I}_n$ and $\mathbf{X}_{t1} = \text{diag}(d_{t-1}^1, \ldots, d_{t-1}^n)$.

In Chapter 4 we found that assuming rainfall occurrences to be conditionally independent in time prevented the hidden Markov model from predicting long duration wet and dry spells as frequently as they occurred in the observed data. Therefore we do not make this assumption here. In the model presented above, each $\mathbf{Z}_t$ node has a single child, $\mathbf{D}_t$. Therefore, if we omitted the latent vectors $\{\mathbf{Z}_t : t = 1, \ldots, T\}$, then $\mathbf{D}_t$ would inherit the parents of $\mathbf{Z}_t$, namely S_t and $\mathbf{D}_{t-1}$. In other words, integrating out the latent vectors $\mathbf{Z}_t$ would lead to a model in which, conditionally on $\{S_t\}$, the rainfall occurrence indicators, $\{\mathbf{D}_t\}$, form a first order Markov chain.

A more sophisticated temporal structure would arise by making $\mathbf{Z}_{t-1}$ a parent of $\mathbf{Z}_t$ for each t. If the latent vectors were then integrated out, $\{\mathbf{D}_t\}$ would neither be independent nor first order Markov, conditionally on $\{S_t\}$. Thinking in terms of MCMC sampling, however, we choose to avoid this approach because we would expect it to be detrimental to the mixing of the chain. This is because, updating $(\mathbf{Z}_1, \ldots, \mathbf{Z}_T)$ one–at–a–time, the full conditional distribution of $\mathbf{Z}_t$ would depend on $\mathbf{Z}_{t-1}$ and $\mathbf{Z}_{t+1}$ in addition to S_t and $\mathbf{D}_t$, leading to high dependence amongst the large number of $\mathbf{Z}_t$ blocks. Standard forward–backward algorithms to update the $\mathbf{Z}_t$ in a single block would not be applicable here. This is because linear combinations of normal and truncated

normal random variables do not have standard distributions. Therefore in the (forward) filtering algorithm, storing enough information to summarise the (non-standard) filtered distributions would become increasingly impractical.

In this highly parameterised model, there may be some within-state parameters about which the data are not particularly informative. Therefore it is especially important that our prior conveys genuine initial beliefs. Using methods discussed in Germain *et al.* (2010*b*), which will be outlined in Section 6.4, we can specify a prior for the variance matrix that conveys substantive prior information. In contrast, as discussed in Section 6.2.3, it is difficult to express genuine beliefs in a prior for a correlation matrix. Therefore, we choose to handle the non-identifiability problem in the MVP model by introducing constraints on the parameters $\boldsymbol{\beta}_k$, $k \in \mathcal{S}_r$ (rather than $\boldsymbol{\Sigma}_k$, $k \in \mathcal{S}_r$) in the manner described in Section 6.2.3. Specifically, we restrict the coefficients, $\boldsymbol{\beta}_{1k}$, for the lag-1 rainfall occurrence indicators to be such that $\beta_{1k}^i \in \{-1, 1\}$ for each $i = 1, \ldots, n$ and each $k \in \mathcal{S}_r$. Although we would generally expect the probability of rain following rain to exceed the probability of rain following dry, this also allows the opposite to be true.

6.3 Jointly modelling rainfall occurrences and amounts

The data applications in Chapters 4 and 5 suggested the need to relax the assumptions of conditional spatial independence in both rainfall occurrences and non-zero amounts, given the weather state. In constructing a spatial model for rainfall, one of the main challenges is the mixed nature of rainfall distributions, combining a point mass at zero with a continuous, positively skewed density function on the positive real line. Clearly, rainfall data are highly non-normal. However, through the introduction of latent multivariate normal random variables, the Gaussian dependence structure can easily be used to build spatial dependence between both the discrete and continuous components of rainfall distributions.

Section 6.3.1 provides a general framework illustrating how this approach can be used to capture the spatial dependence in rainfall data. We show that many models from the literature are particular cases of this general structure and provide some criticism of each. Section 6.3.2 then outlines how temporal dependence can be built into these models. One possibility is through hidden Markov models and in Section 6.3.3 we describe an NHMM for rainfall in which spatial dependence within weather states is captured via latent multivariate normal variables.

6.3.1 Using latent normal variables to build spatial dependence

In this section we consider rainfall models whose spatial structure is generated through latent, or partially latent, multivariate normal random variables, which we denote by $\mathbf{Z}_0 = (Z_0^1, \ldots, Z_0^n)^T$ and $\mathbf{Z}_1 = (Z_1^1, \ldots, Z_1^n)^T$. We associate $\mathbf{Z}_0$ with rainfall occurrences and $\mathbf{Z}_1$ with non-zero rainfall amounts. Many models from the literature are based on the DAG in Figure 6.2 where $\mathbf{W} = (W^1, \ldots, W^n)^T$ is a vector of observed rainfall amounts at n sites. Occasionally the node $\mathbf{Z}_1$ will be omitted in which case the vector $\mathbf{Z}_0$ is responsible for modelling both rainfall occurrences and non-zero amounts. Note that the DAG does not include a vector of rainfall occurrence indicators, $\mathbf{D}$. This is because, in general, models of this type are not based on

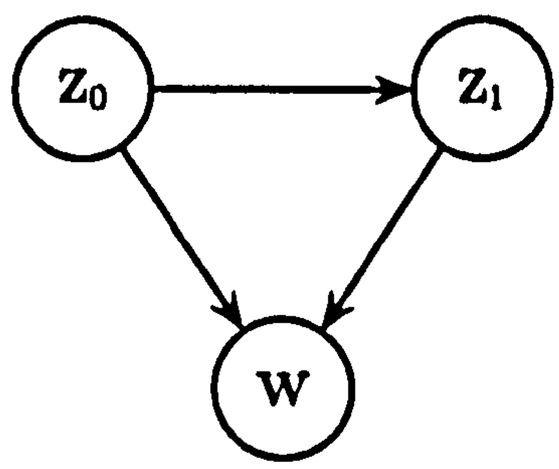


Figure 6.2: A DAG showing the factorisation of the joint density for $(\mathbf{W}, \mathbf{Z}_0, \mathbf{Z}_1)$ where $\mathbf{Z}_0$ and $\mathbf{Z}_1$ are latent multivariate normal random vectors and $\mathbf{W}$ denotes rainfall.

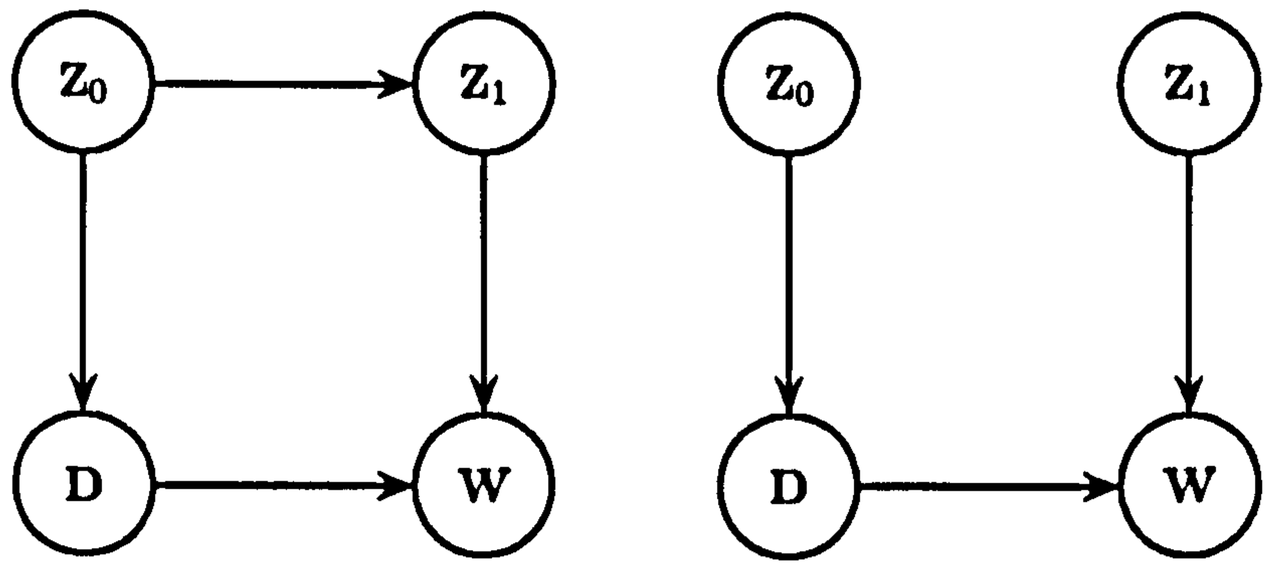


Figure 6.3: A DAG illustrating the two–stage model specification for rainfall amounts $\mathbf{W}$ and occurrences $\mathbf{D}$ when the latent variables $\mathbf{Z}_0$ and $\mathbf{Z}_1$ (a) are not and (b) are assumed to be independent.

the two stage approach of previous chapters which first provides a model for occurrences and then a model for amounts when rainfall occurs. However, the use of latent multivariate normal variables does not preclude analysis by the two–stage procedure, and we begin by considering this kind of model.

Under a two–stage model specification, it is convenient to introduce a vector of rainfall occurrence indicators, $\mathbf{D}$, and to consider a DAG of the form illustrated in Figure 6.3(a). This is a special case of Figure 6.2. Consider the hierarchical model for rainfall occurrence characterised by equations (6.1) and (6.2). Writing $\mathbf{Z}_0$ in place of $\mathbf{Z}$, a latent variable two–stage specification would arise by extending this occurrence model to include a generalized linear spatial process model for rainfall amounts when rain occurs. For example, we might assume that, given model parameters, say ϕ, and latent variables Z_1^i, the non–zero rainfall amounts W^i are conditionally independent gamma random variables

$$W^i \mid D^i = 1, Z_1^i, \phi, \gamma \overset{iid}{\sim} \text{Ga}\left(\frac{1}{\gamma}, \frac{1}{\gamma g_1^{-1}(\eta_1^i)}\right), \qquad i \in \{1, \ldots, n : D^i = 1\} \tag{6.6}$$

where γ is a dispersion parameter, $\eta_1^i = (\mathbf{x}_1^i)^T\phi + Z_1^i$ and $g_1(\cdot)$ is some suitable link function, for example, the log link. The Z_1^i would then be modelled as spatial random effects. We

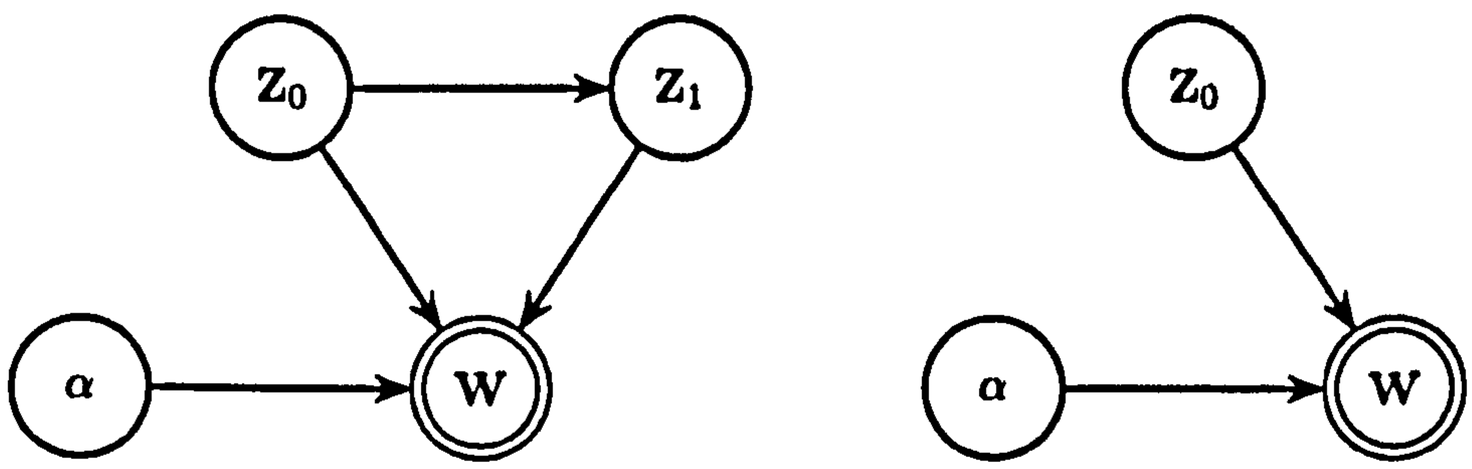


Figure 6.4: Modified versions of the DAG in Figure 6.2 in which rainfall $\mathbf{W}$ depends deterministically on some model parameters $\boldsymbol{\alpha}$ and (a) latent variables $\mathbf{Z}_0$ and $\mathbf{Z}_1$, or (b) the single latent variable $\mathbf{Z}_0$, having omitted the node $\mathbf{Z}_1$.

might expect the same physical factors to be responsible for the dependence between rainfall occurrences and between non–zero rainfall amounts. Therefore at the second stage specification we could define $\mathbf{Z}_1$ conditionally on $\mathbf{Z}_0$ or, more in the spirit of hierarchical spatial modelling, $\mathbf{Z}_1$ and $\mathbf{Z}_0$ could be conditionally independent at the second stage, but marginally correlated through lower level prior specifications.

Velarde *et al.* (2004) adopt a two–stage modelling approach for rainfall. The occurrence model was described in Section 6.2.1 and, in this case, the spatial random effects, Z_0^i, were modelled as CAR effects. When rainfall was observed, the amounts were modelled according to (6.6) with $\gamma = 1$ (giving the exponential distribution), a log link and CAR effects, Z_1^i, of the form (6.3). The CAR effects in the link function for the occurrence and amounts processes were assumed to be independent, so the DAG reduced to that in Figure 6.3(b).

In this two–stage approach, $\mathbf{D}$ and $\mathbf{W}$ depend stochastically on the latent variables $\mathbf{Z}_0$ and $\mathbf{Z}_1$. However, there is an alternative one–stage procedure in which the value of each rainfall amount, W^i, depends deterministically on $\mathbf{Z}_0$ and $\mathbf{Z}_1$ and possibly some unknown model parameters, say, $\boldsymbol{\alpha}$. The DAG corresponding to this kind of model is illustrated in Figure 6.4(a) which shows that, in effect, $\mathbf{W} = f(\mathbf{Z}_0, \mathbf{Z}_1, \boldsymbol{\alpha})$ for some known function f. For example, non–occurrence of rain at site i could correspond to Z_0^i falling below some threshold, whilst rainfall amounts on wet days might arise as some known transformation of Z_1^i and $\boldsymbol{\alpha}$. The generalized linear spatial process models discussed previously can often be computationally awkward because the latent variables typically have non–standard full conditional distributions. A benefit of this transformation–type approach is that the full conditional distributions for the latent variables tend to be either normal or truncated normal which is convenient in terms of MCMC sampling, especially when there are many latent variables, corresponding to many time points.

When the $\mathbf{Z}_1$ node is omitted, Figure 6.4(a) reduces to Figure 6.4(b). Bardossy & Plate (1992), Sansó & Guenni (1999), Sansó & Guenni (2000), Allcroft & Glasbey (2003) and Ailliot *et al.* (2009) all base their models for rainfall on the DAG in Figure 6.4(b), where $\boldsymbol{\alpha} = (\alpha_A, \boldsymbol{\alpha}_B^T)^T$ and W^i is related to Z_0^i through truncation and transformation,

$$w^i = \mathbb{I}(z_0^i > \alpha_A) g(z_0^i, \boldsymbol{\alpha}), \qquad i = 1, \ldots, n. \tag{6.7}$$

The function g is necessary to provide a distribution with heavier tails than the normal distribution in order to accommodate the possibility of (moderately) extreme rainfall. Allcroft & Glasbey (2003) use $\alpha_B = (\alpha_1, \alpha_2, \gamma)^T$ and assume that $\mathbf{Z}_0$ has a multivariate normal distribution with zero mean and unit marginal variances but unknown correlation matrix. The transformation g is defined through the inverse mapping

$$g^{-1}(w^i, \alpha) = \alpha_A + \alpha_1 (w^i)^\gamma + \alpha_2 (w^i)^{2\gamma} \tag{6.8}$$

which is found to be monotonic in the region of interest. In each of the other examples cited above, $\alpha_A = 0$, $\alpha_B = (\alpha^1, \ldots, \alpha^n)^T$, Z_0 is assumed to have unknown mean vector and unknown variance matrix and the transformation g is given by

$$g(z_0^i, \alpha) = (z_0^i)^{\alpha^i}, \tag{6.9}$$

often setting $\alpha^i = \alpha$ for all $i = 1, \ldots, n$. Here, the induced distribution for $\mathbf{W}$ is called a truncated, power transformed Gaussian distribution. An advantage of this type of model is its parsimonious description of spatial dependence in rainfall occurrences and amounts through a single latent variable. However, by dropping the node $\mathbf{Z}_1$ some flexibility in modelling non-zero rainfall amounts is lost. To illustrate, consider the univariate ($n = 1$) version of the truncated, power transformed Gaussian distribution defined by (6.7) and (6.9). Suppose that $Z_0 \mid \mu, \sigma^2 \sim \mathrm{N}(\mu, \sigma^2)$. Conditional on rainfall occurring, the density function for W is given up to proportionality as

$$p(w \mid w > 0, \mu, \sigma^2, \alpha) \propto w^{(1-\alpha)/\alpha} \exp\left\{-\frac{1}{2\sigma^2}(w^{1/\alpha} - \mu)^2\right\}, \qquad w > 0,$$

and the p–th quantile in its conditional distribution lies at

$$w_p = \sigma^\alpha \left[m + \Phi^{-1}\left\{1 - (1-p)\Phi(m)\right\}\right]^\alpha,$$

where $m = \mu/\sigma$. For the power transformation to produce a distribution with longer tails than the corresponding truncated normal distribution ($\alpha = 1$), it is therefore necessary to take $\alpha > 1$. In this case the conditional density has an asymptote at $w = 0$. Consequently, long-tailed density functions with modes greater than 0 are not possible and this lack of flexibility immediately provides a criticism of the truncated, power transformed Gaussian distribution as a model for rainfall. A more noteworthy criticism, however, is that the model does not allow independent changes in the probability of rain and the distribution of non–zero rainfall amounts. As an illustration, consider the distributions for W conditional on two sets of parameters, $(\mu_0, \sigma_0^2, \alpha_0)$ and $(\mu_1, \sigma_1^2, \alpha_1)$. Let $\mu_0 = 1.225$, $\sigma_0^2 = 1.5$ and $\alpha_0 = 2.5$ so that $m_0 = \mu_0/\sigma_0 = 1$ and the probability of rain given the first set of parameters is $\Phi(m_0) = 0.841$. The corresponding probability given the second set of parameters will be smaller if we choose $m_1 = \mu_1/\sigma_1$ to be less than m_0. Suppose we make this the case by taking $m_1 = -m_0 = -1$, that is, $\mu_1 = -\sigma_1$, which leads to a probability equal to $\Phi(m_1) = 0.159$. Now suppose that we want to choose the parameters $(\mu_1, \sigma_1^2, \alpha_1)$, where $\mu_1 = -\sigma_1$, such that the lower and upper quartiles in the conditional distribution of $(W \mid W > 0)$ are the same, given both sets of parameters. Solving two equations in two unknowns, this requires $\mu_1 = -4.141$, $\sigma_1^2 = 17.15$ and $\alpha_1 = 1.741$ leading to lower and upper quartiles of $w_{0.25} = 0.600$ and $w_{0.75} = 7.269$, respectively, in both conditional distributions. For each set of parameters, Figure 6.5 plots the survivor function for non–zero rainfall amounts, $\Pr(W > w \mid W > 0, \mu, \sigma^2, \alpha)$, on the log scale. It is clear that the tail in the

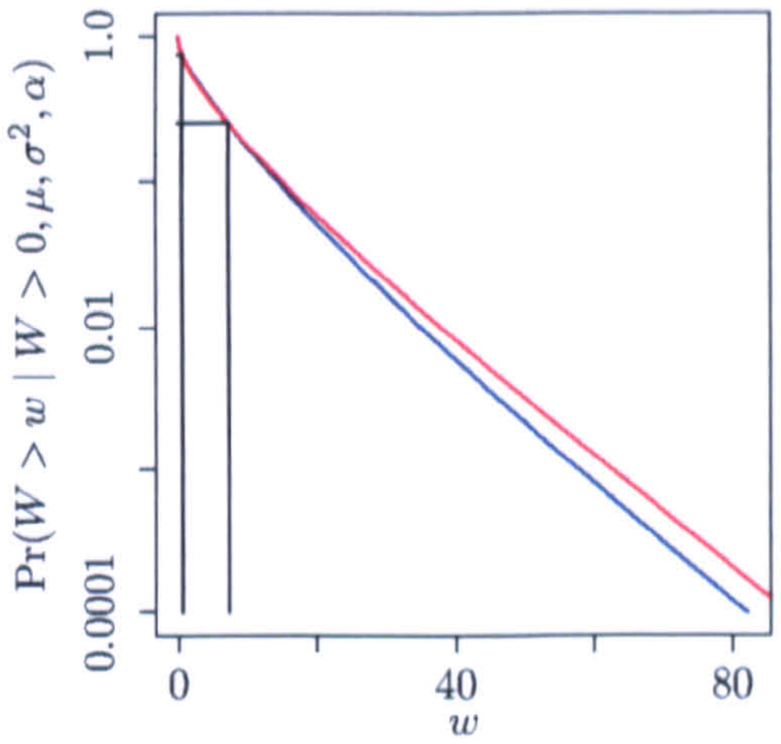


Figure 6.5: Survivor function for non–zero rainfall amounts, on the log scale, conditional on the first parameter set, $(\mu_0, \sigma_0^2, \alpha_0)$, (———) and on the second parameter set, $(\mu_1, \sigma_1^2, \alpha_1)$, (———). Also indicated are the matched lower and upper quartiles (———).

distribution conditional on the second set of parameters is heavier than that in the distribution conditional on the first set.

The introduction of a relationship between the probability of rain and the distribution of non–zero rainfall amounts can be avoided by using two latent variables, $\mathbf{Z}_0$ and $\mathbf{Z}_1$. Rappold *et al.* (2008) consider spatio–temporal models for wet mercury deposition which occurs when dissolved gaseous, aerosol or particulate mercury species is transferred to the Earth's surface by precipitation. Therefore without precipitation there can be no wet deposition. When wet mercury samples are available, the amount of mercury is related to the amount and type of precipitation, so many of the spatio–temporal patterns in deposition data can be explained by the spatio–temporal autocorrelation in rainfall. Regarding precipitation as a driver of deposition, Rappold *et al.* (2008) model the two processes jointly in a dynamic hierarchical structure, in which the model for rainfall alone could be represented through the DAG in Figure 6.4(a), with

$$w^i = \mathbb{I}(z_0^i > 0) \exp(z_1^i), \qquad i = 1, \ldots, n, \tag{6.10}$$

and the $\boldsymbol{\alpha}$ node omitted. Effectively this allows the parameter $\mathbf{Z}_0$ to control the probabilities of rain at each site, whilst $\mathbf{Z}_1$ models log rainfall amounts via a normal distribution. The mean and correlation matrix in the distribution for $\mathbf{Z}_0$ are assumed unknown, the variances being fixed at 1 to ensure identifiability. A multivariate linear regression then models $\mathbf{Z}_1$ conditionally on $\mathbf{Z}_0$. Precipitation is modelled in a similar way by Sahu *et al.* (2010) as part of a hierarchical model for chemical deposition. This model extends that of Rappold *et al.* (2008) by allowing (observed) point data and (simulated) areal data to be modelled jointly.

Although it would not naturally be represented by a DAG such as that in Figure 6.2, another model from the literature which essentially uses multivariate normal random variables to build spatial dependence is the regional weather state model proposed by Thompson *et al.* (2007). This is based on the marginal specification of a separate (three state) partially hidden Markov model at each site, in which non–zero rainfall amounts have an exponential distribution with state and site specific parameters. The term "partially" refers to the fact that the "dry" weather state

at each site is observable, corresponding to days with no precipitation. Latent multivariate normal random variables are then used to combine the marginal specifications and produce a joint model with spatial dependence structure. For the non-zero rainfall amounts, this is achieved by introducing a Gaussian copula function (see, for example, Nelsen, 2006) to export the Gaussian dependence structure to the exponential marginals. On day t, spatial dependence between the site–specific weather states (and hence rainfall occurrences) is modelled through an unobserved multivariate normal random variable, $\mathbf{Z}_{0t} = (Z_{0t}^1, \ldots, Z_{0t}^n)^T$ whose i–th component, Z_{0t}^i, determines the weather state at site i by a set of threshold specifications. The thresholds depend on the site and the weather state at that site on the previous day. Therefore although the partially observed Markov chain has 3^n states, their joint probability mass function makes some combinations very unlikely. Note that if the two (unobserved) non–dry weather states at each site were combined to produce a single (observable) wet state, this would be equivalent to modelling rainfall occurrences using a multivariate probit model with mean dependent on the vector of rainfall occurrences the previous day.

The main criticism of this regional weather state model relates to its use of copula functions. Although copulas are a very flexible way of building multivariate distributions with non–normal marginal distributions, they typically produce complicated likelihood functions. In Bayesian analyses this leads to non–standard full conditional distributions which can be computationally inconvenient.

6.3.2 Incorporating temporal dependence

We have described how the dependence structure of latent, or partially latent, multivariate normal random variables can be used to generate spatial autocorrelation in multi–site rainfall distributions. Time series of rainfall data also exhibit temporal dependence which can be modelled in a variety of ways. If models use both the latent vectors $\mathbf{Z}_0$ and $\mathbf{Z}_1$, a parsimonious spatio–temporal model might describe temporal dependence through $\{\mathbf{Z}_{0t}\}$ whilst assuming $\{\mathbf{Z}_{1t}\}$ to be conditionally independent in time given $\{\mathbf{Z}_{0t}\}$. The reasoning behind this is that the same physical processes which create temporal autocorrelation between rainfall occurrences, modelled using $\mathbf{Z}_{0t}$, will also be responsible for temporal autocorrelation between non–zero rainfall amounts, modelled using $\mathbf{Z}_{1t}$.

Allcroft & Glasbey (2003) use the truncated and transformed multivariate normal distribution summarised by the DAG in Figure 6.4(b) and equations (6.7)–(6.8) to model hourly rainfall data on a two–dimensional lattice. Denoting by $Z_{0,ijt}$ the normal random variable associated with the observed rainfall W_{ijt} at the (i,j)–th spatial location at time t, the partially latent normal variables $\{Z_{0,ijt}\}$ are modelled using a Gaussian Markov random field (GMRF) with a particular spatio–temporal neighbourhood structure. This means that $Z_{0,ijt}$ and $Z_{0,k\ell s}$ are assumed to be conditionally independent unless (i,j,t) and (k,ℓ,s) are neighbours; see Rue & Held (2005) for a thorough examination of GMRFs. Allcroft & Glasbey (2003) were interested in using Gibbs sampling to disaggregate coarse resolution observed rainfall, that is, generating data at a finer spatial resolution, conditional on the coarse resolution observations. The conditional independence assumptions underpinning GMRFs makes them particularly amenable to Gibbs sampling and this was the motivation for modelling the transformed rainfall data using GMRFs. However, for forecasting, time series models are typically more convenient.

A variety of time series models are available for multivariate normal data and can be applied to the latent vectors $\{\mathbf{Z}_{0t}\}$. Bardossy & Plate (1992) use the truncated and transformed multivariate normal distribution summarised by the DAG in Figure 6.4(b) and equations (6.7) and (6.9) with $\alpha_A = 0$. The rainfall distribution is specified conditionally on an observed weather state and temporal dependence is modelled by supposing the multivariate normal random vector, $\mathbf{Z}_{0t}$, to follow a vector autoregressive VAR(1) process, in the case of a persisting weather state. On days when the weather state differs from that on the previous day, $\mathbf{Z}_{0t}$ is reinitialised at the stationary distribution of the particular VAR(1) model which corresponds to the current weather state. This resetting of the likelihood each time the weather state changes significantly complicates the likelihood function. Rappold *et al.* (2008) jointly modelled precipitation and mercury deposition using latent multivariate normal random vectors in which the model for precipitation was based on the DAG in Figure 6.4(a) and equation (6.10). Again, $\mathbf{Z}_{0t}$ is modelled as a mean centred VAR(1) process, with the simplifying assumption that the $n \times n$ autoregressive coefficient matrix is diagonal, with a common diagonal element, $\phi \in (0, 1)$, which is constant over time. Given $\mathbf{Z}_{0t}$, the process for $\mathbf{Z}_{1t}$ is assumed to be conditionally independent across time. Then $\mathbf{Z}_{1t}$ is related to $\mathbf{Z}_{0t}$, $t = 1, \ldots, T$, by a multivariate normal linear regression in which the regression coefficients and conditional variances are time–dependent. The correlations between the error terms in both the marginal $\mathbf{Z}_{0t}$ process and the conditional $(\mathbf{Z}_{0t} \mid \mathbf{Z}_{1t})$ process are modelled using exponential correlation functions which do not vary over time. Viewing $\mathbf{Z}_{0t}$ as the "state" and $\mathbf{Z}_{1t}$ as the "data", Rappold *et al.* (2008) regard their model for $(\mathbf{Z}_{0t}, \mathbf{Z}_{1t})$ as a spatio–temporal dynamic linear model.

In both of the above examples, first order Markovian dependence is assumed for the $\{\mathbf{Z}_{0t}\}$ process. A simple way of inducing non–Markovian dependence in a time series model for $\{\mathbf{Z}_{0t}\}$ is to assume that $\mathbf{Z}_{0t}$ are conditionally independent given another latent process, say $\{\theta_t\}$, which forms a first order Markov chain. Then, the $\{\theta_t\}$ can either be continuous random vectors, suggesting a dynamic linear model formulation, or discrete random variables, suggesting a hidden Markov model formulation. This is precisely the approach adopted by Sansó & Guenni (2000) and Ailliot *et al.* (2009) who use the truncated and transformed multivariate normal model to relate the observed rainfall, $\mathbf{W}_t$, to the single latent vector $\mathbf{Z}_{0t}$, $t = 1, \ldots, T$. Ailliot *et al.* (2009) use a hidden Markov model, in which the parameters of the truncated and transformed multivariate normal distribution depend on the hidden state. Although this provides a simple stochastic framework for modelling the temporal persistence in rainfall, Chapter 4 showed that vesting all the dynamic structure in the hidden states sometimes fails to account for the strong autocorrelation in time series of rainfall occurrences. Further comments on this model will be provided in Section 6.9. Sansó & Guenni (2000) adopt a dynamic linear model for $\{\mathbf{Z}_{0t}\}$, set up in such a way as to capture spatial structure and seasonal trends. Compared to the (stationary) hidden Markov model of Ailliot *et al.* (2009), an advantage of the dynamic linear model approach is the easy incorporation of (time–varying) seasonal trends.

6.3.3 An NHMM for rainfall occurrence and amount

The NHMM for rainfall which we consider in this chapter only differs from the model of Chapter 5 in its parameterisation of the precipitation process. We therefore continue to assume that the dynamics of the weather state process, $\{S_t\}$, conditional on the observed atmospheric data,

$\{\mathbf{X}_t\}$, can be summarised through assumption A3 of Section 6.2.4 which, in turn, was carried over from Chapter 5. The available atmospheric data comprise Lamb weather types, $\mathbf{X}_t = X_t \in \mathcal{Q} = \{1, \ldots, 27\}$, and we continue to parameterise the weather state process according to the model described in Section 5.3.2.

Introducing the latent multivariate normal random variables, $(\mathbf{Z}_{0t}, \mathbf{Z}_{1t})$, for $t = 1, \ldots, T$, we build our conditional model for precipitation, given the weather state, by defining the observed precipitation at site i on day t as

$$W_t^i = \mathbb{I}(Z_{0t}^i > 0) \exp(Z_{1t}^i) \tag{6.11}$$

so that W_t^i will be equal to $\exp(Z_{1t}^i)$ if $Z_{0t}^i > 0$ and equal to 0 otherwise. The model, outlined below, describes the joint distribution for $\{(\mathbf{W}_t, \mathbf{Z}_{0t}, \mathbf{Z}_{1t})\}$ conditional on the weather states, $\{S_t\}$. Notationally, it will also be convenient to introduce the rainfall occurrence indicators, implicitly defined as $D_t^i = \mathbb{I}(Z_{0t}^i > 0)$.

In standard hidden Markov models, all of the temporal dynamics are vested in the hidden states which, in this case, would correspond to an assumption that the bivariate latent process $(\mathbf{Z}_{0t}, \mathbf{Z}_{1t})$ is conditionally independent across time, t, given the weather state. However, in Chapter 4 we found that a model with this temporal structure predicted wet and dry spells which tended to be of shorter duration than those observed. Conditional on the weather state, $S_t = k$, we suppose that $\mathbf{Z}_{0t}$ also depends on $\{\mathbf{Z}_{0s} : s = 1, \ldots, t-1\}$ through the signs of $(Z_{0,t-1}^1, \ldots, Z_{0,t-1}^n)$, that is, through the rainfall occurrence indicator $\mathbf{D}_{t-1}$ (since $D_{t-1}^i = \mathbb{I}(Z_{0,t-1}^i > 0)$). Again, let $\theta = (\theta_{\text{hid}}, \theta_{\text{obs}})$ comprise the parameters of the NHMM where θ_{hid} parameterises the weather state process and θ_{obs} denotes the parameters which do not relate directly to the weather state process. We specify

$$\mathbf{Z}_{0t} \mid \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, S_t = k, \theta_{\text{obs}} \sim \mathrm{N}_n(\mathbf{X}_t \beta_k, \Sigma_k), \tag{6.12}$$

where Σ_k is an $n \times n$ symmetric positive definite matrix, $\beta_k = (\beta_{0k}^T, \beta_{1k}^T)^T$ in which β_{0k} and β_{1k} are n–vectors with real entries. Further, $\mathbf{X}_t = (\mathbf{X}_{t0}, \mathbf{X}_{t1})$ where $\mathbf{X}_{t0} = \mathbf{I}_n$ and $\mathbf{X}_{t1} = \mathrm{diag}(d_{t-1}^1, \ldots, d_{t-1}^n)$. Next, given $\{\mathbf{Z}_{0t}\}$ and $\{S_t\}$, we specify the (partially) latent amounts process, $\{\mathbf{Z}_{1t}\}$, to be conditionally independent across time with

$$\mathbf{Z}_{1t} \mid \mathbf{Z}_{0t}, S_t = k, \theta_{\text{obs}} \sim \mathrm{N}_n(\mu_k + \gamma_k \mathbf{Z}_{0t}, \Omega_k). \tag{6.13}$$

Here, μ_k is an n–vector and γ_k is an $n \times n$ matrix, both with real entries, and Ω_k is a $n \times n$ symmetric positive definite matrix.

Overall, the temporal structure of the NHMM can be summarised by combining assumption A3 from Section 6.2.4 with

A6. (a) $W_t^i = \mathbb{I}(Z_{0t}^i > 0) \exp(Z_{1t}^i)$ (which means $D_t^i = \mathbb{I}(Z_{0t}^i > 0)$) for $i = 1, \ldots, n$ and $t = 1, \ldots, T$.

(b) $p(\mathbf{w}_t, \mathbf{z}_{0t}, \mathbf{z}_{1t} \mid \mathbf{w}_{1:t-1}, \mathbf{z}_{0,1:t-1}, \mathbf{z}_{1,1:t-1}, D_0, S_{0:T}, \mathbf{X}_{1:T}, \theta)$
$= p(\mathbf{z}_{0t} \mid \mathbf{d}_{t-1}, S_t = k, \theta_{\text{obs}}) p(\mathbf{z}_{1t} \mid \mathbf{z}_{0t}, S_t = k, \theta_{\text{obs}})$
for $t = 1, \ldots, T$ with an initial model $\Pr(\mathbf{D}_0 \mid S_{0:T}, \mathbf{X}_{1:T}, \theta) = \Pr(\mathbf{D}_0 \mid S_0, \theta_{\text{obs}})$.

Here, the conditional densities of $(\mathbf{Z}_{0t} \mid \mathbf{D}_{t-1}, S_t)$ and $(\mathbf{Z}_{1t} \mid \mathbf{Z}_{0t}, S_t)$ are parameterised according to equations (6.12) and (6.13), respectively. The simplification in A6(b),

$$\begin{aligned} &p(\mathbf{w}_t, \mathbf{z}_{0t}, \mathbf{z}_{1t} \mid \mathbf{w}_{1:t-1}, \mathbf{z}_{0,1:t-1}, \mathbf{z}_{1,1:t-1}, D_0, S_{0:T}, \mathbf{X}_{1:T}, \theta) \\ &\qquad = p(\mathbf{z}_{0t} \mid \mathbf{w}_{t-1}, S_t = k, \theta_{\text{obs}}) p(\mathbf{z}_{1t} \mid \mathbf{z}_{0t}, S_t = k, \theta_{\text{obs}}) p(\mathbf{w}_t \mid \mathbf{z}_{0t}, \mathbf{z}_{1t}) \\ &\qquad = p(\mathbf{z}_{0t} \mid \mathbf{d}_{t-1}, S_t = k, \theta_{\text{obs}}) p(\mathbf{z}_{1t} \mid \mathbf{z}_{0t}, S_t = k, \theta_{\text{obs}}) \end{aligned}$$

arises because assumption A6(a) means that $p(\mathbf{w}_t \mid \mathbf{z}_{0t}, \mathbf{z}_{1t}) = 1$ for any actually observed $\mathbf{w}_t$.

The set of conditional independence assumptions A3 and A6 can be represented graphically through the DAG in Figure 6.6, where the nodes $\{\mathbf{D}_t\}$ are included for clarity. If we were only modelling rainfall occurrences and therefore omitted the latent vectors $\{\mathbf{Z}_{1t}\}$, the model would reduce exactly to the NHMM for rainfall occurrence described in Section 6.2.4. We therefore regard our model as an extension of this combined NHMM and MVP model, which additionally incorporates non-zero rainfall amounts. Note that introducing the latent vector $\mathbf{Z}_{1t}$ does not remove the non-identifiability problem from the underlying MVP model because changes in the scale of $\mathbf{Z}_{0t}$ could be exactly compensated for by changes in the scale of γ_k. Therefore to ensure identifiability, we continue to constrain the coefficients β_{1k} such that $\beta_{1k} \in \{-1, 1\}^n$ for all $k \in \mathcal{S}_r$.

As explained in Section 6.2.4, assuming $\mathbf{Z}_{0t}$ to follow a VAR(1) process, conditionally on the weather state, may significantly impair the performance of the MCMC sampler. This is the motivation for the simpler temporal model for $\mathbf{Z}_{0t}$, summarised through A6. Integrating out the latent vectors $\{(\mathbf{Z}_{0t}, \mathbf{Z}_{1t})\}$ then produces a non-homogeneous Markov switching model, that is, an NHMM in which $\{(\mathbf{W}_t, \mathbf{D}_t)\}$ forms a Markov chain conditionally on $\{S_t\}$. The model studied in Chapter 5 also belonged to this general class although the decision to make $(W_t^i \mid D_t^i = 1, S_t)$ conditionally independent of $\mathbf{D}_{t-1}$ in this earlier model gave it a simpler temporal structure. This observation notwithstanding, the two NHMMs differ fundamentally in the way in which spatial dependence is modelled between non-zero rainfall amounts. The model from Chapter 5 assumed conditional independence across sites, given the weather state, whilst the model presented in this section allows the log rainfall amounts to be correlated within weather states.

We complete our model specification by choosing a distribution for the unobserved rainfall occurrence indicator at time $t = 0$, $\mathbf{D}_0 = (D_0^1, \ldots, D_0^n)^T$. For consistency, we give $\mathbf{D}_0$ the same distribution that was adopted in Chapter 5, that is, $D_0^1, \ldots, D_0^n$ are assumed to be independent of each other and of S_0 with

$$(D_0^i \mid S_0) \equiv D_0^i \sim \text{Bern}(p_0^i), \qquad \text{independently for } i = 1, \ldots, n \tag{6.14}$$

where each $p_0^i \in [0, 1]$ is fixed.

6.3.3.1 A simplification to the spatial structure

The NHMM presented above contains $rn(n+1)/2$ variance and covariance parameters in $(\Sigma_1, \ldots, \Sigma_r)$ and the same number in $(\Omega_1, \ldots, \Omega_r)$. For models with more than one state, it is likely that many of these parameters will be only weakly identified in the likelihood. Marginalising over $\mathbf{Z}_{0t}$ in the joint distribution for $(\mathbf{Z}_{0t}, \mathbf{Z}_{1t} \mid \mathbf{D}_{t-1}, S_t)$ gives

$$\mathbf{Z}_{1t} \mid \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, S_t = k, \theta_{\text{obs}} \sim \text{N}_n(\mu_k + \gamma_k \mathbf{X}_t \beta_k, \; \Omega_k + \gamma_k \Sigma_k \gamma_k^T). \tag{6.15}$$

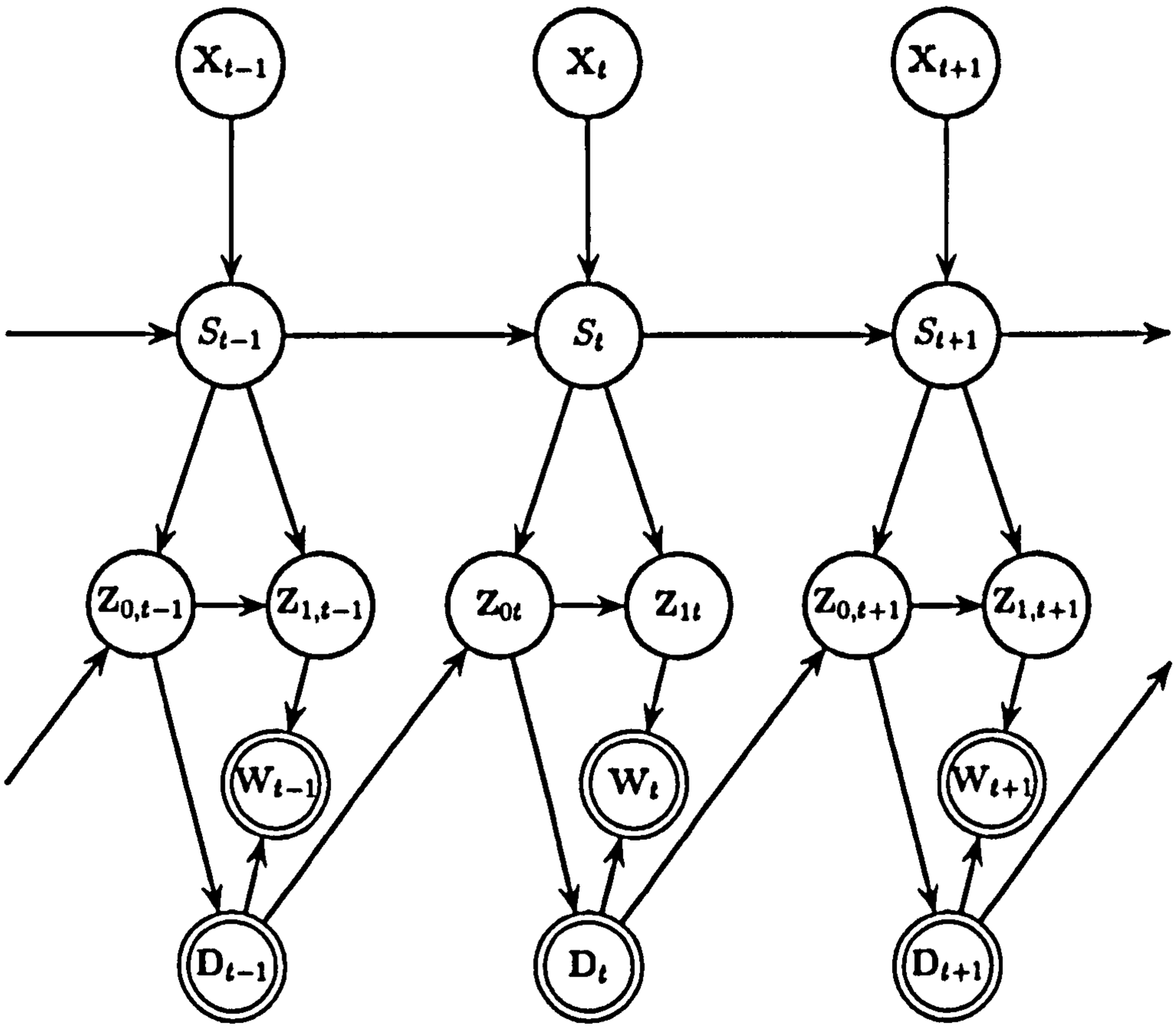


Figure 6.6: A DAG showing the temporal dependence structure in the NHMM described by assumptions A3 and A6.

Therefore, assuming the variance matrices $(\Omega_1, \ldots, \Omega_r)$ to be diagonal would result in a more parsimonious model, in which the components of $\mathbf{Z}_{1t}$ are still correlated when $\mathbf{Z}_{0t}$ is integrated out. Allowing γ_k to be non–diagonal creates a flexible dependence structure for $\mathbf{Z}_{1t}$.

We only observe Z^i_{1t} if there is rain on day t at site i and, in this case, $Z^i_{1t} = \log W^i_t$. However, MCMC becomes much more straightforward if we additionally sample the latent Z^i_{1t} because the full conditional distributions of the model parameters in $\{(\mu_k, \gamma_k, \Omega_k) : k \in S_r\}$ and the latent variables in $\{\mathbf{Z}_{0t} : t = 1, \ldots, T\}$ have much simpler forms when we work with the complete data. An exception is when $Z^1_{1t}, \ldots, Z^n_{1t}$ are conditionally independent, given $\mathbf{Z}_{0t}$ and S_t. In this case the latent Z^i_{1t} are terminal nodes and can be trivially integrated out of the model. It follows that we can avoid sampling the unobserved Z^i_{1t} by assuming each matrix in $(\Omega_1, \ldots, \Omega_r)$ to be diagonal. The resulting model then affords a more straightforward computational analysis, whilst still offering a flexible spatial dependence structure for $\mathbf{Z}_{1t}$. If the latent Z^i_{1t} are integrated out of the model, the only Z^i_{1t} which remain are those corresponding to wet days on which $Z^i_{1t} = \log W^i_t$. Therefore, there is no need to distinguish between the W^i_t and the remaining Z^i_{1t}. Within weather states, this suggests a return to the two–stage model specification from Chapters 4 and 5, where we first modelled rainfall occurrences and then rainfall amounts, conditionally on occurrence. Applying this idea here, we model rainfall occurrences through equation (6.12), which uses the latent Gaussian variables $\{\mathbf{Z}_{0t}\}$. Then we model rainfall amounts, conditionally

on occurrence *and* these latent Gaussian variables, through

$$p(\mathbf{w}_t \mid \mathbf{d}_t, s_t, \mathbf{z}_{0t}, \theta_{\text{obs}}) = \prod_{i=1}^{n} p(w_t^i \mid d_t^i, s_t, \mathbf{z}_{0t}, \theta_{\text{obs}}) \tag{6.16}$$

where

$$\Pr(W_t^i = 0 \mid D_t^i = 0) = 1, \quad W_t^i \mid D_t^i = 1, S_t = k, \mathbf{Z}_{0t} \sim \text{LogN}\left(\mu_k^i + \sum_{j=1}^{n} \gamma_k^{ij} z_{0t}^j, \Omega_k^{ii}\right). \tag{6.17}$$

The DAG corresponding to this formulation of the model is shown in Figure 6.7(a). Correspondingly, the set of conditional independence assumptions defining the temporal structure of the NHMM can be summarised through assumption A3 from Section 6.2.4 and

A7. (a) $D_t^i = \mathbb{I}(Z_{0t}^i > 0)$ for $i = 1, \ldots, n$ and $t = 1, \ldots, T$.

