Aspects of Objective Priors and Computations for Bayesian Modelling

TANG, DOUDOU

Aspects of Objective Priors and Computations for Bayesian Modelling

Doudou Tang

A Thesis presented for the degree of
Doctor of Philosophy

Probability and Statistics
Department of Mathematical Sciences
University of Durham
England

September 2015

Aspects of Objective Priors and Computations for Bayesian Modelling

Doudou Tang

Submitted for the degree of Doctor of Philosophy
July 2015

Abstract

Bayesian statistics is flourishing nowadays not only because it provides ways to combine prior knowledge with statistical models but also because many algorithms have become available to sample from the resulting posterior distributions. However, how to specify a good objective prior can be very difficult. This is largely because ignorance does not have a unique definition. For sampling from posterior distributions, Markov Chain Monte Carlo (MCMC) methods are main tools. However, as statistical models become more and more sophisticated, there is a need for more efficient MCMC methods than the traditional ones.

For objective prior specifications, we present a new principle to express ignorance through the global distance structure. This principle allows us to assign the prior weight to points in parameter space according to their correspondences to the statistical models displayed in the structure of the global distance. This method is applied to simple problems such as location family, scale family and location-scale family. It is also applied to the one-way random effect model which attracts considerable interest from many researchers. The method considered here allows us to avoid the dependency of the priors on the experimental design, which has been seriously disputed, and enables the resulting prior to reflect how the models change with respect to the population and not the collected samples.

Of MCMC methods for sampling from posterior distributions, the Hamiltonian Monte Carlo (HMC) method is one that has the potential to avoid random-walk behaviour. It does so by exploiting ideas from Hamiltonian dynamics. Its performance,

however, depends on the choice of step-size which is required by this method when numerically solving the Hamiltonian equations. We propose an algorithm, which we call HMC with stochastic step-size, to automatically tune the step-size by exploiting the local curvature information. We also present a meta-algorithm which includes HMC, HMC with stochastic step-size and the ordinary Metropolis-Hastings algorithm as a special case.

Finally, we come to a sophisticated hierarchical model developed for analysing the exco-toxicology data. We present ways to obtain more informative posterior samples by embedding the marginalized approach and advanced samplers into the entire Gibbs structure of the modified MCMCglmm algorithm provided by Craig (2013). The combination of the marginalized approach and HMC with stochastic step-size is found to be the best choice among a range of methods for the challenging problem of sampling the hyper-parameters in the model.

Declaration

The work in this thesis is based on research carried out at the Group of Probability and Statistics, the Department of Mathematical Sciences, Durham University, England. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor Dr Peter Craig for the continuous support of my Ph.D study and related research. Being supervised by Peter has been a tremendous privilege. He have helped me grow enormously in thinking and pursuing ideas, and have been constantly encouraging, patient and positive. I am also very thankful to Dr Ian Jermyn, Prof Frank Coolen, Dr Jochen Einbeck for their insightful comments, encouragement and supports.

There have been very few dull or lonely moments in the department, and almost always fun and entertaining people around; Ben, Tom, Thomai and Lewis to name a few. I thank all my friends in the departments for the stimulating discussions, for the time we were working together, and for all the fun we have had.

I thank Lingling, Lucia and Fechesca who have been good housemates and friends, and a great source of comfort and advices.

My parents have encouraged me throughout my PhD and my general life. They have been supportive of whatever I have chosen to do, and I am very grateful to them for never putting me under pressure to do something 'proper' and their constantly reminding of that, while a thesis is an interesting and enjoyable endeavour, it is only happiness that will endure forever during the life, and for that I am most thankful.

Finally, I am thankful to Yafeng who has loved, amused and looked after me through all manner of fretting, ranting and confusion.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Bayesian methods now have extensive applications in a wide range of fields. Whatever the problem is, a prior and an efficient computation method for posterior distributions must be involved to drive the Bayesian engine. In this thesis, we will discuss some topics in these two areas.

1.1 Prior Distributions

A prior density is a probabilistic representation of our beliefs about model parameters of interest. Rather than considering parameters as fixed unknown values as in the frequentist approach, Bayesian methods take model parameters as uncertain values and specify a probability distribution for them. With adequate expert opinions or historical data, a subjective prior could be determined accordingly. However, many statisticians admit that a default prior is needed if little prior knowledge is available. Such a prior is usually called an objective or non-informative prior.

Determining an objective prior is not easy even for some basic models. The one-way random effect model is the basic hierarchical model but it turns out that selecting an objective prior for such a basic model is notoriously difficult. Our research was actually started from selecting an objective prior for the one-way random effect model for which various objective priors have been suggested. We first looked at the Half-t prior distribution (Gelman et al., 2006) which is particularly designed for the one-way random effect model. This Half-t prior distribution, however, re-

quires the users