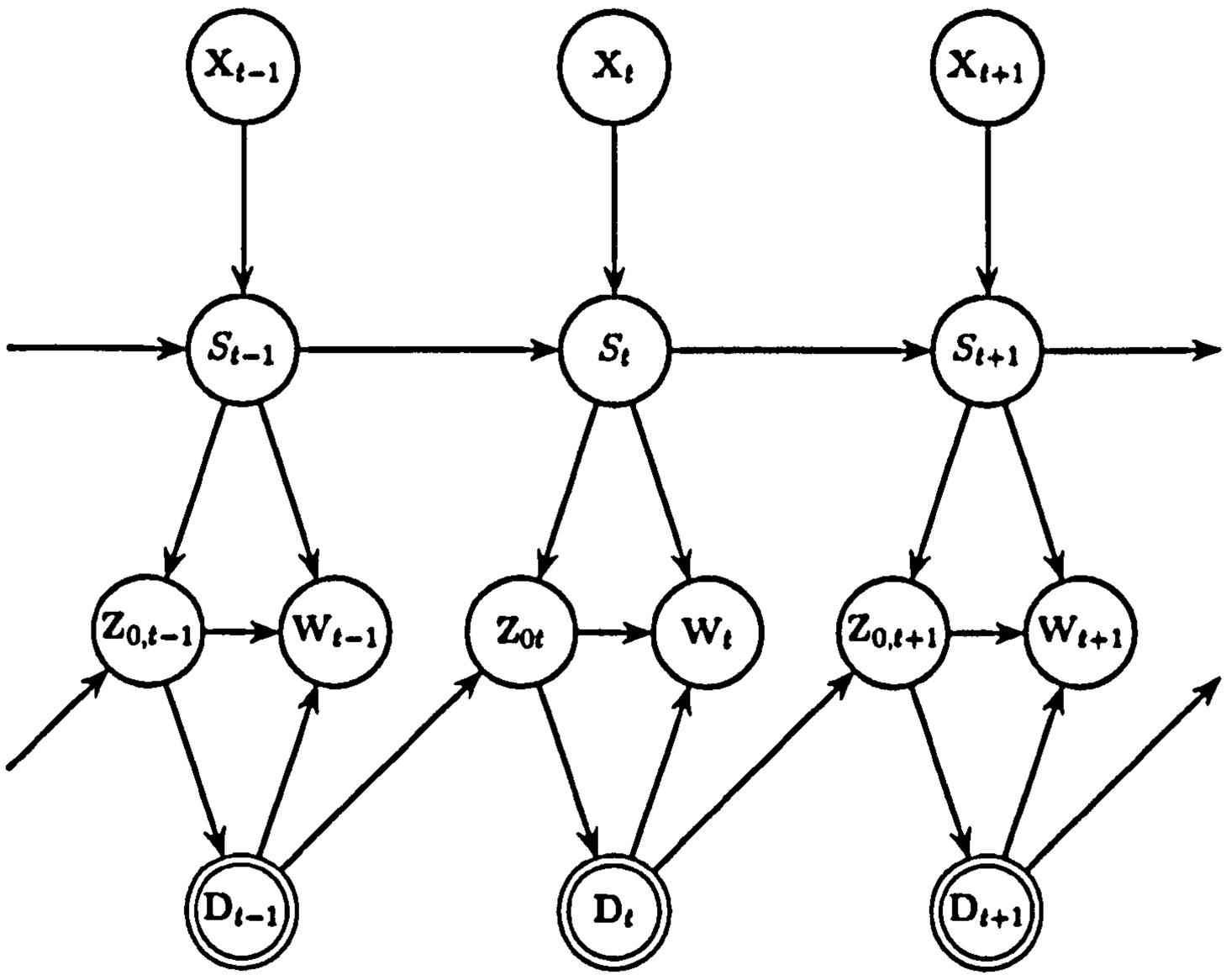
(b) $p(\mathbf{w}_t, \mathbf{d}_t, \mathbf{z}_{0t} \mid \mathbf{w}_{1:t-1}, \mathbf{d}_{0:t-1}, \mathbf{z}_{0,1:t-1}, S_{0:T}, \mathbf{X}_{1:T}, \theta)$
$= p(\mathbf{w}_t \mid \mathbf{d}_t, s_t, \mathbf{z}_{0t}, \theta_{\text{obs}}) p(\mathbf{z}_{0t} \mid \mathbf{d}_{t-1}, s_t, \theta_{\text{obs}})$
for $t = 1, \ldots, T$ with an initial model $\Pr(\mathbf{D}_0 \mid S_{0:T}, \mathbf{X}_{1:T}, \theta) = \Pr(\mathbf{D}_0 \mid S_0, \theta_{\text{obs}})$.

in place of assumption A6 outlined earlier in this section. Here the conditional densities of $(\mathbf{Z}_{0t} \mid \mathbf{D}_{t-1}, S_t)$ and $(\mathbf{W}_t \mid \mathbf{D}_t, \mathbf{Z}_{0t}, S_t)$ follow from equations (6.12) and (6.16)–(6.17), respectively. The initial rainfall occurrence distribution, $\Pr(\mathbf{D}_0 \mid S_0, \theta_{\text{obs}})$, was specified in equation (6.14). Figure 6.7(b) shows the factorisation of the joint distribution for $\{(\mathbf{W}_t, \mathbf{D}_t, S_t)\}$ that arises after marginalising over $\{\mathbf{Z}_{0t}\}$. Comparing this with the DAG for the Chapter 5 model (see Figure 5.2), it is clear that $\mathbf{D}_{t-1}$ is now a parent of $\mathbf{W}_t$ as well as $\mathbf{D}_t$. This illustrates our earlier observation that the model presented here offers a more sophisticated spatial *and* temporal structure for non–zero rainfall amounts.

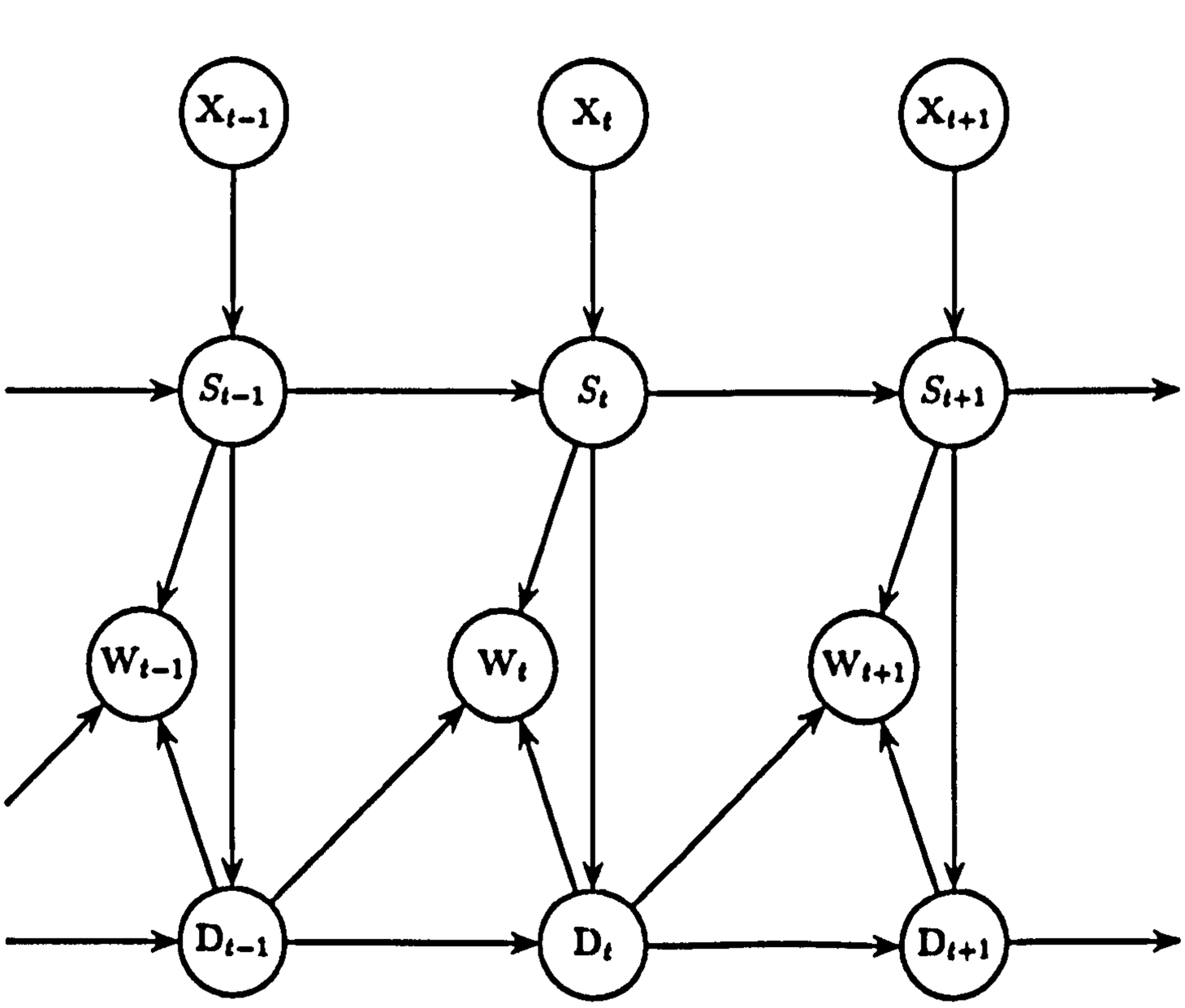
The conditional model defined through assumption A7 shares something in common with the truncated, power transformed Gaussian distribution in that it only contains a single time series of latent multivariate normal variables, $\{\mathbf{Z}_{0t}\}$. However, in contrast to the latter model, it does not prevent independent changes in the probability of rain and the distribution of non–zero rainfall amounts. This is because the model outlined above does not assume the non–zero rainfall amounts to be deterministically dependent on the latent variables.

Section 6.3.1 discussed the use of generalized linear spatial process models as a means of capturing spatial dependence between non–zero rainfall amounts. With $\mathbf{Z}_{0t}$ playing the role of the vector of spatial random effects, the model defined through equations (6.16)–(6.17) is similar to a generalized linear spatial process model with normally distributed observables (log non–zero rainfall) and an identity link. An alternative NHMM might be based on A3 and A7, but could model non–zero rainfall amounts (or their logs) using a different distribution from the class of exponential family models. The mean in this distribution may then be linked to the latent process, $\mathbf{Z}_{0t}$, using a different link. For example, the gamma distribution could be used, together with a log link.

For the remainder of this chapter we choose to represent the model through the DAG in Figure 6.7(a), that is, the characterisation summarised by assumptions A3 and A7.



(a)



(b)

Figure 6.7: DAGs showing the temporal dependence structure in the NHMM described by assumptions A3 and A7 (a) before and (b) after omitting the $\mathbf{Z}_{0t}$ nodes.

6.4 Prior distribution

In this section, we outline the prior distribution chosen for the model parameters, θ. Many of the components of θ are assigned priors which encourage borrowing of strength between sites or, for some of the parameters of higher dimension, between states.

The set of all model parameters is denoted by $\theta = (\theta_{\text{hid}}, \theta_{\text{obs}})$, where $\theta_{\text{obs}} = (\theta_{\text{obs},1}, \ldots, \theta_{\text{obs},r})$ and

$$\theta_{\text{obs},k} = (\beta_{0k}, \beta_{1k}, \mu_k, \gamma_k, \Sigma_k, \Omega_k) \in \mathbb{R}^n \times \{-1, 1\}^n \times \mathbb{R}^n \times \mathbb{R}^{n^2} \times \mathcal{D}^n \times \mathbb{R}_+^n$$

for each $k \in \mathcal{S}_r$. Above, $\mathcal{D}^n$ denotes the set of all $n \times n$ symmetric, positive definite matrices. The parameters of the weather state process are

$$\theta_{\text{hid}} = (\mathbf{A}, \nu, \mathcal{E}) \in \mathscr{S}_r^{27r} \times \mathscr{S}_r \times \mathscr{S}_r^r$$

where $\mathbf{A} = (\mathbf{A}_1^1, \ldots, \mathbf{A}_1^{27}, \ldots, \mathbf{A}_r^1, \ldots, \mathbf{A}_r^{27})$ in which $\mathbf{A}_j^x = (A_{j1}^x, \ldots, A_{jr}^x)$ and $A_{jk}^x = \Pr(S_t = k \mid S_{t-1} = j, X_t = x, \theta_{\text{hid}})$; $\nu = (\nu_1, \ldots, \nu_r)$ where $\nu_k = \Pr(S_0 = k \mid \theta_{\text{hid}})$; and $\mathcal{E} = (\xi_1, \ldots, \xi_r)$ in which $\xi_j = (\xi_{j1}, \ldots, \xi_{jr})$ is the variable mean in the hierarchical prior for $(\mathbf{A}_j^1, \ldots, \mathbf{A}_j^{27})$.

Uncertainty about the model parameters, *a priori*, is expressed through a prior of the form

$$\begin{aligned}\pi(\theta) &= \pi(\theta_{\text{obs}}) \times \pi(\theta_{\text{hid}}) \\ &= \{\pi(\beta_{01}, \ldots, \beta_{0r})\pi(\beta_{11}, \ldots, \beta_{1r})\pi(\mu_1, \ldots, \mu_r)\pi(\gamma_1, \ldots, \gamma_r)\pi(\Sigma_1, \ldots, \Sigma_r) \\ &\quad \times \pi(\Omega_1, \ldots, \Omega_r)\} \times \{\pi(\mathbf{A} \mid \mathcal{E})\pi(\mathcal{E})\pi(\nu)\} \qquad (6.18)\end{aligned}$$

which is assumed to be exchangeable across weather states. Note that the parameters of the precipitation and weather state processes, θ_{obs} and θ_{hid}, respectively, are assumed to be independent *a priori*.

The parameters of the weather state process are given an identical prior specification to that outlined in Section 5.4 (see equations (5.13) and (5.20)) and we refer to Section 5.4.2 for details regarding the elicitation of this prior.

For the latent process $\{(\mathbf{Z}_{0t} \mid S_t, \mathbf{D}_{t-1})\}$, the coefficients $\beta_{01}, \ldots, \beta_{0r}$ and $\beta_{11}, \ldots, \beta_{1r}$ are assumed to be independent across weather states, *a priori*, and similarly for the parameters $\mu_1, \ldots, \mu_r$ arising from the process $\{(\mathbf{W}_t \mid \mathbf{D}_t, S_t, \mathbf{Z}_{0t})\}$. These parameters appear in the conditional means $\mathrm{E}(\mathbf{Z}_{0t} \mid \mathbf{D}_{t-1}, S_t, \theta_{\text{obs}})$ and $\mathrm{E}(\log W_t^i \mid D_t^i = 1, S_t, \mathbf{Z}_{0t}, \theta_{\text{obs}})$ and so influence first order properties of rainfall, such as the probability of rain and the mean rainfall amount on wet days. We would expect these properties to show different patterns within different weather states and so the independence assumption does not seem unreasonable. However, because we expect broadly similar behaviour at each site within any particular weather state, we might not wish to assume *a priori* independence across sites. For example, learning that β_{0k}^i was greater (less) than its expected value would lead to an upward (downward) revision of our beliefs about the expectation of β_{0k}^j, $j \neq i$, for each $k \in \mathcal{S}_r$. The same is true of the parameters in β_{1k} and μ_k. To account for these relationships in our prior beliefs, for each $k \in \mathcal{S}_r$, we adopt hierarchical priors with first level specifications given by

$$\beta_{0k}^i \mid \beta_{0k}, \sigma_{\beta_0,k}^2 \sim \mathrm{N}(\beta_{0k}, \sigma_{\beta_0,k}^2) \qquad \text{independently for } i \in \{1, \ldots, n\},$$

$$\begin{aligned} \beta_{1k}^i \mid p_k &\sim \text{ScBern}(p_k) && \text{independently for } i \in \{1,\ldots,n\}, \\ \mu_k^i \mid \mu_k, \sigma_{\mu,k}^2 &\sim \text{N}(\mu_k, \sigma_{\mu,k}^2) && \text{independently for } i \in \{1,\ldots,n\}, \end{aligned}$$

where the notation $X \sim \text{ScBern}(p)$ means that the random variable X has a Bernoulli distribution with parameter p, scaled to have support on $\{-1,1\}$, that is, $X = 2Y - 1$ where $Y \sim \text{Bern}(p)$. The associated probability mass function is given in Appendix E.

The second level prior specification is

$$\begin{aligned} \beta_{0k} &\sim \text{N}(a_{0,\beta_0}, a_{1,\beta_0}^2), & \sigma_{\beta_0,k}^2 &\sim \text{IG}(h_{0,\beta_0}, h_{1,\beta_0}), \\ p_k &\sim \text{Beta}(b_{0,\beta_1}, b_{1,\beta_1}), & & \\ \mu_k &\sim \text{N}(a_{0,\mu}, a_{1,\mu}^2), & \sigma_{\mu,k}^2 &\sim \text{IG}(h_{0,\mu}, h_{1,\mu}). \end{aligned}$$

These priors have the advantage of being (semi)–conjugate to the form of the likelihood function. They also allow borrowing of strength between sites which is likely to be helpful if, in a particular weather state, the likelihood is not very informative about the parameters associated with certain sites. The effect of regarding the first level variances, $(\sigma_{\beta_0,k}^2, \sigma_{\mu,k}^2)$, as random variables is to allow the data to influence the extent to which strength is borrowed between sites. More details and formulae for the marginal moments in hierarchical Gaussian priors can be found in Section 5.4. Concerning the hierarchical scaled Bernoulli prior, using the law of total expectation it can readily be shown that, marginally

$$\begin{aligned} \text{E}(\beta_{1k}^i) &= \frac{b_{0,\beta_1} - b_{1,\beta_1}}{b_{0,\beta_1} + b_{1,\beta_1}}, \qquad \text{Var}(\beta_{1k}^i) = \frac{4 b_{0,\beta_1} b_{1,\beta_1}}{(b_{0,\beta_1} + b_{1,\beta_1})^2}, \qquad i = 1,\ldots,n \\ \text{Corr}(\beta_{1k}^i, \beta_{1k}^j) &= \frac{1}{b_{0,\beta_1} + b_{1,\beta_1} + 1} \qquad i,j = 1,\ldots,n,\ i \neq j \end{aligned}$$

for each $k \in \mathcal{S}_r$.

As described, the model contains $rn(n+1)/2$ variance and covariance parameters in $(\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_r)$ and rn variance parameters in $(\boldsymbol{\Omega}_1,\ldots,\boldsymbol{\Omega}_r)$. Consider first the set of unstructured variance matrices $(\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_r)$ and the dependence in their joint prior. On the one extreme, the variance matrices could be assumed to be independent, but in this case, it is unlikely that all of the $rn(n+1)/2$ distinct parameters would be properly identified in the posterior. This would be especially true if some of the states occurred infrequently so that the data provided little information about the associated variance matrix. Therefore, pragmatically, assuming *a priori* independence between $\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_r$ is not an attractive option. The other extreme choice would be to assume perfect positive dependence between the variance matrices, that is, $\boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}$ for all $j \in \mathcal{S}_r$, but this prevents the data from being able to suggest that a common matrix is, in fact, untenable.

The problem of simultaneously modelling several variance matrices has been considered by, for example, Bensmail *et al.* (1997), Barnard *et al.* (2000) and Daniels (2006). In pursuit of a simple compromise between the two extreme prior assumptions, a common approach is to decompose the variance matrix into components, of which some are constrained to be common across states (or more generally, across groups) and others are unrestricted. For a problem in cluster analysis, Bensmail *et al.* (1997) consider choosing a joint prior for several variance matrices. Using the geometric interpretation of the spectral decomposition of the variance matrix (see,

for example, Banfield & Raftery, 1993) they elect to assume constancy across clusters of certain features, such as the shape, volume or orientation, whilst allowing the parameters controlling other features to vary. Barnard *et al.* (2000) consider modelling related variance matrices across cells in general location–scale models. Their approach is to decompose the matrix in terms of the corresponding correlation matrix and standard deviations. Next, they assume a common correlation matrix across cells, but leave the vectors of standard deviations unrestricted, although possibly correlated in their prior through a log–linear regression model.

A more flexible compromise between the two extreme dependence assumptions does not fix any of the components to be constant across groups but allows the parameters in the variance matrices $(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_r)$, or some reparameterisation thereof, to be positively correlated *a priori*. The two extreme cases are then recovered in the limit as these correlations tend to one (common variance matrix) or as they tend to zero (independence). This approach borrows strength from the assumption of a common variance matrix whilst allowing the data to inform the posterior when this is not tenable.

In applications in longitudinal analyses where patients have been partitioned into multiple groups, Daniels (2006) considers reparameterising the variance matrix for each group in terms of its spectral decomposition or in terms of the modified Cholesky decomposition of the precision matrix. The spectral decomposition of the variance matrix, $\boldsymbol{\Sigma}$, is given by $\boldsymbol{\Sigma} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^T$ where $\boldsymbol{\Lambda}$ is a diagonal matrix of (ordered) eigenvalues and $\mathbf{P}$ is the corresponding orthogonal matrix of eigenvectors. If $\boldsymbol{\Sigma}$ is an $n \times n$ matrix, then $\mathbf{P}$ can be reparameterised in terms of the $n(n-1)/2$ *Givens angles*, which represent rotations in the plane spanned by pairs of components of the multivariate normal vector; see Daniels (2006) for further details. The modified Cholesky decomposition of the precision matrix reparameterises the variance matrix in terms of the slope coefficients and the conditional variances in the regression of each component of the underlying multivariate normal vector on its predecessors. In these reparameterisations, the Givens angles and slope coefficients can be interpreted as "dependence" parameters, whilst the eigenvalues and conditional variances can be interpreted as "variance" parameters. Based on either decomposition, Daniels (2006) suggests placing two or three stage hierarchical priors on the $n(n-1)/2$ sets of dependence parameters, whilst allowing the variance parameters to be independent *a priori*. The idea is to encourage shrinkage within each set of dependence parameters to a common value. The dependence parameters are targeted simply because there are more of them. Intuitively, however, this might not be unreasonable. Consider the geometry of the underlying multivariate normal distribution for each group (or state, in our case). In some sense, variance–type parameters provide a measure of the overall size and shape of the multi–dimensional density functions, whilst dependence–type parameters roughly measure their orientation. In fact, this is precisely the interpretation which Bensmail *et al.* (1997) assign to the eigenvalues and the orthogonal matrix of eigenvectors in the spectral decomposition of the variance matrix. It is conceivable that the size and shape of the density functions for each group might differ according to whether that group represents a broad or narrow range of behaviours in each of its dimensions. In this case, learning the value of a variance parameter in one group would not necessarily lead us to revise our beliefs about the value of that parameter in other groups. However, we might expect the general orientation of the density functions to be reasonably similar across groups. When the dependence and variance parameters do not correspond to the relevant components of the spectral decomposition of the variance matrix, there is clearly less support for this argument but it nevertheless provides an intuitive motivation for introducing *a priori* correlation between

dependence–type parameters.

The inverse Wishart distribution is commonly chosen as a prior for the variance matrix $\boldsymbol{\Sigma}$ of a multivariate normal distribution because it is conjugate and therefore convenient. However, it is very inflexible in terms of prior elicitation, having only $n(n+1)/2+1$ hyperparameters. This means that once the expectation of the variance matrix has been chosen, there is only one hyperparameter to set all the prior variances and covariances, $\mathrm{Cov}(\Sigma_{jk}, \Sigma_{\ell m})$. Motivated by the inadequacy of the standard conjugate prior, in Germain *et al.* (2010*b*) we developed priors for the variance matrix that were capable of conveying genuine initial beliefs, and proposed elicitation strategies for these priors. Our favoured method followed a similar approach to Daniels & Pourahmadi (2002) by first reparameterising the variance matrix in terms of the modified Cholesky decomposition of its inverse. Omitting full details for brevity, suppose that $\mathbf{Y} = (Y_1, \ldots, Y_n)^T \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathrm{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then write $\mathbf{R} = \mathbf{Y} - \boldsymbol{\mu}$ so that $\mathbf{R} \mid \boldsymbol{\Sigma} \sim \mathrm{N}_n(0, \boldsymbol{\Sigma})$. The elicitation method we propose requires a tentative estimate, $\boldsymbol{\Sigma}_0$, of the variance matrix which is used to define a more natural order amongst the variables in $\mathbf{R}$. This is important for the parameters in the transformed variance matrix to be meaningful. The variables are then reordered through $\mathbf{Q} = \mathbf{M}\mathbf{R}$ where the fixed matrix $\mathbf{M}$ is just an $n \times n$ identity matrix whose rows have been permuted appropriately. Now $\mathbf{Q} \mid \tilde{\boldsymbol{\Sigma}} \sim \mathrm{N}_n(0, \tilde{\boldsymbol{\Sigma}})$ where $\tilde{\boldsymbol{\Sigma}} = \mathbf{M}\boldsymbol{\Sigma}\mathbf{M}^T$. In the next step, we reparamerise the variance matrix in terms of the modified Cholesky decomposition of $\tilde{\boldsymbol{\Sigma}}^{-1}$. This is given by $\tilde{\boldsymbol{\Sigma}}^{-1} = \tilde{\mathbf{T}}^T \tilde{\mathbf{D}}^{-1} \tilde{\mathbf{T}}$ where $\tilde{\mathbf{D}}$ is a diagonal matrix with (j,j)–th entry $\tilde{\sigma}_j^2 > 0$ and $\tilde{\mathbf{T}}$ is a unit lower triangular matrix with (j,k)–th entry $-\tilde{\phi}_{jk} \in \mathbb{R}$. Let $\tilde{\boldsymbol{\phi}}_j = (\tilde{\phi}_{j1}, \ldots, \tilde{\phi}_{j,j-1})^T$ and $\mathbf{Q}_{1:j-1} = (Q_1, \ldots, Q_{j-1})^T$, then the marginal/conditional decomposition of the joint density of $\mathbf{Q}$ is given by

$$p(\mathbf{Q} \mid \tilde{\boldsymbol{\Sigma}}) = p(Q_1 \mid \tilde{\boldsymbol{\Sigma}}) \prod_{j=2}^{n} p(Q_j \mid \mathbf{Q}_{1:j-1}, \tilde{\boldsymbol{\Sigma}})$$

where $Q_1 \mid \tilde{\boldsymbol{\Sigma}} \sim \mathrm{N}(0, \tilde{\sigma}_1^2)$ and $Q_j \mid \mathbf{Q}_{1:j-1}, \tilde{\boldsymbol{\Sigma}} \sim \mathrm{N}(\tilde{\boldsymbol{\phi}}_j^T \mathbf{Q}_{1:j-1}, \tilde{\sigma}_j^2)$ for $j = 2, \ldots, n$. The $\tilde{\phi}_{jk}$ are therefore slope coefficients in the best linear predictor of Q_j based on its predecessors $Q_1, \ldots, Q_{j-1}$, whilst $\tilde{\sigma}_j^2$ is the conditional variance. A semi–conjugate prior is available by assuming *a priori* independence between $\tilde{\boldsymbol{\phi}} = (\tilde{\boldsymbol{\phi}}_2^T, \ldots, \tilde{\boldsymbol{\phi}}_n^T)^T$ and $\tilde{\boldsymbol{\sigma}}_j^2 = (\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_n^2)$, and then by giving $\tilde{\boldsymbol{\phi}}$ a multivariate normal distribution and the $\tilde{\sigma}_j^2$ inverse gamma distributions.

Based on this approach, we use a tentative estimate of the variance matrix, $\boldsymbol{\Sigma}_k$, for state k to reorder the variables in $(\mathbf{Z}_{0t} - \mathbf{X}_t \boldsymbol{\beta}_k)$, then reparameterise the permuted variance matrix, $\boldsymbol{\Sigma}_k$, in terms of the slope coefficients, $\tilde{\boldsymbol{\phi}}_k = (\tilde{\phi}_{k,21}, \tilde{\phi}_{k,31}, \ldots, \phi_{k,n,n-1})^T$, and the conditional variances, $\tilde{\boldsymbol{\sigma}}_k^2 = (\tilde{\sigma}_{k,1}^2, \ldots, \tilde{\sigma}_{k,n}^2)$, arising from the modified Cholesky decomposition of $\tilde{\boldsymbol{\Sigma}}_k^{-1}$. Motivated by the practical and intuitive considerations outlined earlier in this section, we adopt a prior in which the slope coefficients, $\tilde{\boldsymbol{\phi}}_1, \ldots, \tilde{\boldsymbol{\phi}}_r$, are correlated *a priori* but the conditional variances, $\tilde{\boldsymbol{\sigma}}_1^2, \ldots, \tilde{\boldsymbol{\sigma}}_r^2$, are independent across weather states. As stated previously, our prior is to be exchangeable across weather states and so the slope coefficients, $(\tilde{\boldsymbol{\phi}}_1, \ldots, \tilde{\boldsymbol{\phi}}_r)$, can be given a prior

$$(\tilde{\boldsymbol{\phi}}_1, \ldots, \tilde{\boldsymbol{\phi}}_r) \sim \mathrm{N}_{rn(n-1)/2}(\mathbf{m}_{\tilde{\boldsymbol{\phi}},0}, \mathbf{V}_{\tilde{\boldsymbol{\phi}},0}), \tag{6.19}$$

in which $\mathbf{m}_{\tilde{\Phi},0}$ and $\mathbf{V}_{\tilde{\Phi},0}$ have block structures: $\mathbf{m}_{\tilde{\Phi},0} = (\mathbf{m}_{\tilde{\phi},0}^T, \ldots, \mathbf{m}_{\tilde{\phi},0}^T)^T$ and

$$\mathbf{V}_{\tilde{\Phi},0} = \begin{pmatrix} \tilde{\mathbf{V}}_{\tilde{\phi},0} & \mathbf{C}_{\tilde{\phi},0} & \cdots & \mathbf{C}_{\tilde{\phi},0} \\ \mathbf{C}_{\tilde{\phi},0} & \tilde{\mathbf{V}}_{\tilde{\phi},0} & \cdots & \mathbf{C}_{\tilde{\phi},0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_{\tilde{\phi},0} & \mathbf{C}_{\tilde{\phi},0} & \cdots & \tilde{\mathbf{V}}_{\tilde{\phi},0} \end{pmatrix}.$$

Here $\mathbf{m}_{\tilde{\phi},0}$ is an $\{n(n-1)/2\}$–vector and $\tilde{\mathbf{V}}_{\tilde{\phi},0}$ and $\mathbf{C}_{\tilde{\phi},0}$ are $\{n(n-1)/2 \times n(n-1)/2\}$ matrices. For simplicity, we take $\mathbf{C}_{\tilde{\phi},0} = \rho_{\tilde{\phi}}\tilde{\mathbf{V}}_{\tilde{\phi},0}$, where $\rho_{\tilde{\phi}} \in (0,1)$ is fixed. This means that for $k, \ell \in \mathcal{S}_r$, with $k \neq \ell$, $\text{Corr}(\tilde{\phi}_{k,st}, \tilde{\phi}_{\ell,uv}) = \rho_{\tilde{\phi}}\text{Corr}(\tilde{\phi}_{k,st}, \tilde{\phi}_{k,uv}) = \rho_{\tilde{\phi}}\text{Corr}(\tilde{\phi}_{\ell,st}, \tilde{\phi}_{\ell,uv})$. To avoid working with a normal distribution of such high dimension, we found it easier, computationally, to adopt the hierarchical prior specification

$$\tilde{\phi}_k \mid \tilde{\phi} \sim \text{N}_{n(n-1)/2}(\tilde{\phi}, \mathbf{V}_{\tilde{\phi},0}),$$

independently for $k = 1, \ldots, r$, where $\mathbf{V}_{\tilde{\phi},0} = (\tilde{\mathbf{V}}_{\tilde{\phi},0} - \mathbf{C}_{\tilde{\phi},0}) = (1 - \rho_{\tilde{\phi}})\tilde{\mathbf{V}}_{\tilde{\phi},0}$, then

$$\tilde{\phi} \sim \text{N}_{n(n-1)/2}(\mathbf{m}_{\tilde{\phi},0}, \mathbf{C}_{\tilde{\phi},0}).$$

The variable mean, $\tilde{\phi}$, can then be appended to the set of unknown parameters, θ_{obs}, and sampled as part of the MCMC scheme. Formally, this should be equivalent to adopting the one–stage prior in (6.19), since marginalising over $\tilde{\phi}$ produces precisely this distribution. Note that we prefer to avoid a hierarchical prior in which the first level variance matrix, $\mathbf{V}_{\tilde{\phi},0}$, is assigned a prior at the second level of the specification. Pragmatically, it would be difficult to learn about an unknown first level variance matrix if the number of states, r, was small. Conceptually, too, there is no real justification for this added complication. Hierarchical Gaussian priors with unknown variance parameters are the most meaningful when the parameters, here $\tilde{\phi}_1, \ldots, \tilde{\phi}_r$, can be regarded as a sample from some infinite population. This is appropriate when, for example, the collection of parameters are indexed by site and we can think of an infinite collection of sites. However, because the weather states are just artifacts of the model, and not real in any physical sense, this concept seems less credible when the parameters are indexed by weather state.

As remarked previously, the conditional variances $(\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_r^2)$ are assumed to be independent across weather states. However, within any particular weather state, we believe it to be more likely that we will under/over estimate the general variability at *all* sites, rather than only at a few. Therefore, we choose to correlate the conditional variances across sites. To this end, we adopt a hierarchical prior in which, for each $k \in \mathcal{S}_r$,

$$\tilde{\sigma}_{k,i}^2 \mid \tilde{\sigma}_k^2 \sim \text{IG}\left(\frac{1}{v_{\tilde{\sigma}^2,i}^2} + 2, C_i\tilde{\sigma}_k^2\left(\frac{1}{v_{\tilde{\sigma}^2,i}^2} + 1\right)\right),$$

independently for $i = 1, \ldots, n$ with

$$\tilde{\sigma}_k^2 \sim \text{Ga}(c_{0,\tilde{\sigma}^2}, c_{1,\tilde{\sigma}^2}).$$

It follows from the law of total expectation that, marginally

$$\begin{aligned} \mathrm{E}(\tilde{\sigma}^2_{k,i}) &= \frac{C_i c_{0,\tilde{\sigma}^2}}{c_{1,\tilde{\sigma}^2}}, \quad \mathrm{Var}(\tilde{\sigma}^2_{k,i}) = \frac{C_i^2 c_{0,\tilde{\sigma}^2}\{1 + v^2_{\tilde{\sigma}^2,i}(1 + c_{0,\tilde{\sigma}^2})\}}{c^2_{1,\tilde{\sigma}^2}}, \quad i = 1, \ldots, n \\ \mathrm{Corr}(\tilde{\sigma}^2_{k,i}, \tilde{\sigma}^2_{k,j}) &= \frac{1}{\sqrt{\{1 + v^2_{\tilde{\sigma}^2,i}(1 + c_{0,\tilde{\sigma}^2})\}\{1 + v^2_{\tilde{\sigma}^2,j}(1 + c_{0,\tilde{\sigma}^2})\}}}, \quad i, j = 1, \ldots, n, \; i \neq j \end{aligned} \tag{6.20}$$

for each $k \in \mathcal{S}_r$. The fixed hyperparameters, C_i, allow the marginal mean to be different at each site.

The variances $\{(\Omega^{11}_k, \ldots, \Omega^{nn}_k) : k \in \mathcal{S}_r\}$ in the conditional distributions for $(\log W^i_t \mid D^i_t = 1, S_t, \mathbf{Z}_{0t})$ will also be assumed to be independent across weather states. In any particular state, however, the distributions of non–zero rainfall amounts are likely to display a similar level of variability at all sites. As such we assume the parameters $(\Omega^{11}_k, \ldots, \Omega^{nn}_k)$ to be correlated *a priori*. To retain the computational benefits of (semi)–conjugacy, we adopt a hierarchical prior such that, for each $k \in \mathcal{S}_r$,

$$\Omega^{ii}_k \mid \Omega_k \sim \mathrm{IG}\left(\frac{1}{v^2_\Omega} + 2, \Omega_k\left(\frac{1}{v^2_\Omega} + 1\right)\right), \tag{6.21}$$

independently for $i = 1, \ldots, n$ with

$$\Omega_k \sim \mathrm{Ga}(c_{0,\Omega}, c_{1,\Omega}).$$

The expressions in (6.20) give the marginal moments in a similar inverse gamma hierarchical prior. The coefficient of variation parameter, v_Ω, in equation (6.21), could also have been made state specific and assigned a distribution at the second level of the prior specification. Although this would allow the data to influence the degree of borrowing of strength between sites, we experimented with variable $v_{\Omega,k}$ parameters and found that, for some weather states, they could not be identified in the posterior, yielding MCMC algorithms which failed to converge.

Finally, the parameters $\gamma_1, \ldots, \gamma_r$ arising from the process $\{(\mathbf{W}_t \mid \mathbf{D}_t, S_t, \mathbf{Z}_{0t})\}$ are each $n \times n$ matrices containing coefficients in the regressions of the $\log W^i_t$ on $\mathbf{Z}_{0t}$. Although they do not formally contain variance or covariance parameters, it is clear from equation (6.15) that if we integrated out the latent Gaussian vectors, $\{\mathbf{Z}_{0t}\}$, the coefficients, γ^{ij}_k, would appear in expressions for the variances and covariances amongst the log rainfall amounts. Therefore, analogously to the slope coefficients, $\tilde{\phi}_1, \ldots, \tilde{\phi}_r$, we are motivated by both pragmatic and intuitive considerations to make the γ_k correlated between states. We choose to adopt the following hierarchical prior

$$\mathrm{vec}(\gamma_k) \mid \boldsymbol{\gamma} \sim \mathrm{N}_{n^2}(\boldsymbol{\gamma}, \mathbf{V}_{\gamma,0}),$$

independently for $k = 1, \ldots, r$, with

$$\boldsymbol{\gamma} \sim \mathrm{N}_{n^2}(\mathbf{m}_{\gamma,0}, \mathbf{C}_{\gamma,0}),$$

Here $\mathrm{vec}(\gamma_k) = (\gamma^{11}_k, \ldots, \gamma^{1n}_k, \ldots, \gamma^{n1}_k, \ldots, \gamma^{nn}_k)^T$ and we choose $\mathbf{V}_{\gamma,0} = (1 - \rho_\gamma)\tilde{\mathbf{V}}_{\gamma,0}$ and $\mathbf{C}_{\gamma,0} = \rho_\gamma \tilde{\mathbf{V}}_{\gamma,0}$ where the fixed hyperparameter $\tilde{\mathbf{V}}_{\gamma,0}$ contains elicited values for the marginal variances and covariances amongst $\mathrm{vec}(\gamma_k)$. Note that γ^{ij}_k is the coefficient of Z^j_{0t} in the regression of $\log W^i_t$

on $\mathbf{Z}_{0t}$, given $(D_t^i = 1, S_t = k)$. We would expect the effect of Z_{0t}^j on $\log W_t^i$ to differ when $j = i$ compared to when $j \neq i$ and so it might be reasonable to assume *a priori* independence between the on and off diagonal elements in γ_k.

The random variables introduced at the first level of the hierarchical prior specifications $\{(\beta_{0k}, \sigma^2_{\beta_0,k}, p_k, \mu_k, \sigma^2_{\mu,k}, \tilde{\sigma}^2_k, \Omega_k) : k \in \mathcal{S}_r\}$, $\tilde{\phi}$ and γ are appended to the set of unknown model parameters θ_{obs}. Its prior, $\pi(\theta_{\text{obs}})$, in (6.18) must then be replaced with an expression which factorises as

$$\begin{aligned}\pi(\theta_{\text{obs}}) = \prod_{k=1}^{r} \Bigg\{ & \pi(\beta_{0k})\pi(\sigma^2_{\beta_0,k}) \prod_{i=1}^{n} \pi(\beta^i_{0k} \mid \beta_{0k}, \sigma^2_{\beta_0,k}) \times \pi(p_k) \prod_{i=1}^{n} \pi(\beta^i_{1k} \mid p_k) \\ & \times \pi(\mu_k)\pi(\sigma^2_{\mu,k}) \prod_{i=1}^{n} \pi(\mu^i_k \mid \mu_k, \sigma^2_{\mu,k}) \times \pi(\tilde{\sigma}^2_k) \prod_{i=1}^{n} \pi(\tilde{\sigma}^2_{k,i} \mid \tilde{\sigma}^2_k) \\ & \times \pi(\Omega_k) \prod_{i=1}^{n} \pi(\Omega^{ii}_k \mid \Omega_k) \Bigg\} \times \pi(\tilde{\phi})\pi(\tilde{\phi}_1, \ldots, \tilde{\phi}_r \mid \tilde{\phi}) \times \pi(\gamma)\pi(\gamma_1, \ldots, \gamma_r \mid \gamma).\end{aligned}$$

Providing general guidelines on the elicitation of this prior is, in general, very difficult because the latent Gaussian variables, $\mathbf{Z}_{0t}$, are not observable. Nevertheless, in an application to the Yorkshire dataset in Section 6.8.1, we explain the justifications for our choice of fixed hyperparameters.

6.5 Likelihood

Denote by $\mathbf{z}_0 = (\mathbf{z}_{01}, \ldots, \mathbf{z}_{0T})$ the complete time series of latent multivariate normal variables underlying the MVP model for rainfall occurrence. The Yorkshire dataset that is analysed in Section 6.8 divides naturally into Y sub–series, one for each of the Y winter periods. As such, conditionally on the atmospheric data, $\mathbf{x}$, the sub–series $\{(\mathbf{w}_{T^{y}+1}, \mathbf{d}_{T^{y}+1}, \mathbf{z}_{0,T^{y}+1}, \mathbf{s}_{T^{y}+1}), \ldots, (\mathbf{w}_{T^{y+1}}, \mathbf{d}_{T^{y+1}}, \mathbf{z}_{0,T^{y+1}}, \mathbf{s}_{T^{y+1}})\}$ for $y = 1, \ldots, Y$ are modelled as independent realisations of the same NHMM, where the T^y notation was explained in Section 4.4. Denote by $s_{0,y}$ and $\mathbf{d}_{0,y} = (d^1_{0,y}, \ldots, d^n_{0,y})^T$ the initial weather state and occurrence vector for the y–th sub–series then write $\mathbf{s}_0 = (s_{0,1}, \ldots, s_{0,Y})$ and $\mathbf{d}_0 = (\mathbf{d}_{0,1}, \ldots, \mathbf{d}_{0,Y})$. Posterior inference is via MCMC with data augmentation, and so derivation of the full conditional distributions will require the complete data likelihood,

$$p(\mathbf{w}, \mathbf{d}, \mathbf{d}_0, \mathbf{s}, \mathbf{s}_0, \mathbf{z}_0 \mid \theta, \mathbf{x}) = p(\mathbf{w}, \mathbf{d}, \mathbf{d}_0, \mathbf{z}_0 \mid \mathbf{s}, \mathbf{s}_0, \theta_{\text{obs}}) p(\mathbf{s}, \mathbf{s}_0 \mid \theta_{\text{hid}}, \mathbf{x}) \tag{6.22}$$

where $p(\mathbf{s}, \mathbf{s}_0 \mid \theta_{\text{hid}}, \mathbf{x})$ was computed in Section 5.5 (see equation (5.28)) and

$$\begin{aligned}& p(\mathbf{w}, \mathbf{d}, \mathbf{d}_0, \mathbf{z}_0 \mid \mathbf{s}, \mathbf{s}_0, \theta_{\text{obs}}) \\ & \qquad = \prod_{y=1}^{Y} p(\mathbf{w}_{T^{y}+1:T^{y+1}}, \mathbf{d}_{T^{y}+1:T^{y+1}}, \mathbf{d}_{0,y}, \mathbf{z}_{0,T^{y}+1:T^{y+1}} \mid \mathbf{s}_{T^{y}+1:T^{y+1}}, s_{0,y}, \theta_{\text{obs}}).\end{aligned}$$

This can be factorised as

$$p(\mathbf{w}, \mathbf{d}, \mathbf{d}_0, \mathbf{z}_0 \mid \mathbf{s}, \mathbf{s}_0, \theta_{\text{obs}}) = \prod_{y=1}^{Y} \left\{ \prod_{i=1}^{n} p(d_{0,y}^i) \times \prod_{t=T^{y}+1}^{T^{y+1}} \prod_{i=1}^{n} p(d_t^i \mid z_{0t}^i) \times p(\mathbf{z}_{0,T^{y}+1} \mid \mathbf{d}_{0,y}, s_{T^{y}+1}, \theta_{\text{obs}}) \right.$$
$$\left. \times \prod_{t=T^{y}+2}^{T^{y+1}} p(\mathbf{z}_{0t} \mid \mathbf{d}_{t-1}, s_t, \theta_{\text{obs}}) \times \prod_{t=T^{y}+1}^{T^{y+1}} p(\mathbf{w}_t \mid \mathbf{d}_t, s_t, \mathbf{z}_{0t}, \theta_{\text{obs}}) \right\}. \quad (6.23)$$

Denote by $\text{Bern}(d \mid p)$ the probability mass function of the Bernoulli distribution, $\text{Bern}(p)$, evaluated at d, and by $\phi_k(\mathbf{z} \mid \mu, \Sigma)$ the density function of the k–dimensional multivariate normal distribution, $N_k(\mu, \Sigma)$, evaluated at $\mathbf{z}$. Then the components in (6.23) are given by

$$p(d_{0,y}^i) = \text{Bern}(d_{0,y}^i \mid p_{0i}),$$
$$p(d_t^i \mid z_{0t}^i) = \mathbb{I}\{d_t^i = \mathbb{I}(z_{0t}^i > 0)\},$$
$$p(\mathbf{z}_{0t} \mid \mathbf{d}_{t-1}, s_t, \theta_{\text{obs}}) = \phi_n\left(\mathbf{z}_{0t} \mid \{\beta_{0,s_t} + \text{diag}(d_{t-1}^1, \ldots, d_{t-1}^n)\beta_{1,s_t}\}, \Sigma_{s_t}\right),$$
$$p(\mathbf{w}_t \mid \mathbf{d}_t, s_t, \mathbf{z}_{0t}, \theta_{\text{obs}}) = \prod_{i=1}^{n} \phi_1\left(\log w_t^i \mid \left(\mu_{s_t}^i + \sum_{j=1}^{n} \gamma_{s_t}^{ij} z_{0t}^j\right), \Omega_{s_t}^{ii}\right)^{\mathbb{I}(d_t^i=1)}, \quad (6.24)$$
$$p(\mathbf{z}_{0,T^{y}+1} \mid \mathbf{d}_{0,y}, s_{T^{y}+1}, \theta_{\text{obs}}) = \phi_n\left(\mathbf{z}_{0,T^{y}+1} \mid \{\beta_{0,s_{T^{y}+1}} + \text{diag}(d_{0,y}^1, \ldots, d_{0,y}^n)\beta_{1,s_{T^{y}+1}}\}, \Sigma_{s_{T^{y}+1}}\right).$$

Recall that $\mathbf{X}_t = (\mathbf{I}_n, \mathbf{X}_{t1})$ where $\mathbf{X}_{t1} = \text{diag}(d_{t-1}^1, \ldots, d_{t-1}^n)$. If, in addition, we write $\mathbf{X}_{t1} = \text{diag}(d_{0,y}^1, \ldots, d_{0,y}^n)$ at times $t = T^y + 1$ for $y = 1, \ldots, Y$, then we have

$$p(\mathbf{z}_{0t} \mid \mathbf{d}_{t-1}, s_t, \theta_{\text{obs}}) = \phi_n\left(\mathbf{z}_{0t} \mid \mathbf{X}_t\beta_{s_t}, \Sigma_{s_t}\right), \quad (6.25)$$

for any time, $t = 1, \ldots, T$.

6.6 Posterior inference via MCMC

Augmenting the observed data $(\mathbf{w}, \mathbf{d})$ with the latent Gaussian variables $\mathbf{z}_0$, the weather states $(\mathbf{s}, \mathbf{s}_0)$ and the initial rainfall occurrence vectors $\mathbf{d}_0$, the joint posterior distribution of interest is then $\pi(\theta, \mathbf{s}, \mathbf{s}_0, \mathbf{d}_0, \mathbf{z}_0 \mid \mathbf{w}, \mathbf{d}, \mathbf{x})$, which we can write, via Bayes theorem, as

$$\pi(\theta, \mathbf{s}, \mathbf{s}_0, \mathbf{d}_0, \mathbf{z}_0 \mid \mathbf{w}, \mathbf{d}, \mathbf{x}) \propto p(\mathbf{w}, \mathbf{d}, \mathbf{d}_0, \mathbf{s}, \mathbf{s}_0, \mathbf{z}_0 \mid \theta, \mathbf{x})\pi(\theta)$$
$$= p(\mathbf{w}, \mathbf{d}, \mathbf{d}_0, \mathbf{z}_0 \mid \mathbf{s}, \mathbf{s}_0, \theta_{\text{obs}})p(\mathbf{s}, \mathbf{s}_0 \mid \mathbf{x}, \theta_{\text{hid}})\pi(\theta_{\text{hid}})\pi(\theta_{\text{obs}}). \quad (6.26)$$

We can easily sample from this distribution using a straightforward Gibbs scheme which repeatedly iterates through the following four steps. Note that only 1(a) involves generating draws from a non–standard distribution.

1. Sample θ from its conditional posterior distribution, $\pi(\theta \mid \mathbf{w}, \mathbf{d}, \mathbf{d}_0, \mathbf{s}, \mathbf{s}_0, \mathbf{z}_0, \mathbf{x})$, given the latent vectors, $\mathbf{z}_0$ and $\mathbf{d}_0$, and the hidden states, $(\mathbf{s}, \mathbf{s}_0)$. This step is broken down further as

   (a) Sample θ_{hid} from $\pi(\theta_{\text{hid}} \mid \mathbf{s}, \mathbf{s}_0, \mathbf{x})$. This posterior is not affected by the conditional model chosen for precipitation, given the weather state, and so is consistent with the distribution derived in Chapter 5. Sampling θ_{hid} therefore proceeds according to Section 5.6.1.

   (b) Sample θ_{obs} from $\pi(\theta_{\text{obs}} \mid \mathbf{w}, \mathbf{d}, \mathbf{d}_0, \mathbf{s}, \mathbf{z}_0)$ in a series of Gibbs steps. These will be described in Section 6.6.1.

2. Sample $(\mathbf{s}, \mathbf{s}_0)$ from its conditional posterior, $\pi(\mathbf{s}, \mathbf{s}_0 \mid \mathbf{w}, \mathbf{d}, \mathbf{d}_0, \mathbf{z}_0, \theta, \mathbf{x})$, given the model parameters, θ, and the latent vectors, $\mathbf{z}_0$ and $\mathbf{d}_0$. We provide further details below.

3. Sample $\mathbf{z}_0$ from its conditional posterior, $\pi(\mathbf{z}_0 \mid \mathbf{w}, \mathbf{d}, \mathbf{d}_0, \mathbf{s}, \theta_{\text{obs}})$, given the model parameters, θ_{obs}, the initial occurrence vectors, $\mathbf{d}_0$, and the weather states, $\mathbf{s}$. This step will be described in Section 6.6.2

4. Sample $\mathbf{d}_0$ from its conditional posterior, $\pi(\mathbf{d}_0 \mid \mathbf{z}_0, \mathbf{s}, \theta_{\text{obs}})$, given the model parameters, θ_{obs}, the latent Gaussian vectors, $\mathbf{z}_0$, and the weather states, $\mathbf{s}$. Section 6.6.3 outlines how to simulate from this distribution in a series of Gibbs steps.

This scheme can be regarded as an extension to Algorithm 3.3.2, which described Gibbs sampling with data augmentation for more standard hidden Markov models, in which the hidden states were the only latent variables.

Algorithm 3.3.3 outlined a generic forward backward scheme for sampling the hidden states from their joint conditional posterior distribution, given the model parameters, in a single block. In step 2, above, the corresponding posterior distribution for $(\mathbf{s}, \mathbf{s}_0)$ also involves conditioning on the latent vectors, $\mathbf{z}_0$ and $\mathbf{d}_0$. For the purposes of sampling the hidden states, these latent vectors are treated no differently from the observed data and we can think of the NHMM as being characterised by a DAG with the structure of Figure 5.1, in which $\mathbf{R}_t = (\mathbf{Z}_{0t}^T, \mathbf{W}_t^T, \mathbf{D}_t^T)^T$. This is just a non-homogeneous Markov switching model and so the conditions needed to apply Algorithm 3.3.3 are satisfied. Conditionally on the atmospheric data, $\mathbf{x}$, each sub-series is modelled as an independent realisation of the same NHMM and so we apply the forward backward algorithm separately to each sub-series. However, both the (forward) filtering and backward recursions need to be modified so that the first time point is $t = 0$, rather than $t = 1$. The assumption that $\mathbf{D}_0$ and S_0 are independent then allows the initialisation of the filtering algorithm to be simplified so that

$$\Pr(S_0 = \ell \mid \mathbf{d}_0, \theta) = \Pr(S_0 = \ell \mid \theta) = \nu_\ell.$$

In the current notation, within the one step-ahead predictive probabilities and the filtered probabilities in equations (3.14)–(3.16), we have

$$\Pr(S_t = \ell \mid S_{t-1} = k, \mathbf{w}_{1:t-1}, \mathbf{d}_{0:t-1}, \mathbf{z}_{0,1:t-1}, \theta, x_t) = \Pr(S_t = \ell \mid S_{t-1} = k, \theta, x_t) = A_{k\ell}^{x_t} \quad (6.27)$$

and

$$\begin{aligned} p(\mathbf{w}_t, \mathbf{d}_t, \mathbf{z}_{0t} \mid S_t = \ell, \mathbf{w}_{1:t-1}, \mathbf{d}_{0:t-1}, \mathbf{z}_{0,1:t-1}, \theta) \\ &= p(\mathbf{w}_t, \mathbf{d}_t, \mathbf{z}_{0t} \mid \mathbf{d}_{t-1}, S_t = \ell, \theta) \\ &= p(\mathbf{w}_t \mid \mathbf{d}_t, \mathbf{z}_{0t}, S_t = \ell, \theta) p(\mathbf{d}_t \mid \mathbf{z}_{0t}) p(\mathbf{z}_{0t} \mid \mathbf{d}_{t-1}, S_t = \ell, \theta) \\ &= p(\mathbf{w}_t \mid \mathbf{d}_t, \mathbf{z}_{0t}, S_t = \ell, \theta) p(\mathbf{z}_{0t} \mid \mathbf{d}_{t-1}, S_t = \ell, \theta) \end{aligned} \tag{6.28}$$

with $p(\mathbf{w}_t \mid \mathbf{d}_t, \mathbf{z}_{0t}, S_t = \ell, \theta)$ and $p(\mathbf{z}_{0t} \mid \mathbf{d}_{t-1}, S_t = \ell, \theta)$ given in equations (6.24) and (6.25), respectively. Equation (6.27) also holds in (3.18) within the backward sweep.

As in earlier chapters, we address the problem of label switching using the online relabelling algorithm (Algorithm 3.3.4).

6.6.1 Sampling from the complete data posterior $\pi(\theta_{\text{obs}} \mid \mathbf{w}, \mathbf{d}, \mathbf{d}_0, \mathbf{s}, \mathbf{z}_0)$

From the expression for the joint posterior distribution of $(\theta, \mathbf{s}, \mathbf{s}_0, \mathbf{d}_0, \mathbf{z}_0)$ in equation (6.26), it follows that the full conditional distribution for any of the parameters in θ_{obs} can be derived by combining the relevant components from the likelihood expression, $p(\mathbf{w}, \mathbf{d}, \mathbf{d}_0, \mathbf{z}_0 \mid \mathbf{s}, \mathbf{s}_0, \theta_{\text{obs}})$, and the prior, $\pi(\theta_{\text{obs}})$. In the following subsections we derive the full conditional distributions for the parameters in θ_{obs}. They are all standard distributions from which sampling is straightforward.

6.6.1.1 Full conditional distribution for $(\beta_{01}, \ldots, \beta_{0r})$

To deduce the full conditional distribution for the coefficients $(\beta_{01}, \ldots, \beta_{0r})$, the contribution from the likelihood expression, (6.23), is

$$\begin{aligned} &\prod_{y=1}^{Y} \prod_{t=T^{y}+1}^{T^{y+1}} \phi_n \left(\mathbf{z}_{0t} \mid \mathbf{X}_t \beta_{s_t}, \Sigma_{s_t} \right) \\ &\qquad \propto \prod_{k=1}^{r} \exp \left[-\frac{1}{2} \sum_{t:S_t=k} \{(\mathbf{z}_{0t} - \mathbf{X}_{t1}\beta_{1k}) - \beta_{0k}\}^T \Sigma_k^{-1} \{(\mathbf{z}_{0t} - \mathbf{X}_{t1}\beta_{1k}) - \beta_{0k}\} \right]. \end{aligned}$$

The relevant expression from the prior, $\pi(\theta_{\text{obs}})$, is

$$\prod_{k=1}^{r} \pi(\beta_{0k} \mid \beta_{0k}, \sigma^2_{\beta_0,k})$$

where $(\beta_{0k} \mid \beta_{0k}, \sigma^2_{\beta_0,k}) \sim \mathrm{N}_n(\beta_{0k}\mathbf{1}_n, \sigma^2_{\beta_0,k}\mathrm{I}_n)$ in which $\mathbf{1}_n$ is an n-vector of 1's. The expressions for the likelihood and prior as products of factors over the weather state index means that $\beta_{01}, \ldots, \beta_{0r}$ are conditionally independent in their joint full conditional distribution. In Appendix B, we derive the full conditional distribution for the regression coefficients in a multivariate normal linear regression model, assuming a multivariate normal prior for the regression

coefficients. This result is directly applicable here, and we use it to deduce the full conditional distribution for β_{0k} as

$$\beta_{0k} \mid \cdots \sim \mathrm{N}_n(\mathrm{m}_{\beta_{0k},p}, \mathrm{V}_{\beta_{0k},p})$$

where, letting $N_k = \sum_{t=1}^{T} \mathbb{I}(s_t = k)$, the posterior variance and mean are

$$\mathrm{V}_{\beta_{0k},p} = \left(\frac{1}{\sigma^2_{\beta_0,k}}\mathrm{I}_n + N_k\Sigma_k^{-1}\right)^{-1}, \quad \mathrm{m}_{\beta_{0k},p} = \mathrm{V}_{\beta_{0k},p}\left\{\frac{\beta_{0k}}{\sigma^2_{\beta_0,k}}1_n + \Sigma_k^{-1}\sum_{t:S_t=k}(\mathrm{z}_{0t} - \mathrm{X}_{t1}\beta_{1k})\right\}.$$

6.6.1.2 Full conditional distribution for $(\beta_{11}, \ldots, \beta_{1r})$

The contribution from the likelihood expression in equation (6.23) to the full conditional distribution for the coefficients $(\beta_{11}, \ldots, \beta_{1r})$ is given by

$$\prod_{y=1}^{Y}\prod_{t=T^{y}+1}^{T^{y+1}} \phi_n\left(\mathrm{z}_{0t} \mid \mathrm{X}_t\beta_{s_t}, \Sigma_{s_t}\right)$$

$$\propto \prod_{k=1}^{r} \exp\left[-\frac{1}{2}\sum_{t:S_t=k}\{(\mathrm{z}_{0t} - \beta_{0k}) - \mathrm{X}_{t1}\beta_{1k}\}^T\Sigma_k^{-1}\{(\mathrm{z}_{0t} - \beta_{0k}) - \mathrm{X}_{t1}\beta_{1k}\}\right]. \qquad (6.29)$$

Using the results from Appendix B, equation (6.29) can be written as

$$\prod_{k=1}^{r} \exp\left\{-\frac{1}{2}(\beta_{1k} - \hat{\beta}_{1k})^T\mathrm{W}_{\beta_{1k}}(\beta_{1k} - \hat{\beta}_{1k})\right\}$$

where

$$\mathrm{W}_{\beta_{1k}} = \sum_{t:s_t=k}\mathrm{X}_{t1}^T\Sigma_k^{-1}\mathrm{X}_{t1} \qquad \text{and} \qquad \hat{\beta}_{1k} = \mathrm{W}_{\beta_{1k}}^{-1}\sum_{t:s_t=k}\mathrm{X}_{t1}^T\Sigma_k^{-1}(\mathrm{z}_{0t} - \beta_{0k}).$$

The prior distribution, $\pi(\theta_{\text{obs}})$, involves the expression

$$\prod_{k=1}^{r}\pi(\beta_{1k} \mid p_k) = \prod_{k=1}^{r}\prod_{i=1}^{n}\pi(\beta_{1k}^i \mid p_k),$$

where $\Pr(\beta_{1k}^i = j \mid p_k) = p_k^{(1+j)/2}(1-p_k)^{(1-j)/2}$ for $j = -1, 1$. The factorisation of the likelihood and prior contributions as products over the weather state index means that $\beta_{11}, \ldots, \beta_{1r}$ are conditionally independent in their joint full conditional distribution, which will be a product of r independent discrete probability mass functions with support on $\{-1, 1\}^n$. However, the normalising constant in each of these probability mass functions is a sum over 2^n terms, making it computationally expensive to evaluate. It will therefore be easier to simulate $\beta_{1k}^1, \ldots, \beta_{1k}^n$, $k \in \mathcal{S}_r$, one–at–a–time from their (univariate) full conditional distributions,

$$\Pr(\beta_{1k}^i = j \mid \beta_{1k}^{-i}, \cdots) \propto p_k^{(1+j)/2}(1-p_k)^{(1-j)/2}\exp\left\{-\frac{1}{2}\left(\beta_{1k}^{i,\{j\}} - \hat{\beta}_{1k}\right)^T\mathrm{W}_{\beta_{1k}}\left(\beta_{1k}^{i,\{j\}} - \hat{\beta}_{1k}\right)\right\},$$

for $j = -1, 1$, where $\beta_{1k}^{i,\{j\}} = (\beta_{1k}^1, \ldots, \beta_{1k}^{i-1}, j, \beta_{1k}^{i+1}, \ldots, \beta_{1k}^n)^T$ and β_{1k}^{-i} denotes β_{1k} with the i–th element omitted. It can readily be verified that this simplifies so that the full conditionals are scaled Bernoulli distributions

$$\beta_{1k}^i \mid \beta_{1k}^{-i}, \cdots \sim \text{ScBern}\left(p_k\left[p_k + (1-p_k)\exp\left\{2\left(\sum_{j\neq i} W_{\beta_{1k}}^{ij}\beta_{1k}^j - \hat{\beta}_{1k}^T \mathbf{W}_{\beta_{1k}}\mathbf{c}_i\right)\right\}\right]^{-1}\right),$$

where $W_{\beta_{1k}}^{ij}$ denotes the (i,j)–th entry of $\mathbf{W}_{\beta_{1k}}$ and $\mathbf{c}_i$ is an n–vector with a 1 in the i–th place and zero everywhere else.

6.6.1.3 Full conditional distributions for $(\tilde{\phi}_1, \ldots, \tilde{\phi}_r)$ and $(\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_r^2)$

In deriving the full conditional distributions for the parameters $(\tilde{\phi}_1, \ldots, \tilde{\phi}_r)$, the contribution from the likelihood arises through the terms

$$\prod_{y=1}^{Y}\prod_{t=T^{y}+1}^{T^{y+1}} \phi_n\left(\mathbf{z}_{0t} \mid \mathbf{X}_t\beta_{s_t}, \Sigma_{s_t}\right).$$

We then reparameterise the variance matrix according to $\Sigma_k^{-1} = \mathbf{M}^T\tilde{\mathbf{T}}_k^T\tilde{\mathbf{D}}_k^{-1}\tilde{\mathbf{T}}_k\mathbf{M}$ for each $k \in S_r$. Here $\mathbf{M}$ is an $n \times n$ known matrix which transforms $\mathbf{R}_t = (\mathbf{Z}_{0t} - \mathbf{X}_t\beta_{s_t})$ to $\mathbf{Q}_t = \mathbf{M}\mathbf{R}_t$ so that $\text{Var}(\mathbf{Q}_t \mid S_t = k, \tilde{\Sigma}_k) = \tilde{\Sigma}_k = \mathbf{M}\Sigma_k\mathbf{M}^T$. The $n \times n$ matrices $\tilde{\mathbf{T}}_k$ and $\tilde{\mathbf{D}}_k$ then arise from the modified Cholesky decomposition of $\tilde{\Sigma}_k^{-1}$, where $\tilde{\mathbf{T}}_k$ is a unit lower triangular matrix with $-\tilde{\phi}_{k,ij}$ in the (i,j)–th position and $\tilde{\mathbf{D}}_k = \text{diag}(\tilde{\sigma}_{k,1}^2, \ldots, \tilde{\sigma}_{k,n}^2)$. To derive the full conditional distribution for $(\tilde{\phi}_1, \ldots, \tilde{\phi}_r)$, given the transformed data, $\mathbf{Q}_1, \ldots, \mathbf{Q}_T$, the likelihood can be written as

$$\prod_{k=1}^{r}\prod_{t:S_t=k} |\tilde{\mathbf{T}}_k^T\tilde{\mathbf{D}}_k^{-1}\tilde{\mathbf{T}}_k|^{1/2}\exp\left\{-\frac{1}{2}(\tilde{\mathbf{T}}_k\mathbf{Q}_t)^T\tilde{\mathbf{D}}_k^{-1}(\tilde{\mathbf{T}}_k\mathbf{Q}_t)\right\}$$

$$= \prod_{k=1}^{r}\prod_{t:S_t=k} |\tilde{\mathbf{D}}_k|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{Q}_t - \mathbf{U}_t\tilde{\phi}_k)^T\tilde{\mathbf{D}}_k^{-1}(\mathbf{Q}_t - \mathbf{U}_t\tilde{\phi}_k)\right\}, \qquad (6.30)$$

where $\mathbf{U}_t$ is a $\{n \times n(n-1)/2\}$ matrix given by

$$\mathbf{U}_t = \begin{pmatrix} 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ Q_t^1 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ 0 & Q_t^1 & Q_t^2 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 0 & Q_t^1 & Q_t^2 & \ldots & Q_t^{n-1} \end{pmatrix}.$$

A hierarchical prior was adopted for $(\tilde{\phi}_1, \ldots, \tilde{\phi}_r)$ in which

$$\prod_{k=1}^{r} \pi(\tilde{\phi}_k \mid \tilde{\phi}),$$

where $(\tilde{\phi}_k \mid \tilde{\phi}) \sim \mathrm{N}_{n(n-1)/2}(\tilde{\phi}, \mathbf{V}_{\tilde{\phi},0})$. Since the likelihood and prior can both be written as a product of factors over the weather state index, $\tilde{\phi}_1, \ldots, \tilde{\phi}_r$ are conditionally independent in their joint full conditional distribution. It follows from the results in Appendix B that

$$\tilde{\phi}_k \mid \cdots \sim \mathrm{N}_{n(n-1)/2}(\mathbf{m}_{\tilde{\phi}_k,p}, \mathbf{V}_{\tilde{\phi}_k,p})$$

where

$$\mathbf{V}_{\tilde{\phi}_k,p} = \left(\mathbf{V}_{\tilde{\phi},0}^{-1} + \sum_{t:s_t=k} \mathbf{U}_t^T \tilde{\mathbf{D}}_k^{-1} \mathbf{U}_t\right)^{-1}, \quad \mathbf{m}_{\tilde{\phi}_k,p} = \mathbf{V}_{\tilde{\phi}_k,p}\left(\mathbf{V}_{\tilde{\phi},0}^{-1}\tilde{\phi} + \sum_{t:s_t=k} \mathbf{U}_t^T \tilde{\mathbf{D}}_k^{-1} \mathbf{Q}_t\right).$$

In deriving the full conditional distribution for $(\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_r^2)$, since $\tilde{\mathbf{D}}_k$ is a diagonal matrix, we can rewrite the likelihood expression (6.30) as

$$\prod_{k=1}^{r}\prod_{i=1}^{n}(\tilde{\sigma}_{k,i}^2)^{-N_k/2} \exp\left\{-\frac{1}{2\tilde{\sigma}_{k,i}^2}\sum_{t:s_t=k}\left(Q_t^i - \sum_{j<i}\tilde{\phi}_{k,ij}Q_t^j\right)^2\right\}.$$

Combining this with the prior contribution,

$$\prod_{k=1}^{r}\prod_{i=1}^{n}\pi(\tilde{\sigma}_{k,i}^2 \mid \tilde{\sigma}_k^2),$$

where $(\tilde{\sigma}_{k,i}^2 \mid \tilde{\sigma}_k^2) \sim \mathrm{IG}\left(v_{\tilde{\sigma}^2,i}^{-2}+2, C_i\tilde{\sigma}_k^2\left(v_{\tilde{\sigma}^2,i}^{-2}+1\right)\right)$, it is clear that $\tilde{\sigma}_{1,1}^2, \ldots, \tilde{\sigma}_{1,n}^2, \ldots, \tilde{\sigma}_{r,1}^2, \ldots, \tilde{\sigma}_{r,n}^2$ are conditionally independent in their joint full conditional distribution with

$$\tilde{\sigma}_{k,i}^2 \mid \cdots \sim \mathrm{IG}\left(\frac{1}{v_{\tilde{\sigma}^2,i}^2}+2+\frac{N_k}{2}, C_i\tilde{\sigma}_k^2\left(\frac{1}{v_{\tilde{\sigma}^2,i}^2}+1\right)+\frac{1}{2}\sum_{t:s_t=k}\left(Q_t^i - \sum_{j<i}\tilde{\phi}_{k,ij}Q_t^j\right)^2\right).$$

6.6.1.4 Full conditional distribution for $\{(\mu_1, \gamma_1), \ldots, (\mu_r, \gamma_r)\}$

The parameters $(\mu_1, \ldots, \mu_r)$ and $(\gamma_1, \ldots, \gamma_r)$ from the model for non-zero rainfall amounts can be updated jointly. This is likely to produce an MCMC scheme with better mixing properties than a two-block scheme. To deduce the joint full conditional distribution, the relevant components from the likelihood expression, (6.23), can be written as

$$\prod_{y=1}^{Y}\prod_{t=T^{y+1}}^{T^{y+1}}\prod_{i=1}^{n}\phi_1\left(\log w_t^i \middle| \left(\mu_{s_t}^i + \sum_{j=1}^{n}\gamma_{s_t}^{ij}z_{0t}^j\right), \Omega_{s_t}^{ii}\right)^{\mathrm{I}(d_t^i=1)}$$

$$\propto \prod_{k=1}^{r}\exp\left\{-\frac{1}{2}\sum_{t:s_t=k}\sum_{i:d_t^i=1}\frac{1}{\Omega_k^{ii}}(\log w_t^i - \tilde{\mathbf{X}}_t^i\alpha_k)^2\right\},$$

where

$$\alpha_k = (\mu_k^1, \ldots, \mu_k^n, \gamma_k^{11}, \ldots, \gamma_k^{1n}, \ldots, \gamma_k^{n1}, \ldots, \gamma_k^{nn})^T$$
$$\tilde{\mathbf{X}}_t^i = \left(\mathbb{I}(i=1), \ldots, \mathbb{I}(i=n), \mathbb{I}(i=1)z_{0t}^1, \ldots, \mathbb{I}(i=1)z_{0t}^n, \ldots, \mathbb{I}(i=n)z_{0t}^1, \ldots, \mathbb{I}(i=n)z_{0t}^n\right).$$

The relevant terms from the prior distribution, $\pi(\boldsymbol{\theta}_{\text{obs}})$, are

$$\prod_{k=1}^{r} \pi(\alpha_k \mid \mu_k, \gamma, \sigma_{\mu,k}^2)$$

where $(\alpha_k \mid \mu_k, \gamma, \sigma_{\mu,k}^2) \sim \mathrm{N}_{n(n+1)}(\mathbf{m}_{\alpha_k,0}, \mathbf{V}_{\alpha_k,0})$. Here $\mathbf{m}_{\alpha_k,0}$ is an $n(n+1)$-vector with μ_k in the first n positions and the elements of γ in the remaining n^2 positions. Similarly, $\mathbf{V}_{\alpha_k,0}$ is an $\{n(n+1) \times n(n+1)\}$ matrix with $\sigma_{\mu,k}^2$ in the first n diagonal positions, $\mathbf{V}_{\gamma,0}$ occupying the $n \times n$ submatrix in the bottom right hand corner and zero everywhere else. The factorisation of the likelihood and prior as products over the weather state index mean that $\alpha_1, \ldots, \alpha_r$ are conditionally independent in their joint full conditional distribution. Using Appendix B, the component corresponding to α_k is given by

$$\alpha_k \mid \cdots \sim \mathrm{N}_{n(n+1)}(\mathbf{m}_{\alpha_k,p}, \mathbf{V}_{\alpha_k,p})$$

where

$$\mathbf{V}_{\alpha_k,p} = \left(\mathbf{V}_{\alpha_k,0}^{-1} + \sum_{t:s_t=k} \sum_{i:d_t^i=1} \frac{1}{\Omega_k^{ii}} \tilde{\mathbf{X}}_t^{iT} \tilde{\mathbf{X}}_t^i\right)^{-1},$$
$$\mathbf{m}_{\alpha_k,p} = \mathbf{V}_{\alpha_k,p}\left(\mathbf{V}_{\alpha_k,0}^{-1}\mathbf{m}_{\alpha_k,0} + \sum_{t:s_t=k} \sum_{i:d_t^i=1} \frac{1}{\Omega_k^{ii}} \tilde{\mathbf{X}}_t^{iT} \log w_t^i\right).$$

6.6.1.5 Full conditional distribution for $(\Omega_1, \ldots, \Omega_r)$

The full conditional distribution for the diagonal elements $\{(\Omega_k^{11}, \ldots, \Omega_k^{nn}) : k \in \mathcal{S}_r\}$ can be deduced from the likelihood and prior contributions

$$\prod_{y=1}^{Y} \prod_{t=T^y+1}^{T^{y+1}} \prod_{i=1}^{n} \phi_1\left(\log w_t^i \middle| \left(\mu_{s_t}^i + \sum_{j=1}^{n} \gamma_{s_t}^{ij} z_{0t}^j\right), \Omega_{s_t}^{ii}\right)^{\mathrm{I}(d_t^i=1)}$$
$$\propto \prod_{k=1}^{r} \prod_{i=1}^{n} \left[(\Omega_k^{ii})^{-T_{ik}^1/2} \exp\left\{-\frac{1}{2\Omega_k^{ii}} \sum_{\substack{\{t:s_t=k,\\ d_t^i=1\}}} (\log w_t^i - \mu_k^i - \sum_{j=1}^{n} \gamma_k^{ij} z_{0t}^j)^2\right\}\right]$$

and

$$\prod_{k=1}^{r} \prod_{i=1}^{n} \pi(\Omega_k^{ii} \mid \Omega_k),$$

respectively, where $(\Omega_k^{ij} \mid \Omega_k) \sim \text{IG}\left(v_\Omega^{-2} + 2, \Omega_k\left(v_\Omega^{-2} + 1\right)\right)$. Since both the likelihood and prior can be expressed as a product of factors over the weather state and site indices, it follows that $\Omega_1^{11}, \ldots, \Omega_1^{nn}, \ldots, \Omega_r^{11}, \ldots, \Omega_r^{nn}$ are conditionally independent in their joint full conditional distribution with

$$\Omega_k^{ij} \mid \cdots \sim \text{IG}\left(\frac{1}{v_\Omega^2} + 2 + \frac{1}{2}T_{ik}^1, \Omega_k\left(\frac{1}{v_\Omega^2} + 1\right) + \frac{1}{2}\sum_{\substack{\{t:s_t=k,\\ d_t^i=1\}}}\left(\log w_t^i - \mu_k^i - \sum_{j=1}^{n}\gamma_k^{ij}\tilde{z}_{0t}^j\right)^2\right).$$

6.6.1.6 Full conditional distributions for second stage prior parameters

Consider the parameters $\{(\beta_{0k}, \mu_k) : k \in \mathcal{S}_r\}$ and $\{(\sigma_{\beta_0,k}^2, \sigma_{\mu,k}^2) : k \in \mathcal{S}_r\}$ which were given distributions at the second level in the hierarchical prior specifications for $\{(\beta_{0k}, \mu_k) : k \in \mathcal{S}_r\}$. Omitting details, for brevity, the full conditional distributions are

$$\theta_k \mid \theta_k^1, \ldots, \theta_k^{n_\theta}, \sigma_{\theta,k}^2 \sim \text{N}\left(\frac{\sigma_{\theta,k}^2 a_{0,\theta} + a_{1,\theta}^2 \sum_{i=1}^{n_\theta}\theta_k^i}{\sigma_{\theta,k}^2 + n_\theta a_{1,\theta}^2}, \frac{\sigma_{\theta,k}^2 a_{1,\theta}^2}{\sigma_{\theta,k}^2 + n_\theta a_{1,\theta}^2}\right)$$

and

$$\sigma_{\theta,k}^2 \mid \theta_k^1, \ldots, \theta_k^{n_\theta}, \theta_k \sim \text{IG}\left(h_{0,\theta} + \frac{n_\theta}{2}, h_{1,\theta} + \frac{1}{2}\sum_{i=1}^{n_\theta}(\theta_k^i - \theta_k)^2\right),$$

independently for $k \in \mathcal{S}_r$, where θ represents β_0 or μ and $n_\theta = n$.

For the parameters $(p_1, \ldots, p_r)$, which were given distributions at the second level in the hierarchical prior specifications for each β_{1k}, the full conditional distributions are

$$p_k \mid \beta_{1k}^1, \ldots, \beta_{1k}^n \sim \text{Beta}\left(b_{0,\beta_1} + \sum_{i=1}^{n}\mathbb{I}(\beta_{1k}^i = 1), b_{1,\beta_1} + \sum_{i=1}^{n}\mathbb{I}(\beta_{1k}^i = -1)\right),$$

independently for $k \in \mathcal{S}_r$.

From the hierarchical prior specification for $(\gamma_1, \ldots, \gamma_r)$ and $(\tilde{\phi}_1, \ldots, \tilde{\phi}_r)$, the full conditional distributions for the variable means, γ and $\tilde{\phi}$, are given by

$$\gamma \mid \gamma_1, \ldots, \gamma_r \sim \text{N}_{n^2}(\text{m}_{\gamma,p}, \text{V}_{\gamma,p}) \quad \text{and} \quad \tilde{\phi} \mid \tilde{\phi}_1, \ldots, \tilde{\phi}_r \sim \text{N}_{n(n-1)/2}(\text{m}_{\tilde{\phi},p}, \text{V}_{\tilde{\phi},p})$$

where

$$\text{V}_{\gamma,p} = \left(\text{C}_{\gamma,0}^{-1} + r\text{V}_{\gamma,0}^{-1}\right)^{-1}, \quad \text{m}_{\gamma,p} = \text{V}_{\gamma,p}\left\{\text{C}_{\gamma,0}^{-1}\text{m}_{\gamma,0} + \text{V}_{\gamma,0}^{-1}\sum_{k=1}^{r}\text{vec}(\gamma_k)\right\}$$

and

$$\text{V}_{\tilde{\phi},p} = \left(\text{C}_{\tilde{\phi},0}^{-1} + r\text{V}_{\tilde{\phi},0}^{-1}\right)^{-1}, \quad \text{m}_{\tilde{\phi},p} = \text{V}_{\tilde{\phi},p}\left(\text{C}_{\tilde{\phi},0}^{-1}\text{m}_{\tilde{\phi},0} + \text{V}_{\tilde{\phi},0}^{-1}\sum_{k=1}^{r}\tilde{\phi}_k\right).$$

The full conditional distributions for the variable mean components $(\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_r^2)$ in the hierarchical priors for the conditional variances $\{(\tilde{\sigma}_{k,1}^2, \ldots, \tilde{\sigma}_{k,n}^2) : k \in \mathcal{S}_r\}$ are given by

$$\tilde{\sigma}_k^2 \mid \tilde{\sigma}_{k,1}^2, \ldots, \tilde{\sigma}_{k,n}^2 \sim \mathrm{Ga}\left(c_{0,\tilde{\sigma}^2} + \sum_{i=1}^{n}\left(\frac{1}{v_{\tilde{\sigma}^2,i}^2} + 2\right), c_{1,\tilde{\sigma}^2} + \sum_{i=1}^{n} \frac{C_i}{\tilde{\sigma}_{k,i}^2}\left(\frac{1}{v_{\tilde{\sigma}^2,i}^2} + 1\right)\right),$$

independently for $k \in \mathcal{S}_r$.

Finally, from the hierarchical priors for the variance parameters $(\Omega_1, \ldots, \Omega_r)$, the full conditional distributions for the variable means $(\Omega_1, \ldots, \Omega_r)$ are

$$\Omega_k \mid \Omega_k^{11}, \ldots, \Omega_k^{nn} \sim \mathrm{Ga}\left(c_{0,\Omega} + n\left(\frac{1}{v_\Omega^2} + 2\right), c_{1,\Omega} + \left(\frac{1}{v_\Omega^2} + 1\right)\sum_{i=1}^{n} \frac{1}{\Omega_k^{ii}}\right),$$

independently for $k \in \mathcal{S}_r$.

6.6.2 Sampling the latent Gaussian vectors from $\pi(\mathbf{z}_0 \mid \mathbf{w}, \mathbf{d}, \mathbf{d}_0, \mathbf{s}, \theta_{\mathrm{obs}})$ and handling missing data

The joint posterior distribution, $\pi(\theta, \mathbf{s}, \mathbf{s}_0, \mathbf{d}_0, \mathbf{z}_0 \mid \mathbf{w}, \mathbf{d}, \mathbf{x})$, was defined in equation (6.26). In this distribution, the latent Gaussian vectors, $\mathbf{z}_0$, appear only in the component, $p(\mathbf{w}, \mathbf{d}, \mathbf{d}_0, \mathbf{z}_0 \mid \mathbf{s}, \mathbf{s}_0, \theta_{\mathrm{obs}})$, of the complete data likelihood, through which their contribution to the joint posterior is given by

$$\prod_{v=1}^{Y} \prod_{t=T^v+1}^{T^{v+1}} \left\{ \phi_n\left(\mathbf{z}_{0t} \mid \mathbf{X}_t\beta_{s_t}, \Sigma_{s_t}\right) \times \prod_{i=1}^{n} \phi_1\left(\log w_t^i \mid \left(\mu_{s_t}^i + \sum_{j=1}^{n} \gamma_{s_t}^{ij} z_{0t}^j\right), \Omega_{s_t}^{ii}\right)^{\mathrm{I}(d_t^i=1)} \times \mathrm{I}(\mathbf{z}_{0t} \in B_t) \right\}$$

which can be written, up to proportionality, as

$$\prod_{t=1}^{T}\left(\exp\left\{-\frac{1}{2}(\mathbf{z}_{0t} - \mathbf{X}_t\beta_{s_t})^T\Sigma_k^{-1}(\mathbf{z}_{0t} - \mathbf{X}_t\beta_{s_t})\right\} \times \exp\left[-\frac{1}{2}\sum_{i:d_t^i=1}\frac{1}{\Omega_{s_t}^{ii}}\{(\log w_t^i - \mu_{s_t}^i) - \gamma_{s_t}^{i\cdot}\mathbf{z}_{0t}\}^2\right] \times \mathrm{I}(\mathbf{z}_{0t} \in B_t)\right), \tag{6.31}$$

where $\gamma_k^{i\cdot}$ denotes the i–th row of γ_k and $B_t = B_t^1 \times \cdots \times B_t^n$, in which

$$B_t^i = \begin{cases} (0, \infty), & \text{if } d_t^i = 1 \\ (-\infty, 0], & \text{if } d_t^i = 0 \\ (-\infty, \infty), & \text{if } d_t^i \text{ is unobserved.} \end{cases} \tag{6.32}$$

It follows from (6.31) that when $\mathbf{d}_1, \ldots, \mathbf{d}_T$ are all observed, the latent variables $\mathbf{Z}_{01}, \ldots, \mathbf{Z}_{0T}$ are independent in their joint full conditional distribution. For a random vector $\mathbf{X}$, the notation

$\mathbf{X} \sim \mathrm{TN}_n(\mu, \Sigma, B)$ means that $\mathbf{X}$ has a multivariate normal distribution, $\mathrm{N}_n(\mu, \Sigma)$, truncated to the region $B \subseteq \mathbb{R}^n$. At any time, t, we can regard the first term in (6.31) as the "prior" and the second term as the "likelihood", then Appendix B can be used to deduce the full conditional distribution for $\mathbf{Z}_{0t}$ as

$$\mathbf{Z}_{0t} \mid \cdots \sim \mathrm{TN}_n(\mathbf{m}_{Z_{0t},p}, \mathbf{V}_{Z_{0t},p}, B_t)$$

where

$$\mathbf{V}_{Z_{0t},p} = \left\{ \Sigma_{s_t}^{-1} + \sum_{i:d_t^i=1} \frac{1}{\Omega_{s_t}^{ii}} (\gamma_{s_t}^{i,:})^T \gamma_{s_t}^{i,:} \right\}^{-1},$$

$$\mathbf{m}_{Z_{0t},p} = \mathbf{V}_{Z_{0t},p} \left\{ \Sigma_{s_t}^{-1} \mathbf{X}_t \beta_{s_t} + \sum_{i:d_t^i=1} \frac{1}{\Omega_{s_t}^{ii}} (\gamma_{s_t}^{i,:})^T (\log w_t^i - \mu_{s_t}^i) \right\},$$

and B_t was defined in (6.32).

Each truncated multivariate normal distribution is sampled according to the Gibbs scheme in Appendix C. This comprises a cycle of n steps to successively simulate the components of $\mathbf{Z}_{0t}$ one-at-a-time from the distributions $(Z_{0t}^i \mid \mathbf{Z}_{0t}^{-i}, \mathbf{W}_t, \mathbf{D}_t, \mathbf{D}_{t-1}, S_t, \theta_{\mathrm{obs}})$ for $i = 1, \ldots, n$. Each of these conditional distributions is truncated univariate normal.

As discussed in Section 4.5.2, we assume that the missing data mechanism is ignorable. Non-zero rainfall amounts, W_t^i, are assumed to be conditionally independent in space and time given the latent Gaussian vectors, $\mathbf{Z}_{0t}$, and the weather states, S_t. As such we can proceed according to Section 3.3.6 in order to marginalise analytically over the missing rainfall amounts. If the arrows from the nodes $\mathbf{D}_{t-1}$ to $\mathbf{Z}_{0t}$ were omitted, then we could easily marginalise over the missing rainfall occurrences, expressing the probability of rain at the observed sites through a latent vector from the relevant (marginal) multivariate normal distribution of lower dimension. This is possible because, in this case, the $\mathbf{D}_t$ nodes would not have any children. However, since $\mathbf{D}_{t-1}$ is a regressor for $\mathbf{Z}_{0t}$ in our model, analytic marginalisation over the missing data is very difficult. We therefore numerically average over the missing rainfall occurrences by drawing them jointly with the latent vectors $\mathbf{Z}_{0t}$. That is, if the rainfall occurrence indicator at site i on day t is missing, we jointly draw (z_{0t}^i, d_t^i) by sampling z_{0t}^i and then $(d_t^i \mid z_{0t}^i)$ which is given deterministically by $d_t^i = \mathbb{I}(z_{0t}^i > 0)$.

6.6.3 Sampling the initial rainfall occurrence indicators from $\pi(\mathbf{d}_0 \mid \mathbf{s}, \mathbf{z}_0, \theta)$

For the sake of notational concision in this section, we write t_y in place of $T^y + 1$. Now, the initial rainfall occurrence indicators contribute to the joint posterior only through the component, $p(\mathbf{w}, \mathbf{d}, \mathbf{d}_0, \mathbf{z}_0 \mid \mathbf{s}, \mathbf{s}_0, \theta_{\mathrm{obs}})$, more precisely through the terms

$$\prod_{y=1}^{Y} \left\{ \prod_{i=1}^{n} \mathrm{Beta}(d_{0,y}^i \mid p_{0i}) \times \phi_n \left(z_{0,t_y} \mid \{\beta_{0,s_{t_y}} + \mathrm{diag}(d_{0,y}^1, \ldots, d_{0,y}^n)\beta_{1,s_{t_y}}\}, \Sigma_{s_{t_y}} \right) \right\}.$$

This is proportional to

$$\prod_{y=1}^{Y}\left(\prod_{i=1}^{n} p_{0i}^{d_{0,y}^i}(1-p_{0i})^{1-d_{0,y}^i} \times \exp\left[-\frac{1}{2}\{(\mathbf{z}_{0,t_y}-\boldsymbol{\beta}_{0,s_{t_y}})-\mathbf{B}_{1,s_{t_y}}\mathbf{d}_{0,y}\}^T\boldsymbol{\Sigma}_{s_{t_y}}^{-1}\{(\mathbf{z}_{0,t_y}-\boldsymbol{\beta}_{0,s_{t_y}})-\mathbf{B}_{1,s_{t_y}}\mathbf{d}_{0,y}\}\right]\right),$$

where $\mathbf{B}_{1,s_{t_y}} = \text{diag}(\beta_{1,s_{t_y}}^1,\ldots,\beta_{1,s_{t_y}}^n)$. It follows that $\mathbf{d}_{0,1},\ldots,\mathbf{d}_{0,Y}$ are conditionally independent in their joint full conditional distribution which is a product of Y independent discrete probability mass functions, each with support on $\{0,1\}^n$. However, evaluation of the normalising constant for each posterior mass function requires summation over 2^n terms and so, computationally, it is more convenient to sample the elements $d_{0,y}^1,\ldots,d_{0,y}^n$, $y=1,\ldots,Y$, one-at-a-time from their full conditional distributions,

$$\begin{aligned}&\Pr(D_{0,y}^i = j \mid \mathbf{D}_{0,y}^{-i} = \mathbf{d}_{0,y}^{-i},\cdots)\\ &\quad\propto p_{0i}^j(1-p_{0i})^{1-j}\\ &\qquad\times\exp\left[-\frac{1}{2}\left\{(\mathbf{z}_{0,t_y}-\boldsymbol{\beta}_{0,s_{t_y}})-\mathbf{B}_{1,s_{t_y}}\mathbf{d}_{0,y}^{i,\{j\}}\right\}^T\boldsymbol{\Sigma}_{s_{t_y}}^{-1}\left\{(\mathbf{z}_{0,t_y}-\boldsymbol{\beta}_{0,s_{t_y}})-\mathbf{B}_{1,s_{t_y}}\mathbf{d}_{0,y}^{i,\{j\}}\right\}\right]\end{aligned}$$

for $j=0,1$, where $\mathbf{d}_{0,y}^{i,\{j\}} = (d_{0,y}^1,\ldots,d_{0,y}^{i-1},j,d_{0,y}^{i+1},\ldots,d_{0,y}^n)^T$. These Bernoulli distributions can be written more concisely as

$$D_{0,y}^i \mid \mathbf{D}_{0,y}^{-i} = \mathbf{d}_{0,y}^{-i},\cdots \sim \text{Bern}(p_{p,y}^i)$$

where

$$p_{p,y}^i = p_{0i}\left[p_{0i}+(1-p_{0i})\exp\left\{\sum_{j\neq i}B_{2,s_{t_y}}^{ij}d_{0,y}^j + B_{2,s_{t_y}}^{ii}/2 - (\mathbf{Z}_{0,t_y}-\boldsymbol{\beta}_{0,s_{t_y}})^T\boldsymbol{\Sigma}_{s_{t_y}}^{-1}\mathbf{B}_{1,s_{t_y}}\mathbf{c}_i\right\}\right]^{-1}$$

in which $B_{2,s_{t_y}}^{ij}$ denotes the (i,j)-th entry of $(\mathbf{B}_{1,s_{t_y}}^T\boldsymbol{\Sigma}_{s_{t_y}}^{-1}\mathbf{B}_{1,s_{t_y}})$ and $\mathbf{c}_i$ is an n-vector with a 1 in the i-th place and zero everywhere else.

6.7 Posterior inference for r

In this section, we begin by reviewing the power posterior approach in order to demonstrate that a correction is needed when the method is applied to models, such as that presented in this chapter, with certain properties. We then explain why difficulties in approximating the correction term for the latent Gaussian variable NHMM prohibit the use of power posteriors. Chib's original method was outlined in Section 3.5.1.3. In this section, we provide further details, focusing on the extension (due to Chib & Jeliazkov, 2001) which allows the method to be applied when some parameter blocks have full conditional distributions with unknown normalising constants. We then explain how this method can be applied to the latent Gaussian variable NHMM.

6.7.1 The power posterior approach

Consider data $\mathbf{y}$, unknowns θ' (which might include auxiliary variables) and a temperature parameter $t \in [0, 1]$. Friel & Pettitt (2008) define the power posterior at temperature t as

$$p_t(\theta' \mid \mathbf{y}) \propto p(\mathbf{y} \mid \theta')^t p(\theta').$$

Let Θ_0 and Θ_1 denote the sets of values of θ' which have non-zero probability/density in the prior and posterior, respectively, and note that $\Theta_1 \subseteq \Theta_0$. Let us define the normalising constant in the power posterior as

$$z(\mathbf{y} \mid t) = \int_{\Theta_1} p(\mathbf{y} \mid \theta')^t p(\theta') d\theta'$$

when $t \neq 0$ and

$$z(\mathbf{y} \mid t = 0) = \int_{\Theta_0} p(\theta') d\theta'$$

when $t = 0$. Regarding $z(\mathbf{y} \mid t)$ as a function of t, if Θ_1 is a strict subset of Θ_0, then

$$\lim_{t \to 0^+} z(\mathbf{y} \mid t) = \int_{\Theta_1} p(\theta') d\theta' \neq \int_{\Theta_0} p(\theta') d\theta' = z(\mathbf{y} \mid t = 0)$$

and so, in this case, $z(\mathbf{y} \mid t)$ has a discontinuity at $t = 0$. We now denote

$$z^*(\mathbf{y}) = \lim_{t \to 0^+} z(\mathbf{y} \mid t) = \int_{\Theta_1} p(\theta') d\theta'.$$

When $\Theta_1 \subset \Theta_0$, the discontinuity at $t = 0$ means that the function $z(\mathbf{y} \mid t)$ is not differentiable at $t = 0$. Following the derivation from Friel & Pettitt (2008), but restricting attention to $0 < t \leq 1$, we have

$$\begin{aligned}
\frac{d}{dt} \log\{z(\mathbf{y} \mid t)\} &= \frac{1}{z(\mathbf{y} \mid t)} \frac{d}{dt} z(\mathbf{y} \mid t) \\
&= \frac{1}{z(\mathbf{y} \mid t)} \frac{d}{dt} \int_{\Theta_1} p(\mathbf{y} \mid \theta')^t p(\theta') d\theta' \\
&= \frac{1}{z(\mathbf{y} \mid t)} \int_{\Theta_1} p(\mathbf{y} \mid \theta')^t \log\{p(\mathbf{y} \mid \theta')\} p(\theta') d\theta' \\
&= \int_{\Theta_1} \frac{p(\mathbf{y} \mid \theta')^t p(\theta')}{z(\mathbf{y} \mid t)} \log\{p(\mathbf{y} \mid \theta')\} d\theta' \\
&= \mathrm{E}_{\theta' \mid \mathbf{y}, t}[\log\{p(\mathbf{y} \mid \theta')\}] = z^*(\mathbf{y} \mid t), \quad \text{say.}
\end{aligned} \tag{6.33}$$

Integrating with respect to $t \in [\epsilon, 1]$, where $0 < \epsilon \leq 1$, then gives

$$\log\{z(\mathbf{y} \mid t = 1)\} - \log\{z(\mathbf{y} \mid t = \epsilon)\} = \int_{\epsilon}^{1} \mathrm{E}_{\theta' \mid \mathbf{y}, t}[\log\{p(\mathbf{y} \mid \theta')\}] dt,$$

where $z(\mathbf{y} \mid t = 1)$ is equal to the marginal likelihood, $z(\mathbf{y} \mid t = 1) = p(\mathbf{y})$. Finally, letting $\epsilon \to 0^+$ and rearranging, yields the following expression for the log marginal likelihood

$$\begin{aligned}\log\{p(\mathbf{y})\} &= \lim_{\epsilon \to 0^+} \int_{\epsilon}^{1} \mathrm{E}_{\theta'|\mathbf{y},t}[\log\{p(\mathbf{y} \mid \theta')\}]dt + \log\{z^*(\mathbf{y})\} \\ &= \int_0^1 z^*(\mathbf{y} \mid t)dt + \log\{z^*(\mathbf{y})\}. \end{aligned} \tag{6.34}$$

Friel & Pettitt (2008) only considered problems for which $\Theta_1 = \Theta_0$. In this case $z(\mathbf{y} \mid t)$ does not have a discontinuity at $t = 0$ and $z^*(\mathbf{y})$ is simply the integral of the prior over the set of values with non–zero prior density. As such, $\log\{z^*(\mathbf{y})\}$ is equal to zero and we recover

$$\log\{p(\mathbf{y})\} = \int_0^1 \mathrm{E}_{\theta'|\mathbf{y},t}[\log\{p(\mathbf{y} \mid \theta')\}]dt, \tag{6.35}$$

which is the expression for the log marginal likelihood presented by Friel & Pettitt.

For most problems, the prior and posterior have support over the same set of values of θ' and so equation (6.35) can be used directly. This is even true of many problems in which θ' includes auxiliary variables, for example, the hidden Markov models considered in Chapters 4 and 5. However, there are some problems involving auxiliary variables for which Θ_1 is a strict subset of Θ_0. Let $\theta' = (\theta, \mathbf{z})$ where $\mathbf{z}$ are auxiliary variables and θ are "parameters". In auxiliary variable models with $\Theta_1 \subset \Theta_0$, the likelihood is often such that for any actually observed $\mathbf{y}$, $p(\mathbf{y} \mid \theta') = p(\mathbf{y} \mid \mathbf{z})$ where $p(\mathbf{y} \mid \mathbf{z})$ is a constant with respect to $\mathbf{z}$, say, $p(\mathbf{y} \mid \mathbf{z}) = C(\mathbf{y})$. For example, this will be the case with a simple probit model, in which each observed binary indicator variable depends deterministically on the sign of an (auxiliary) Gaussian variable. If this is the case then equation (6.34) becomes

$$\log\{p(\mathbf{y})\} = \log\{C(\mathbf{y})\} + \log\{z^*(\mathbf{y})\}, \quad \text{or equivalently,} \quad p(\mathbf{y}) = C(\mathbf{y})z^*(\mathbf{y}), \tag{6.36}$$

and estimation of the marginal likelihood reduces to evaluating or, more likely, approximating the integral $z^*(\mathbf{y}) = \int_{\Theta_1} p(\theta')d\theta'$. In other models with $\Theta_1 \subset \Theta_0$, where $p(\mathbf{y} \mid \theta')$ is *not* a constant with respect to θ' for any actually observed $\mathbf{y}$, the first term in (6.34) can be approximated using the techniques discussed in Friel & Pettitt (2008) for approximating the right–hand–side of (6.35), but limiting sampling to Θ_1 when $t = 0$. To obtain an estimate of the log marginal likelihood, this must then be corrected by adding $\log\{z^*(\mathbf{y})\}$ which, itself, may need to be approximated. The model discussed in this chapter is of the latter type, requiring estimation in two stages.

In general, an area for future research is the development of efficient ways of computing the correction term, $z^*(\mathbf{y})$, when it is not available analytically.

6.7.1.1 Application to the latent Gaussian variable NHMM

Let $\theta' = (\theta_{\text{obs}}, \theta_{\text{hid}}, \mathbf{s}, \mathbf{s}_0, \mathbf{d}_0, \mathbf{z}_0)$. For the parameters $\theta = (\theta_{\text{obs}}, \theta_{\text{hid}})$ and the latent variables $(\mathbf{s}, \mathbf{s}_0, \mathbf{d}_0)$, the set of values which has non–zero density in the prior is the same as that in the posterior. However, for each $t = 1, \ldots, T$, the latent Gaussian variable $\mathbf{Z}_{0t}$ has non–zero density

over $\mathbb{R}$ in the prior, but only over B_t in the posterior, where the n–dimensional region B_t was defined in (6.32). It follows that $\Theta_1 \subset \Theta_0$. For any actually observed set of rainfall occurrences, d, the likelihood is given by

$$p(\mathbf{w}, \mathbf{d} \mid \mathbf{s}, \mathbf{s}_0, \mathbf{d}_0, \mathbf{z}_0, \theta) = p(\mathbf{w} \mid \mathbf{d}, \mathbf{s}, \mathbf{z}_0, \theta_{\text{obs}})$$
$$= \prod_{y=1}^{Y} \prod_{t=T^{y}+1}^{T^{y+1}} \prod_{i=1}^{n} \phi_1 \left(\log w_t^i \middle| \left(\mu_{s_t}^i + \sum_{j=1}^{n} \gamma_{s_t}^{ij} z_{0t}^j \right), \Omega_{s_t}^{ii} \right)^{\mathbf{I}(d_t^i = 1)}, \qquad (6.37)$$

whilst $p(\mathbf{w}, \mathbf{d} \mid \mathbf{s}, \mathbf{s}_0, \mathbf{d}_0, \mathbf{z}_0, \theta) = 0$ if $\theta' \notin \Theta_1$. Equation (6.37) is clearly not a constant with respect to θ'. Therefore, estimation of the log marginal likelihood could, in theory, proceed by adding an estimate of the integral of $z^*(\mathbf{w}, \mathbf{d} \mid t)$, defined in (6.33), over $t \in [0, 1]$, to an estimate of $\log z^*(\mathbf{w}, \mathbf{d})$.

Let S_θ, S_{s,s_0} and S_{d_0} denote the sets of values of θ, $(\mathbf{s}, \mathbf{s}_0)$ and $\mathbf{d}_0$ which have non–zero probability/density in both the prior and the posterior. Now consider the integral of the "prior" over the support of the posterior,

$$z^*(\mathbf{w}, \mathbf{d}) = \int_{S_\theta} \sum_{S_{d_0}} \sum_{S_{s,s_0}} \int_{B_T} \cdots \int_{B_1} p(\theta, \mathbf{s}, \mathbf{s}_0, \mathbf{d}_0, \mathbf{z}_0 \mid \mathbf{x}) dz_{01}, \ldots, dz_{0T} d\theta$$
$$= \int_{S_\theta} \sum_{S_{d_0}} \sum_{S_{s,s_0}} p(\theta, \mathbf{s}, \mathbf{s}_0, \mathbf{d}_0, \mathbf{d} \mid \mathbf{x}) d\theta$$
$$= \int_{S_\theta} p(\mathbf{d} \mid \theta, \mathbf{x}) \pi(\theta) d\theta. \qquad (6.38)$$

For a given value of θ, estimation of the observed data likelihood $p(\mathbf{w}, \mathbf{d} \mid \theta, \mathbf{x})$, will be discussed in Section 6.7.2.1, where we show how it can be approximated at a single point, $\theta = \theta^*$. The observed data likelihood for rainfall occurrences, $p(\mathbf{d} \mid \theta, \mathbf{x})$, appearing in equation (6.38) could be evaluated in an analogous fashion. Essentially this would involve enumerating $p(\mathbf{d}, \mathbf{d}_0 \mid \theta, \mathbf{x}) = p(\mathbf{d}_0) p(\mathbf{d} \mid \mathbf{d}_0, \theta, \mathbf{x})$ over all possible values of $\mathbf{d}_0$. For each value of $\mathbf{d}_0$ in turn, $p(\mathbf{d} \mid \mathbf{d}_0, \theta, \mathbf{x})$ would be computed using forward filtering to marginalise the joint density $p(\mathbf{d}, \mathbf{s}, \mathbf{s}_0 \mid \mathbf{d}_0, \theta, \mathbf{x})$ over the hidden states. As part of the filtering recursion, probabilities of the form $\Pr(\mathbf{D}_t = \mathbf{d}_t \mid S_t = k, \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, \theta) = \Pr(\mathbf{Z}_t \in B_t \mid S_t = k, \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, \theta)$ would have to be estimated using a numerical algorithm for approximating multivariate normal probabilities. In principle, therefore, Monte Carlo integration could be used to approximate $z^*(\mathbf{w}, \mathbf{d})$ by averaging (estimates of) $p(\mathbf{d} \mid \theta, \mathbf{x})$ over draws from the prior of the model parameters, $\pi(\theta)$. However, even at a single value of θ, estimation of $p(\mathbf{d} \mid \theta, \mathbf{x})$ requires a large amount of computing time, rendering this approach to approximating $z^*(\mathbf{w}, \mathbf{d})$ infeasible.

Although it is possible that better methods for estimating the integral $z^*(\mathbf{w}, \mathbf{d})$ could be found, we choose to pursue an alternative approach to approximating the marginal likelihood for this model.

6.7.2 Chib's method

For general data $\mathbf{y}$ and a model characterised by the parameter θ, Chib's method (Chib, 1995; Chib & Jeliazkov, 2001) is based on the identity

$$p(\mathbf{y}) = \frac{p(\mathbf{y} \mid \theta)\pi(\theta)}{\pi(\theta \mid \mathbf{y})}.$$

Denoting by θ^* a point of high posterior density, an estimate of the marginal likelihood, on the log scale, is given by

$$\log p(\mathbf{y}) = \log p(\mathbf{y} \mid \theta^*) + \log \pi(\theta^*) - \log \pi(\theta^* \mid \mathbf{y}),$$

and so, as long as the prior and likelihood can be evaluated or approximated at the high density point θ^*, the marginal likelihood can be estimated by finding an approximation to the posterior ordinate, $\pi(\theta^* \mid \mathbf{y})$.

Suppose $\theta = (\theta_1, \ldots, \theta_B)$ for some partitioning of θ that is convenient to Gibbs or Metropolis within Gibbs sampling. If data augmentation is used when drawing MCMC samples from the posterior, denote the auxiliary variables by $\mathbf{z}$. Under Chib's method, the posterior ordinate is estimated by the marginal/conditional decomposition of the joint posterior density

$$\pi(\theta^* \mid \mathbf{y}) = \prod_{i=1}^{B} \pi(\theta_i^* \mid \mathbf{y}, \theta_1^*, \ldots, \theta_{i-1}^*).$$

Chib's original method (Chib, 1995) requires that all the parameter blocks $\theta_1, \ldots, \theta_B$ have full conditional densities with known normalising constants. The first ordinate $\pi(\theta_1^* \mid \mathbf{y})$ is estimated by averaging the full conditional density of θ_1, evaluated at θ_1^*, over the draws from a full MCMC run. The second ordinate is then obtained in a reduced run in which we fix $\theta_1 = \theta_1^*$ and sample all other blocks $\theta_2, \ldots, \theta_B$ and auxiliary variables $\mathbf{z}$ from their full conditional densities. This leads to an estimate of the conditional ordinate given by

$$\pi(\theta_2^* \mid \mathbf{y}, \theta_1^*) = \frac{1}{M} \sum_{g=1}^{M} \pi(\theta_2^* \mid \mathbf{y}, \theta_1^*, \theta_3^{[g]}, \ldots, \theta_B^{[g]}, \mathbf{z}^{[g]}).$$

Subsequent conditional ordinates are estimated by the same procedure, fixing additional parameter blocks at their high density points.

Clearly, this technique cannot be used if some of the parameter blocks have unknown normalising constants. However, using the detailed balance property of the Metropolis Hastings algorithm, Chib & Jeliazkov (2001) provided an extension to overcome this problem. Suppose that the i-th parameter block, θ_i, has a full conditional distribution, $\pi(\theta_i \mid \mathbf{y}, \theta_{-i}, \mathbf{z})$, whose normalising constant is unknown. Let $\theta_{j:k} = (\theta_j, \ldots, \theta_k)$ and suppose that θ_i is updated by a Metropolis Hastings step in which $q(\theta_i, \theta_i' \mid \mathbf{y}, \theta_{1:i-1}, \theta_{i+1:B}, \mathbf{z})$ is the proposal density and $\alpha(\theta_i, \theta_i' \mid \mathbf{y}, \theta_{1:i-1}, \theta_{i+1:B}, \mathbf{z})$ is the acceptance probability of the proposed move from θ_i to θ_i'. In this case, it can be shown that the posterior ordinate $\pi(\theta_i^* \mid \mathbf{y}, \theta_1^*, \ldots, \theta_{i-1}^*)$ is given by

$$\pi(\theta_i^* \mid \mathbf{y}, \theta_1^*, \ldots, \theta_{i-1}^*) = \frac{\mathrm{E}_1\left\{\alpha(\theta_i, \theta_i^* \mid \mathbf{y}, \theta_{1:i-1}^*, \theta_{i+1:B}, \mathbf{z}) q(\theta_i, \theta_i^* \mid \mathbf{y}, \theta_{1:i-1}^*, \theta_{i+1:B}, \mathbf{z})\right\}}{\mathrm{E}_2\left\{\alpha(\theta_i^*, \theta_i \mid \mathbf{y}, \theta_{1:i-1}^*, \theta_{i+1:B}, \mathbf{z})\right\}} \quad (6.39)$$

where the expectation in the numerator, $\mathrm{E}_1(\cdot)$, is with respect to the conditional posterior $\pi(\theta_i, \theta_{i+1:B}, \mathbf{z} \mid \mathbf{y}, \theta^*_{1:i-1})$ and the expectation in the denominator, $\mathrm{E}_2(\cdot)$, is with respect to the distribution $\pi(\theta_{i+1:B}, \mathbf{z} \mid \mathbf{y}, \theta^*_{1:i}) \times q(\theta^*_i, \theta_i \mid \mathbf{y}, \theta^*_{1:i-1}, \theta_{i+1:B}, \mathbf{z})$. From (6.39), an estimate of the conditional posterior ordinate is available as

$$\hat{\pi}(\theta^*_i \mid \mathbf{y}, \theta^*_1, \ldots, \theta^*_{i-1}) = \frac{\frac{1}{M}\sum_{g=1}^{M} \alpha(\theta^{[g]}_i, \theta^*_i \mid \mathbf{y}, \theta^*_{1:i-1}, \theta^{[g]}_{i+1:B}, \mathbf{z}^{[g]}) q(\theta^{[g]}_i, \theta^*_i \mid \mathbf{y}, \theta^*_{1:i-1}, \theta^{[g]}_{i+1:B}, \mathbf{z}^{[g]})}{\frac{1}{J}\sum_{j=1}^{J} \alpha(\theta^*_i, \theta^{[j]}_i \mid \mathbf{y}, \theta^*_{1:i-1}, \theta^{[j]}_{i+1:B}, \mathbf{z}^{[j]})}$$

where the draws $\{\theta^{[g]}_i, \theta^{[g]}_{i+1:B}, \mathbf{z}^{[g]}\}$ in the numerator are obtained by setting $\theta_{1:i-1} = \theta^*_{1:i-1}$ and sampling from $\pi(\theta_k \mid \mathbf{y}, \theta_{-k}, \mathbf{z})$ for $k = i, \ldots, B$ and from $\pi(\mathbf{z} \mid \mathbf{y}, \theta)$ in a reduced MCMC run. The draws $\{\theta^{[j]}_{i+1:B}, \mathbf{z}^{[j]}\}$ in the denominator are obtained in a second reduced run by setting $\theta_{1:i-1} = \theta^*_{1:i-1}$ *and* $\theta_i = \theta^*_i$ and sampling from $\pi(\theta_k \mid \mathbf{y}, \theta_{-k}, \mathbf{z})$ for $k = i+1, \ldots, B$ and from $\pi(\mathbf{z} \mid \mathbf{y}, \theta)$. At every step of this second reduced run, a draw $\theta^{[j]}_i$ is also required from the proposal density $q(\theta^*_i, \theta_i \mid \mathbf{y}, \theta^*_{1:i-1}, \theta^{[j]}_{i+1:B}, \mathbf{z}^{[j]})$ in order to evaluate each summand in the denominator. Note that the draws $\{\theta^{[j]}_{i+1:B}, \mathbf{z}^{[j]}\}$ obtained in the second reduced run can also be used in the evaluation of the next conditional posterior ordinate $\pi(\theta^*_{i+1} \mid \mathbf{y}, \theta^*_{1:i})$.

6.7.2.1 Application to the latent Gaussian variable NHMM

In Chapters 4 and 5, many of the parameters of the observed process were updated in single parameter blocks using Metropolis Hastings steps, for example, each of the coefficient of variation parameters, v_{ik}, from the gamma distributions of non-zero rainfall amounts. Therefore, the number of parameter blocks, B, would have been large, necessitating many reduced MCMC runs to estimate the marginal likelihood by Chib's method. This made it less attractive than the power posterior approach. For the latent Gaussian variable model, however, we explained in Section 6.7.1.1 that the power posterior approach is not viable. In contrast, Chib's method actually provides a satisfactory solution as long as we fix each β^i_{1k} to be either 1 or -1. Setting $\beta^i_{1k} = 1$ for all $i = 1, \ldots, n$ and all $k = 1, \ldots, r$ does not seem like an unreasonable simplification to the model since it corresponds to an assumption that the probability of rain following rain at site i in state k exceeds the probability of rain following no rain, which we would intuitively expect. Furthermore, in ordinary MCMC analyses when the β^i_{1k} were allowed to be either 1 or -1, their marginal posteriors generally offered zero or negligible support for $\beta^i_{1k} = -1$; see Section 6.8.3.1, which concerns implementation of the MCMC scheme, for more details. When the β^i_{1k} are fixed, the MCMC scheme (see Section 6.6) is such that θ can be sampled in a modest number of blocks which are all relatively high in dimension. In particular, the only parameters requiring Metropolis Hastings updates are the weather state transition probabilities $(\xi_j, \mathbf{A}^1_j, \mathbf{A}^2_j, \ldots, \mathbf{A}^{27}_j)$, $j \in \mathcal{S}_r$, and for each j these are updated jointly.

For the purposes of marginal likelihood estimation, we chose to partition θ as

$$\theta_1 = (\beta_{01}, \ldots, \beta_{0r})$$
$$\theta_2 = \{(\mu_1, \gamma_1), \ldots, (\mu_r, \gamma_r)\}$$

$$
\begin{aligned}
\theta_3 &= (\xi_1, \mathbf{A}_1^1, \mathbf{A}_1^2, \ldots, \mathbf{A}_1^{27}) \\
&\vdots \\
\theta_{2+r} &= (\xi_r, \mathbf{A}_r^1, \mathbf{A}_r^2, \ldots, \mathbf{A}_r^{27}) \\
\theta_{3+r} &= (\tilde{\phi}_1, \ldots, \tilde{\phi}_r) \\
\theta_{4+r} &= (\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_r^2) \\
\theta_{5+r} &= (\Omega_1, \ldots, \Omega_r) \\
\theta_{6+r} &= \nu \\
\theta_{7+r} &= (\beta_{01}, \ldots, \beta_{0r}, \tilde{\phi}, \tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_r^2, \mu_1, \ldots, \mu_r, \gamma, \Omega_1, \ldots, \Omega_r) \\
\theta_{8+r} &= (\sigma_{\beta_0,1}^2, \ldots, \sigma_{\beta_0,r}^2, \sigma_{\mu,1}^2, \ldots, \sigma_{\mu,r}^2).
\end{aligned}
$$

Following the recommendation of Chib & Jeliazkov (2001), the parameter blocks are arranged so that those of higher dimension appear earlier in the list. Note that the final ordinate $\pi(\theta_{8+r}^* \mid \mathbf{w}, \mathbf{d}, \theta_{1:7+r}^*, \mathbf{x})$ is available directly, since

$$
\begin{aligned}
\pi(\theta_{8+r}^* \mid \mathbf{w}, \mathbf{d}, \theta_{1:7+r}^*, \mathbf{x}) = \prod_{k=1}^{r} \mathrm{IG}\left(\sigma_{\mu,k}^{2*} \middle| h_{0,\beta_0} + \frac{n}{2}, h_{1,\beta_0} + \frac{1}{2}\sum_{i=1}^{n}(\beta_{0k}^{i*} - \beta_{0k}^*)^2\right) \\
\times \prod_{k=1}^{r} \mathrm{IG}\left(\sigma_{\beta_0,k}^{2*} \middle| h_{0,\mu} + \frac{n}{2}, h_{1,\mu} + \frac{1}{2}\sum_{i=1}^{n}(\mu_k^{i*} - \mu_k^*)^2\right),
\end{aligned}
$$

where $\mathrm{IG}(\theta \mid a, b)$ denotes the inverse gamma $\mathrm{IG}(a, b)$ density evaluated at θ. Therefore a reduced run is not necessary for its computation. Moreover, although the auxiliary variables $(\mathbf{s}, \mathbf{s}_0, \mathbf{d}_0, \mathbf{z}_0)$, need to be sampled during the first $6 + r$ MCMC runs, they are not required for estimation of $\pi(\theta_{7+r}^* \mid \mathbf{w}, \mathbf{d}, \theta_{1:6+r}^*, \mathbf{x})$ in the $(7 + r)$-th run. Consequently, estimation of the final two conditional posterior ordinates can be performed very quickly.

When Chib's method is used to estimate marginal likelihoods for hidden Markov models, Section 3.5.1.3 explained that the existence of $r!$ posterior modes (for an r–state model) needs to be accounted for in order to avoid introducing bias in approximations. This is only a consideration for the marginal posterior ordinate $\pi(\theta_1^* \mid \mathbf{w}, \mathbf{d}, \mathbf{x})$ because the conditional posterior distributions $\pi(\theta_i \mid \mathbf{w}, \mathbf{d}, \theta_{1:i-1}^*, \mathbf{x})$ will not, in general, be exchangeable with respect to the hidden state labels, although they might not be unimodal. However, if the states are clearly identified by the conditioning set $\theta_{1:i-1}^*$, multimodality in the conditional posterior distributions is less likely to occur. For the dataset considered in the following section, the parameters $(\beta_{01}, \ldots, \beta_{0r})$ had marginal posteriors which were clearly separated across states, and were therefore chosen to appear first in this list, $\theta_1 = (\beta_{01}, \ldots, \beta_{0r})$. To force exploration of all $r!$ posterior modes in the first (full) MCMC run, which is used to estimate the marginal posterior ordinate $\pi(\theta_1^* \mid \mathbf{w}, \mathbf{d}, \mathbf{x})$, we employed a random permutation sampler (Frühwirth-Schnatter, 2001) in which each draw from the posterior is concluded with a random permutation of the parameter state labels and the hidden state labels; see Section 3.3.5 for more details.

For the NHMM described in this chapter, Chib's estimate of the marginal likelihood, on the log scale, arises by approximating

$$
\log p(\mathbf{w}, \mathbf{d} \mid \theta^*, \mathbf{x}) = \log p(\mathbf{w}, \mathbf{d} \mid \theta^*, \mathbf{x}) + \log \pi(\theta^*) - \log \pi(\theta^* \mid \mathbf{w}, \mathbf{d}, \mathbf{x}).
$$

Estimation of the posterior ordinate $\pi(\theta^* \mid \mathbf{w}, \mathbf{d}, \mathbf{x})$ was explained above and the prior ordinate $\pi(\theta^*)$ can be evaluated directly. Although there is no closed form expression for the observed data likelihood, $p(\mathbf{w}, \mathbf{d} \mid \theta, \mathbf{x})$, it can be approximated at the high density point as follows. Conditionally on the atmospheric data, $\mathbf{x}$, the data from the Y winter periods in the Yorkshire dataset, $\{(\mathbf{w}_{T^{y}+1}, \mathbf{d}_{T^{y}+1}, \mathbf{z}_{0,T^{y}+1}, s_{T^{y}+1}), \ldots, (\mathbf{w}_{T^{y}+1}, \mathbf{d}_{T^{y}+1}, \mathbf{z}_{0,T^{y}+1}, s_{T^{y}+1})\}$ for $y = 1, \ldots, Y$ are modelled as independent realisations of the same NHMM. Therefore the log of the likelihood ordinate can be expressed as

$$\log p(\mathbf{w}, \mathbf{d} \mid \theta^*, \mathbf{x}) = \sum_{y=1}^{Y} \log p(\mathbf{w}_{T^{y}+1:T^{y+1}}, \mathbf{d}_{T^{y}+1:T^{y+1}} \mid \theta^*, \mathbf{x}_{T^{y}+1:T^{y+1}}).$$

For each year, y, the likelihood contribution $p(\mathbf{w}_{T^{y}+1:T^{y+1}}, \mathbf{d}_{T^{y}+1:T^{y+1}} \mid \theta^*, \mathbf{x}_{T^{y}+1:T^{y+1}})$ can be expressed as

$$\begin{aligned} &p(\mathbf{w}_{T^{y}+1:T^{y+1}}, \mathbf{d}_{T^{y}+1:T^{y+1}} \mid \theta^*, \mathbf{x}_{T^{y}+1:T^{y+1}}) \\ &\qquad = \sum_{\mathbf{d}_{0,y}} p(\mathbf{w}_{T^{y}+1:T^{y+1}}, \mathbf{d}_{T^{y}+1:T^{y+1}} \mid \mathbf{d}_{0,y}, \theta^*, \mathbf{x}_{T^{y}+1:T^{y+1}}) p(\mathbf{d}_{0,y}) \end{aligned}$$

where the summation is over all 2^n possible initial rainfall occurrence indicators $\mathbf{d}_{0,y}$ for the y–th year. Note that if there were any missing data in the y–th year, we could simply evaluate the density of the observed data by summing the joint distribution of observed data and missing occurrences over all possible values for the missing rainfall occurrences.

Each density $p(\mathbf{w}_{T^{y}+1:T^{y+1}}, \mathbf{d}_{T^{y}+1:T^{y+1}} \mid \mathbf{d}_{0,y}, \theta^*, \mathbf{x}_{T^{y}+1:T^{y+1}})$ can be approximated using the forward filtering recursion (Algorithm 3.3.1). Dropping the T^{y} notation for clarity, and initialising at $t = 0$ with $\Pr(S_0 = \ell \mid \mathbf{d}_0, \theta^*, \mathbf{x}) = \Pr(S_0 = \ell \mid \theta^*) = \nu_\ell^*$, each density $p(\mathbf{w}, \mathbf{d} \mid \mathbf{d}_0, \theta^*, \mathbf{x})$ is expressed as a product of the normalising constants in the filtered probabilities. The conditional densities of the observables in the filtered probabilities (equation (3.15)) can be written as

$$p(\mathbf{w}_t, \mathbf{d}_t \mid S_t = \ell, \mathbf{d}_{0:t-1}, \mathbf{w}_{1:t-1}, \theta^*, \mathbf{x}) = p(\mathbf{d}_t \mid \mathbf{d}_{t-1}, S_t = \ell, \theta^*) p(\mathbf{w}_t \mid \mathbf{d}_t, \mathbf{d}_{t-1}, S_t = \ell, \theta^*). \tag{6.40}$$

In equation (6.40),

$$p(\mathbf{d}_t \mid \mathbf{d}_{t-1}, S_t = \ell, \theta^*) = \int_{B_t^n} \cdots \int_{B_t^1} \phi_n(\mathbf{Z}_{0t} \mid \mathbf{X}_t \beta_\ell^*, \Sigma_\ell^*) d\mathbf{Z}_{0t} \tag{6.41}$$

in which $\beta_\ell^* = \{(\beta_{0\ell}^*)^T, (\beta_{1\ell}^*)^T\}^T$, $\mathbf{X}_t = \{\mathbf{I}_n, \mathrm{diag}(\mathbf{d}_{t-1})\}$ and the B_t^i were defined in (6.32). Although there is no closed form solution to the integral in (6.41), the multivariate normal probability can be computed numerically. For the application presented in Section 6.8, we used the `pmvnorm()` function in the R package `mvtnorm` (Genz *et al.*, 2010; Genz & Bretz, 2009). This in based on the algorithms of Genz (1992) which transform the integrals into integrals over unit hypercubes, this form being more suitable for numerical integration.

The term $p(\mathbf{w}_t \mid \mathbf{d}_t, \mathbf{d}_{t-1}, S_t = \ell, \theta^*)$ in equation (6.40) can be evaluated directly as

$$p(\mathbf{w}_t \mid \mathbf{d}_t, \mathbf{d}_{t-1}, S_t = \ell, \theta^*) = p_{\widetilde{\mathbf{W}}_t}(\widetilde{\mathbf{w}}_t \mid \mathbf{d}_{t-1}, S_t = \ell, \theta^*)$$

where $\widetilde{\mathbf{W}}_t$ denotes the subset of $\{W_t^1, \ldots, W_t^n\}$ with corresponding $d_t^i = 1$. If $d_t^i = 1$ for all $i = 1, \ldots, n$ then $\widetilde{\mathbf{W}}_t = \mathbf{W}_t$ and by marginalising over $\mathbf{Z}_{0t}$ in the joint distribution for

$(\log \mathbf{W}_t, \mathbf{Z}_{0t} \mid S_t, \mathbf{D}_{t-1})$, it can easily be verified that

$$\log \mathbf{W}_t \mid \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, S_t = \ell, \theta^* \sim \mathrm{N}_n(\mu_\ell^* + \gamma_\ell^* \mathbf{X}_t \beta_\ell^* , \Omega_\ell^* + \gamma_\ell^* \Sigma_\ell^* \gamma_\ell^{*T}), \tag{6.42}$$

where $\log \mathbf{W}_t = (\log W_t^1, \ldots, \log W_t^n)$. If, instead, $\sum_{i=1}^n d_t^i = k$ where $1 \leq k < n$, then $\widetilde{\mathbf{W}}_t$ is a proper (non-empty) subset of $\mathbf{W}_t$ and the joint distribution for $\log \widetilde{\mathbf{W}}_t$ has a marginal distribution that is k-dimensional multivariate normal and easily deduced from (6.42). For example, if $k = 1$ and $d_t^1 = 1$ then $\log \widetilde{\mathbf{W}}_t$ would be univariate normal with mean equal to the first component of $\mu_\ell^* + \gamma_\ell^* \mathbf{X}_t \beta_\ell^*$ and variance equal to the first diagonal element of $\Omega_\ell^* + \gamma_\ell^* \Sigma_\ell^* \gamma_\ell^{*T}$. Finally, if $d_t^i = 0$ for all $i = 1, \ldots, n$ then $\widetilde{\mathbf{W}}_t$ is an empty set and $p_{\widetilde{\mathbf{W}}_t}(\tilde{\mathbf{w}}_t \mid \mathbf{d}_{t-1}, S_t = \ell, \theta^*) = 1$.

6.8 Application to Yorkshire winter rainfall data

In this section we illustrate application of the model and inferential procedures through an analysis of the Yorkshire winter dataset. Interest lies in the joint posterior distribution of the model parameters, the weather states *and* the number of states, r. To identify the model (that is, the number of states it contains) from which the different parameters and hyperparameters arise, we adopt the convention from earlier chapters of attaching r as their first subscript.

This section begins by explaining our prior specification for (r, θ_r). We then use Chib's extended method to estimate the posterior distribution for r and compare the log marginal likelihoods to those obtained for the models in Chapters 4 and 5. This is followed by a discussion of the convergence and mixing problems that arose during MCMC sampling from the posterior distributions $\pi(\theta_r, \mathbf{s}, \mathbf{s}_0, \mathbf{d}_0, \mathbf{z}_0 \mid \mathbf{w}, \mathbf{d}, \mathbf{x}, r)$ where $r = 1, \ldots, r_{\max}$. Next we present summaries of the posterior distributions for the parameters and weather states in the model with $\hat{r}$ states, where $\hat{r}$ is the mode of the posterior distribution for r. The section concludes with an assessment of the fit of the model, comparing the posterior predictive distributions for various test quantities to corresponding observed statistics. For this final chapter, we also perform an out-of-sample assessment in which observed test quantities for 6 years of data not used in model fitting are compared to their posterior predictive distributions.

6.8.1 Prior specification

As in Chapters 4 and 5, it was initially our intention to consider models with $r = 1, \ldots, r_{\max}$ states where $r_{\max} = 5$. However, non-convergence of the MCMC chain for the model with $r = 5$ states ultimately dictated the choice $r_{\max} = 4$. Such MCMC problems will be explored in detail in Section 6.8.3.1. We continued to adopt a truncated Poisson Po(3) prior for r, now truncated to the set $\{1, 2, 3, 4\}$. Note, however, that the posterior was very insensitive to this distributional choice, once the sample space for r was decided.

Conditional on each value of r, the parameters of the weather state process, $\theta_{r,\mathrm{hid}}$, were given an identical prior specification to that outlined in Section 5.7.1 for the application of the NHMM from Chapter 5 to the Yorkshire dataset. In common with earlier chapters, the prior specified for the parameters of the observed process, $\pi(\theta_{r,\mathrm{obs}} \mid r)$ was exchangeable with respect to

the weather state labels, and identical for each value of r. In general, eliciting the prior was difficult because the latent Gaussian variables, $\mathbf{Z}_0$, are not observable. In addition, the MCMC problems which prohibited convergence of the model with $r = 5$ states were also manifest in models with fewer states, although to a lesser extent. This meant parts of our prior specification were influenced by computational considerations, in order to facilitate convergence of the MCMC chains. Of course, this is not ideal, and finding reparameterisations or modifications to the model which would make the posterior better behaved is an area for future work. In the remainder of this section we detail our choice of hyperparameters in the priors, $\pi(\theta_{r,\text{obs}} \mid r)$. Where possible, we try to indicate the reasoning behind our specifications.

The prior chosen for the $\beta^i_{r,0k}$ was pivotal in determining the convergence of the MCMC chains. As we explain in Section 6.8.3.1, it was necessary to make the prior very concentrated and shorter-tailed than its original form, in the latter case, by fixing the variance parameters $\sigma^2_{r,\beta_0,k}$ in the hierarchical priors, rather than modelling them as inverse gamma random variables. After much experimentation, we arrived at the following hyperparameter specification in the prior for the $\beta^i_{r,0k}$,

$$a_{r,0,\beta_0} = -0.5, \qquad a^2_{r,1,\beta_0} = 0.19, \qquad \sigma^2_{r,\beta_0,k} = 0.01,$$

leading to a prior in which, marginally, $\mathrm{E}(\beta^i_{r,0k} \mid r) = -0.5$, $\mathrm{Var}(\beta^i_{r,0k} \mid r) = 0.2$ and $\mathrm{Corr}(\beta^i_{r,0k}, \beta^j_{r,0k} \mid r) = 0.95$ for each site, $i = 1, \ldots, n$, or each pair of sites, in every weather state, $k \in \mathcal{S}_r$. To understand the reasoning behind the choice of prior mean, consider the marginal probability of rain at site i in weather state k,

$$\begin{aligned}\Pr(D^i_t = 1 \mid D^i_{t-1} = d^i_{t-1}, S_t = k, \theta_{r,\text{obs}}, r) &= \Pr(Z^i_{0t} > 0 \mid D^i_{t-1} = d^i_{t-1}, S_t = k, \theta_{r,\text{obs}}, r) \\ &= \Phi\left(\frac{\beta^i_{r,0k} + \beta^i_{r,1k} d^i_{t-1}}{\sqrt{\Sigma^{ii}_{r,k}}}\right), \end{aligned} \tag{6.43}$$

where $\Sigma^{ii}_{r,k}$ is the i-th diagonal element in $\mathbf{\Sigma}_{r,k}$. For a typical winter day in Yorkshire, it seems reasonable to think that the probability of rain following a dry day should be less than 1/2 whilst the probability of rain following a wet day should be greater than 1/2. From equation (6.43), this would be the case if $\beta^i_{r,0k} < 0$ and $\beta^i_{r,0k} + \beta^i_{r,1k} > 0$. The $\beta^i_{r,1k}$ are constrained to be either -1 or 1, and so, because we expect that $\beta^i_{r,1k} = 1$, taking $a_{r,0,\beta_0} = -0.5$ ties in with our beliefs about the probabilities of rain. Note that, for a simple probit model where $\Pr(D = 1 \mid \beta) = \Phi(\beta)$, if β is given a normal prior, the prior distribution for the probability $\Pr(D = 1 \mid \beta)$ becomes U-shaped as soon as the variance in the prior for β exceeds 1. This can easily be demonstrated by determining the nature of the turning points in the induced prior density function for $\Pr(D = 1 \mid \beta)$. Therefore, a marginal prior variance of 1.0 for the $\beta^i_{r,0k}$ had originally seemed sensible. Unfortunately, this value had to be reduced to moderate the convergence problems during MCMC sampling.

In the hierarchical prior for the $\beta^i_{r,1k}$ we chose

$$b_{r,0,\beta_1} = 0.2, \qquad b_{r,1,\beta_0} = 0.05,$$

so that, marginally, $\mathrm{E}(\beta^i_{r,1k} \mid r) = 0.6$, $\mathrm{Var}(\beta^i_{r,1k} \mid r) = 0.64$ and $\mathrm{Corr}(\beta^i_{r,1k}, \beta^j_{r,1k} \mid r) = 0.8$ for each site, $i = 1, \ldots, n$, or each pair of sites, in every weather state, $k \in \mathcal{S}_r$. The prior specification

above was chosen to express the belief that $\beta^i_{r,1k} = 1$ is more likely than $\beta^i_{r,1k} = -1$, and that the $\beta^i_{r,1k}$ are highly positively correlated, although still allowed to differ.

To help in eliciting the prior for $(\mu_{r,k} \mid r)$, we found it helpful to think of the conditions under which $\mu^i_{r,k}$ would correspond to a more interpretable parameter. Recall that the prior for $(\mu_{r,k} \mid r)$ treats sites exchangeably. At site i, consider hypothetical days associated with latent variables $S_t = k$ and $Z^j_{0t} = 0$ for all $j \neq i$. As Z^i_{0t} tends to zero (from above), $\mu^i_{r,k}$ corresponds to the mean log rainfall amount at site i in weather state k. We can therefore think of $\mu^i_{r,k}$ as the mean log rainfall amount on days of indeterminate weather which are just on the threshold of rainfall occurrence. We thought that the median rainfall amount on such days should be small, say 0.55 mm, and because of the symmetry of the normal distribution, this corresponds to a mean log rainfall amount of $\log 0.55 = -0.6$. The hyperparameters chosen in the hierarchical priors for the $\mu^i_{r,k}$ were

$$a_{r,0,\mu} = -0.6, \qquad a^2_{r,1,\mu} = 2.0, \qquad h_{r,0,\mu} = 2.1, \qquad h_{r,1,\mu} = 0.1155$$

leading to marginal means of -0.6, marginal variances of just over 2 and marginal correlations between sites equal to 0.950. This tallies with our beliefs about the mean, above, and indicates that we think the $\mu^i_{r,k}$ are highly correlated between sites.

The i-th row in the matrix $\gamma_{r,k}$ comprises coefficients in the regression of $\log W^i_t$ on the latent variables Z^j_{0t}, conditional on the weather state, $S_t = k$. However, it is very difficult to think about likely values for these coefficients because we have little understanding about the scale of the latent variables, Z^j_{0t}. To assist in prior elicitation, therefore, we considered regressing the $\log W^i_t$ on $\mathbf{Z}^*_{0t} = \mathbf{S}^{-1}_{r,k}\mathbf{Z}_{0t}$ where $\mathbf{S}_{r,k}$ is a diagonal matrix of standard deviations, specifically, the square roots of the diagonal elements in $\mathbf{\Sigma}_{r,k}$. The advantage of this modification is that the variables $\mathbf{Z}^*_{0t}$ have unit scale. Unfortunately, rescaling is achieved at the expense of complicating the likelihood, so that the priors for $\tilde{\phi}_{r,k}$ and the $\tilde{\sigma}^i_{r,k}$ would no longer be semi-conjugate. As such, we did not adopt this reparameterisation.

In practice, the mean hyperparameter, $\mathbf{m}_{r,\gamma,0}$, in the hierarchical prior for $(\gamma_{r,1}, \ldots, \gamma_{r,r} \mid r)$ was chosen so that on-diagonal elements in the coefficient matrices had means equal to 1.0 and off-diagonal elements had means equal to 0.0. This was based on the idea that a large value of Z^i_{0t}, corresponding to a day with a high probability of rain at site i, would be suggestive of a large rainfall amount at site i, if it rained. However, we were indifferent to whether an increase in Z^j_{0t}, $j \neq i$, would increase or decrease the log rainfall amount at site i. Therefore it did not seem unreasonable to specify 0's and 1's for the off- and on-diagonal elements, respectively, in the mean, $\mathbf{m}_{r,\gamma,0}$. Our prior specification was then completed by taking $\mathbf{V}_{r,\gamma,0} = (1 - \rho_{r,\gamma})\mathbf{V}_{r,\gamma,0}$ and $\mathbf{C}_{r,\gamma,0} = \rho_{r,\gamma}\tilde{\mathbf{V}}_{r,\gamma,0}$, where $\rho_{r,\gamma} = 0.999$, whilst $\tilde{\mathbf{V}}_{r,\gamma,0}$ was chosen so that, for any matrix $\gamma_{r,k}$, the marginal variances of both on and off-diagonal elements were 2.0, the marginal correlations amongst on-diagonal elements and amongst off-diagonal elements were 0.95, and the on- and off- diagonal elements were uncorrelated. The specification $\rho_{r,\gamma} = 0.999$ meant that, *a priori*, the correlation between the (i,j)-th elements of $\gamma_{r,k}$ and $\gamma_{r,\ell}$ was 0.999. Such high correlations were necessary to ensure parameter identifiability in the posterior. However, they did not prevent the data from informing the posterior of a difference amongst certain $\gamma^{ij}_{r,1}, \ldots, \gamma^{ij}_{r,r}$, in particular, those for which $i = j$. For example, conditional on $r = 4$, the posterior means (standard deviations) for $\gamma^{44}_{4,1}$ and $\gamma^{44}_{4,3}$ were 0.119 (0.059) and 0.242 (0.071), respectively. Moreover, when

we assumed $\gamma_{r,1} = \cdots = \gamma_{r,r}$, or equivalently, $\rho_{r,\gamma} = 1.0$, the marginal likelihoods for models with $r \geq 2$ states were smaller than those for the corresponding models where $\rho_{r,\gamma} = 0.999$, for example, -27930.84 compared with -27902.27 for models with $r = 4$ states.

Our strategy for handling the non-identifiability problem in MVP models initially seemed to offer advantages in terms of prior elicitation because it allowed us to assess a prior for a full variance matrix, rather than a correlation matrix. In the case of observable data, Germain *et al.* (2010*b*) outlined a prior elicitation strategy for the slope coefficients and conditional variances arising from the modified Cholesky decomposition of the precision matrix in a multivariate normal distribution. However, applying this strategy in the prior specifications for $(\tilde{\phi}_{r,1}, \ldots, \tilde{\phi}_{r,r} \mid r)$ and $(\tilde{\sigma}^2_{r,1}, \ldots, \tilde{\sigma}^2_{r,r} \mid r)$ is substantially complicated by the fact that the variables $\mathbf{Z}_{0t}$ are not observable and, as remarked above, we do not have a good understanding of their scale. The latent variables $\mathbf{Z}_{0t}$ are related to the observed log (non-zero) rainfall amounts through the regression of each $\log W^i_t$ on $\mathbf{Z}_{0t}$. We chose to elicit the priors for each pair $(\tilde{\phi}_{r,k}, \tilde{\sigma}^2_{r,k} \mid r)$ by supposing that the $\mathbf{Z}_{0t}$ were log non-zero daily rainfall amounts. Elicitation then followed along the same lines as the example in Germain *et al.* (2010*b*), where we specified a prior for log monthly rainfall totals at the same network of Yorkshire sites. Space does not permit a full explanation, but an outline of this elicitation procedure is as follows. The first stage involved choosing a tentative estimate, $\boldsymbol{\Sigma}_0$, of the variance matrix, $\boldsymbol{\Sigma}_{r,k}$, for the hypothetical log non-zero rainfall amounts in state k. This was used to reorder the variables in $\mathbf{R}_t = (\mathbf{Z}_{0t} - \mathbf{X}_t\boldsymbol{\beta}_{r,k})$ by successively choosing the site R^i_t which, under the tentative model $\boldsymbol{\Sigma}_{r,k} = \boldsymbol{\Sigma}_0$, minimised the variance of the remaining sites R^j_t the most. The first site in the reordered list was therefore chosen to be the site which we believed to be the most informative about the others. This procedure led to a matrix $\mathbf{M}$, given in Appendix D, for reordering the sites according to $\mathbf{Q}_t = \mathbf{M}\mathbf{R}_t$. The choices for the marginal prior means of the slope coefficients $\tilde{\phi}_{r,k,i}$ and of the conditional variances $\tilde{\sigma}^2_{r,k,i}$ were based on a transformation of the tentative estimate $\boldsymbol{\Sigma}_0$, then the marginal prior variances were chosen to reflect confidence in these estimates. Finally, the marginal prior correlations amongst the $\tilde{\phi}_{r,k,ij}$ were chosen in a manner which induced positive *a priori* correlations between covariances, $\Sigma_{r,k,ij}$ and $\Sigma_{r,k,\ell m}$, associated with pairs of sites, (i, j) and (ℓ, m), separated by similar distances.

The parameter $\rho_{r,\tilde{\phi}}$, representing the marginal correlation between $\tilde{\phi}_{r,k,ij}$ and $\tilde{\phi}_{r,\ell,ij}$, $k \neq \ell$, was chosen to be reasonably large, $\rho_{r,\tilde{\phi}} = 0.95$, to assist parameter identifiability in the posterior. As a result of applying the above procedure, the hyperparameters in the gamma prior for each $\tilde{\sigma}^2_{r,k}$ were set at $c_{r,0,\tilde{\sigma}^2} = 0.15$ and $c_{r,1,\tilde{\sigma}^2} = 0.15$. The choices for the other hyperparameters, $\mathbf{m}_{r,\tilde{\phi},0}$, $\mathbf{C}_{r,\tilde{\phi},0}$, $\mathbf{V}_{r,\tilde{\phi},0}$, $(C_{r,1}, \ldots, C_{r,n})$ and $(v_{r,\tilde{\sigma}^2,1}, \ldots, v_{r,\tilde{\sigma}^2,n})$ are provided in Appendix D. Although the strategy outlined above is not entirely satisfactory, developing elicitation methods in complicated latent variable models is a difficult problem in which further research would be beneficial.

From the conditional model specification

$$\log W^i_t \mid D^i_t = 1, S_t = k, \mathbf{Z}_{0t}, \boldsymbol{\theta}_{r,\text{obs}}, r \sim \text{N}\left(\mu^i_{r,k} + \sum_{j=1}^{n} \gamma^{ij}_{r,k} z^j_{0t}\,,\ \Omega^{ii}_{r,k}\right) \tag{6.44}$$

it is clear that $\Omega^{ii}_{r,k}$ is a conditional variance parameter. For the purposes of prior elicitation, suppose that the latent variables S_t and $\mathbf{Z}_{0t}$ were observed. For each site, independently, equation (6.44) would then represent a multiple linear regression of the log non-zero rainfall amounts

on covariates $Z^1_{0t}, \ldots, Z^n_{0t}$. In this case the least squares estimate of $\Omega^{ii}_{r,k}$ would be the sum of squared residuals at site i in weather state k, divided by $T^1_{ik} - n - 1$, where T^1_{ik} is the number of wet days for this site/weather state pair. This quantity is related to the amount of variation in log rainfall amounts that is not explained by the regression on $\mathbf{Z}_{0t}$, and we can think of $\mathbf{Z}_{0t}$ as a measure of the general propensity for rain at the sites. Therefore, in choosing the prior means, $\mathrm{E}(\Omega^{ii}_{r,k}) = c_{r,0,\Omega}/c_{r,1,\Omega}$, for $\Omega^{ii}_{r,k}$, we found it helpful to start by thinking about the unconditional variances of log (non–zero) rainfall amounts. To this end, we judged that a value of 0.55 would be appropriate, and that approximately 1/2 of this variance could not be explained by regression on $\mathbf{Z}_{0t}$. This led to a point estimate of $\Omega^{ii}_{r,k}$ equal to 0.275, and so we originally chose $\mathrm{E}(\Omega^{ii}_{r,k}) = 0.275$. However, as we explain in Section 6.8.3.1, to ensure convergence of the MCMC chains, we had to adapt the prior for the $\Omega^{ii}_{r,k}$ so that less mass was assigned close to 0. Therefore we eventually selected the following values for the hyperparameters

$$v_{r,\Omega} = 0.2, \qquad c_{r,0,\Omega} = 0.2, \qquad c_{r,1,\Omega} = 0.3$$

giving prior expectations, variances and correlations between sites of 0.666, 2.32 and 0.954, respectively.

6.8.2 Posterior inference for r

In this and subsequent sections we make references to models from earlier chapters. For notational brevity, we refer to the latent Gaussian variable NHMM from this chapter by the acronym LG–NHMM. The simple hidden Markov model from Chapter 4, based on assumptions of conditional independence in space and time, is referred to as the CI–HMM, and the NHMM from Chapter 5, based on a Markov chain of autologistic models, is referred to as the MCA–NHMM.

The Yorkshire dataset includes two years that contain missing values. For the MCA–NHMM, when estimating the marginal likelihood using power posteriors, we were unable to find a satisfactory means of handling these missing data and so chose to estimate the marginal likelihood for the 28 years of complete data only. Section 6.7.2.1 explained how Chib's method can handle missing data in the marginal likelihood calculation for the LG–NHMM. However, for comparability with the MCA–NHMM, we, again, compute the marginal likelihood using only the 28 years of complete data.

The high density points for models with $r = 1, \ldots, 4$ states, $\boldsymbol{\theta}^*_r$, were taken to be the posterior means from an initial MCMC run of length 500,000 iterations after a burn-in comprising the same number of samples. The full and reduced MCMC runs needed to estimate the marginal and conditional posterior ordinates each consisted of 250,000 posterior draws, thinning to every 10–th iterate to reduce computational overheads.

In approximating the log of the marginal likelihood by Chib's method, we can think of the numerical standard error as the variation that would be expected in the estimate if the simulation were to be repeated. Chib & Jeliazkov (2001) suggest a method for estimating the numerical standard error in the approximation of the log posterior ordinate. This is based on an estimate of the sample variance matrix of the summands in the approximation of each marginal/conditional posterior ordinate, for example, $\{\pi(\theta^*_{r,1} \mid \mathbf{w}, \mathbf{d}, \boldsymbol{\theta}^{[g]}_{r,2:8+r}, \mathbf{s}^{[g]}, \mathbf{s}^{[g]}_0, \mathbf{d}^{[g]}_0, \mathbf{z}^{[g]}_0, r, \mathbf{x})\}$ for $g = 1, \ldots, M$

r	1	2	3	4
Log marginal likelihood	-28726.68	-28373.44	-27927.40	-27902.27
Numerical standard error	1.11	1.12	1.57	1.69
Posterior probability	0.00	2.36×10^{-205}	1.22×10^{-11}	1.00

Table 6.1: Estimates of the log marginal likelihood, the numerical standard error of the log posterior ordinate estimate and the posterior distribution for r for the 28 years in the Yorkshire dataset with no missing values. The estimates of the log marginal likelihoods were computed using the extended version of Chib's method.

whose mean value provides an estimate of the marginal ordinate, $\pi(\theta^*_{r,1} \mid \mathbf{w}, \mathbf{d}, r, \mathbf{x})$. The proposed estimate of the sample variance matrix adjusts for autocorrelation in the series of summands. For full details, see Chib & Jeliazkov (2001). However, this only accounts for the standard error in the approximation of the log posterior ordinate. It does not account for the fact that the log likelihood ordinates were themselves only approximated. To fully quantify the numerical standard error in the estimate of the log marginal likelihood, therefore, the variances of the approximations to the log likelihood ordinates should also be incorporated. Unfortunately, it is difficult to estimate the variance of the log likelihood ordinate for this model because of the way in which it is approximated in a forward recursion. However, we can gauge, roughly, the error in the estimates of the log likelihoods by repeating the estimation procedure multiple times and assessing the variability of the estimates.

Table 6.1 presents estimates of the log marginal likelihoods together with the estimated posterior distribution for r. It also displays the numerical standard errors in the estimates of the log posterior ordinates, calculated according to the method outlined above. For each value of r, we assessed the variability in the log likelihood estimates by repeating the calculations five times. The log likelihood estimates generally agreed to two decimal places, with the maximum difference being 0.01. It would therefore appear that the variation arising from the approximation of the likelihood is negligible in comparison to that from the approximation of the posterior ordinate. This, in turn, is very small in comparison to the differences between the log marginal likelihood estimates for models with different numbers of states.

From Table 6.1 it is clear that virtually all the posterior mass lies at $r_{\max} = 4$. As in previous chapters this is likely to be because the hidden Markov model is only a simple approximation of the very complex data generating mechanism. Therefore it could be that increasing the number of states from $r = 1$ to $r = 4$ substantially increases the likelihood by compensating for the disparities between the shape of the "actual" distribution of rainfall and that allowed under the models with fewer states. Figure 6.8 displays the estimated log marginal likelihoods together with the corresponding values for the CI-HMM and the MCA-NHMM. The CI-HMM assumed that both rainfall occurrences and rainfall amounts (given occurrences) were conditionally independent in time and space, given the weather state. The MCA-NHMM upheld this assumption for rainfall amounts, but it was relaxed for rainfall occurrences, allowing them to follow a Markov chain of autologistic models, conditional on the weather state. The LG-NHMM relaxed the conditional independence assumption for *both* rainfall occurrences and amounts. Therefore, if we compare log marginal likelihoods for the version of each of the three models with $r = 1$ state, it is not surprising that the most complex model, the LG-NHMM, has the largest marginal

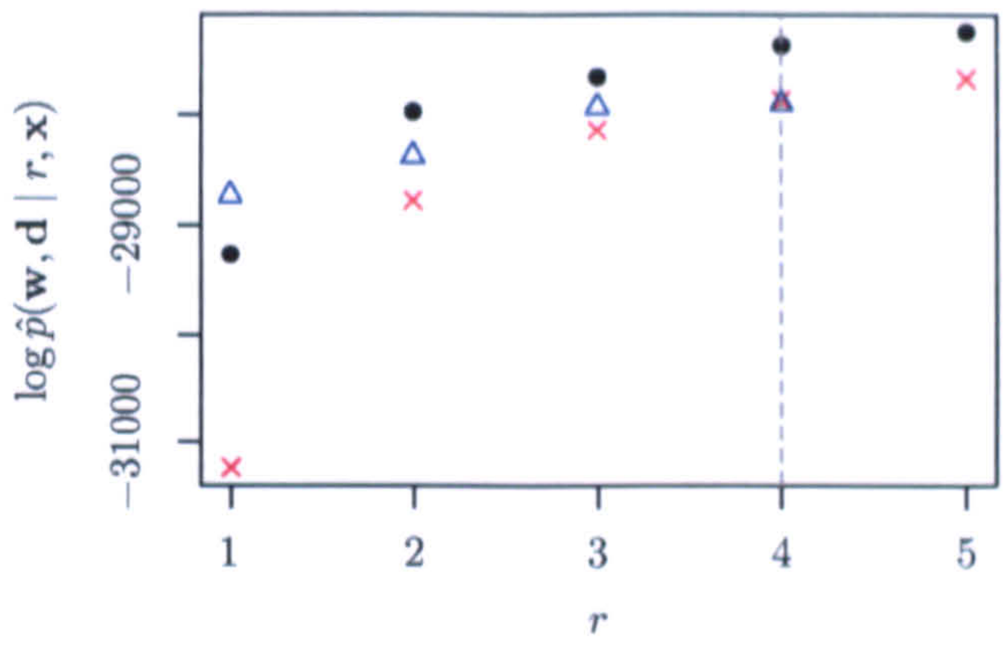


Figure 6.8: Estimates of the log marginal likelihood for the Yorkshire data calculated using Chib's extended method for the LG–NHMM (△). Also shown are the estimates calculated using the power posterior approach for the MCA–NHMM (•) and the CI–HMM (×).

likelihood. However, comparing models with $r = 4$ states, the log marginal likelihood for the LG–NHMM coincides with that for the simple CI–HMM. If we were to discriminate between models purely on the basis of their marginal likelihoods, therefore, we would have no reason to favour the 4–state LG–NHMM over the 4–state CI–HMM. Given that the LG–NHMM is a substantially more complex model than the CI–HMM, we might consider this to mean that any improvement in fit offered by the LG–NHMM does not compensate for the increase in model complexity. However, this overstates the interpretation of the marginal likelihood as a means of trading off model fit and model complexity. It is important to appreciate that the marginal likelihood compares model and prior *combinations*, and so we could only begin to draw such conclusions if our priors for the two models conveyed information that was, in some sense, equivalent, perhaps in terms of the prior predictive distribution for rainfall on a single day.

In Section 6.8.1 we remarked that, for the LG–NHMM, convergence of the MCMC sampler could only be achieved by making the priors for certain parameters very concentrated. In particular, this was true of the $\boldsymbol{\beta}_{r,0k}$. To help in understanding the effect this had on the marginal likelihood, it is helpful to consider the value of the prior ordinate evaluated at the high density point in Chib's estimator. Recall that Chib's estimator of the log marginal likelihood for an r–state model is

$$\log p(\mathbf{w}, \mathbf{d} \mid r, \mathbf{x}) = \log \pi(\boldsymbol{\theta}_r^* \mid r) + \log p(\mathbf{w}, \mathbf{d} \mid \boldsymbol{\theta}_r^*, r, \mathbf{x}) - \log \pi(\boldsymbol{\theta}_r^* \mid \mathbf{w}, \mathbf{d}, r, \mathbf{x}),$$

and so decreasing the value of the prior ordinate, whilst keeping the likelihood and posterior ordinates unchanged, decreases the marginal likelihood. The contribution to the log prior ordinate from the parameters $(\boldsymbol{\beta}_{4,01}, \ldots, \boldsymbol{\beta}_{4,04}, \beta_{4,01}, \ldots, \beta_{4,04})$ in the $r = 4$ state model was given by

$$\begin{aligned}
&\log\{\pi(\boldsymbol{\beta}_{4,01}^*, \ldots, \boldsymbol{\beta}_{4,04}^*, \beta_{4,01}^*, \ldots, \beta_{4,04}^*)\} \\
&\qquad = \sum_{k=1}^{4} \left[\log\{\phi_n(\boldsymbol{\beta}_{4,0k}^* \mid \boldsymbol{\beta}_{4,0k}^*, \sigma_{4,\beta_{0k},k}^2)\} + \log\{\phi_1(\beta_{4,0k}^* \mid a_{4,0,\beta_0}, a_{4,1,\beta_0}^2)\}\right] \\
&\qquad = -14.58
\end{aligned}$$

meaning the high density point (posterior mean) was very unlikely under the prior. Contributions from most other parameter blocks were positive. The prior means and variances for each (variable mean) parameter $\beta_{4,0k}$ were -0.5 and 0.19, respectively. In the wettest and driest states in the $r = 4$ state model, the high density points for the $\beta_{4,0k}$ were 1.47 and -1.80, respectively. This illustrates that, for these two "extreme" states, the prior assigns very little support to the values of $\beta_{4,0k}$ championed by the posterior. The marginal likelihood is simply the expectation of the likelihood with respect to the prior and so, in some dimensions, the values of the model parameters $\boldsymbol{\theta}_4$ which produced large likelihood values were heavily downweighted by the prior. This likelihood–prior conflict will certainly have contributed to the relatively small log marginal likelihood values for the LG–NHMM. We explore this issue further in Section 6.9 after taking into consideration the posterior predictive performance of the model.

6.8.3 Posterior inference for $(\boldsymbol{\theta}_r, \mathbf{s} \mid r)$ using MCMC samples

In this section, we begin by discussing implementation of the MCMC scheme used to obtain posterior samples from models with $r = 1, \ldots, r_{\max}$ states, exploring the convergence and mixing problems that were experienced. We then present summaries of the posterior distribution for the model with $r = \hat{r} = 4$ states where $\hat{r}$ is the mode of the posterior distribution for r.

6.8.3.1 Implementation, convergence and mixing

Fixing the number of states at $r = 1, \ldots, 4$, in turn, the MCMC algorithm was used to generate 2,500,000 draws from the posterior, omitting the first 500,000 as burn-in and thinning the remaining output to retain every 200-th iterate. This gave posterior samples of size $N = 10000$. The usual graphical diagnostic checks were employed to inspect the convergence and mixing properties of the chains.

In earlier MCMC runs, using a more diffuse prior specification than that detailed in Section 6.8.1, the MCMC chains for models with $r = 3$ and $r = 4$ states failed to converge. For $r \geq 2$, the model always identified one state associated with dry conditions at all sites, and another associated with wet conditions at all sites. As the number of states increased, these wet and dry states came to represent more clear-cut versions of their namesake conditions, for example, the marginal probabilities of rain in the wet state moved closer to 1. Consequently, in these extreme states, we only observed data from a small part of its total sample space. This made it difficult to identify the parameters associated with these states in the likelihood. For models where $r = 3$ and $r = 4$, two specific problems arose; one for the parameters in the wet state, and another for the parameters in the dry state.

In these earlier runs (which failed to converge), by partitioning the data according to the marginal posterior mode estimate of the weather state sequence, $\hat{\mathbf{s}}$, we found that the wet state in the $r = 3$ and $r = 4$ state model was such that rain was observed on almost all days at some sites. Consider equation (6.43), which gives the expression for the probability of rain at site i, given that $S_t = k$ and $D^i_{t-1} = d^i_{t-1}$. In the extreme case when the data suggest that it rains on all of the days, t, allocated to state k at site i when $D^i_{t-1} = d$, the likelihood would be maximised by this probability taking the value 1. This can be achieved in the limit as $(\beta^i_{r,0k} + \beta^i_{r,1k} d)/\sqrt{\Sigma^{ii}_{r,k}} \to \infty$.

The $\beta^i_{r,1k}$ are constrained to belong to the set $\{-1, 1\}$ whilst prior correlations between the variance matrices $\Sigma_{r,1}, \ldots, \Sigma_{r,r}$ should discourage the diagonal elements from tending to zero. Therefore, unless $\beta^i_{r,0k}$ has a prior that makes large values very unlikely, its posterior will have substantial density at arbitrarily large values of $\beta^i_{r,0k}$ in order to support the likelihood's bid to increase the probability of rain. Consequently, when the prior for each $\beta^i_{r,0k}$ was not very concentrated about its mean, -0.5, trace plots showed some of the $\beta^i_{r,0k}$ parameters increasing without bound over the course of the MCMC run. To prevent this from happening we had to reduce our initial choices for the prior variances of the $\beta^i_{r,0k}$, in addition to shortening the tails of the prior by fixing the variance parameters $\sigma^2_{r,\beta_0,k}$, $k = 1, \ldots, r$, in their hierarchical priors.

In the dry weather state, for models with $r = 4$ states, a different problem arose. Consider the conditional distribution for non–zero rainfall amounts at site i in state k,

$$W^i_t \mid D^i_t = 1, S_t = k, \mathrm{Z}_{0t}, \theta_{r,\mathrm{obs}} \sim \mathrm{logN}\left(\mu^i_{r,k} + \sum_{j=1}^{n} \gamma^{ij}_{r,k} z^j_{0t}\,,\ \Omega^{ii}_{r,k}\right).$$

In the dry weather state on days when rain was observed, typically these rainfall amounts were small and displayed little variation. Consequently, the variation in the Z^j_{0t} could explain a substantial proportion of the variation in the W^i_t. Therefore, when the prior for the $\Omega^{ii}_{r,k}$ was chosen to be relatively diffuse, zero values were not implausible in the posterior, leading to the $\Omega^{ii}_{r,k}$ tending to zero in the dry state. To avoid the associated numerical problems, therefore, we had to modify our initially chosen prior so that it assigned less density near zero.

The problems described in the preceding paragraphs ultimately forced us to set $r_{\max} = 4$, rather than $r_{\max} = 5$, which had been the limit chosen in earlier chapters. Specifically, for the $r = 5$ state version of the LG–NHMM, even after making the modifications above, the problems persisted and the MCMC did not converge. For models with $r \leq 4$ states, however, these adjustments seemed to remove the MCMC problems just described. Convergence was assessed by initialising chains at a variety of starting points and comparing trace plots from the different runs. For models with $r = 1$ and $r = 2$ states, the various runs produced essentially the same results. The same was true for models with $r = 3$ and $r = 4$ states when each $\beta^i_{r,1k} \in \{-1, 1\}$ was initialised at 1. However, when these parameters were initialised at -1, some, associated with the clear–cut wet and dry states, remained at -1, whilst the rest switched to 1 almost immediately, where they stayed for the rest of the MCMC run. It could be argued that this was a sign of non–convergence.

Judging whether an MCMC sample has converged is a problem when analysing any complex model. This is especially true if the posterior is multimodal because the sampler may not have run long enough to establish the mixing weights, or even to visit all of the modes. In our case, the lack of convergence indicated by non–convergent chains, seems to have been caused by a combination of (i) weak likelihood identifiability of some parameters in the wet and dry states and (ii) deficiencies in the MCMC sampler, creating barriers to moving between modes.

Consider a wet weather state, say, state k. Within such wet states there was always one site, say, site i, at which it was wet on most days. This means we make a number of observations that are informative about the probability of rain after rain, but few which are informative about the probability of rain after no rain. Therefore, the value of $\beta^i_{r,1k} \in \{-1, 1\}$, whose sign determines

the relative risk of rainfall occurrence between days following wet versus dry days, will be only weakly identified in the likelihood. Unless the data are very informative that $\beta^i_{r,1k} = 1$ it will be difficult for the sampler to jump from -1 to 1 because, if it were to make this transition, similar large jumps would be necessary for many other parameters. However, precisely because the data tell us little about whether $\beta^i_{r,1k} = \pm 1$, the value of $\beta^i_{r,1k}$ will be relatively unimportant in terms of providing a statistical description of the data. Indeed, the probability density function $p(\mathbf{w}, \mathbf{d}, \mathbf{z}_0, \mathbf{d}_0 \mid \theta_r, r, \mathbf{x})$, which is computed as a by-product of the forward backward algorithm, provides an overall measure of the "likelihood" and trace plots of this combined parameter from runs with different starting points overlapped completely. If parameters are only weakly identified in the likelihood, then it is likely that it will only be their value in *combination* with other parameters that is important in describing the data. Correspondingly, it will be these parameter combinations that are important in producing posterior predictive inferences. In support of this argument, we found that samples from the posterior predictive distributions of various test quantities were insensitive to the starting point of the MCMC sampler used to generate the original MCMC output. For these reasons, we were not concerned about the convergence of the chains analysed in subsequent sections.

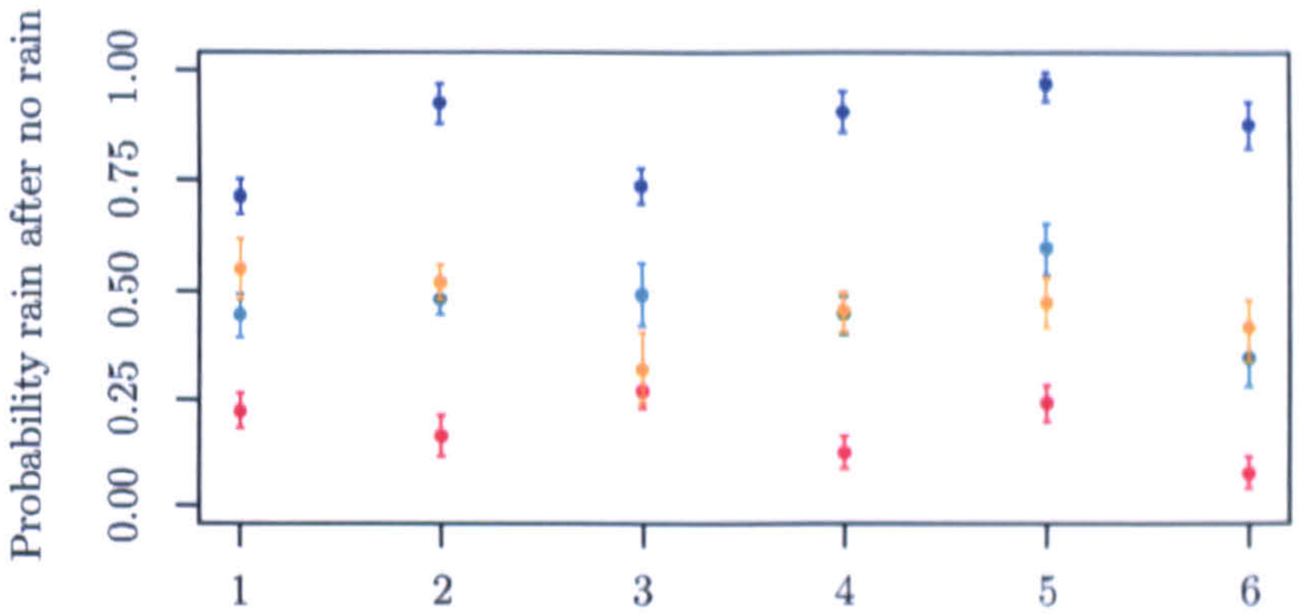
In addition to causing convergence difficulties, weak identifiability of some of the parameters in the likelihood is also likely to have contributed to poor mixing of the chains for models with $r = 3$ and $r = 4$ states. ACF plots of the MCMC output revealed that the autocorrelation for certain parameters, particularly in the clear-cut wet and dry weather states, decayed very slowly. For example, an additional thin to every 40 iterations would have been needed to eliminate the autocorrelation in some series.

As we remarked in Section 6.2.3, the performance of the MCMC sampler can be affected by the way in which the non-identifiability problem is handled in the MVP model. We therefore considered an alternative approach to making the model identifiable, by fixing the conditional variances $\tilde{\sigma}^2_{r,k}$, $k \in \mathcal{S}_r$, at the values selected for their prior means. The restriction on the sample space of the $\beta_{r,1k}$ could then be removed, assigning them multivariate normal priors. However, this led to chains in which the autocorrelation decayed even more slowly and so we did not consider this approach further.

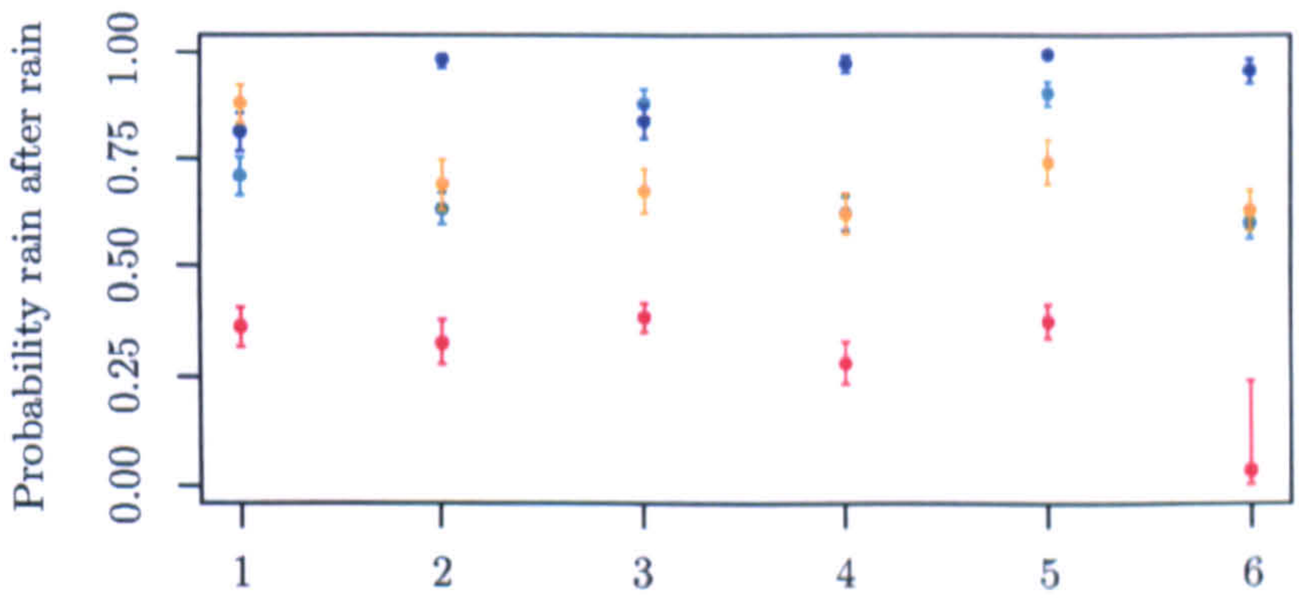
6.8.3.2 Posterior for $(\theta_4 \mid r = 4)$

Conditional on the posterior mode, $r = 4$, Figure 6.9 displays the posterior distribution for the conditional probability of rainfall at each site, given its rainfall status the previous day and the current weather state, that is, the posteriors for the combinations of parameters equal to $\Pr(D^i_t = 1 \mid D^i_{t-1} = d, S_t = k, \theta_{4,\text{obs}}, r = 4)$, where $d = 0, 1$ and $k = 1, \ldots, r$ for each $i = 1, \ldots, n$. The posterior distributions are visualised through their means and 95% equi-tailed Bayesian credible regions.

In Section 6.7.2.1, by showing how the observed data likelihood could be approximated at a single point, we illustrated that the univariate distribution for $(W^i_t \mid D^i_t = 1, \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, S_t = k, \theta_{4,\text{obs}}, r = 4)$ was lognormal with location and scale parameters given by the i-th and (i,i)-th elements of $\mu_{4,k} + \gamma_{4,k}\mathbf{X}_t\beta_{4,k}$ and $\Omega_{4,k} + \gamma_{4,k}\Sigma_{4,k}\gamma^T_{4,k}$, respectively. For each possible value of $\mathbf{d}_{t-1} \in \{0,1\}^n$ and for each state, $k \in \mathcal{S}_4$, the posteriors for the means in these lognormal



(a)



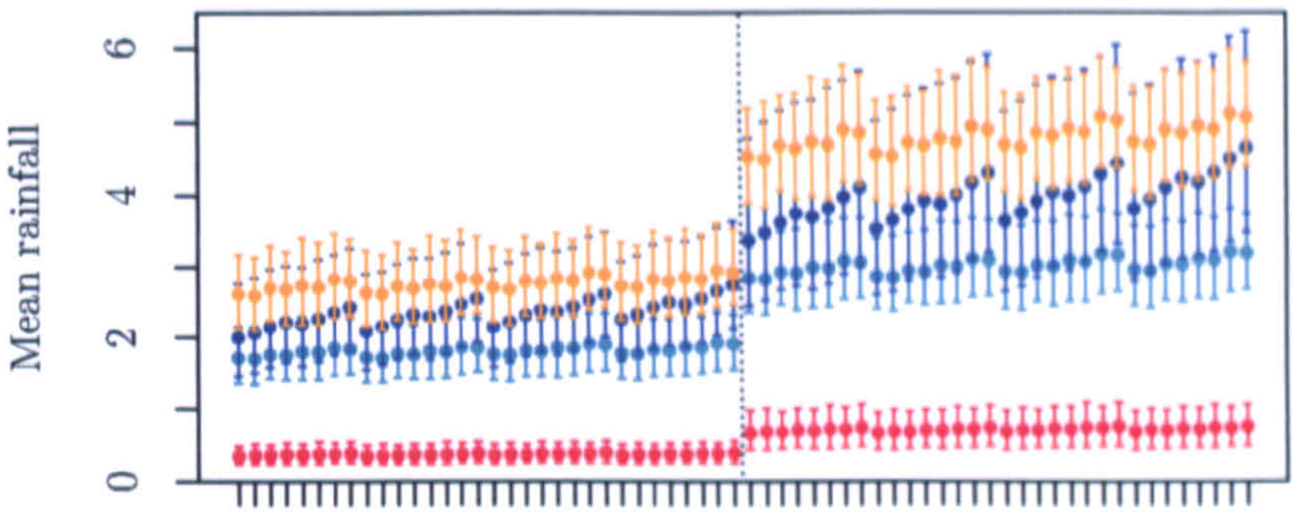
(b)

Figure 6.9: Conditional on $r = 4$, posterior means with 95% equi–tailed Bayesian credible intervals for the probabilities (a) $\Pr(D_t^i = 1 \mid S_t = k, D_{t-1}^i = 0, \boldsymbol{\theta}_{4,\text{obs},k}, r = 4)$ and (b) $\Pr(D_t^i = 1 \mid S_t = k, D_{t-1}^i = 1, \boldsymbol{\theta}_{4,\text{obs},k}, r = 4)$ at all sites, $i = 1, \ldots, 6$ in weather states $k = 1$ (———), 2 (———), 3 (———) and 4 (———).

distributions are displayed in Figure 6.10 for sites 1–3 and Figure 6.11 for sites 4–6. At each site, the numerical labellings of the 2^n possible values of $\mathbf{D}_{t-1}$, $\mathcal{I}(\mathbf{d}_{t-1})$, are ordered so that the first 2^{n-1} posteriors, to the left of the dotted line, correspond to no rain at the site in question on the previous day, and conversely for the posteriors to the right of the dotted line.

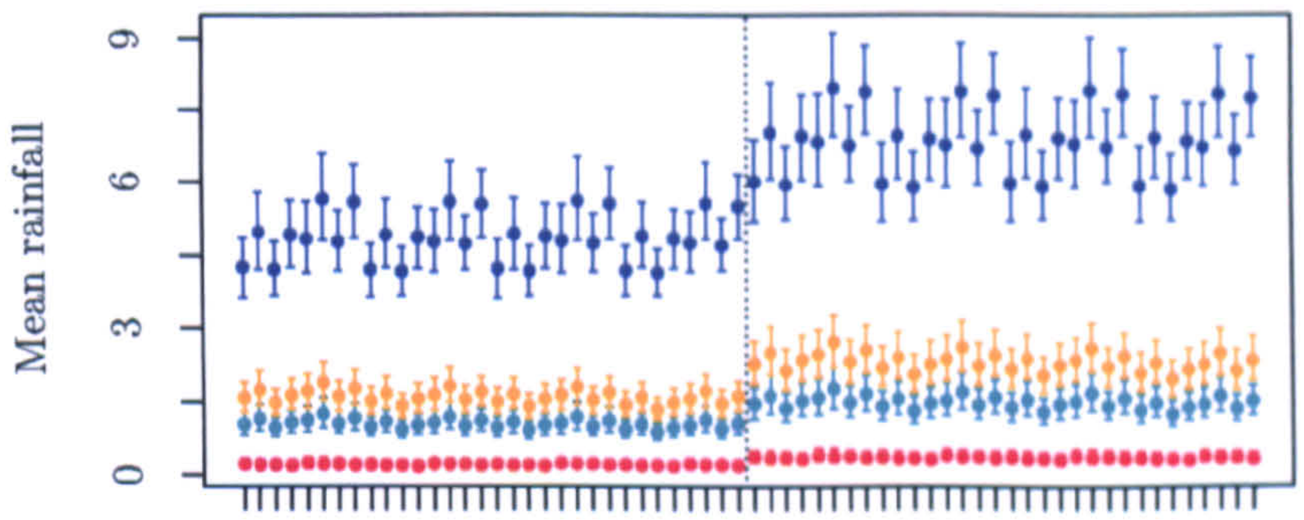
Based on the plots in Figures 6.9–6.11, it appears that the weather state labelled 2 is clear–cut wet, and that labelled 3 is clear–cut dry. For the low elevation sites (sites 2, 4 and 6), there appears to be little distinction between states 1 and 4, both of which represent days on which the probability of rain is close to 1/2 and with mean rainfall amounts on wet days of around 1 mm. At site 1, state 4 is the state associated with the greatest probability of rain following rain and with the largest mean rainfall amounts on wet days. For site 3, the same is true of state 1. The states identified by the models in Chapters 4 and 5 had similar characteristics.

Comparing Figures 6.9(a) and 6.9(b), it can be seen that at most sites, in the majority of states, there is a clear separation between the posteriors for the conditional probabilities of



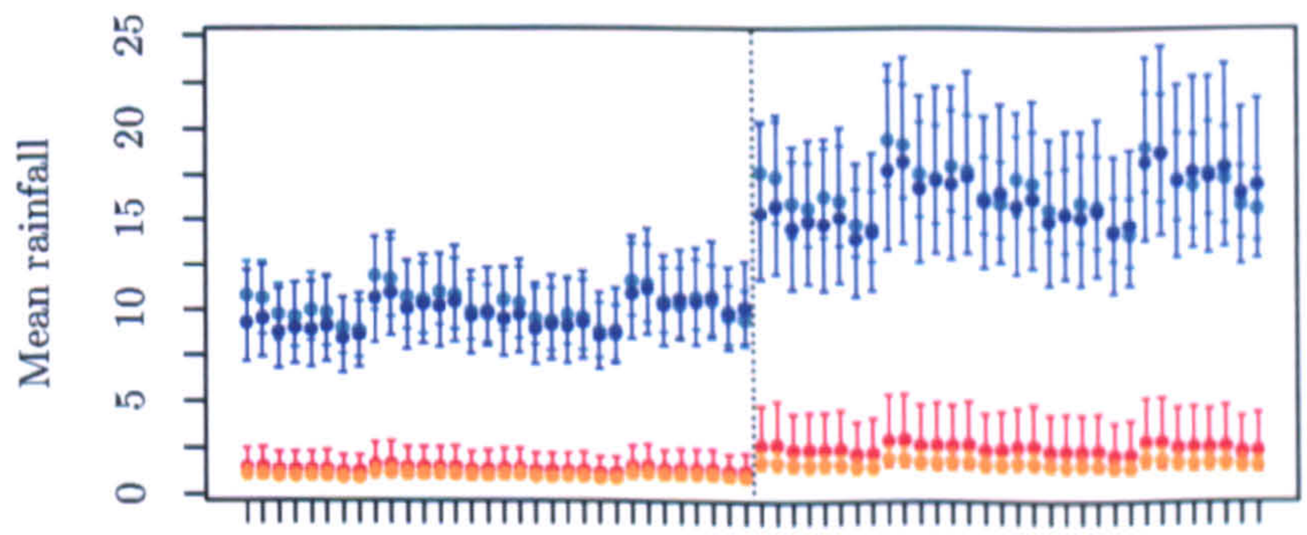
$\mathcal{I}(\mathbf{d}_{t-1})$

(a)



$\mathcal{I}(\mathbf{d}_{t-1})$

(b)



$\mathcal{I}(\mathbf{d}_{t-1})$

(c)

Figure 6.10: Conditional on $r = 4$, posterior means with 95% equi–tailed Bayesian credible intervals for the means in the lognormal distributions for $(W_t^i \mid D_t^i = 1, \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, S_t = k, \boldsymbol{\theta}_{4,\text{obs},k}, r = 4)$, where $\mathcal{I}(\mathbf{d}_{t-1}) = 0, \ldots, 2^n - 1$, at sites $i =$ (a) 1 (b) 2 and (c) 3, in weather states $k = 1$ (———), 2 (———), 3 (———) and 4 (———). The values $\mathcal{I}(\mathbf{d}_{t-1})$ are ordered so that the first 2^{n-1} correspond to $d_{t-1}^i = 0$, in ascending order, and the last 2^{n-1} correspond to $d_{t-1}^i = 1$, in ascending order.

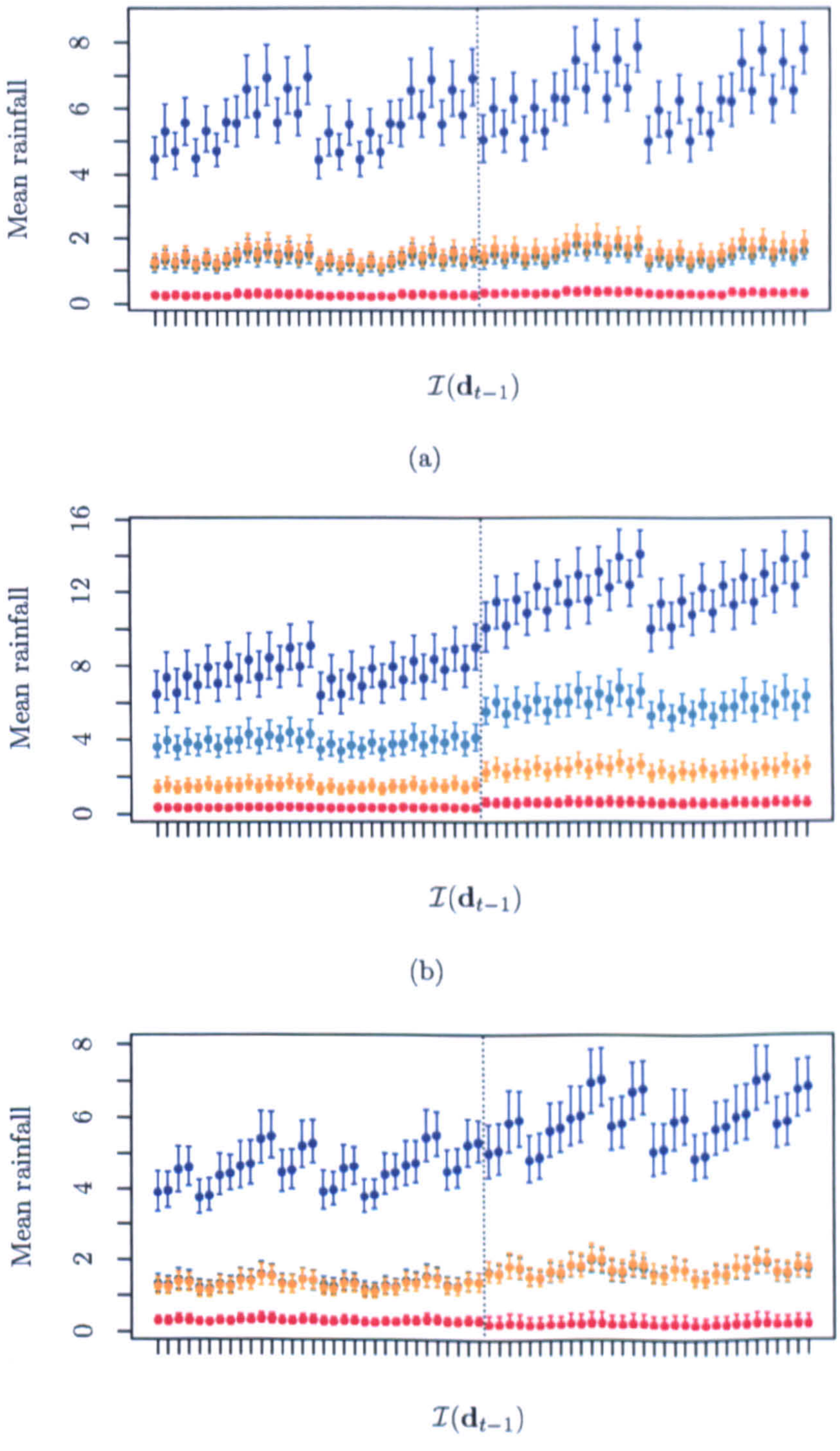


Figure 6.11: Conditional on $r = 4$, posterior means with 95% equi–tailed Bayesian credible intervals for the means in the lognormal distributions for $(W_t^i \mid D_t^i = 1, \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, S_t = k, \boldsymbol{\theta}_{4,\text{obs},k}, r = 4)$, where $\mathcal{I}(\mathbf{d}_{t-1}) = 0, \ldots, 2^n - 1$, at sites $i =$ (a) 4 (b) 5 and (c) 6, in weather states $k = 1$ (———), 2 (———), 3 (———) and 4 (———). The values $\mathcal{I}(\mathbf{d}_{t-1})$ are ordered so that the first 2^{n-1} correspond to $d_{t-1}^i = 0$, in ascending order, and the last 2^{n-1} correspond to $d_{t-1}^i = 1$, in ascending order.

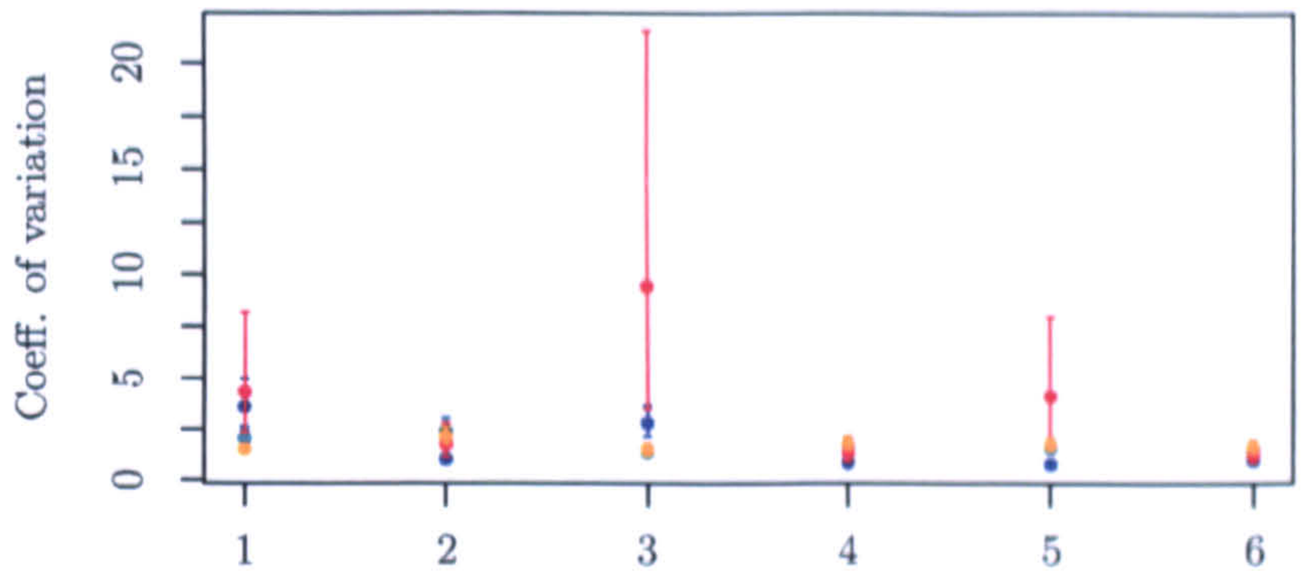


Figure 6.12: Conditional on $r = 4$, posterior means with 95% equi–tailed Bayesian credible intervals for the coefficients of variation in the lognormal distributions for $(W_t^i \mid D_t^i = 1, \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, S_t = k, \boldsymbol{\theta}_{4,\text{obs},k}, r = 4)$ at each site, i, in weather states $k = 1$ (———), 2 (———), 3 (———) and 4 (———). Note that the coefficients of variation do not depend on $\mathbf{D}_{t-1}$.

rain depending on whether the conditioning argument D_{t-1}^i is equal to 0 or 1. Similarly, in the plots for each site in Figures 6.10 and 6.11, in most states, at least some of the posteriors corresponding to different $\mathbf{d}_{t-1}$ do not overlap, especially in states associated with large rainfall amounts. This illustrates that, after conditioning on the weather state, there is still evidence, *a posteriori*, to identify the effect of the previous day's rainfall occurrence indicator in both the process of rainfall occurrence and the process of non–zero rainfall amounts. Given the weather state, the CI–HMM and MCA–NHMM modelled non–zero rainfall amounts as conditionally independent gamma random variables and so the effect of the rainfall occurrence indicator on the previous day was ignored.

For each site, i, in each state, k, Figure 6.12 shows the posterior for the coefficient of variation in the conditional lognormal distribution for $(W_t^i \mid D_t^i = 1, \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, S_t = k, \boldsymbol{\theta}_{4,\text{obs}}, r = 4)$. Note that in forming the coefficient of variation, dependence on $\mathbf{d}_{t-1}$ in the mean and standard deviation of the lognormal distribution cancel, and so the coefficient of variation does not depend on $\mathbf{D}_{t-1}$. For the 5–state CI–HMM (see Figure 4.10(c)) and the 5–state MCA–NHMM, we found that the coefficients of variation in the gamma distributions for non–zero rainfall amounts showed little variation between states, with considerable overlap in their posterior distributions. Moreover, the means in the posteriors were typically less than 1.2. It is immediately clear, therefore, that the coefficients of variation in the lognormal distributions, here, are considerably larger than those associated with the gamma distributions in the CI–HMM and the MCA–NHMM. This may, in part, be because the model whose posterior is summarised in this section had 4, rather than 5 states, meaning that each state had to represent a broader range of precipitation patterns. At the low elevation sites (sites 2, 4 and 6), the posteriors for the coefficients of variation in the clear–cut dry state (state 3) are concentrated about small means, but the converse is true for the high elevation sites (sites 1, 3 and 5). At the latter sites, therefore, the dry state is associated with a lot of variation in the non–zero rainfall amounts, relative to the mean. We explore this observation further in the following section as it appears to be related to posterior uncertainty regarding the weather state sequence.

The properties of the weather states, highlighted above, clearly illustrate the source of the convergence problems during MCMC sampling. Figure 6.9(b) shows that, at some sites, in the wet weather state (state 2), the posterior for the probability of rain following rain is very concentrated near one. As explained in the previous section, this leads to weak identifiability of the $\beta^i_{4,02}$ in the likelihood. Similarly, we remarked that the coefficients of variation in the dry state at the low elevation sites were very small, implying little variation in non-zero rainfall amounts relative to the (small) mean. This made it possible for the variation in the Z^i_{0t} to explain much of the variation in the rainfall amounts.

Plots (not shown) based on the posterior distributions for the weather state transition probabilities, $A^x_{4,jk} = \Pr(S_t = k \mid S_{t-1} = j, X_t = x, \theta_{4,\text{hid}}, r = 4)$, led to the same kinds of conclusions as those reached for the MCA–NHMM in Chapter 5. For example, there was very little separation between the posteriors for $A^1_{4,jk}, \ldots, A^{27}_{4,jk}$ for transitions *from* the clear–cut wet weather state, but much more for transitions *from* the clear–cut dry weather state. Therefore, for brevity, we do not repeat a similar analysis of the posterior for $\theta_{4,\text{hid}}$ here.

6.8.3.3 Posterior for $(\mathbf{s} \mid r = 4)$

In Chapters 4 and 5, plots of the marginal posterior mode estimate of the weather state sequence, $\hat{\mathbf{s}}$, indicated that the dry state was the most persistent and, on average, more days were allocated to this state towards the end of February. The same comments can be made about the corresponding plot for the 4–state LG–NHMM, and so, we omit the plot and further details.

Figure 6.13 shows the marginal posterior probabilities $\Pr(S_t = j \mid \mathbf{w}, \mathbf{d}, \mathbf{x}, r = 4)$, $j \in S_4$, for the days t in the first and last winter seasons. Comparing these plots to Figures 4.11(b)–(c) and 5.10(b)–(c) for the 5–state CI–HMM and the 5–state MCA–NHMM, it is clear that the 4–state LG–NHMM leads to considerably more posterior uncertainty in the allocation of days to weather states. Focusing on the last winter season, for example, within the first 15 days there are two days on which the posterior probability that $S_t = j$ is roughly equally shared between states $j = 3$ and $j = 1$ and another two days for which the same is true of states $j = 3$ and $j = 4$. This is consistent with our earlier observation, where we highlighted the support for large values in the posteriors for the coefficients of variation at sites 1, 3 and 5 in the dry state (state 3). Elucidating further, at site 1, Figure 6.10 showed that the means in the conditional lognormal distributions for rainfall amounts were large in state 4. Similarly, the posteriors for the means at sites 3 and 5 supported large values in state 1. Therefore, for there to exist this kind of uncertainty, *a posteriori*, in whether a day should be allocated to a state characterised by small rainfall amounts (state 3), or large rainfall amounts at some sites (states 1 or 4), the coefficients of variation at these sites, in the dry state, would have to be large.

6.8.4 Model checking

This section begins by comparing the posterior predictive distributions for the test quantities introduced in Chapter 4 with the observed statistics, focusing on differences in the performance of the LG–NHMM compared with the simpler CI–HMM and MCA–NHMM. Using data from outside of the sample used to fit the model, we then compare observed values of the test quantities

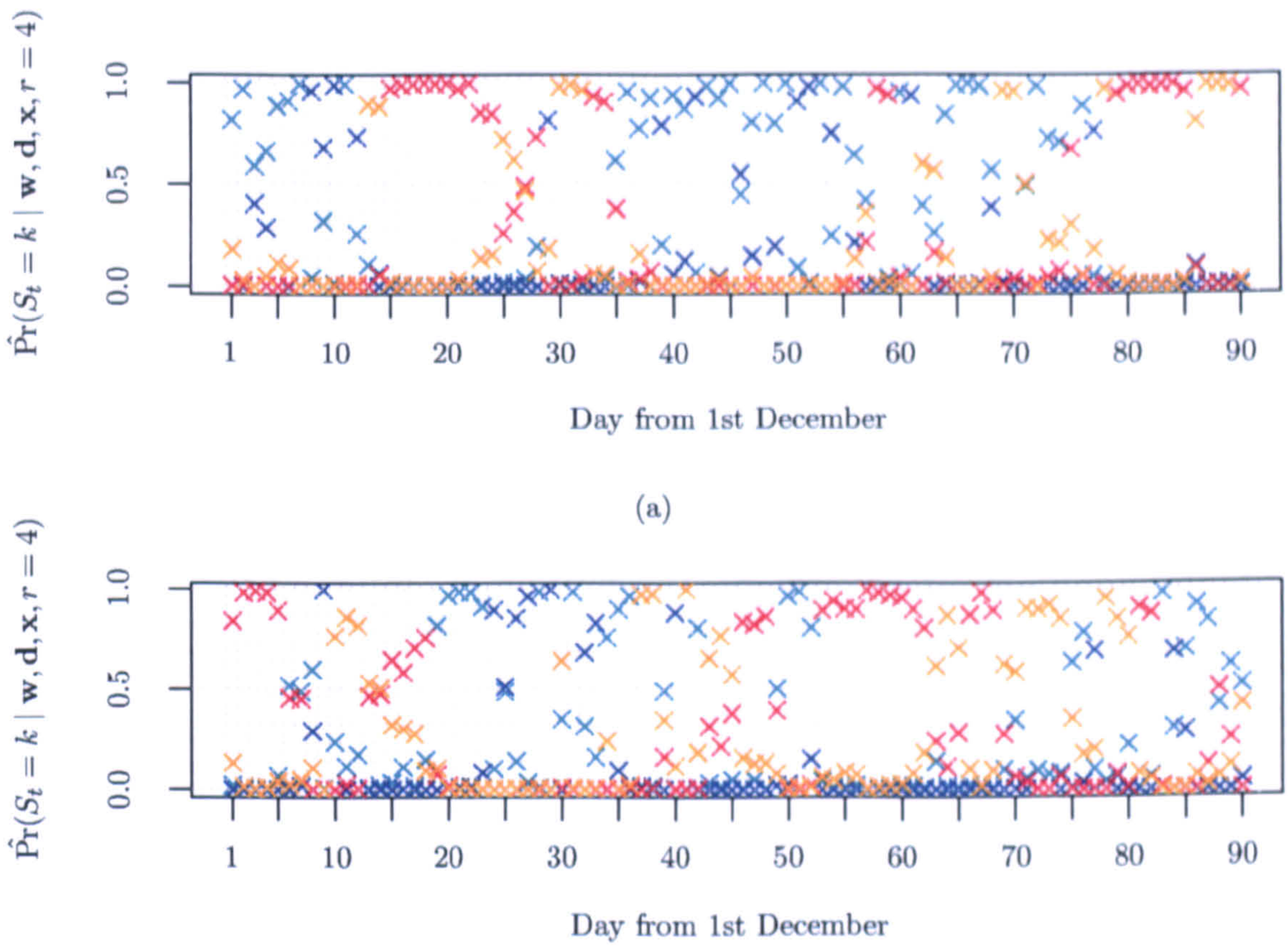


Figure 6.13: Conditional on $r = 4$, posterior weather state probabilities $\hat{\Pr}(S_t = k \mid \mathbf{w}, \mathbf{d}, \mathbf{x}, r = 4)$ for $k = 1$ (———), $k = 2$ (———), $k = 3$ (———) and $k = 4$ (———) in the winter (a) 1961/62 and (b) 1990/91.

to their posterior predictive distributions, conditioning the model on Lamb weather type data from the out–of–sample period.

Only a negligibly small proportion of the posterior for r is shared by values $r = 1, \ldots, 3$ and so averaging the posterior predictive distribution for any particular test quantity over the posterior for r is essentially equivalent to using the predictive distribution conditioned on $r = 4$.

6.8.4.1 Within sample

Figure 6.14 shows the observed relative frequencies of rainfall occurrence at each site and their posterior predictive distributions. Consider the corresponding plot in Figure 4.12(a), based on the CI–HMM, but also representative of the plot for the MCA–NHMM. Compared with Figure 4.12(a), it appears that the LG–NHMM slightly overestimates the proportion of wet days at the site (site 6) where this proportion is the lowest, and *vice versa* for the site (site 5) where this proportion is highest. This was also true for the posterior predictive distribution of a model with just $r = 1$ state. It is likely to be due to the very concentrated, highly correlated prior for the $\beta^i_{r,0k}$ whose variance was made very small to facilitate convergence of the MCMC

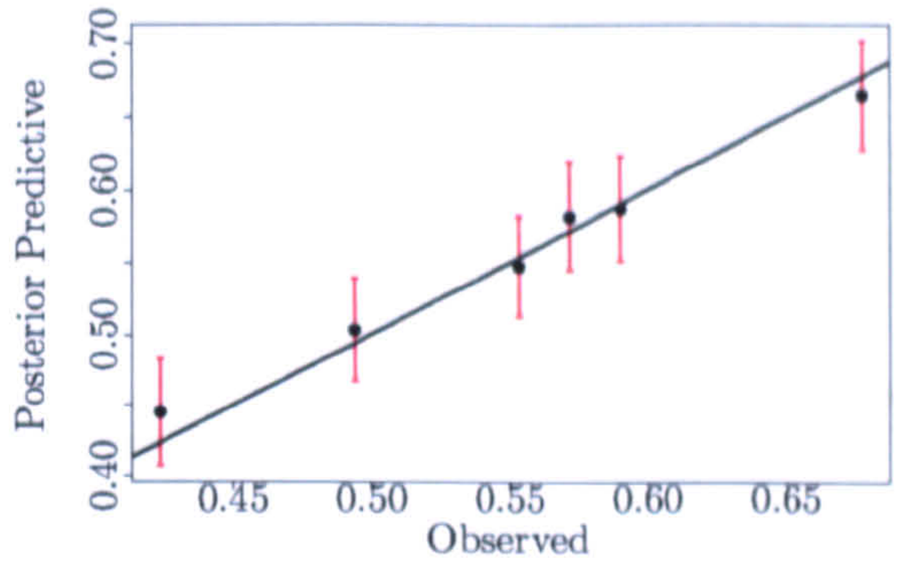


Figure 6.14: Observed values versus posterior predictive means for precipitation occurrence relative frequencies at each Yorkshire site. (———) indicate the posterior predictive 95% Bayesian credible regions.

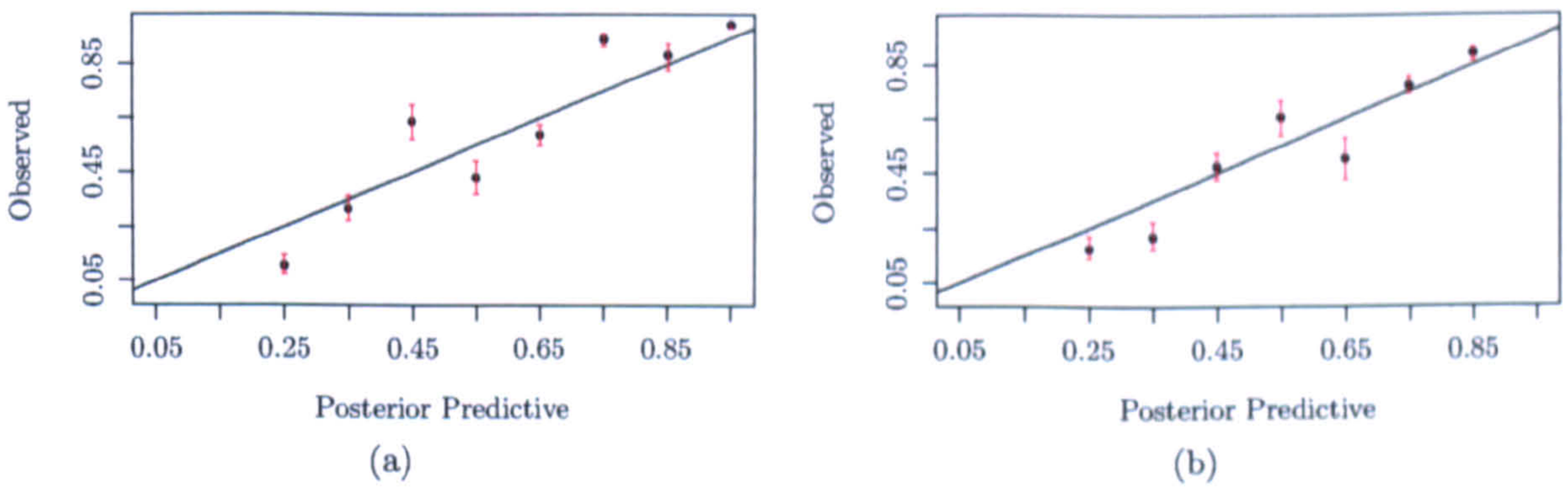


Figure 6.15: Calibration curves for the posterior predictive probability of rain at Lockwood Reservoir (site 1) obtained by modelling data according to the (a) 5–state MCA–NHMM and (b) 4–state LG–NHMM. (———) is a posterior 95% Bayesian interval for the "true" probability based on the observed sample (assumed binomial) and a uniform prior on the "true" probability.

sampler. This prior will have limited the allocation of density in the tails of the posterior for the $\beta^i_{r,0k}$, possibly preventing the prediction of relative frequencies which matched the high and low observed proportions exactly. A similar plot for the relative frequencies of each rainfall occurrence vector (not shown) revealed that, like the MCA–NHMM, the LG–NHMM gave rise to posterior predictive distributions with means very close to the observed statistics.

Figure 6.15(b) shows the calibration curve for the posterior predictive probability of rain at site 1, $\Pr(D_t^{*\,1} = 1 \mid D^1_{t-1} = d^1_{t-1}, \mathbf{w}, \mathbf{d}, \mathbf{x})$, where $D_t^{*\,1}$ is a hypothetical replication of D^1_t and d^1_{t-1} is the observed rainfall occurrence indicator on day $t-1$. Figure 6.15(a) shows a corresponding plot obtained using the MCA–NHMM. Although the patterns displayed in the two plots do not imply better fit by one model or the other, Figure 6.15(a) is the only one to include a point with x–coordinate equal to 0.95. In other words, the LG–NHMM was unable to predict any probabilities in the interval $[0.9, 1.0]$, but there were some days on which the MCA–NHMM was able to do this. At other sites we also observed similarity between the plots (not shown) for the two NHMMs, with the MCA–NHMM occasionally predicting probabilities in the intervals $[0.0, 0.1)$ or $[0.9, 1.0]$ when the LG–NHMM did not. Again, this may be because of the restrictive

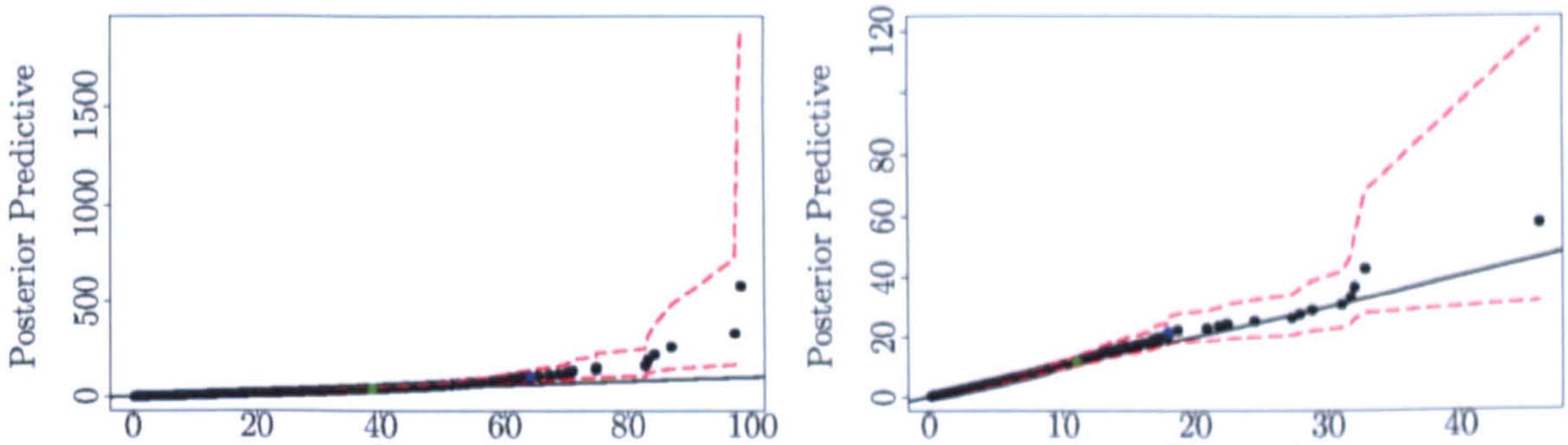


Figure 6.16: Quantile–quantile plots for the observed versus posterior predictive mean rainfall amounts (in mm) at (a) Moorland Cottage (site 3) and (b) the Retreat, York (site 4). (------) indicate the posterior predictive 95% Bayesian credible regions. For reference, (•) and (•) indicate the 95–th and 99–th quantiles.

prior chosen for the $\beta^i_{r,0k}$.

The sample quantiles of the distribution of non–zero rainfall amounts and summaries of the corresponding posterior predictive distributions are shown in Figure 6.16 for sites 3 and 4. Figure 6.16(a) is representative of those for the other high elevation sites (sites 1 and 5) whilst Figure 6.16(b) is representative of the other low elevation sites (sites 2 and 6). Figure 4.14 showed the same kind of plots for the CI–HMM and the corresponding figure for the MCA–NHMM was not discernibly different. For all three models, the agreement between observed and predicted quantiles is good, with only the very highest observed quantiles lying outside of the central 95% of their posterior predictive distribution. However, in replacing the gamma distributions for non–zero rainfall amounts with lognormal distributions, the slight tendency for the CI–HMM and MCA–NHMM to underestimate the highest quantiles has been replaced with a slight tendency for overestimation by the LG–NHMM. This can be seen by comparing Figures 4.14(c)–(d) with Figure 6.16. The large width of the credible region for site 3 may be due to the large coefficient of variation in the dry state, making very large rainfall amounts plausible in the posterior predictive distribution.

The means and 95% equi–tailed Bayesian credible regions in the posterior predictive distributions for the log odds ratios between rainfall occurrences and Spearman's rank correlation coefficients between non–zero rainfall amounts at all pairs of sites are displayed in Figure 6.17, along with the observed statistics. Corresponding plots for the CI–HMM and MCA–NHMM are displayed in Figures 4.15 and 5.12, respectively. The CI–HMM generally underestimated the larger log odds ratios and Spearman's rank correlation coefficients. The former problem was largely eliminated by the MCA–NHMM through the introduction of within–state spatial dependence between rainfall occurrences. However, the latter problem remained. Figure 6.17(a) shows that when the data are modelled according to the LG–NHMM, the larger log odds ratios lie further into the right–hand tails of their posterior predictive distributions than they had done when modelling with the MCA–NHMM. This indicates that the MCA–NHMM is more able to predict strong spatial dependence between rainfall occurrences than the LG–NHMM (although

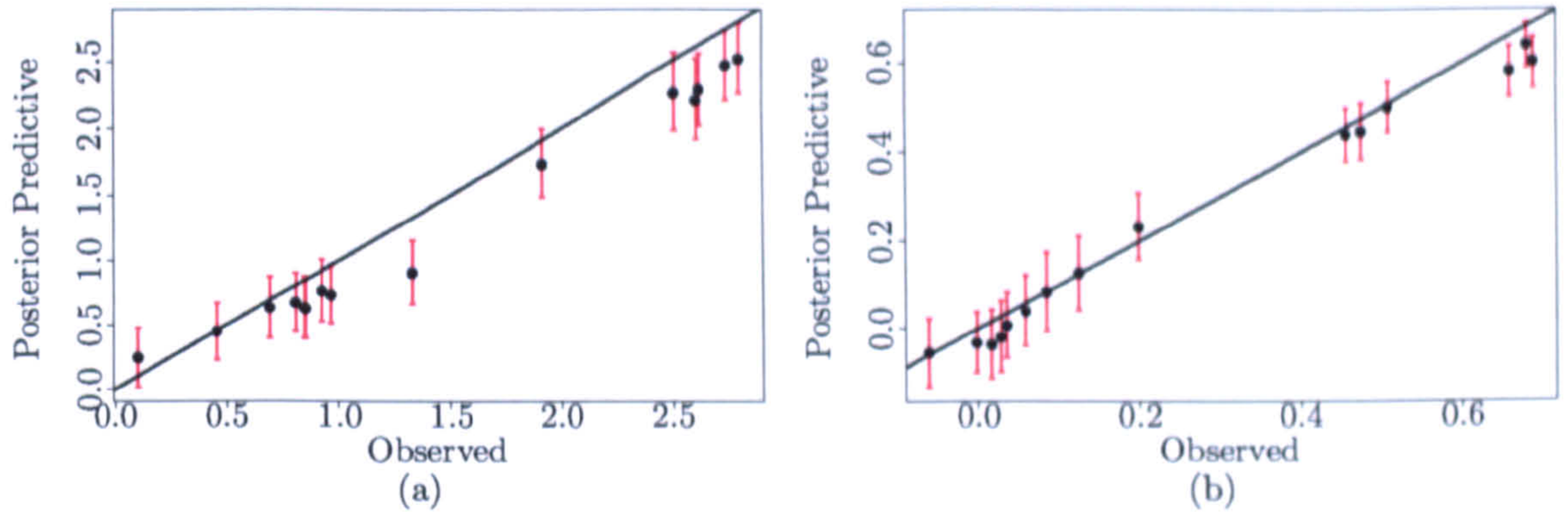


Figure 6.17: Observed values versus posterior predictive means for (a) log odds ratios between rainfall occurrences; and (b) Spearman's rank correlation coefficients between non–zero rainfall amounts at each pair of sites in the Yorkshire network. (———) indicate the posterior predictive 95% Bayesian credible regions.

the LG–NHMM still offers a considerable improvement over the CI–HMM). This may be due to the restrictive prior for the $\beta^i_{r,0k}$, or possibly because of differences between the shapes of the autologistic and multivariate–probit link functions, which are likely to be the most noticeable in regions where joint probabilitites are very small or large.

The benefit of introducing within–state spatial dependence between the non–zero rainfall amounts in the LG–NHMM is clear from Figure 6.17(b), in which nearly all the observed Spearman's rank correlation coefficients lie within the central 95% of their posterior predictive distributions. The problem associated with earlier models of underestimating the higher correlations has been substantially reduced.

The MCA–NHMM assumes rainfall occurrences to be conditionally Markov, given the weather state. The same is true of the LG–NHMM after marginalising over the latent Gaussian variables, $\mathbf{Z}_0$. The posterior predictive distributions for the survival functions of wet and dry spells obtained for the two models were not discernibly different, and so, plots showing the comparison with the observed distributions are not shown. We can therefore repeat our conclusions from Chapter 5 and surmise that the LG–NHMM captures the persistence of wet and dry spells well, although at two sites (sites 3 and 6), there is a tendency to underestimate the proportions of long duration wet spells.

For the CI–HMM, Figure 4.18 showed the observed Spearman's rank correlation coefficients between rainfall amounts (within wet spells) at lags 0–8, together with summaries of the corresponding posterior predictive distributions. Plots based on the MCA–NHMM were similar. From these figures, we concluded that the models underestimated the correlations at the high elevation sites (sites 1, 3 and 5) where the temporal dependence was strong. This is likely to have been because both models adopt a simple temporal structure for rainfall amounts, assuming them to be conditionally independent in time given occurrences and the weather state.

In contrast, after marginalising over $\mathbf{Z}_0$ in the LG–NHMM, each node, $\mathbf{W}_t$, has parents $\mathbf{D}_t$, S_t *and* $\mathbf{D}_{t-1}$. Figure 6.18 displays plots for the lagged correlations at sites 3 and 5. The plot for

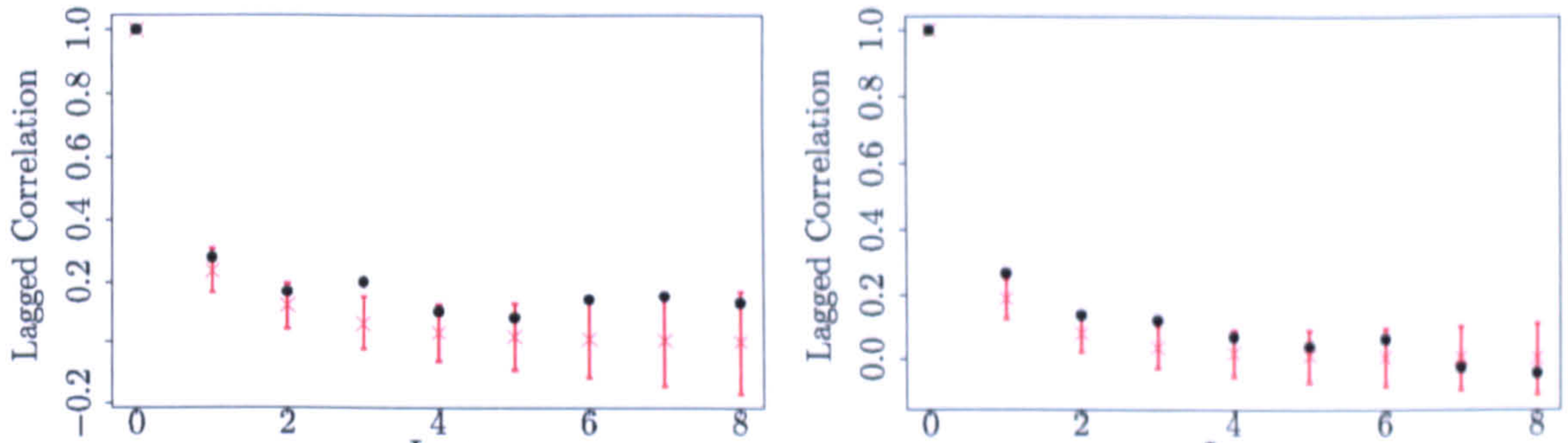


Figure 6.18: Observed (•), posterior predictive mean (×) and posterior predictive 95% Bayesian credible region (———) for the Spearman's rank correlation coefficient between wet days (within runs of consecutive wet days) at various lags at (a) Moorland Cottage (site 3) and (b) Great Walden Edge (site 5).

site 1 was very similar to that for site 5 and is omitted. Compared with Figures 4.18(c) and 4.18(e), the observed statistics are now much more plausible under their posterior predictive distributions, so the incorporation of dependence on the previous day's rainfall occurrence indicator seems to have improved modelling of the temporal dependence between rainfall amounts within wet spells.

6.8.4.2 Out–of–sample

The method of model checking from the previous section is sometimes criticised on the grounds that the data have been used twice, both in the model fitting and model checking stages. For example, see Bayarri & Berger (2000). We can go some way towards addressing these concerns by comparing the posterior predictive distribution to data that were not included in the model. This might give more insight into which aspects of the rainfall process are not captured by the model, as well as an indication of how useful the model might be as a means of predicting rainfall in Yorkshire.

Precipitation and Lamb weather type data are available for the six winter (December–February) seasons from 1991/2 to 1996/7 which follow the 30 winter seasons used to fit the model. With two leap years in this period, the overall length of the dataset is 542 days. At site 2 there are fewer data available for model checking because the 213 values between December 1st 1991 and January 1st 1994 (inclusive) are missing. The purpose of this section is to compare the observed test quantities from this out–of–sample period to their posterior predictive distributions, conditioned on the out–of–sample atmospheric data.

Figure 6.19 displays summaries of the posterior predictive distributions for a variety of test quantities. Subfigures (a), (e) and (f) show the observed and posterior predictive distributions for relative frequencies of rainfall occurrence at each site, log odds ratios between rainfall occurrences at all pairs of sites and Spearman's rank correlation coefficients between non–zero rainfall amounts at all pairs of sites, respectively. From these plots, it seems that agreement between the

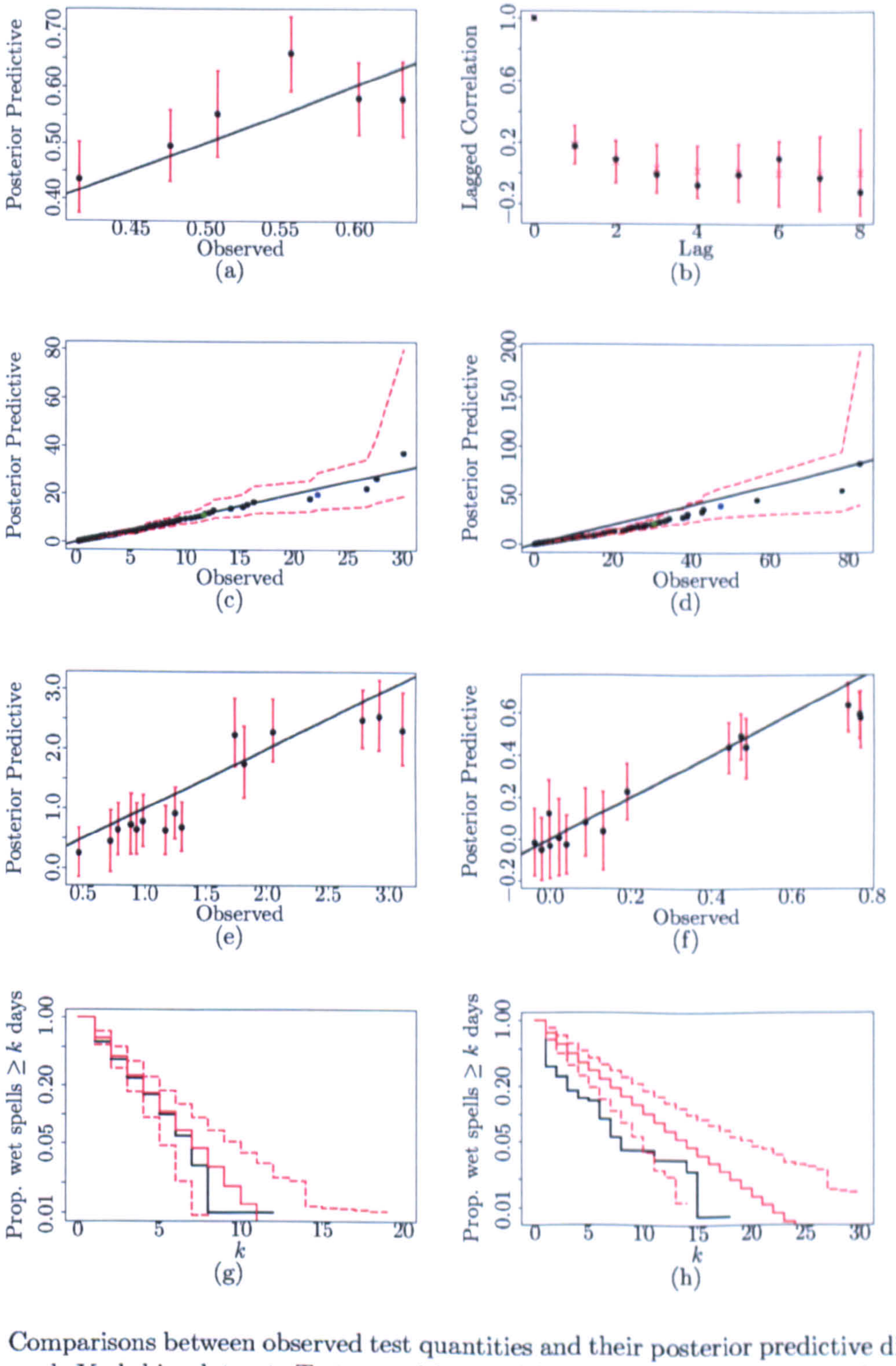


Figure 6.19: Comparisons between observed test quantities and their posterior predictive distributions for the out–of–sample Yorkshire dataset. Test quantities are (a) precipitation occurrence relative frequencies; (b) Spearman's rank correlation coefficients between wet days (within wet spells) at various lags at site 5; sample quantiles for rainfall amounts at (c) site 4 and (d) site 5; (e) log odds ratios between rainfall occurrences; (f) Spearman's rank correlation coefficients between rainfall amounts; survival distributions of wet spells at (g) site 4 and (h) site 5. (• / ×) and (——/ ——) indicate observed statistics/posterior predictive means in (b) and (g)–(h). (——) indicate posterior predictive 95% Bayesian credible regions in (a), (b), (e) and (f). (------) indicate posterior predictive 95% Bayesian credible regions in (c), (d), (g) and (h). (•) and (•) indicate the 95–th and 99–th quantiles in (c) and (d).

observed statistics and corresponding posterior predictive means is not as good as it had been for the within-sample test quantities, but this is to be expected since the latter data were used to fit the model. Moreover, the 95% Bayesian credible regions appear to be wider here than they had been in the previous section, perhaps because the out-of-sample dataset is smaller than that used in model fitting. However, in each of subfigures (a), (e) and (f), most of the observed statistics lie within the central 95% of their posterior predictive distributions. The only noticeable exception is the proportion of wet days at site 5 in subfigure (a). A similar plot (not shown) for the relative frequencies of all rainfall occurrence vectors showed comparable patterns, with only the largest proportion (rain at all sites) lying slightly beyond the 97.5% point in its posterior predictive distribution. At site 5, the proportion of wet days in the within-sample dataset was 0.677, but in the out-of-sample dataset, this proportion was only 0.557. There appears to have been a medium-term shift in the precipitation behaviour at site 5, which has not been captured by conditioning on the Lamb weather type data. This shift is also evident in the comparison of some of the other test quantities for site 5 with their posterior predictive distributions.

For example, subfigures (d) and (h), respectively, compare the observed sample quantiles in the distribution of non-zero rainfall amounts at site 5 and the empirical survival function for wet spells at site 5, to the posterior predictive distributions. Subfigures (c) and (g) show corresponding plots for site 4, in which the observed test quantities look highly plausible under their posterior predictive distributions. The latter plots were representative of those for sites 1, 2, 3 and 6. For all sites, plots showing the empirical survival functions for dry spells (not shown) and the Spearman's rank correlation coefficients between non-zero rainfall amounts at various lags revealed that the observed test quantities lay within the central 95% of their posterior predictive distributions. This was even true for site 5. For example, the plot for the lagged correlations within wet spells is shown in subfigure (b).

In general, these checks reinforce the observation from the previous section that the LG-NHMM seems able to predict rainfall data with similar spatio-temporal characteristics as sets of observed data. However the medium-term shift in precipitation patterns at site 5 was not predicted by the model, in spite of conditioning on Lamb weather type data.

6.9 Summary

The main objective of this chapter was to develop an NHMM for rainfall such that, conditional on the weather state, *both* rainfall occurrences *and* non-zero rainfall amounts were correlated in space and time. This was achieved by adopting a hierarchical model in which the weather states were separated from the observable quantities by a time series of latent multivariate normal random variables. These played the dual roles of thresholding variables, whose sign governed the occurrence or otherwise of rain, and of spatially-varying regressors in the conditional distributions of the log non-zero rainfall amounts. By making the previous day's rainfall occurrence indicator a parent of the current latent multivariate normal random variable, additional temporal structure was introduced, beyond that incorporated by the weather state process.

In studying this model, we have been able to develop useful ideas of more general applicability.

These are highlighted in the following paragraphs, before we summarise the contents of the chapter and our main conclusions.

Our first contribution concerns the method of handling the non-identifiability problem of multivariate probit models. The usual approach is to constrain the variance matrix to be a correlation matrix, but this leads to the difficult problem of sampling a correlation matrix during MCMC. We proposed two alternatives, both of which remove this obstacle and present the opportunity to choose semi-conjugate priors. These were; (i) placing restrictions on the vectors of regression coefficients, in order to prevent arbitrary rescaling of the linear predictor and (ii) fixing the conditional variances arising from the modified Cholesky decomposition of the precision matrix (see Section 6.8.3.1). In the former case, we proposed constraining the coefficients of one (suitable) covariate to be ± 1. Simulation experiments suggested that this led to improved MCMC mixing compared with some of the approaches from the literature which rely on random walk Metropolis Hastings steps to update a correlation matrix. In (ii), we proposed fixing the conditional variances at the values that would otherwise have been chosen for their prior means, but this did not offer improved mixing compared with suggestion (i).

In addition to the computational benefit, we had hoped that both (i) and (ii) would offer the opportunity to elicit a more meaningful prior because we could work directly with the variance matrix, rather than the correlation matrix, which has a more complicated sample space. However, choosing the hyperparameters in the priors was not easy because of our poor understanding of the scale of the latent multivariate normal random variables. Further work in developing elicitation strategies for these priors is therefore required. It is possible that fixing the conditional variances at some other value in (ii) would simultaneously simplify prior elicitation and improve MCMC mixing.

Our second contribution was the discovery that the power posterior approximation to the marginal likelihood needs a correction when the support of the posterior differs from that of the prior. Developing techniques to approximate the correction term in complicated models, such as the hidden Markov model in this chapter, is another area for future work.

In this chapter, we identified a general framework, based on (partially) latent Gaussian variables, for modelling spatial dependence amongst rainfall occurrences and amongst non-zero rainfall amounts. We then showed that many of the rainfall models from the literature are grounded in this general framework. Amongst these is a hidden Markov model, proposed by Ailliot *et al.* (2009), which was fitted to daily rainfall data collected at seven sites in South Island, New Zealand over 26 consecutive Aprils. Conditional on the weather state, daily rainfall totals were assumed to be independent in time, but correlated in space through the dependence structure of the truncated and transformed multivariate normal distribution, which had state dependent parameters. Compared with this model, the LG-NHMM offers a more sophisticated temporal structure in which, jointly, rainfall occurrences and amounts are conditionally Markov given the weather state. Moreover, by modelling non-zero rainfall amounts conditionally on the latent Gaussian variables, rather than through some deterministic function thereof, we do not prevent independent changes in the probability of rain and the distribution of non-zero rainfall amounts. In addition to offering a less flexible model, Ailliot *et al.* (2009) also took a frequentist approach, whereas we offer a fully Bayesian treatment. Philosophical superiority aside, the Bayesian approach to inference also offers practical advantages. To stabilise their estimates

of the variance matrix, Ailliot *et al.* (2009) consider models that assume various parametric forms, for example, basing their variance matrices on the exponential covariance function. By correlating the variance matrices in the prior, we can borrow strength between states, and so, do not need to impose any kind of structure on the variance matrix.

In spite of these advantages, the model presented in this chapter is not without fault. Introducing latent variables into models often simplifies computation, however, by creating further separation between the parameters and observable quantities, prior elicitation can often become more difficult. This issue was not fully addressed in this chapter and so, as commented earlier, eliciting the prior was not easy. Furthermore, in Chapter 5 we noted a practical problem of using hidden Markov models with highly parameterised within-state distributions. Essentially, in certain states, when we only observe data from a small subset of its sample space, it is inevitable that some parameters will be only weakly identified in the likelihood. The susceptibility of hidden Markov models to this problem was reinforced by the mixing and convergence problems we experienced in this chapter when implementing MCMC. To avoid such problems, therefore, one solution might be to adopt different within-state models for states representing particularly extreme conditions. For example, we might consider different conditional distributions for rainfall, with fewer parameters, given classification into very wet or very dry states. We could then choose the priors for the parameters in these states to encourage the intended classification of weather conditions. Alternatively, we could introduce two states, one where it always rains at all sites, and another where it is always dry at all sites. We provide further comments on this modification to the model in Chapter 7.

Finally, the application of the model to the Yorkshire dataset highlighted an interesting question concerning the utility of the marginal likelihood as a means of comparing models. The marginal likelihood for the LG-NHMM with $r = 4$ states was smaller than that for the very simple 4-state CI-HMM. Almost certainly this is because the prior for the LG-NHMM was very concentrated in some directions of the parameter space, and the regions with the greatest prior and likelihood support did not always coincide. This, in turn, was largely because computational considerations had necessitated the choice of very tight priors. However, after comparing observed data to their posterior predictive distribution, it was clear that the spatial and temporal characteristics were much better predicted by the LG-NHMM than the CI-HMM. Rationalising in this way, if asked to pick the "best" model, we would probably disregard the information from the marginal likelihood and choose the LG-NHMM.

Posterior model probabilities are obtained by combining the marginal likelihood with the prior model probabilities using Bayes Theorem. Theoretically, these probabilities provide a complete summary of our posterior uncertainty about the models under consideration and are often used to compare them. However, we know that the marginal likelihood really compares model and prior combinations, and not just models. Unfortunately, when attempting to discriminate between complex models, it is generally difficult to balance the information in the priors because of computational considerations or simply because eliciting the priors is not easy. In these cases, therefore, there is certainly an argument that the posterior model probabilities are not the correct metric for making the comparison. Very often, models are constructed in order to make predictions, and so, perhaps the real metric for comparing them should focus on their posterior predictive performance, although deciding how this should be measured is not a trivial problem. For example, see Gelfand & Ghosh (1998).

Typically, as statisticians, we are likely to prefer models that are convenient to use. As such, we might like our posterior–predictive based mediator to include a penalty for inconvenience which would be application–specific. For example, we might penalise models associated with excessively slow computing times. An interesting area for future work would involve building these ideas into a principled framework for model comparison, for example, using a multi–attribute value function. We provide further comments along these lines in Chapter 7.

Chapter 7

Conclusions and future work

7.1 Introduction

In this chapter we highlight our contributions to the literature (Section 7.2) and then review our overall conclusions (Section 7.3). Section 7.4 then summarises our findings in applying the models and inferential procedures from earlier chapters to a larger, spatially diffuse network of sites. The objectives of this section are to both demonstrate how our inferential procedures scale up to handle larger networks of sites and to examine the performance of the models in applications involving sites which are spatially well separated. In Section 7.5 we conclude with a discussion of some possible directions for future work.

7.2 Objectives and contributions of the thesis

The primary objective of this thesis was to develop homogeneous and non-homogeneous hidden Markov models for rainfall, within a Bayesian framework. To this end, we have investigated three hidden Markov models with increasingly complex distributions for rainfall, given the weather state. We have also shown how categorical atmospheric data, namely Lamb weather types, can be incorporated in the weather state process.

Taking a broad view, the main contribution of Chapters 4, 5 and 6 has been towards an improved understanding of the potential of hidden Markov models to describe rainfall, identifying strengths and limitations. Some of the problems with models from earlier chapters were addressed subsequently. For example, the simple model from Chapter 4 failed to capture the spatial dependence amongst rainfall occurrences, and this was remedied in Chapters 5 and 6 by explicitly modelling spatial association between rainfall occurrences, given the weather state. Other problems remain unresolved and require further investigation, two examples of which are provided in Section 7.5.

By investigating the practicalities of fitting hidden Markov models in a Bayesian framework, we have made methodological contributions in several specific areas, three of which are detailed here.

The simulation experiment and discussion in Chapter 4 showed that the recently proposed power posterior approach provides a good approximation to the marginal likelihood for hidden Markov models. Also, unlike many competing methods from the literature, it does not require that fully conjugate priors are chosen or that all full conditional distributions have known normalising constants. In Chapter 6, we then showed that the power posterior approximation needs a correction term when applied to models for which the prior and posterior do not have support over the same sets of values.

Chapter 6 contained two alternative strategies for handling the identifiability problem in multivariate probit models. The first involved constraining the coefficients in the linear predictor so that the coefficient of one covariate could only take the value 1 or -1. The second involved fixing the conditional variances in the marginal/conditional decomposition of the joint density function for the latent multivariate normal random variable. Under both approaches, a particular choice of prior leads to full conditional distributions with known normalising constants, and this presents a computational benefit. Although problems regarding prior elicitation remain, the second approach shows potential and would be worthy of further investigation.

Finally, rainfall modelling is an example of a problem where we need to use a mixture distribution with a degenerate component at zero. Chapter 6 described a hierarchical modelling framework, based on two (partially) latent multivariate normal variables, to account for the spatial dependence between rainfall occurrences and between rainfall amounts, given occurrences. However, the model does not force us to prescribe the extent to which the dependences are linked. Similar ideas could be applied in other spatial problems involving such mixtures; see, for example, Boys *et al.* (2011), where we consider a zero-inflated Poisson model for counts of fish at various spatio-temporal locations.

The secondary objective of this project was to develop and demonstrate techniques to assist in the task of prior elicitation. Throughout Chapters 4, 5 and 6, we explained how prior knowledge could be used to specify priors subjectively. This was the most straightforward for the simple model in Chapter 4, where we could employ standard techniques, such as the equivalent prior sample approach and the quantile method, but became more difficult as model complexity increased. Nevertheless, in Chapter 5, we suggested a novel elicitation strategy for a hierarchical (two-stage) Dirichlet prior. Having chosen values for the marginal means and variances, prior correlations between stochastic vectors were fixed by quantifying the value, in terms of the size of a hypothetical number of observations, of learning that all but one parameter at the first stage of the prior was equal to the common and unknown mean.

Although space did not allow full details to be provided in this thesis, we also made a contribution to the elicitation literature through work on building genuine beliefs into a prior for the variance matrix in multivariate normal distributions. A brief outline of some of this work was provided in Chapter 6, where it was applied in specifying a prior for the variance matrix of the latent Gaussian variables in the hierarchical NHMM for rainfall. This research is ongoing and further details can be found in a technical report, Germain *et al.* (2010*b*).

7.3 Conclusions

Chapter 2 was based on an exploratory analysis of the Yorkshire dataset which was analysed further in later chapters. We found clear relationships between the Lamb weather types and precipitation, for example, smaller (larger) proportions of wet days tended to be associated with the anticyclonic (cyclonic) types. In addition, different relationships were observed at the Pennine sites where westerly Lamb weather types were linked with much larger rainfall amounts on wet days than they had been at eastern sites. Exploration of the spatial characteristics of each dataset revealed that dependence between rainfall occurrences generally decreased with increasing distance between sites, and likewise for non-zero rainfall amounts. Temporally, there were no obvious long-term or within season trends, and no effort was made to model such effects subsequently. By fitting separate binary Markov models to rainfall occurrences at each site, we found that chains of order 1 or 2 were generally favoured over those of higher or lower order. Similarly, by fitting autoregressive models to mean-centred log rainfall amounts within wet spells, we found that AR(1) models usually had more support from the data than AR(0) (independence) or AR(2) models. Higher orders were often favoured at high elevation sites in both the Markov and autoregressive models.

After introducing Bayesian inference for hidden Markov models in Chapter 3, Chapters 4, 5 and 6 investigated increasingly sophisticated hidden Markov models for rainfall, in which the hidden states were interpreted as states of the weather. In Chapter 4 we studied a simple homogeneous hidden Markov model whose temporal structure was defined by "standard" assumptions: (i) the hidden states evolve as a homogeneous first order Markov chain and (ii) jointly, rainfall occurrences and amounts are conditionally independent in time, given the weather state. The precipitation process was then factorised so that both rainfall occurrences and rainfall amounts, given occurrences, were conditionally independent in space, given the weather state. Rainfall occurrences and non-zero amounts were modelled as Bernoulli and gamma random variables, respectively, with site (and state) specific parameters. The advantages of this model were borne out of its simplicity. Analysis by MCMC was fast and did not lead to any convergence or mixing problems. Moreover, the parameters of the observed process, for example, probabilities of rain and expected non-zero rainfall amounts, were natural quantities about which to solicit prior beliefs. This meant we could use standard elicitation techniques, such as the quantile method, to specify priors. However, model checks revealed that the strong spatial associations observed between rainfall occurrences and between non-zero amounts at some pairs of sites would be very unlikely under the posterior predictive distribution. The same was true for the long duration wet and dry spells that were observed at some high elevation sites.

The problem of approximating the posterior distribution for the number of hidden states was also addressed in Chapter 4. For the models we considered, modelling data at multiple sites led to within-state distributions with a large number of parameters per state. Consequently, the parameter spaces for models with different numbers of states differed substantially in dimension. We therefore judged that it would be difficult to design an across model sampler which mixed well over the joint space of the model indicators and the model parameters, and instead focused on within model simulation techniques. This was a viable option because we limited the maximum number of states, and hence the number of marginal likelihood calculations, to only $r_{max} = 5$. Amongst the methods from the literature for approximating the marginal likelihood, many of

the more sophisticated techniques rely on choosing conjugate priors, or at least priors that lead to full conditional distributions with known normalising constants. They were therefore inappropriate for the models investigated in this thesis for which this condition did not hold. Two exceptions were the recently proposed power posterior approach and Chib's (extended) method, although the latter is less convenient when many parameters are updated singly in Metropolis Hastings steps, as was the case for the models in Chapters 4 and 5. In a simulation experiment, we compared the performance of various marginal likelihood estimators, including the power posterior approximation, Chib's estimator and other simple estimators, such as the harmonic mean. This revealed that the power posterior approach outperformed all others, with negligible bias and small variance. We used this approach in an application to the Yorkshire dataset and found that plots of the expected half deviance against temperature were not smooth, but exhibited sharp changes. We conjectured that these occurred at temperatures at which the likelihood had enough weight to allow additional hidden states to be recognised in the power posterior.

The model from Chapter 4 did not incorporate any atmospheric information and could not, therefore, respond to non–stationary shifts in atmospheric conditions, or be used in statistical downscaling. To remedy this deficiency, the models analysed in Chapters 5 and 6 were non–homogeneous hidden Markov models (NHMMs), in which the Markov assumption for the weather state process was modified to allow the Lamb weather types to influence the transition probabilities. This was achieved by introducing a different stochastic vector for each pair, (j, x), of lag–one weather state and current Lamb weather type, then specifying a prior which encouraged borrowing of strength between them. In the Yorkshire dataset, the posteriors for some state j to state k transition probabilities showed marked differences depending on the Lamb weather type. The patterns identified could be explained by meteorological knowledge of the relationships between precipitation and Lamb weather types in Yorkshire.

We criticised the model from Chapter 4 on the grounds that it failed to capture some of the spatial and temporal patterns in the data. The models from Chapters 5 and 6 were attempts to remedy these problems by successively relaxing the assumptions of conditional independence between observables. In Chapter 5, rainfall occurrences were modelled as a Markov chain of autologistic models, given the weather state. We continued to model non–zero rainfall amounts as conditionally independent gamma random variables. A significant drawback with autologistic models is the need to compute the normalising constants. For the Yorkshire dataset, with only $n = 6$ sites, this was not particularly challenging, but for larger networks of sites, exact computation would not have been feasible. We provide further comments about how this problem can be addressed in the following section.

Model checks revealed good agreement between observed and posterior predictive log odds ratios between the rainfall occurrences at most pairs of sites in the Yorkshire dataset. Similarly, the long duration wet and dry spells at the high elevation sites were much more plausible under their posterior predictive distributions than they had been using the Chapter 4 model. However, conditional independence in space and time was still assumed between rainfall amounts, given occurrences and the weather state. This meant that the larger spatial associations between non–zero rainfall amounts continued to be underestimated. There did not appear to be any natural extension to introduce dependence between non–zero rainfall amounts, given the weather state, and this can be regarded as another drawback of this model.

The model in Chapter 6 was a hierarchical NHMM, relying on the introduction of latent multivariate normal random variables, $\{\mathbf{Z}_{0t}\}$. The rainfall occurrence indicators were assumed to arise deterministically through a threshold specification on the underlying latent Gaussian vectors, with the time $(t-1)$ rainfall occurrence indicator and the time t weather state both taken to be parents of $\mathbf{Z}_{0t}$. In effect, rainfall occurrences were modelled as a Markov chain of multivariate probit models, given the weather state. The latent Gaussian variables also induced dependence amongst non–zero rainfall amounts, which were assumed to be conditionally independent log-normal random variables, given the weather state and $\mathbf{Z}_{0t}$, with location parameter expressed as a linear combination of the terms in $\mathbf{Z}_{0t}$. Marginalising over the latent Gaussian variables gave rise to an NHMM in which neither rainfall occurrences nor non–zero rainfall amounts were conditionally spatially or temporally independent, given the weather state. This model was grounded in a more general framework, in which spatial structure is modelled through the incorporation of two (partially) latent Gaussian variables, one responsible for dependence amongst occurrences, and the other for dependence amongst non–zero amounts. We showed that many other rainfall models from the literature also have their foundations in this framework.

Approximation of the marginal likelihood for this model presented difficulties. The latent Gaussian variables had non–zero support over the whole real line *a priori*, but only over the positive or negative half real line *a posteriori*. The power posterior approach (used previously) was found to require a correction term when applied to such models. In light of the problems in approximating this correction term, we focused on finding a different technique, and Chib's extended method provided a viable alternative. Although model checks based on the posterior predictive distribution suggested that the latent Gaussian variable NHMM provided a good fit to the Yorkshire data, the 4–state version of this model had a smaller marginal likelihood than the 4–state versions of each of the simpler models. We judged this comparison to be unfair, however, because it was necessary to specify a much more concentrated prior for the latent Gaussian variable model in order to achieve parameter identifiability in the posterior; see Section 7.5.2, where Bayesian model choice is discussed further.

Although the complexity of the latent Gaussian variable model facilitated good fit to the data, it was also responsible for its two main limitations. First, as will be explained in Section 7.5.1, problems arose during MCMC sampling when certain parameters were only weakly identified in the likelihood. When analysing the Yorkshire dataset, these problems were so severe that the MCMC for a model with $r = 5$ states simply did not converge. In order to force parameter identifiability in the posterior, therefore, it was necessary to choose a very strong prior. This led to a specification which was too concentrated to truly represent our prior beliefs and, of course, affected our posterior inferences. For example, very little posterior predictive density could be assigned to probabilities of rain near zero or one, compared with earlier models. However, even if computational considerations had not influenced our prior specification, incorporation of genuine initial beliefs would have remained problematic because the latent Gaussian variables are not observable. This made it difficult to put into practice the ideas from the technical report Germain *et al.* (2010*b*) regarding prior elicitation for the variance matrix in multivariate normal distributions. We therefore regard difficulties in eliciting the prior as the second main criticism of the latent Gaussian variable model.

7.4 Application to UK winter rainfall data

The Yorkshire dataset analysed in this thesis comprised a small number of sites with little spatial separation between them. Two important questions are (i) how well the inferential procedures for the models in Chapters 4, 5 and 6 scale up to handle larger networks of sites and (ii) the performance of these models in describing data from networks in which sites are spatially well separated. To investigate these questions we applied the three hidden Markov models to a larger, spatially diffuse network of $n = 12$ sites located throughout the entire UK. The locations of these sites can be seen on the map presented in Chapter 2 (Figure 2.1). Note that the distances between sites range from 113.6 km to 813.1 km, with a mean of 344.3 km. This is compared with minimum, maximum and mean distances of 39.8 km, 133.2 km and 82.8 km, respectively, for the six sites in the Yorkshire network. The UK dataset comprises winter (December to February) rainfall observations over the 28 years from 1961/2 to 1988/9 and does not contain any missing values. In this section, we begin by discussing the computational implications of modelling a larger network of sites with each of the three hidden Markov models. Next we summarise the results of applying the models to the UK dataset, including assessments of the fit of each model. For notational convenience, we adopt the abbreviations from Chapter 6 and refer to the models in Chapters 4 and 5, and a homogeneous version of the model from Chapter 6, using the acronyms CI–HMM, MCA–NHMM and LG–HMM, respectively.

7.4.1 Scalability of inferential procedures and model simplifications

The simple CI–HMM from Chapter 4 could be applied to the UK dataset directly and without any simplification to the model. Analysis via MCMC generated draws from the posterior distribution which converged quickly and mixed well. In general, the time taken to generate a particular number of MCMC samples was around 1.5 times larger than the time required in the analysis of the Yorkshire dataset.

For the MCA–NHMM introduced in Chapter 5, computation of the normalising constants in the Markov chain of autologistic models presented a difficult challenge, with each calculation involving a sum over $2^{12} = 4096$ terms. Correspondingly, the time taken to obtain 1,000 draws from the posterior was around 800 times greater than the time required when analysing the Yorkshire dataset. One way of dealing with this problem would have been to approximate the normalising constants using one of the techniques discussed in Section 5.2.2, for example, path sampling. However, to avoid such approximations we chose to partition the sites into non–overlapping groups of neighbours which were conditionally independent, given the weather state. Conditional on the weather state, this allowed the rainfall occurrences in each group to be modelled through independent Markov chains of autologistic models. As a consequence, the overall normalising constants were just products of the normalising constants for the conditionally independent groups. As long as all of the groups contained substantially fewer than twelve sites, this made it feasible to compute the normalising constants exactly.

Letting $\mathbf{d}^i = (d_1^i, \ldots, d_T^i)^T$ denote the time series of rainfall occurrences at the i–th site, the general problem of partitioning the sites, or more precisely the data $\{\mathbf{d}^1, \ldots, \mathbf{d}^n\}$, into k disjoint groups is one of cluster analysis; see, for example, Everitt (1993) or Hartigan (1975) for an

introduction. We chose to select simultaneously the number of groups $k \in \{1,\ldots,12\}$ and the set of k groups $C(k) = \{C_1(k),\ldots,C_k(k)\}$ by maximising the multi-attribute value function given by

$$V\{C(k),k\} = \left[\frac{A_0 - A\{C(k),k\}}{A_0}\right]\left[\frac{B_1+1}{B(k)+1}\right]^{\epsilon_1}\left[\frac{E_1}{E\{C(k),k\}}\right]^{\epsilon_2}.$$

A value function is essentially the opposite of a loss function; see Keeney & Raiffa (1976) for a formal definition. In this case $A\{C(k),k\}$, $B(k)$ and $E\{C(k),k\}$ are loss functions given by

$$A\{C(k),k\} = \sum_{j=1}^{k}\sum_{\mathbf{d}^i \in C_j(k)} \|\mathbf{d}^i - \bar{\mathbf{d}}_j(k)\|^2, \quad B(k) = (k-1), \quad E\{C(k),k\} = \sum_{i=1}^{k} 4^{m_i(k)},$$

in which $\bar{\mathbf{d}}_j(k)$ is the mean of the j–th group when there are k groups, $m_i(k)$ is the number of sites in the i–th group when there are k groups and $\epsilon_1, \epsilon_2 > 0$ are fixed constants. Further, A_0 is the maximum ("worst") value of $A\{C(k),k\}$, given by

$$A_0 = \sum_{i=1}^{n} \|\mathbf{d}^i - \bar{\mathbf{d}}_1(1)\|,$$

whilst $B_1 = 0$ and $E_1 = 4n$ are the minimum ("best") values of $B(k)$ and $E\{C(k),k\}$, respectively. For fixed k, the loss function $A\{C(k),k\}$ is just the usual k–means loss function. It assigns small losses to partitions with a high degree of within–group similarity. Allowing k to vary, this loss function favours larger values of k since its (fixed k) minimum cannot increase as k increases. The second loss function $B(k)$ is intended to penalise partitions which lead to a large number of assumptions of conditional independence between sites and favours small values of k. Finally, $E\{C(k),k\}$ is designed to penalise partitions which lead to Markov chains of autologistic models whose normalising constants cannot be calculated without excessive computational expense. This loss function generally favours large values of k.

The loss functions are combined through a multiplicative multi–attribute value function so that partitions are only valued highly if they perform well in all three attributes. The exponents ϵ_1 and ϵ_2 will clearly affect the results so they were chosen in a principled manner by matching preferences over partitions $\{C(k),k\}$, leading to $\epsilon_1 = 2.45$ and $\epsilon_2 = 0.15$. The resulting value function was then maximised by a partition into $k = 2$ equally sized groups. These represented the six sites in the west of the UK and the six sites in the east. The computational burden for this model was reduced further by assuming that the site and state specific temporal trend parameters in the Markov chain of autologistic models were constant across sites, that is, $\gamma_{1k} = \cdots = \gamma_{nk}$ for each weather state, $k \in \mathcal{S}_r$. This reduced the number of times the complete data likelihood had to be evaluated per sweep through the MCMC scheme since parameters were updated one–at–a–time from their full conditional distributions. Following these simplifications, the computing time for the MCMC analysis was no longer unmanageable.

When applying the LG–HMM to the UK dataset, we found that even with *a priori* correlations as high as 99.9% between the coefficient matrices $\gamma_1,\ldots,\gamma_r$, the MCMC algorithm generated posterior samples which converged slowly and were highly autocorrelated. These problems were substantially reduced by assuming a constant coefficient matrix $\gamma_1 = \cdots = \gamma_r$. The time taken to obtain a particular number of draws from the resulting model was then around 1.5 times longer than the time required for the Yorkshire analysis.

7.4.2 Posterior inference and model checking

For each of the three models we computed the posterior distribution for the number of states r, where $r \in \{1, \ldots, 5\}$, and found that the posterior probability for $r = 5$ was approximately equal to one. Conditional on there being $r = 5$ states, analysis of the posterior distributions for the model parameters in each of the three hidden Markov models revealed there to be one clear–cut wet weather state, one clear–cut dry weather state and three intermediate states, the properties of which differed between the three models. The simple CI–HMM identified one weather state with higher than average probabilities of rain at the eastern sites and lower than average probabilities at the western sites and another state characterised by the opposite conditions. Similarly, the MCA–NHMM identified two states which represented a north/south division of the sites. When rainfall occurrences are assumed to be conditionally independent, given the weather state, Section 4.2.3 showed that by assigning probabilities of rain which are similar within regions but different across regions, within–region positive correlation can be induced, with negative or no correlation between sites in different regions. Conditional on the weather state, the simplified version of the MCA–NHMM from Chapter 5 assumed eastern and western sites to be internally correlated but mutually independent. Therefore it seems that by identifying pairs of states which represent opposite conditions in eastern/western or northern/southern parts of the UK, the CI–HMM and MCA–NHMM were able to induce correlations between sites that were conditionally independent in space, given the weather state. The LG–HMM did not make this assumption for any pair of sites which may explain why the states identified by this model did not represent any kind of geographical division of the sites.

The MCA–NHMM introduced atmospheric data in the form of Lamb weather types. In the application to the Yorkshire dataset, our elicitation strategy for the parameters of the weather state process $A^x_{r,jk} = \Pr(S_t = k \mid S_{t-1} = j, X_t = x, \theta_{r,\text{hid}}, r)$ led to *a priori* correlations of around 80% between the stochastic vectors $\mathbf{A}^1_{r,j}, \ldots, \mathbf{A}^{27}_{r,j}$ for every $j \in \mathcal{S}_r$ and each $r \in \{1, \ldots, r_{\max}\}$. In the UK application we had to make the *a priori* correlations very high, ultimately around 98%, because smaller values led to overestimation of the overall proportion of wet days at most sites. This was because the data suggested that the wetter weather states had a substantially shorter sojourn time than that predicted by the prior. When there was less borrowing of strength between Lamb weather types, the probabilities of self transition $A^x_{r,jj}$ corresponding to less common Lamb weather types were pulled more strongly towards the prior mean, increasing the length of time spent in the wetter weather states. An analysis of the posterior distributions for $A^1_{5,jk}, \ldots, A^{27}_{5,jk}$ for each pair (j, k), $j, k \in \mathcal{S}_5$, showed there to be very little difference between the transition probabilities for different Lamb weather types. This suggests that the Lamb weather types are not a rich enough source of atmospheric information when the NHMM is applied to datasets with large geographical coverage. Although this result will have been influenced by the choice of high *a priori* correlations amongst the transition probabilities, the data were not sufficiently informative about some of the parameters to allow a weaker specification to be used effectively. Due to the similarity in the posterior distributions for $\mathbf{A}^1_{r,j}, \ldots, \mathbf{A}^{27}_{r,j}$ we chose not to condition the model from Chapter 6 on the atmospheric data.

In order to assess the fit of each model, we compared the posterior predictive distributions for the test quantities introduced in Chapter 4 with the observed statistics. We focus on spatial characteristics here as our conclusions regarding other test quantities did not differ appreciably

from those based on the Yorkshire applications. Initially it seemed plausible that the assumption of conditional spatial independence in the observed process, given the weather state, may be a more realistic conjecture for the UK dataset than it had been for the dense Yorkshire network of sites. However the simple CI–HMM still underestimated the strongest spatial autocorrelations between both rainfall occurrences and non–zero rainfall amounts. Prediction of the log odds ratios between rainfall occurrences was improved by the MCA–NHMM, but some of the observed statistics lay beyond the 97.5% point in their posterior predictive distribution. Each of these corresponded to a pair of sites (i, j) in which site i came from Group 1 and site j came from Group 2 or *vice versa*. This highlights the limitations of handling the normalising constant in the Markov chain of autologistic distribution in the manner described in Section 7.4.1. In comparison with these earlier models, the LG–HMM predicted log odds ratios and Spearman's rank correlation coefficients which matched those in the observed data much more closely. Correspondingly, for each $r \in \{1, \ldots, 5\}$, the marginal likelihood for an r–state model was largest for the LG–HMM and smallest for the CI–HMM.

In summary, our inferential techniques for the CI–HMM and LG–HMM scaled up well to handle a larger dataset. This was not true of the MCA–NHMM for which computation of the normalising constants in the Markov chain of autologistic models was problematic. Although MCMC analysis of the simple CI–HMM was straightforward, this model failed to provide a good explanation of the spatial autocorrelation between sites, in spite of the large distances between them. For a network of sites covering a large geographical region, we found that the Lamb weather types were not a rich enough source of atmospheric information to be helpful in explaining the transition probabilities between weather states. It is possible that continuous atmospheric data, such as sea level pressure or air temperature measurements, could be a more useful source of information. Future work could therefore attempt to incorporate this kind of atmospheric data into the weather state process. Alternatively, denoting these continuous covariates by $\mathbf{X}_1, \ldots, \mathbf{X}_T$, we could factorise the joint distribution for $\{(\mathbf{W}_t, \mathbf{D}_t, S_t)\}$, given $\{\mathbf{X}_t\}$, such that $\mathbf{X}_t$ and S_t were *both* parents of $(\mathbf{W}_t, \mathbf{D}_t)$, but the weather states evolved as a *homogeneous* first order Markov chain.

7.5 Future work

Throughout this thesis, we have highlighted areas in which further work would be valuable. In retrospect, there are two directions that might be particularly worthy of exploration. These are discussed below.

7.5.1 Modifications to within–state models in "extreme" states

During MCMC simulation for the models in Chapters 5 and 6, the Markov chains for some of the parameters in the clear–cut wet and dry states mixed poorly and prevented the sampler from converging (within a reasonable time), unless strong priors were chosen. We attributed this to the inherent capacity of hidden Markov models to partition data into homogeneous segments, each associated with realisations from only a small subspace of the space of observable outcomes. To guard against this problem in future work, it may be useful to specify different within–state

models for clear–cut wet and dry states, with fewer, if any, parameters.

To this end, suppose there are more than $r = 2$ states. Let us specify, *a priori*, that states 1 and 2 can only be associated with days when it rains at no sites or at all sites, respectively. In other words

$$\Pr(\mathbf{D}_t = \mathbf{d}_t \mid S_t = 1) = \begin{cases} 1, & \text{if } \mathbf{d}_t = \mathbf{0}_n \\ 0, & \text{otherwise} \end{cases}, \quad \Pr(\mathbf{D}_t = \mathbf{d}_t \mid S_t = 2) = \begin{cases} 1, & \text{if } \mathbf{d}_t = \mathbf{1}_n \\ 0, & \text{otherwise} \end{cases}$$

where $\mathbf{0}_n$ and $\mathbf{1}_n$ are n–vectors of 0's and 1's, respectively. We would then choose a joint distribution for the rainfall amounts at all n sites in the case of rain everywhere (state 2). For example, we might model them using the multivariate lognormal distribution, with location parameter depending on the rainfall occurrence indicator the previous day. Under this model, days with rain at no or all sites could, in principle, be explained by other states, but we would expect most days with these properties to be assigned to states 1 and 2.

As an alternative, we could make weather states 1 and 2 observable. This would reduce the state space of the sampler, which could lead to improved mixing. However, we would then have to assign zero probability to the occurrence of $\mathbf{0}_n$ or $\mathbf{1}_n$ in other states, and this may complicate the analysis.

7.5.2 Model choice through pragmatic posterior predictive loss

The marginal likelihood and therefore posterior model probabilities are sensitive to the choice of prior for the model parameters. In analyses involving complex models, such as the NHMM in Chapter 6, computational considerations or difficulties in conveying prior information can make it difficult to construct a prior which is truly representative of our initial beliefs. If this is the case, and we cannot balance the prior information across models, then we have an argument against discriminating between them using posterior model probabilities.

G.E.P. Box famously wrote, "All models are wrong but some are useful" (Box, 1979). If we do not really believe that any model in the set under consideration is the "true" model, then perhaps we should abandon the idea of assigning prior probabilities to models and dedicate ourselves to the more pragmatic pursuit of the "most useful" model. To do this, we need some justifiable means of quantifying how useful the models under consideration are, in relative terms. If prediction is the main goal of the analysis, as is typically the case in rainfall modelling, then it would seem sensible to devise a criterion which is based on the posterior predictive performance of a model (for example, see Gelfand & Ghosh, 1998) but which also reflects application–specific practical considerations. This is in the same spirit as the clustering analysis outlined in Section 7.4.1 which was based on a multi–attribute value function, some of whose components reflected practical preferences.

Gelfand & Ghosh (1998) adopt a formal utility maximisation approach to model choice. Replacing utilities with losses, a criterion is obtained by minimising posterior predictive loss for a given model. Out of the models under consideration, the one which minimises this criterion is then selected. Formally this involves calculating

$$\min_m \left\{ \min_{\mathbf{a}} \mathrm{E}_{\mathbf{y}_{\text{rep}} \mid \mathbf{y}_{\text{obs}}, m} L(\mathbf{y}_{\text{rep}}, \mathbf{a} \mid \mathbf{y}_{\text{obs}}) \right\} \tag{7.1}$$

and choosing the model $m \in \{1, \ldots, M\}$ to which this minimum corresponds. Here, $\mathbf{y}_{rep}$ is a hypothetical replicate of the observed data $\mathbf{y}_{obs}$ (with the same first stage distribution as $\mathbf{y}_{obs}$), $\mathbf{a}$ is the action vector (in this case an estimate trying to accommodate the partially observed "state of nature" $(\mathbf{y}_{obs}, \mathbf{y}_{rep})$) and the loss function $L(\mathbf{y}_{rep}, \mathbf{a} \mid \mathbf{y}_{obs})$ quantifies the loss for guessing $\mathbf{a}$, when $\mathbf{y}_{rep}$ obtains and $\mathbf{y}_{obs}$ was observed. Note that the expectation is with respect to the posterior predictive distribution for $\mathbf{y}_{rep}$ under model m. For various choices of loss function, the authors show that their criterion partitions into what can be interpreted as a goodness-of-fit term and a penalty term. In contrast to criteria such as the AIC or BIC, the penalty term arises without having to specify model dimension or appeal to asymptotic theory.

If we are prepared to accept that we are simply seeking to advise on the "most useful" model, then our model choice criterion should quantify the practical considerations which would cause us to favour some models over others. The loss function, $L(\mathbf{y}_{rep}, \mathbf{a} \mid \mathbf{y}_{obs})$, proposed by Gelfand & Ghosh (1998) cannot do this because it does not depend on the model. We might solve this problem by modifying the criterion in (7.1) so that the loss function is replaced with $L(m, \mathbf{y}_{rep}, \mathbf{a} \mid \mathbf{y}_{obs})$. If we think that capturing the consequences of m have nothing in common with capturing those of $(\mathbf{y}_{rep}, \mathbf{a})$, given $\mathbf{y}_{obs}$, we could then partition the loss function as $L_1(\mathbf{y}_{rep}, \mathbf{a} \mid \mathbf{y}_{obs}) + L_2(m \mid \mathbf{y}_{obs})$. It might be reasonable to assume that $L_2(m \mid \mathbf{y}_{obs}) = L_2(m)$, then $L_2(m)$ could represent the loss of advising m as the model of choice. If, based on this advice, computer simulations would need to be re-run (perhaps in light of different or new data) then this might include some measure of the computing time. As another example, if the chosen model would need to be explained to some scientists with limited statistical understanding, then $L_2(m)$ might include some measure of the interpretability of model m. The structure of $L_2(m)$ should be formulated with consideration to the problem at hand.

Developing these ideas further would be an interesting direction for future research.

Appendix A

MCMC scheme for Chapter 5

Denote the collection of missing rainfall occurrences by $\mathbf{d}_{\text{miss}}$, then the MCMC scheme can proceed as follows. Initialise the algorithm with a sequence of weather states $(\mathbf{s}^{[0]}, \mathbf{s}_0^{[0]})$, missing data $(\mathbf{d}_0^{[0]}, \mathbf{d}_{\text{miss}}^{[0]})$ and model parameters $\theta^{[0]}$. Then at each iteration $\ell = 1, 2, \ldots$ perform a fixed sweep of the following steps:

1. Simulate $\theta^{[\ell]}$ from $\pi(\theta \mid \mathbf{s}^{[\ell-1]}, \mathbf{s}_0^{[\ell-1]}, \mathbf{w}, \mathbf{d}, \mathbf{d}_0^{[\ell-1]}, \mathbf{x})$:

   (a) Simulate θ_{hid} from $\pi(\theta_{\text{hid}} \mid \mathbf{s}^{[\ell-1]}, \mathbf{s}_0^{[\ell-1]}, \mathbf{x})$ by successively passing through the following Gibbs (or Metropolis–within–Gibbs) steps::

      (i) Perform Metropolis Hastings updates of $(\xi_j, \mathbf{A}_j)$, for each $j \in S_r$, where $\mathbf{A}_j = (\mathbf{A}_j^1, \ldots, \mathbf{A}_j^{27})$:

         I. Generate a proposal value by first drawing

$$\xi_j^* \mid \xi_j^{[\ell-1]} \sim q_1(\xi_j^* \mid \xi_j^{[\ell-1]}) \equiv \mathscr{D}_r(\omega_d \xi_j^{[\ell-1]} + \epsilon \mathbf{1}_r),$$

         and then simulating $\mathbf{A}_j^* \mid \xi_j^* \sim q_2(\mathbf{A}_j \mid \xi_j^*)$ by drawing

$$(\mathbf{A}_j^x)^* \mid \xi_j^*, \mathbf{s}^{[\ell-1]}, \mathbf{s}_0^{[\ell-1]} \sim \mathscr{D}_r\left(\Xi_j \xi_j^* + \mathbf{n}_j^x(\mathbf{s}_0^{[\ell-1]}, \mathbf{s}^{[\ell-1]})\right)$$

         for each $x \in \mathcal{Q}$, independently.

         II. Evaluate the acceptance probability of the proposed move, $\alpha\{(\xi_j^{[\ell-1]}, \mathbf{A}_j^{[\ell-1]}), (\xi_j^*, \mathbf{A}_j^*)\}$, as defined in equations (5.36) and (5.37).

         III. Set $\xi_j^{[\ell]} = \xi_j^*$ and $(\mathbf{A}_j^x)^{[\ell]} = (\mathbf{A}_j^x)^*$, $x \in \mathcal{Q}$, with probability $\alpha\{(\xi_j^{[\ell-1]}, \mathbf{A}_j^{[\ell-1]}), (\xi_j^*, \mathbf{A}_j^*)\}$ and set $\xi_j^{[\ell]} = \xi_j^{[\ell-1]}$ and $(\mathbf{A}_j^x)^{[\ell]} = (\mathbf{A}_j^x)^{[\ell-1]}$, $x \in \mathcal{Q}$, otherwise.

      (ii) Simulate $\nu \mid \mathbf{s}^{[\ell-1]}, \mathbf{s}_0^{[\ell-1]} \sim \mathscr{D}_r\{G\mathbf{g} + \mathbf{m}(\mathbf{s}_0^{[\ell-1]})\}$.

   (b) Simulate θ_{obs} from $\pi(\theta_{\text{obs}} \mid \mathbf{s}^{[\ell-1]}, \mathbf{w}, \mathbf{d}, \mathbf{d}_0^{[\ell-1]})$ by successively passing through the following Gibbs (or Metropolis–within–Gibbs) steps:

(i) Simulate $\alpha_k \mid \cdots \sim \mathrm{N}\left(\dfrac{a_{1,\alpha}^2 \sum_{i=1}^{n} \alpha_{ik}^{[\ell-1]} + a_{0,\alpha}(\sigma_{\alpha,k}^2)^{[\ell-1]}}{n a_{1,\alpha}^2 + (\sigma_{\alpha,k}^2)^{[\ell-1]}}, \dfrac{a_{1,\alpha}^2 (\sigma_{\alpha,k}^2)^{[\ell-1]}}{n a_{1,\alpha}^2 + (\sigma_{\alpha,k}^2)^{[\ell-1]}}\right)$ for each $k \in \mathcal{S}_r$.

(ii) Simulate $\sigma_{\alpha,k}^2 \mid \cdots \sim \mathrm{IG}\left(\frac{1}{2}n + h_{0,\alpha}, \frac{1}{2}\sum_{i=1}^{n}(\alpha_{ik}^{[\ell-1]} - \alpha_k^{[\ell]})^2 + h_{1,\alpha}\right)$ for each $k \in \mathcal{S}_r$.

(iii) Simulate $\beta_k \mid \cdots \sim \mathrm{N}(m_{\beta_k,p}, V_{\beta_k,p})$ for each $k \in \mathcal{S}_r$, where

$$m_{\beta_k,p} = \frac{a_{1,\beta}^2 \sum_{i=2}^{n} \sum_{j=1}^{i-1} \beta_{ijk}^{[\ell-1]} + a_{0,\beta}(\sigma_{\beta,k}^2)^{[\ell-1]}}{\frac{1}{2}n(n-1)a_{1,\beta}^2 + (\sigma_{\beta,k}^2)^{[\ell-1]}} \quad \text{and}$$

$$V_{\beta_k,p} = \frac{a_{1,\beta}^2 (\sigma_{\beta,k}^2)^{[\ell-1]}}{\frac{1}{2}n(n-1)a_{1,\beta}^2 + (\sigma_{\beta,k}^2)^{[\ell-1]}}.$$

(iv) Simulate $\sigma_{\beta,1}^2 \mid \cdots \sim \mathrm{IG}\left(\frac{1}{4}n(n-1) + h_{0,\beta}, \frac{1}{2}\sum_{i=2}^{n}\sum_{j=1}^{i-1}(\beta_{ijk}^{[\ell-1]} - \beta_k^{[\ell]})^2 + h_{1,\beta}\right)$ for each $k \in \mathcal{S}_r$.

(v) Simulate $\gamma_k \mid \cdots \sim \mathrm{N}\left(\dfrac{a_{1,\gamma}^2 \sum_{i=1}^{n} \gamma_{ik}^{[\ell-1]} + a_{0,\gamma}(\sigma_{\gamma,k}^2)^{[\ell-1]}}{n a_{1,\gamma}^2 + (\sigma_{\gamma,k}^2)^{[\ell-1]}}, \dfrac{a_{1,\gamma}^2 (\sigma_{\gamma,k}^2)^{[\ell-1]}}{n a_{1,\gamma}^2 + (\sigma_{\gamma,k}^2)^{[\ell-1]}}\right)$ for each $k \in \mathcal{S}_r$.

(vi) Simulate $\sigma_{\gamma,1}^2 \mid \cdots \sim \mathrm{IG}\left(\frac{1}{2}n + h_{0,\gamma}, \frac{1}{2}\sum_{i=1}^{n}(\gamma_{ik}^{[\ell-1]} - \gamma_k^{[\ell]})^2 + h_{1,\gamma}\right)$ for each $k \in \mathcal{S}_r$.

(vii) Perform Metropolis Hastings updates of α_{ik} for each pair $(i,k) \in \{1,2,\ldots,n\} \times \mathcal{S}_r$:

I. Generate a proposal value

$$\alpha_{ik}^* \mid \alpha_{ik}^{[\ell-1]} \sim q(\alpha_{ik}^{[\ell-1]}, \alpha_{ik}^*) \equiv \mathrm{N}(\alpha_{ik}^{[\ell-1]}, \omega_\alpha^t).$$

II. Evaluate the acceptance probability of the proposed move, $\alpha(\alpha_{ik}^{[\ell-1]}, \alpha_{ik}^*)$, as defined in equations (5.41) and (5.42).

III. Set $\alpha_{ik}^{[\ell]} = \alpha_{ik}^*$ with probability $\alpha(\alpha_{ik}^{[\ell-1]}, \alpha_{ik}^*)$ and set $\alpha_{ik}^{[\ell]} = \alpha_{ik}^{[\ell-1]}$ otherwise.

(viii) Perform Metropolis Hastings updates of β_{ijk} for each $(i,j,k) \in \{(i,j,k) : i = 2,\ldots,n, j = 1,\ldots,i-1, k = 1,\ldots,r\}$ using analogous symmetric Gaussian random walks to those in step 1(b)(vii).

(ix) Perform Metropolis Hastings updates of γ_{ik} for each pair $(i,k) \in \{1,2,\ldots,n\} \times \mathcal{S}_r$ using analogous symmetric Gaussian random walks to those in step 1(b)(vii).

(x) Simulate $m_{ik} \mid \cdots \sim \mathrm{IG}\left\{b_{1i} + \dfrac{T_{ik}^1(\mathbf{s}^{[\ell-1]})}{v_{ik}^2}, b_{2i} + \dfrac{T_{ik}^1(\mathbf{s}^{[\ell-1]})\bar{w}_{ik}(\mathbf{s}^{[\ell-1]})}{v_{ik}^2}\right\}$ for each pair $(i,k) \in \{1,2,\ldots,n\} \times \mathcal{S}_r$

(xi) Perform Metropolis Hastings updates of v_{ik} for each pair $(i,k) \in \{1,2,\ldots,n\} \times \mathcal{S}_r$:

I. Generate a proposal value

$$v_{ik}^* \mid v_{ik}^{[\ell-1]} \sim q(v_{ik}^{[\ell-1]}, v_{ik}^*) \equiv \mathrm{Ga}\left(\omega_v^t, \frac{\omega_v^t}{v_{ik}^{[\ell-1]}}\right).$$

II. Evaluate the acceptance probability of the proposed move, $\alpha(v_{ik}^{[\ell-1]}, v_{ik}^{*})$, as defined in equations (4.25) and (4.26).

III. Set $v_{ik}^{[\ell]} = v_{ik}^{*}$ with probability $\alpha(v_{ik}^{[\ell-1]}, v_{ik}^{*})$ and set $v_{ik}^{[\ell]} = v_{ik}^{[\ell-1]}$ otherwise.

2. Simulate $(\mathbf{s}^{[\ell]}, \mathbf{s}_0^{[\ell]})$ from $\pi(\mathbf{s}, \mathbf{s}_0 \mid \theta^{[\ell]}, \mathbf{w}, \mathbf{d}, \mathbf{d}_0^{[\ell-1]}, \mathbf{x})$ by applying the forward backward scheme outlined in Algorithm 3.3.3 (and made specific to the NHMM in Section 5.6) separately to each sub–series.

3. Simulate $\mathbf{d}_0^{[\ell]}$ from $\pi(\mathbf{d}_0 \mid \mathbf{w}, \mathbf{d}, \mathbf{s}^{[\ell]}, \mathbf{s}_0^{[\ell]}, \theta^{[\ell]}, \mathbf{x})$ then, if there are any missing data, simulate $\mathbf{d}_{\text{miss}}^{[\ell]}$ from $\pi(\mathbf{d}_{\text{miss}} \mid \mathbf{w}, \mathbf{d}, \mathbf{d}_0^{[\ell]}, \mathbf{s}^{[\ell]}, \mathbf{s}_0^{[\ell]}, \theta^{[\ell]}, \mathbf{x})$:

(a) Simulate $D_{0,y}^{i} \mid \cdots \sim \text{Bern}(P_{0,iy})$ for each pair $(i, y) \in \{1, \ldots, n\} \times \{1, \ldots, Y\}$, where $P_{0,iy}$ is equal to

$$C_{s_{Tv+1}}\{\mathcal{I}(d_{0,y}^{i} = 0, \mathbf{d}_{0,y}^{-i})\} \exp(\gamma_{is_{Tv+1}} d_{Tv+1}^{i}) p_0^{i} \times \Big[C_{s_{Tv+1}}\{\mathcal{I}(d_{0,y}^{i} = 1, \mathbf{d}_{0,y}^{-i})\}(1 - p_0^{i}) + C_{s_{Tv+1}}\{\mathcal{I}(d_{0,y}^{i} = 0, \mathbf{d}_{0,y}^{-i})\} \exp(\gamma_{is_{Tv+1}} d_{Tv+1}^{i}) p_0^{i} \Big]^{-1}$$

(b) Simulate $D_t^{i} \mid \cdots \sim \text{Bern}(P_{it})$ for each pair $(i, t) \in \{1, \ldots, n\} \times \{1, \ldots, T\}$ such that d_t^{i} is missing where P_{it} is equal to

$$C_{s_{t+1}}\{\mathcal{I}(d_t^{i} = 0, \mathbf{d}_t^{-i}) \mid \theta_{\text{obs},s_{t+1}}\} \exp\left(\alpha_{is_t} + \sum_{\ell \neq i} \beta_{i\ell s_t} d_t^{\ell} + \gamma_{is_t} d_{t-1}^{i} + \gamma_{is_{t+1}} d_{t+1}^{i}\right)$$

$$\times \Big[C_{s_{t+1}}\{\mathcal{I}(d_t^{i} = 1, \mathbf{d}_t^{-i}) \mid \theta_{\text{obs},s_{t+1}}\}$$

$$+ C_{s_{t+1}}\{\mathcal{I}(d_t^{i} = 0, \mathbf{d}_t^{-i}) \mid \theta_{\text{obs},s_{t+1}}\} \exp\left(\alpha_{is_t} + \sum_{\ell \neq i} \beta_{i\ell s_t} d_t^{\ell} + \gamma_{is_t} d_{t-1}^{i} + \gamma_{is_{t+1}} d_{t+1}^{i}\right) \Big]^{-1},$$

if $T^{y} + 1 \leq t \geq T^{y+1} - 1$, $y = 1, \ldots, Y$, and

$$P_{it} = \frac{\exp\left(\alpha_{is_t} + \sum_{\ell \neq i} \beta_{i\ell s_t} d_t^{\ell} + \gamma_{is_t} d_{t-1}^{i}\right)}{1 + \exp\left(\alpha_{is_t} + \sum_{\ell \neq i} \beta_{i\ell s_t} d_t^{\ell} + \gamma_{is_t} d_{t-1}^{i}\right)}$$

if $t = T^{y+1}$, $y = 1, \ldots, Y$.

Appendix B

FCD for the regression coefficients in a multivariate normal linear regression model with a conjugate prior

Suppose the likelihood for a q–dimensional parameter $\boldsymbol{\theta}$ can be written as

$$L(\boldsymbol{\theta} \mid \cdots) \propto \exp\left\{-\frac{1}{2}\sum_{t\in S}(\mathbf{Z}_t - \mathbf{X}_t\boldsymbol{\theta})^T\boldsymbol{\Sigma}_t^{-1}(\mathbf{Z}_t - \mathbf{X}_t\boldsymbol{\theta})\right\}$$

where S is some subset, $S \subseteq \{1,\ldots,T\}$, with $T \in \mathbb{N}$.

Differentiating the loglikelihood, $\ell(\boldsymbol{\theta} \mid \cdots) = \log L(\boldsymbol{\theta} \mid \cdots)$, with respect to $\boldsymbol{\theta}$ gives

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}} = \sum_{t\in S}\mathbf{X}_t^T\boldsymbol{\Sigma}_t^{-1}(\mathbf{Z}_t - \mathbf{X}_t\boldsymbol{\theta}),$$

then equating to zero and solving yields the least squares estimate of $\boldsymbol{\theta}$ as

$$\hat{\boldsymbol{\theta}} = \mathbf{W}^{-1}\sum_{t\in S}\mathbf{X}_t^T\boldsymbol{\Sigma}_t^{-1}\mathbf{Z}_t \tag{B.1}$$

where

$$\mathbf{W} = \sum_{t\in S}\mathbf{X}_t^T\boldsymbol{\Sigma}_t^{-1}\mathbf{X}_t. \tag{B.2}$$

In the likelihood, the argument of the exponential function can be written as

$$\begin{aligned}
&-\frac{1}{2}\sum_{t\in S}(\mathbf{Z}_t - \mathbf{X}_t\boldsymbol{\theta})^T\boldsymbol{\Sigma}_t^{-1}(\mathbf{Z}_t - \mathbf{X}_t\boldsymbol{\theta}) \\
&\qquad = -\frac{1}{2}\sum_{t\in S}\{\mathbf{Z}_t - \mathbf{X}_t\hat{\boldsymbol{\theta}} + \mathbf{X}_t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\}^T\boldsymbol{\Sigma}_t^{-1}\{\mathbf{Z}_t - \mathbf{X}_t\hat{\boldsymbol{\theta}} + \mathbf{X}_t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\},
\end{aligned}$$

that is,

$$-\frac{1}{2}\sum_{t\in S}(\mathbf{Z}_t-\mathbf{X}_t\theta)^T\boldsymbol{\Sigma}_t^{-1}(\mathbf{Z}_t-\mathbf{X}_t\theta)$$

$$=-\frac{1}{2}\left\{\sum_{t\in S}(\hat{\theta}-\theta)^T\mathbf{X}_t^T\boldsymbol{\Sigma}_t^{-1}\mathbf{X}_t(\hat{\theta}-\theta)+2\sum_{t\in S}(\mathbf{Z}_t-\mathbf{X}_t\hat{\theta})^T\boldsymbol{\Sigma}_t^{-1}\mathbf{X}_t(\hat{\theta}-\theta)\right.$$

$$\left.+\sum_{t\in S}(\mathbf{Z}_t-\mathbf{X}_t\hat{\theta})^T\boldsymbol{\Sigma}_t^{-1}(\mathbf{Z}_t-\mathbf{X}_t\hat{\theta})\right\}. \tag{B.3}$$

Expanding the second term in (B.3) and using (B.1) and (B.2) yields

$$\sum_{t\in S}(\mathbf{Z}_t-\mathbf{X}_t\hat{\theta})^T\boldsymbol{\Sigma}_t^{-1}\mathbf{X}_t(\hat{\theta}-\theta)$$

$$=\sum_{t\in S}\mathbf{Z}_t^T\boldsymbol{\Sigma}_t^{-1}\mathbf{X}_t\hat{\theta}-\sum_{t\in S}\mathbf{Z}_t^T\boldsymbol{\Sigma}_t^{-1}\mathbf{X}_t\theta-\sum_{t\in S}\hat{\theta}^T\mathbf{X}_t^T\boldsymbol{\Sigma}_t^{-1}\mathbf{X}_t\hat{\theta}+\sum_{t\in S}\hat{\theta}^T\mathbf{X}_t^T\boldsymbol{\Sigma}_t^{-1}\mathbf{X}_t\theta$$

$$=\hat{\theta}^T\mathbf{W}\hat{\theta}-\hat{\theta}^T\mathbf{W}\theta-\hat{\theta}^T\mathbf{W}\hat{\theta}+\hat{\theta}^T\mathbf{W}\theta$$

$$=0.$$

It follows that

$$L(\theta\mid\cdots)\propto\exp\left\{-\frac{1}{2}\sum_{t\in S}(\mathbf{Z}_t-\mathbf{X}_t\hat{\theta})^T\boldsymbol{\Sigma}_t^{-1}(\mathbf{Z}_t-\mathbf{X}_t\hat{\theta})\right\}$$

$$\times\exp\left\{-\frac{1}{2}\sum_{t\in S}(\hat{\theta}-\theta)^T\mathbf{X}_t^T\boldsymbol{\Sigma}_t^{-1}\mathbf{X}_t(\hat{\theta}-\theta)\right\}$$

$$\propto\exp\left\{-\frac{1}{2}(\theta-\hat{\theta})^T\mathbf{W}(\theta-\hat{\theta})\right\},$$

which means that $\hat{\theta}$ is sufficient for θ.

Now suppose that the prior distribution for θ is given by

$$\theta\mid\mathbf{m}_0,\mathbf{V}_0\sim\mathrm{N}_q(\mathbf{m}_0,\mathbf{V}_0)$$

where $\mathbf{m}_0$ or $(\mathbf{m}_0,\mathbf{V}_0)$ could be given distributions if a second level was added to this prior specification.

Denote by $\mathbf{P}_0=\mathbf{V}_0^{-1}$ the prior precision matrix. Then the full conditional distribution for θ can be derived via Bayes Theorem as

$$\pi(\theta\mid\cdots)\propto\pi(\theta\mid\mathbf{m}_0,\mathbf{V}_0)L(\theta\mid\cdots)$$

$$\propto\exp\left\{-\frac{1}{2}(\theta-\mathbf{m}_0)^T\mathbf{P}_0(\theta-\mathbf{m}_0)\right\}\times\exp\left\{-\frac{1}{2}(\theta-\hat{\theta})^T\mathbf{W}(\theta-\hat{\theta})\right\}$$

$$=\exp\left[-\frac{1}{2}\left\{\theta^T(\mathbf{P}_0+\mathbf{W})\theta-2\theta^T(\mathbf{P}_0\mathbf{m}_0+\mathbf{W}\hat{\theta})+\mathbf{m}_0^T\mathbf{P}_0\mathbf{m}_0+\hat{\theta}^T\mathbf{W}\hat{\theta}\right\}\right]$$

$$\propto\exp\left\{-\frac{1}{2}(\theta-\mathbf{m}_p)^T\mathbf{V}_p^{-1}(\theta-\mathbf{m}_p)\right\}$$

where

$$\mathbf{V}_p = (\mathbf{P}_0 + \mathbf{W})^{-1} = \left(\mathbf{V}_0^{-1} + \sum_{t \in S} \mathbf{X}_t^T \boldsymbol{\Sigma}_t^{-1} \mathbf{X}_t\right)^{-1}$$

and

$$\mathbf{m}_p = (\mathbf{P}_0 + \mathbf{W})^{-1}(\mathbf{P}_0 \mathbf{m}_0 + \mathbf{W}\hat{\boldsymbol{\theta}}) = \mathbf{V}_p\left(\mathbf{V}_0^{-1}\mathbf{m}_0 + \sum_{t \in S} \mathbf{X}_t^T \boldsymbol{\Sigma}_t^{-1} \mathbf{Z}_t\right),$$

and so the full conditional distribution can be recognised as

$$\boldsymbol{\theta} \mid \cdots \sim \mathrm{N}_q(\mathbf{m}_p, \mathbf{V}_p).$$

Appendix C

Simulating from the truncated multivariate normal distribution

C.1 Accept–reject algorithms for simulating from the truncated univariate normal distribution

Denote by $\mathrm{TN}(0, 1, [a, b])$ the standard normal distribution truncated to the interval $[a, b]$, where $a, b, \in \mathbb{R}$ and a or b can be $-\infty$ or ∞, respectively. In this section we describe two accept-reject algorithms for simulating from the left truncated normal distribution, $\mathrm{TN}(0, 1, [\ell, \infty))$, with density

$$fX(x) \propto \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \mathbb{I}(x \geq \ell). \tag{C.1}$$

The symmetry of the standard normal density about zero means we do not need a different algorithm for simulating from right truncated distributions. In order to simulate from the distribution, $\mathrm{TN}(0, 1, (-\infty, r])$, with density

$$fX(x) \propto \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \mathbb{I}(x \leq r),$$

we simply simulate $X^* = x^* \sim \mathrm{TN}(0, 1, [-r, \infty))$ then set $X = -x^*$. For this purposes of this thesis, we do not need to simulate from a standard normal distribution which is truncated on both the left and the right, although algorithms are available in the literature; see, for example Geweke (1991).

The accept–reject algorithm is a technique which uses draws from a distribution that can easily be sampled to generate realisations from another distribution for which direct simulation is difficult. Consider a distribution with density function f, called the *target density*, from which a sample is needed. Then the accept–reject algorithm generates the required sample by using a density function g, under the restriction that $f(x) \leq Mg(x)$, where $M \geq 1$ is an appropriate bound on $f(x)/g(x)$. To simulate a value $X = x \sim f$ using an *instrumental density* g which satisfies the latter restriction, the accept–reject algorithm proceeds as follows

1. Generate $Y = y \sim g$;
2. Generate $U = u \sim \mathrm{U}[0,1]$. Accept $X = y$ if $u \leq f(y)/\{Mg(y)\}$. Otherwise reject and return to step 1.

When evaluated for the properly normalised densities, the probability of acceptance is $1/M$. For more details on accept–reject methods, see, for example, Robert & Casella (2005).

C.1.1 A "naive" accept–reject method

In the "naive" accept–reject algorithm, the non-truncated standard normal distribution provides the instrumental density. The algorithm is as follows

1. Generate $Y = y \sim \mathrm{N}(0,1)$;
2. Accept $X = y$ if $y \geq \ell$. Otherwise reject and return to step 1.

The probability of acceptance can easily be computed as $\Phi(-\ell)$. Clearly, when ℓ is large, many proposals will have to be generated for a single acceptance.

C.1.2 The exponential accept–reject method

Robert & Casella (2005) present the following accept–reject algorithm, which is generally much more efficient than the naive method and the inverse CDF method (details not provided).

1. Generate $Y = y$ from the translated exponential distribution, $\mathrm{Exp}(\alpha, \ell)$, with density
$$g_\alpha(y) = \alpha \exp\{-\alpha(y-\ell)\}\mathrm{I}(y \geq \ell).$$
That is, generate
$$Y^* = y^* \sim \mathrm{Exp}(\alpha),$$
then take
$$y = y^* + \ell.$$
2. Generate $U = u \sim \mathrm{U}[0,1]$. Accept $W = y$ if $u < \exp\{-\frac{1}{2}(y-\alpha)^2\}$. Otherwise reject and return to step 1.

Here, the rate parameter in the instrumental density is taken as $\alpha = \frac{1}{2}(\ell + \sqrt{\ell^2+4})$, this choice being optimal, in the sense of maximising the acceptance rate, $(2\pi)^{1/2}\alpha\Phi(-\ell)\exp\{\alpha(\ell-\alpha/2)\}$.

The acceptance probability of the exponential method will therefore exceed that of the naive method when
$$(2\pi)^{1/2}\alpha\Phi(-\ell)\exp\{\alpha(\ell-\alpha/2)\} > \Phi(-\ell).$$
Solving this equation, numerically, the exponential method has a higher acceptance rate when $\ell \geq -0.470$, to three decimal places. Note, however, that this takes no account of the computation time of one generation from the instrumental density.

C.2 A Gibbs algorithm for simulating from the truncated multivariate normal distribution (Geweke, 1991)

In this section we describe a Gibbs algorithm, due to Geweke (1991), for simulating from the truncated multivariate normal distribution. This only requires the ability to generate samples from truncated versions of the univariate standard normal distribution.

Suppose interest lies in simulating from an n–variate normal distribution, subject to restrictions on the marginal variables

$$\mathbf{X} = (X_1, \ldots, X_n)^T \sim \mathrm{N}_n(\mu, \Sigma), \qquad \mathbf{a} \leq \mathbf{X} \leq \mathbf{b},$$

where any of the components of $\mathbf{a}$ and $\mathbf{b}$ can be $-\infty$ or ∞, respectively. This is equivalent to sampling from

$$\mathbf{Y} \sim \mathrm{N}_n(0, \Sigma), \qquad \alpha \leq \mathbf{Y} \leq \beta,$$

where $\alpha = \mathbf{a} - \mu$, $\beta = \mathbf{b} - \mu$, and then taking $\mathbf{X} = \mu + \mathbf{Y}$.

The method exploits the fact that the distribution of each element of $\mathbf{Y}$, conditional on all the other elements, is truncated univariate normal. A cycle of n Gibbs steps can therefore be composed in order to simulate each of the components in turn.

At iteration j, the previous draw $\mathbf{Y}^{(j-1)}$ is used to generate a new value $\mathbf{Y}^{(j)}$, and hence $\mathbf{X}^{(j)}$, as follows

1. Generate successive values

$$\begin{aligned} Y_1^{[j]} &\sim \pi(Y_1 \mid Y_2^{[j-1]}, Y_3^{[j-1]}, \ldots, Y_n^{[j-1]}) \\ Y_2^{[j]} &\sim \pi(Y_2 \mid Y_1^{[j]}, Y_3^{[j-1]}, \ldots, Y_n^{[j-1]}) \\ &\vdots \\ Y_n^{[j]} &\sim \pi(Y_n \mid Y_1^{[j]}, Y_2^{[j]}, \ldots, Y_{n-1}^{[j]}) \end{aligned}$$

   where the distribution of a general component, Y_i, conditional on $\mathbf{Y}_{-i} = (Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n)^T$ has the construction

$$Y_i = \mathbf{c}^i \mathbf{Y}_{-i} + h_i \epsilon_i, \quad \epsilon_i \sim \mathrm{TN}\left(0, 1, \left[\frac{\alpha_i - \mathbf{c}^i \mathbf{Y}_{-i}}{h_i}, \frac{\beta_i - \mathbf{c}^i \mathbf{Y}_{-i}}{h_i}\right]\right).$$

   Here

$$\mathbf{c}^i = -P_{ii}^{-1} \mathbf{P}_{i,<i} \qquad \text{and} \qquad h_i^2 = P_{ii}^{-1}$$

   where P_{ii} is the (i,i)-th element in the precision matrix, $\mathbf{P} = \Sigma^{-1}$, and $\mathbf{P}_{i,<i}$ is a row vector of length $(n-1)$ composed of the i-th row of $\mathbf{P}$ with P_{ii} omitted.

2. Compute $\mathbf{X}^{(j)} = \mu + \mathbf{Y}^{(j)}$.

For the theoretical development of this algorithm and details concerning its accuracy, see Geweke (1991).

Appendix D

Prior specification for $(\tilde{\phi}_{r,1}, \ldots, \tilde{\phi}_{r,r} \mid r)$ and $(\tilde{\sigma}^2_{r,1}, \ldots, \tilde{\sigma}^2_{r,r} \mid r)$

In the Yorkshire data application in Chapter 6, conditional on there being $r \in \{1, \ldots, 4\}$ states, the matrix M which is responsible for reordering the sites was chosen to be

$$\mathbf{M} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Next, the hyperparameters $(C_{r,1}, \ldots, C_{r,n})$ and $(v_{r,\tilde{\sigma}^2,1}, \ldots, v_{r,\tilde{\sigma}^2,n})$ in the priors for $(\tilde{\sigma}^2_{r,1}, \ldots, \tilde{\sigma}^2_{r,r} \mid r)$, were chosen to be identical for each value of r and equal to

$$C_{r,1} = 0.524, \quad C_{r,2} = 0.174, \quad C_{r,3} = 0.224, \quad C_{r,4} = 0.158, \quad C_{r,5} = 0.158, \quad C_{r,6} = 0.169$$

and

$$v_{r,\tilde{\sigma}^2,1} = 1.85, \quad v_{r,\tilde{\sigma}^2,2} = 1.90, \quad v_{r,\tilde{\sigma}^2,3} = 1.95, \quad v_{r,\tilde{\sigma}^2,4} = 2.0, \quad v_{r,\tilde{\sigma}^2,5} = 2.05, \quad v_{r,\tilde{\sigma}^2,6} = 2.10,$$

respectively.

In the prior for $(\tilde{\phi}_{r,1}, \ldots, \tilde{\phi}_{r,r} \mid r)$, the hyperparameters $\mathbf{m}_{r,\tilde{\phi},0}$, $\mathbf{C}_{r,\tilde{\phi},0}$ and $\mathbf{V}_{r,\tilde{\phi},0}$ were chosen to be identical for each value of r and equal to

$$\mathbf{m}_{r,\tilde{\phi},0} = (0.818, 0.513, 0.277, 0.402, 0.428, 0.058, 0.183, 0.355, 0.346, 0.036, 0.331, 0.023, 0.332, 0.212, 0.008)^T$$

with $\mathbf{C}_{r,\tilde{\phi},0} = \rho_{r,\tilde{\phi}}\tilde{\mathbf{V}}_{r,\tilde{\phi},0}$ and $\mathbf{V}_{r,\tilde{\phi},0} = (1-\rho_{r,\tilde{\phi}})\tilde{\mathbf{V}}_{r,\tilde{\phi},0}$ where $\rho_{r,\tilde{\phi}} = 0.95$ and $\tilde{\mathbf{V}}_{r,\tilde{\phi},0}$ is equal to

$$\frac{1}{100}\begin{pmatrix}
100 & 10 & 10 & 40 & 25 & 10 & 25 & 25 & 10 & 25 & 40 & 40 & 25 & 10 & 40 \\
10 & 100 & 25 & 10 & 10 & 10 & 25 & 25 & 40 & 40 & 10 & 10 & 25 & 40 & 25 \\
10 & 25 & 100 & 10 & 10 & 10 & 25 & 40 & 25 & 40 & 25 & 25 & 40 & 10 & 40 \\
40 & 10 & 10 & 100 & 25 & 10 & 40 & 40 & 10 & 40 & 10 & 10 & 40 & 25 & 40 \\
25 & 10 & 10 & 25 & 100 & 25 & 25 & 40 & 25 & 10 & 10 & 10 & 40 & 25 & 25 \\
10 & 10 & 10 & 10 & 25 & 100 & 40 & 25 & 40 & 25 & 10 & 10 & 25 & 10 & 25 \\
25 & 25 & 25 & 40 & 25 & 40 & 100 & 10 & 10 & 40 & 25 & 10 & 40 & 25 & 25 \\
25 & 25 & 40 & 40 & 40 & 25 & 10 & 100 & 40 & 25 & 40 & 10 & 25 & 40 & 10 \\
10 & 40 & 25 & 10 & 25 & 40 & 10 & 40 & 100 & 10 & 25 & 40 & 40 & 25 & 40 \\
25 & 40 & 40 & 40 & 10 & 25 & 40 & 25 & 10 & 100 & 10 & 10 & 25 & 40 & 40 \\
40 & 10 & 25 & 10 & 10 & 10 & 25 & 40 & 25 & 10 & 100 & 25 & 10 & 40 & 40 \\
40 & 10 & 25 & 10 & 10 & 10 & 10 & 10 & 40 & 10 & 25 & 100 & 40 & 25 & 40 \\
25 & 25 & 40 & 40 & 40 & 25 & 40 & 25 & 40 & 25 & 10 & 40 & 100 & 10 & 25 \\
10 & 40 & 10 & 25 & 25 & 10 & 25 & 40 & 25 & 40 & 40 & 25 & 10 & 100 & 40 \\
40 & 25 & 40 & 40 & 25 & 25 & 25 & 10 & 40 & 40 & 40 & 40 & 25 & 40 & 100
\end{pmatrix}.$$

Appendix E

Glossary of notation

Table E.1 lists the main variables and parameters used in the "modelling" Chapters (4, 5, 6).

Variable/parameter	Definition/usage
Chapters 4, 5 and 6	
$\mathbf{D}_t = (D_t^1, \ldots, D_t^n)^T$	$D_t^i = 1$ if at least c mm of rain on day t at site i, $D_t^i = 0$ otherwise.
$\mathbf{W}_t = (W_t^1, \ldots, W_t^n)^T$	W_t^i is the amount of rain on day t at site i.
$S_t \in \mathcal{S}_r$	Weather state at time t; $\mathcal{S}_r = \{1, \ldots, r\}$.
$\theta = (\theta_{\text{hid}}, \theta_{\text{obs}})$	Model parameters, where θ_{hid} and θ_{obs} contain parameters associated with the weather state and observed processes, respectively.
$\nu = (\nu_1, \ldots, \nu_r) \in \mathcal{S}_r$	$\Pr(S_1 = j \mid \theta_{\text{hid}}) = \nu_j$ in Chapter 4; $\Pr(S_0 = j \mid \theta_{\text{hid}}) = \nu_j$ in Chapters 5 and 6; $\mathcal{S}_r$ is the r–dimensional unit simplex.
Chapters 4 and 5	
$\mathcal{M} = (m_{ik}) \in \mathbb{R}_+^{nr}$ and $\mathcal{V} = (v_{ik}) \in \mathbb{R}_+^{nr}$	$W_t^i \mid D_t^i = 1, S_t = k, \theta_{\text{obs}} \sim \text{Ga}\left(\frac{1}{v_{ik}^2}, \frac{1}{v_{ik}^2 m_{ik}}\right)$.
Chapter 4	
$\Lambda = (\lambda_j)$, $\lambda_j = (\lambda_{j1}, \ldots, \lambda_{jr})$ where $\lambda_j \in \mathcal{S}_r$	$\Pr(S_t = k \mid S_{t-1} = j, \theta_{\text{hid}}) = \lambda_{jk}$.
$\mathcal{P} = (p_{ik}) \in [0,1]^{nr}$	$D_t^i \mid S_t = k, \theta_{\text{obs}} \sim \text{Bern}(p_{ik})$.
$\delta = (\delta_1, \ldots, \delta_r) \in \mathcal{S}_r$	Stationary distribution of the Markov chain $\{S_t : t = 1, \ldots, T\}$.
$\mathcal{B} = (\beta_{ik}) \in \mathbb{R}_+^{nr}$	$W_t^i \mid D_t^i = 1, S_t = k, \theta_{\text{obs}} \sim \text{Exp}(\beta_{ik})$ in the simplified HMM used in the simulation experiment.
Chapters 5 and 6	
$X_t \in \mathcal{Q}$	Lamb weather type on day t; $\mathcal{Q} = \{1, \ldots, 27\}$.
$\mathbf{D}_0 = (D_0^1, \ldots, D_0^n)^T \in \{0,1\}^n$	D_0^i is the initial rainfall occurrence indicator at site i.
$S_0 \in \mathcal{S}_r$	Initial weather state.

Continued on next page

Continued from previous page

Variable/parameter	Definition/usage
$(p_0^1, \ldots, p_0^n) \in [0,1]^n$ $\mathbf{A} = (\mathbf{A}_1, \ldots, \mathbf{A}_r)$ where $\mathbf{A}_j = (\mathbf{A}_j^1, \ldots, \mathbf{A}_j^{27})$ and $\mathbf{A}_j^x = (A_{j1}^x, \ldots, A_{jr}^x) \in \mathscr{S}_r$ $\mathcal{E} = (\xi_j)$, $\xi_j = (\xi_{j1}, \ldots, \xi_{jr})$ where $\xi_j \in \mathscr{S}_r$	$D_0^i \sim \text{Bern}(p_0^i)$ independently for $i \in \{1, \ldots, n\}$. $\Pr(S_t = k \mid S_{t-1} = j, X_t = x, \theta_{\text{hid}}) = A_{jk}^x$. $\mathbf{A}_j^x \mid \xi_j \sim \mathscr{D}_r(\Xi_j \xi_j)$ independently for each $x \in Q$.
Chapter 5	
$\mathcal{A} = (\alpha_{ik}) \in \mathbb{R}^{nr}$, $\mathcal{B} = (\beta_{ijk}) \in \mathbb{R}^{n(n-1)r/2}$ and $\mathcal{G} = (\gamma_{ik}) \in \mathbb{R}^{nr}$	$\Pr(\mathbf{D}_t = \mathbf{d}_t \mid \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, S_t = k, \theta_{\text{obs}})$ $= \dfrac{\exp\left(\sum_{i=1}^n \alpha_{ik} d_t^i + \sum_{i=2}^n \sum_{j=1}^{i-1} \beta_{ijk} d_t^i d_t^j + \sum_{i=1}^n \gamma_{ik} d_t^i d_{t-1}^i\right)}{\sum_{\mathbf{d}} \exp\left(\sum_{i=1}^n \alpha_{ik} d^i + \sum_{i=2}^n \sum_{j=1}^{i-1} \beta_{ijk} d^i d^j + \sum_{i=1}^n \gamma_{ik} d^i d_{t-1}^i\right)}$ in which the denominator is denoted $C_k\{\mathcal{I}(\mathbf{d}_{t-1}) \mid \theta_{\text{obs},k}\}$. Here $\mathcal{I}(\mathbf{d}_{t-1}) = \sum_{i=1}^n d_{t-1}^i 2^{n-i} \in \{0, 1, \ldots, 2^n - 1\}$ represents the numerical labelling of $\mathbf{d}_{t-1}$.
$\mathcal{A}^0 = \{(\alpha_k, \sigma_{\alpha,k}^2) : k \in \mathcal{S}_r\}$, $\alpha_k \in \mathbb{R}$ and $\sigma_{\alpha,k}^2 \in \mathbb{R}^+$. $\mathcal{B}^0$ and $\mathcal{G}^0$ are defined analogously.	$\alpha_{ik} \mid \alpha_k, \sigma_{\alpha,k}^2 \sim \text{N}(\alpha_k, \sigma_{\alpha,k}^2)$ independently for each $i \in \{1, \ldots, n\}$. The parameters in $\mathcal{B}^0$ and $\mathcal{G}^0$ arise through analogous prior specifications.
Chapter 6	
$\mathbf{Z}_{0t} = (Z_{0t}^1, \ldots, Z_{0t}^n)^T \in \mathbb{R}^n$	Latent Gaussian random vector in the multivariate probit model for rainfall occurrence, $D_t^i = \mathbf{I}(Z_{0t}^i > 0)$. Also acts like a vector of spatial random effects in the model for $\log \mathbf{W}_t$.
$\{(\beta_{0k}, \beta_{1k}, \tilde{\phi}_k, \tilde{\sigma}_k^2) : k \in \mathcal{S}_r\}$, $\beta_{0k} \in \mathbb{R}^n$, $\beta_{1k} \in \{-1, 1\}^n$, $\tilde{\phi}_k \in \mathbb{R}^{n(n-1)/2}$, $\tilde{\sigma}_k^2 = \mathbb{R}_+^n$	$\mathbf{Z}_{0t} \mid \mathbf{D}_{t-1} = \mathbf{d}_{t-1}, S_t = k, \theta_{\text{obs}} \sim \text{N}_n(\mathbf{X}_t \beta_k, \Sigma_k)$, where $\beta_k = (\beta_{0k}^T, \beta_{1k}^T)^T$, $\mathbf{X}_t = \{\mathbf{I}_n, \text{diag}(d_{t-1}^1, \ldots, d_{t-1}^n)\}$ and $\Sigma_k^{-1} = \mathbf{M}^T \tilde{\mathbf{T}}_k^T \tilde{\mathbf{D}}_k^{-1} \tilde{\mathbf{T}}_k \mathbf{M}$. $\tilde{\mathbf{T}}_k$ is a unit lower triangular matrix with (i,j)–th entry $-\tilde{\phi}_{k,ij}$, $\tilde{\mathbf{D}}_k = \text{diag}(\tilde{\sigma}_{k,1}^2, \ldots, \tilde{\sigma}_{k,n}^2)$ and $\mathbf{M}$ is an $n \times n$ known matrix.
$\{(\mu_k, \gamma_k, \Omega_k) : k \in \mathcal{S}_r\}$, $\mu_k \in \mathbb{R}^n$, $\gamma_k = (\gamma_k^{ij}) \in \mathbb{R}^{n \times n}$, $\Omega_k = \text{diag}(\Omega_k^{11}, \ldots, \Omega_k^{nn})$ where $\Omega_k^{ii} \in \mathbb{R}^+$	$W_t^i \mid D_t^i = 1, S_t = k, \mathbf{Z}_{0t} \sim \text{LogN}\left(\mu_k^i + \sum_{j=1}^n \gamma_k^{ij} z_{0t}^j, \Omega_k^{ii}\right)$; $\mathbb{R}^{n \times n}$ denotes the set of $n \times n$ matrices with real entries.
$\{(\beta_{0k}, p_k, \mu_k, \tilde{\sigma}_k^2, \Omega_k) : k \in \mathcal{S}_r\}$, $\beta_{0k}, \mu_k \in \mathbb{R}$, $p_k \in [0,1]$, $\tilde{\sigma}_k^2, \Omega_k \in \mathbb{R}^+$	State specific means introduced at the first level of the hierarchical priors for β_{0k}, β_{1k}, μ_k, $\tilde{\sigma}_k^2$ and Ω_k, respectively.
$\{(\sigma_{\beta_0,k}^2, \sigma_{\mu,k}^2) : k \in \mathcal{S}_r\}$, $\sigma_{\beta_0,k}^2, \sigma_{\mu,k}^2 \in \mathbb{R}^+$	State specific variances introduced at the first level of the hierarchical priors for β_{0k} and μ_k, respectively.

Continued on next page

Continued from previous page	
Variable/parameter	**Definition/usage**
$\tilde{\phi} \in \mathbb{R}^{n(n-1)/2}$ and $\gamma \in \mathbb{R}^{n^2}$	Means introduced at the first level of the hierarchical priors for $(\tilde{\phi}_1, \ldots, \tilde{\phi}_r)$ and $(\gamma_1, \ldots, \gamma_r)$, respectively.

Table E.1: The main variables and parameters introduced in the models from Chapters 4, 5 and 6.

Table E.2 presents the notation, probability density/mass functions and the sample and parameter spaces for the distributions used in the "modelling" Chapters (4, 5 and 6).

Family	Notation	Probability/Density	Sample/Parameter Spaces
Poisson	$\text{Po}(\lambda)$	$\Pr(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$	$x \in \{0, 1, \ldots\}$; $\lambda \in \mathbf{R}^+$
Bernoulli	$\text{Bern}(p)$	$\Pr(X = x) = p^x (1-p)^{1-x}$	$x \in \{0, 1\}$; $p \in [0, 1]$
Scaled Bernoulli	$\text{ScBern}(p)$	$\Pr(X = x) = p^{(1+x)/2}(1-p)^{(1-x)/2}$	$x \in \{-1, 1\}$; $p \in [0, 1]$
Exponential	$\text{Exp}(\lambda)$	$p(x) = \lambda e^{-\lambda x}$	$x \in \mathbb{R}^+$; $\lambda \in \mathbb{R}^+$
Gamma	$\text{Ga}(\alpha, \lambda)$	$p(x) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}$	$x \in \mathbb{R}^+$; $\alpha, \lambda \in \mathbb{R}^+$
Normal	$\text{N}(\mu, \sigma^2)$	$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$	$x \in \mathbb{R}$; $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$
Beta	$\text{Beta}(\alpha, \beta)$	$p(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$	$x \in [0, 1]$; $\alpha, \beta \in \mathbb{R}^+$
Lognormal	$\text{LogN}(\mu, \sigma^2)$	$p(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\log x - \mu)^2}{2\sigma^2}\right\}$	$x \in \mathbb{R}^+$; $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$
Inverse Gamma	$\text{IG}(a, b)$	$p(x) = \frac{b^a x^{-a-1} e^{-b/x}}{\Gamma(a)}$	$x \in \mathbb{R}^+$; $a, b \in \mathbb{R}^+$
Dirichlet	$\mathscr{D}_d(\mathrm{a})$	$p(\mathrm{x}) = \frac{\Gamma\left(\sum_{i=1}^d a_i\right)}{\prod_{i=1}^d \Gamma(a_i)} \prod_{i=1}^d x_i^{a_i - 1}$	$\mathrm{x} \in \mathscr{S}_d$; $\mathrm{a} \in \mathbb{R}_+^d$
Multivariate normal	$\text{N}_d(\mu, \Sigma)$	$p(\mathrm{x}) = \frac{1}{(2\pi)^{d/2}\lvert\Sigma\rvert^{1/2}} \times \exp\left\{-\frac{1}{2}(\mathrm{x}-\mu)^T \Sigma^{-1} (\mathrm{x}-\mu)\right\}$	$\mathrm{x} \in \mathbb{R}^d$; $\mu \in \mathbb{R}^d$, $\Sigma \in \mathcal{D}^d$

Table E.2: Probability distributions. $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta)$ denotes the beta function; $\mathcal{D}^d$ denotes the space of positive definite symmetric $d \times d$ matrices; $\mathscr{S}_d$ denotes the space of d-dimensional unit simplices, $\mathscr{S}_r = \{(x_1, \ldots, x_r) : x_i \geq 0 \, \forall i, \sum x_i = 1\}$.

Bibliography

AILLIOT, P., THOMPSON, C. & THOMSON, P. (2009) Space-time modelling of precipitation by using a hidden Markov model and censored Gaussian distributions. *J. Roy. Statist. Soc. C* **58**, 405–426.

AITCHISON, J. (1986) *The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability* **25**. London: Chapman & Hall.

AITKIN, M. (1997) Contribution to the discussion of "On Bayesian analysis of mixtures with an unknown number of components" by S. Richardson and P. J. Green. *J. Roy. Statist. Soc. B* **59**, 764–768.

ALBERT, J. H. & CHIB, S. (1993) Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88**, 669–679.

ALBERT, J. H. & GUPTA, A. K. (1982) Mixtures of Dirichlet distributions and estimation in contingency tables. *Ann. Statist.* **10**, 1261–1268.

ALLCROFT, D. J. & GLASBEY, C. A. (2003) A latent Gaussian Markov random-field model for spatiotemporal rainfall disaggregation. *J. Roy. Statist. Soc. C* **52**, 487–498.

ALQALLAF, F. & GUSTAFSON, P. (2001) On cross-validation of Bayesian models. *Canad. J. Statist.* **29**, 333–340.

ASHFORD, J. R. & SOWDEN, R. R. (1970) Multi-variate probit analysis. *Biometrics* **26**, 535–546.

BANERJEE, S., CARLIN, B. P. & GELFAND, A. E. (2004) *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, Boca Raton, FL.

BANFIELD, J. D. & RAFTERY, A. E. (1993) Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–821.

BARDOSSY, A. & PLATE, E. J. (1992) Space-time model for daily rainfall using atmospheric circulation patterns. *Water Resour. Res.* **28**, 1247–1259.

BARNARD, J., MCCULLOCH, R. & MENG, X.-L. (2000) Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statist. Sinica* **10**, 1281–1311.

BAYARRI, M. J. & BERGER, J. O. (2000) P values for composite null models. *J. Amer. Statist. Assoc.* **95**, 1127–1142.

BELLONE, E., HUGHES, J. P. & GUTTORP, P. (2000) A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts. *Climate Res.* **15**, 1–12.

BENSMAIL, H., CELEUX, G., RAFTERY, A. E. & ROBERT, C. P. (1997) Inference in model-based cluster analysis. *Stat. Comput.* **7**, 1–10.

BERGER, J. O. & BERNARDO, J.-M. (1994) Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.* **89**, 200–207.

BESAG, J. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc. B* **36**, 192–236.

BESAG, J. (1975) Statistical analysis of non-lattice data. *The Statistician* **24**, 179–195.

BESAG, J., YORK, J. & MOLLIÉ, A. (1991) Bayesian image restoration, with two applications in spatial statistics (with discussion). *Ann. Inst. Statist. Math.* **43**, 1–59.

BESAG, J. E. (1972) Nearest-neighbour systems and the auto-logistic model for binary data. *J. Roy. Statist. Soc. B* **34**, 75–83.

BETRO, B., BODINI, A. & COSSU, Q. A. (2008) Using a hidden Markov model to analyse extreme rainfall events in Central-East Sardinia. *Environmetrics* **19**, 702–713.

BOS, C. S. (2002) A comparison of marginal likelihood computation methods. Tinbergen Institute Discussion Papers 02-084/4. Tinbergen Institute.

BOX, G. E. P. (1979) Robustness in the strategy of scientific model building. In *Robustness in Statistics* (ed. R. L. Launer & G. N. Wilkinson), pp. 201–236. New York, London: Academic Press.

BOYS, R. J., FARROW, M. & GERMAIN, S. E. (2011) Discussion of "Modelling multivariate counts varying continuously in space" by A. M. Schmidt and M. A. Rodriguez. To appear in *Bayesian Statistics 9* (ed. J.M. Bernardo, M. J. Bayarri, J. O. Berger, A. P Dawid, D. Heckerman, A. F. M. Smith & M. West). Oxford University Press.

BOYS, R. J. & HENDERSON, D. A. (2002) On determining the order of Markov dependence of an observed process governed by a hidden Markov model. *Scientific Programming* **10**, 241–251.

BOYS, R. J. & HENDERSON, D. A. (2003) Data augmentation and marginal updating schemes for inference in hidden Markov models. *Tech. Rep.*. Department of Mathematics and Statistics, Newcastle University.

BOYS, R. J. & HENDERSON, D. A. (2004) A Bayesian approach to DNA sequence segmentation. *Biometrics* **60**, 573–588.

BROOKS, S. (1998) Markov chain Monte Carlo method and its application. *The Statistician* **47**, 69–100.

BROOKS, S. P., GIUDICI, P. & ROBERTS, G. O. (2003) Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions (with discussion). *J. Roy. Statist. Soc. B* **65**, 3–55.

CAPPÉ, O., MOULINES, E. & RYDÉN, T. (2005) *Inference in Hidden Markov Models*. New York: Springer.

CAPPÉ, O., ROBERT, C. P. & RYDÉN, T. (2003) Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers. *J. Roy. Statist. Soc. B* **65**, 679–700.

CARLIN, B. P. & CHIB, S. (1995) Bayesian model choice via Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. B* **57**, 473–484.

CELEUX, G. (1998) Bayesian inference for mixtures: the label-switching problem. In *COMPSTAT 98* (ed. P. Green & R. Payne), pp. 227–232. Heidelberg: Physica-Verlag.

CELEUX, G., HURN, M. & ROBERT, C. P. (2000) Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Statist. Assoc.* **95**, 957–970.

CHARLES, S. P., BATES, B. C. & HUGHES, J. P. (1999) A spatiotemporal model for downscaling precipitation occurrence and amounts. *J. Geophysical Res.* **104**, 31657–31669.

CHARLES, S. P., BATES, B. C., SMITH, I. A. & HUGHES, J. P. (2004) Statistical downscaling of daily precipitation from observed and modelled atmospheric fields. *Hydrological Processes* **18**, 1373–1394.

CHIB, S. (1995) Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.* **90**, 1313–1321.

CHIB, S. (1998) Estimation and comparison of multiple change-point models. *J. Econometrics* **86**, 221–241.

CHIB, S. & GREENBERG, E. (1995) Understanding the Metropolis-Hastings algorithm. *Amer. Statist.* **49**, 327–335.

CHIB, S. & GREENBERG, E. (1998) Analysis of multivariate probit models. *Biometrika* **85**, 347–361.

CHIB, S. & JELIAZKOV, I. (2001) Marginal likelihood from the Metropolis-Hastings output. *J. Amer. Statist. Assoc.* **96**, 270–281.

CHIPMAN, H., GEORGE, E. I. & MCCULLOCH, R. E. (2001) The practical implementation of Bayesian model selection. In *Model Selection* (ed. P. Lahiri), *IMS Lecture Notes Monogr. Ser.*, vol. 38, pp. 65–134. Beachwood, OH: Inst. Math. Statist.

CONGDON, P. (2005) *Bayesian Models for Categorical Data*. Chichester: John Wiley & Sons Ltd.

CONGDON, P. (2006) *Bayesian Statistical Modelling*, 2nd edn. Chichester: John Wiley & Sons Ltd.

COVER, T. M. & THOMAS, J. A. (1991) *Elements of Information Theory*. New York: John Wiley & Sons Inc.

COWLES, M. K. & CARLIN, B. P. (1996) Markov chain Monte Carlo convergence diagnostics: a comparative review. *J. Amer. Statist. Assoc.* **91**, 883–904.

CRESSIE, N. A. C. (1993) *Statistics for Spatial Data*. New York: John Wiley & Sons Inc.

DANIELS, M. J. (2006) Bayesian modeling of several covariance matrices and some results on propriety of the posterior for linear regression with correlated and/or heterogeneous errors. *J. Multivariate Anal.* **97**, 1185–1207.

DANIELS, M. J. & POURAHMADI, M. (2002) Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika* **89**, 553–566.

DEGROOT, M. H. (2004) *Optimal Statistical Decisions*. Hoboken, NJ: Wiley-Interscience [John Wiley & Sons].

DELLAPORTAS, P. & PAPAGEORGIOU, I. (2006) Multivariate mixtures of normals with unknown number of components. *Stat. Comput.* **16**, 57–68.

DENISON, D. G. T., HOLMES, C. C., MALLICK, B. K. & SMITH, A. F. M. (2002) *Bayesian Methods for Nonlinear Classification and Regression*. Chichester: John Wiley & Sons Ltd.

DICKEY, J. M. (1982) Conjugate families of distributions. In *Encyclopedia of Statistical Sciences, vol. 2* (ed. S. Kotz & N. L. Johnson), pp. 135–145. New York: Wiley.

DIGGLE, P. J., TAWN, J. A. & MOYEED, R. A. (1998) Model-based geostatistics. *J. Roy. Statist. Soc. C* **47**, 299–350.

EDWARDS, Y. D. & ALLENBY, G. M. (2003) Multivariate analysis of multiple response data. *J. Marketing Res.* **40**, 321–334.

EVERITT, B. S. (1993) *Cluster Analysis*, 3rd edn. London: Edward Arnold.

FOWLER, H. G., KILSBY, C. G. & O'CONNELL, P. E. (2000) A stochastic rainfall model for the assessment of regional water resource systems under changed climatic conditions. *Hydrol. Earth Syst. Sci.* **4**, 263–282.

FRIEL, N. & PETTITT, A. N. (2008) Marginal likelihood estimation via power posteriors. *J. Roy. Statist. Soc. B* **70**, 589–607.

FRÜHWIRTH-SCHNATTER, S. (2001) Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *J. Amer. Statist. Assoc.* **96**, 194–209.

FRÜHWIRTH-SCHNATTER, S. (2004) Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *Econom. J.* **7**, 143–167.

FRÜHWIRTH-SCHNATTER, S. (2006) *Finite Mixture and Markov Switching Models*. New York: Springer.

GAMERMAN, D. & LOPES, H. F. (2006) *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, 2nd edn. Chapman & Hall/CRC, Boca Raton, FL.

GARTHWAITE, P. H., KADANE, J. B. & O'HAGAN, A. (2005) Statistical methods for eliciting probability distributions. *J. Amer. Statist. Assoc.* **100**, 680–700.

GELFAND, A. E., DEY, D. K. & CHANG, H. (1992) Model determination using predictive distributions with implementation via sampling-based methods. In *Bayesian Statistics 4* (ed. J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith), pp. 147–167. New York: Oxford Univ. Press.

GELFAND, A. E. & GHOSH, S. K. (1998) Model choice: a minimum posterior predictive loss approach. *Biometrika* **85**, 1–11.

GELMAN, A., CARLIN, J. B., STERN, H. S. & RUBIN, D. B. (1995) *Bayesian Data Analysis*. London: Chapman & Hall.

GELMAN, A. & MENG, X.-L. (1998) Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statist. Sci.* **13**, 163–185.

GELMAN, A. & RUBIN, D. B. (1992) Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.* **7**, 457–511.

GENZ, A. (1992) Computation of multivariate normal probabilities. *J. Comput. Graph. Statist.* **1**, 141–149.

GENZ, A. & BRETZ, F. (2009) *Computation of Multivariate Normal and t Probabilities*. *Lecture Notes in Statistics* **195**. Heidelberg: Springer-Verlag.

GENZ, A., BRETZ, F., MIWA, T., MI, X., LEISCH, F., SCHEIPL, F. & HOTHORN, T. (2010) *mvtnorm: Multivariate Normal and t Distributions*. R package version 0.9-9.

GERMAIN, S. E., BOYS, R. J. & FARROW, M. (2010*a*) Bayesian analysis of hidden Markov models. Newcastle University Statistics Research Report 10-01. Newcastle University.

GERMAIN, S. E., BOYS, R. J. & FARROW, M. (2010*b*) Building genuine beliefs into a prior distribution for the variance matrix of a multivariate normal distribution. Newcastle University Statistics Research Report 10-02. Newcastle University.

GEWEKE, J. (1991) Efficient simulation from the multivariate normal and student-*t* distributions subject to linear constraints and the evaluation of constraint probabilities. *Comp. Sci. and Statist.* **23**, 571–578.

GEWEKE, J. (1992) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4* (ed. J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith), pp. 169–193. New York: Oxford Univ. Press.

GEYER, C. J. (1992) Practical Markov chain Monte Carlo. *Statist. Sci.* **7**, 473–483.

GEYER, C. J. & THOMPSON, E. A. (1992) Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. Roy. Statist. Soc. B* **54**, 657–699.

GLICKMAN, T. S. (2000) *Glossary of Meteorology*, 2nd edn. Boston: American Meteorological Society.

GNEITING, T., BALABDAOUI, F. & RAFTERY, A. E. (2007) Probabilistic forecasts, calibration and sharpness. *J. Roy. Statist. Soc. B* **69**, 243–268.

GODSILL, S. J. (2001) On the relationship between Markov chain Monte Carlo methods for model uncertainty. *J. Comput. Graph. Statist.* **10**, 230–248.

GREEN, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

GREEN, P. J. (2003) Trans-dimensional Markov chain Monte Carlo. In *Highly Structured Stochastic Systems* (ed. P. J. Green, N. L. Hjort & S. Richardson), *Oxford Statist. Sci. Ser.*, vol. 27, pp. 179–206. Oxford: Oxford Univ. Press.

GREGORY, J. M., WIGLEY, T. M. L. & JONES, P. D. (1993) Application of Markov models to area-average daily precipitation series and interannual variability in seasonal totals. *Climate Dyn.* **8**, 299–310.

GUMPERTZ, M. L., GRAHAM, J. M. & RISTAINO, J. B. (1997) Autologistic model of spatial pattern of phytophthora epidemic in bell pepper: effects of soil variables on disease presence. *J. Agric. Biol. Environ. Stat.* **2**, 131–156.

GUYON, X. & HARDOUIN, C. (2002) Markov chain Markov field dynamics: models and statistics. *Statistics* **36**, 339–363.

HARTIGAN, J. A. (1975) *Clustering Algorithms*. John Wiley & Sons, New York–London–Sydney.

HAY, L. E., MCCABE, G. J., WOLOCK, D. M. & AYERS, M. A. (1991) Simulation of precipitation by weather type analysis. *Water Resour. Res.* **27**, 493–501.

HENDERSON, D. A. (1999) Modelling and analysis of non-coding DNA sequence data. PhD thesis, Newcastle University.

HUGHES, J. P. & GUTTORP, P. (1994*a*) A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. *Water Resour. Res.* **30**, 1535–1546.

HUGHES, J. P. & GUTTORP, P. (1994*b*) Incorporating spatial dependence and atmospheric data in a model of precipitation. *J. Appl. Meteorology* **33**, 1503–1515.

HUGHES, J. P., GUTTORP, P. & CHARLES, S. P. (1999) A non-homogeneous hidden Markov model for precipitation occurrence. *J. Roy. Statist. Soc. C* **48**, 15–30.

JENKINSON, A. F. & COLLINSON, B. P. (1977) An initial climatology of gales over the north sea. Synoptic climatology branch memorandum no. 62. Meteorological Office, London.

JENNISON, C. (1997) Contribution to the discussion of "On Bayesian analysis of mixtures with an unknown number of components" by S. Richardson and P. J. Green. *J. Roy. Statist. Soc. B* **59**, 778–779.

JOE, H. (1997) *Multivariate Models and Dependence Concepts*, *Monographs on Statistics and Applied Probability*, vol. 73. London: Chapman & Hall.

JOHNSON, N. L. (1949) Systems of frequency curves generated by methods of translation. *Biometrika* **36**, 149–176.

JONES, G. L., HARAN, M., CAFFO, B. S. & NEATH, R. (2006) Fixed width output analysis for Markov chain Monte Carlo. *J. Amer. Statist. Assoc.* **101**, 1537–1547.

KASS, R. E., CARLIN, B. P., GELMAN, A. & NEAL, R. M. (1998) Markov chain Monte Carlo in practice: a roundtable discussion. *Amer. Statist.* **52**, 93–100.

KASS, R. E. & RAFTERY, A. E. (1995) Bayes factors. *J. Amer. Statist. Assoc.* **90**, 773–795.

KASS, R. E. & WASSERMAN, L. (1996) The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc.* **91**, 1343–1370.

KEENEY, R. L. & RAIFFA, H. (1976) *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. John Wiley & Sons, New York–London–Sydney.

LAMB, H. H. (1972) British Isles weather types and a register of daily sequence of circulation patterns, 1861–1971. Geophysical memoir 116. HMSO, London.

LAURITZEN, S. L., DAWID, A. P., LARSEN, B. N. & LEIMER, H.-G. (1990) Independence properties of directed Markov fields. *Networks* **20**, 491–505.

LAWSON, A. B. & CLARK, A. (1997) Contribution to the discussion of "On Bayesian analysis of mixtures with an unknown number of components" by S. Richardson and P. J. Green. *J. Roy. Statist. Soc. B* **59**, 779.

LAWSON, A. B. & DENISON, D. G. T., ed. (2002) *Spatial Cluster Modelling*. Chapman & Hall/CRC, Boca Raton, FL.

LINDLEY, D. V. (1957) A statistical paradox. *Biometrika* **44**, 187–192.

LIU, C. (2001) Comment on "The art of data augmentation". *J. Comput. Graph. Statist.* **10**, 75–81.

LIU, C., RUBIN, D. B. & WU, Y. N. (1998) Parameter expansion to accelerate EM: the PX–EM algorithm. *Biometrika* **85**, 755–770.

LIU, X. & DANIELS, M. J. (2006) A new algorithm for simulating a correlation matrix based on parameter expansion and reparameterization. *J. Comput. Graph. Statist.* **15**, 897–914.

MACDONALD, I. L. & ZUCCHINI, W. (1997) *Hidden Markov and Other Models for Discrete-Valued Time Series*, *Monographs on Statistics and Applied Probability*, vol. 70. London: Chapman & Hall.

MARIN, J.-M. & ROBERT, C. (2008) Approximating the marginal likelihood in mixture models. *Bulletin of the Indian Chapter of ISBA* **1**, 2–7.

MCCULLAGH, P. & NELDER, J. A. (1989) *Generalized Linear Models*, 2nd edn., *Monographs on Statistics and Applied Probability*, vol. 37. Chapman & Hall.

MCCULLOCH, R. E., POLSON, N. G. & ROSSI, P. E. (2000) A Bayesian analysis of the multinomial probit model with fully identified parameters. *J. Econometrics* **99**, 173–193.

MENG, X.-L. & WONG, W. H. (1996) Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statist. Sinica* **6**, 831–860.

MØLLER, J., PETTITT, A. N., REEVES, R. & BERTHELSEN, K. K. (2006) An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika* **93**, 451–458.

NELSEN, R. B. (2006) *An Introduction to Copulas*, 2nd edn. New York: Springer.

NEWTON, M. A. & RAFTERY, A. E. (1994) Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). *J. Roy. Statist. Soc. B* **56**, 3–48.

NOBILE, A. & FEARNSIDE, A. T. (2007) Bayesian finite mixtures with an unknown number of components: the allocation sampler. *Stat. Comput.* **17**, 147–162.

R DEVELOPMENT CORE TEAM (2008) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RABINER, L. R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257–286.

RAPPOLD, A. G., GELFAND, A. E. & HOLLAND, D. M. (2008) Modelling mercury deposition through latent space–time processes. *J. Roy. Statist. Soc. C* **57**, 187–205.

REEVES, R. & PETTITT, A. N. (2004) Efficient recursions for general factorisable models. *Biometrika* **91**, 751–757.

RICHARDSON, S. & GREEN, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Roy. Statist. Soc. B* **59**, 731–792.

ROBERT, C. P. & CASELLA, G. C. (2005) *Monte Carlo Statistical Methods*, 2nd edn. New York: Springer–Verlag.

ROBERT, C. P., CELEUX, G. & DIEBOLT, J. (1993) Bayesian estimation of hidden Markov chains: a stochastic implementation. *Statist. Probab. Lett.* **16**, 77–83.

ROBERT, C. P., RYDÉN, T. & TITTERINGTON, D. M. (2000) Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *J. Roy. Statist. Soc. B* **62**, 57–75.

ROBERTS, G. O. (1996) Markov chain concepts related to sampling algorithms. In *Markov Chain Monte Carlo in Practice* (ed. W. R. Gilks, S. Richardson & D. J. Spiegelhalter), pp. 45–57. London: Chapman & Hall.

ROBERTSON, A. W., KIRSHNER, S. & SMYTH, P. (2004) Downscaling of daily rainfall occurrence over Northeast Brazil using a hidden Markov model. *J. Climate* **17**, 4407–4424.

ROEDER, K. & WASSERMAN, L. (1997) Practical Bayesian density estimation using mixtures of normals. *J. Amer. Statist. Assoc.* **92**, 894–902.

RUE, H. & HELD, L. (2005) *Gaussian Markov Random Fields: Theory and Applications*, *Monographs on Statistics and Applied Probability*, vol. 104. Chapman & Hall/CRC, Boca Raton, FL.

SAHU, S. K., GELFAND, A. E. & HOLLAND, D. M. (2010) Fusing point and areal level space-time data with application to wet deposition. *J. Roy. Statist. Soc. C* 59, 77–103.

SANSÓ, B. & GUENNI, L. (1999) Venezuelan rainfall data analysed by using a Bayesian space-time model. *J. Roy. Statist. Soc. C* 48, 345–362.

SANSÓ, B. & GUENNI, L. (2000) A nonstationary multisite model for rainfall. *J. Amer. Statist. Assoc.* 95, 1089–1100.

SANSOM, J. (1998) A hidden Markov model for rainfall using breakpoint data. *J. Climate* 11, 42–53.

SHI, J. Q., MURRAY-SMITH, R. & TITTERINGTON, D. M. (2002) Birth-death MCMC methods for mixtures with an unknown number of components. *Tech. Rep.*. Department of Computing Science, University of Glasgow.

SISSON, S. A. (2005) Transdimensional Markov chains: a decade of progress and future perspectives. *J. Amer. Statist. Assoc.* 100, 1077–1089.

SMITH, J. Q. (1988) *Decision Analysis: A Bayesian Approach*. London: Chapman & Hall.

SMITH, R. L. (1994) Spatial modelling of rainfall data. In *Statistics for the Environment 2: Water Related Issues* (ed. V. Barnett & K. F. Turkman), pp. 19–41. Chichester: John Wiley & Sons Ltd.

STEELE, R. J., RAFTERY, A. E. & EMOND, M. J. (2006) Computing normalizing constants for finite mixture models via incremental mixture importance sampling (IMIS). *J. Comput. Graph. Statist.* 15, 712–734.

STEPHENS, M. (1997) Bayesian methods for mixtures of normal distributions. PhD thesis, Oxford University.

STEPHENS, M. (2000) Dealing with label switching in mixture models. *J. Roy. Statist. Soc. B* 62, 795–809.

STERN, R. D. & COE, R. (1984) A model fitting analysis of daily rainfall data. *J. Roy. Statist. Soc. A* 147, 1–34.

TANNER, M. A. & WONG, W. H. (1987) The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* 82, 528–550.

THOMPSON, C. S., THOMSON, P. J. & ZHENG, X. (2007) Fitting a multisite daily rainfall model to New Zealand data. *J. Hydrology* 340, 25–39.

VELARDE, L. G. C., MIGON, H. S. & PEREIRA, B. DE B. (2001) Space-time modeling of rainfall data. *Environmetrics* 15, 561–576.

VIALLEFONT, V., RICHARDSON, S. & GREEN, P. J. (2002) Bayesian analysis of Poisson mixtures. *J. Nonpara. Stat.* 14, 181–202.

WHITTAKER, J. (1990) *Graphical Models in Applied Multivariate Statistics*. Chichester: John Wiley & Sons Ltd.

Woolhiser, D. A. & Roldán, J. (1982) Stochastic daily precipitation models 2. A comparison of distributions of amounts. *Water Resour. Res.* **18**, 1461–1468.

Woolhiser, D. A. & Roldán, J. (1986) Seasonal and regional variability of parameters for stochastic daily precipitation models: South Dakota, U.S.A. *Water Resour. Res.* **22**, 965–978.

Wu, H. & Huffer, F. W. (1997) Modelling the distribution of plant species using the autologistic regression model. *Environ. Ecol. Statist.* **4**, 49–64.

Zhang, X., Boscardin, W. J. & Belin, T. R. (2006) Sampling correlation matrices in Bayesian models with correlated latent variables. *J. Comput. Graph. Statist.* **15**, 880–896.

Zheng, Y. & Zhu, J. (2008) Markov chain Monte Carlo for a spatial-temporal autologistic regression model. *J. Comput. Graph. Statist.* **17**, 123–137.

Zhu, J., Huang, H.-C. & Wu, J. (2005) Modeling spatial-temporal binary data using Markov random fields. *J. Agric. Biol. Environ. Stat.* **10**, 212–225.

Zhu, J., Zheng, Y., Carroll, A. L. & Aukema, B. H. (2008) Autologistic regression analysis of spatial-temporal binary data via Monte Carlo maximum likelihood. *J. Agric. Biol. Environ. Stat.* **13**, 84–98.

Zucchini, W. & Guttorp, P. (1991) A hidden Markov model for space-time precipitation. *Water Resour. Res.* **27**, 1917–1923.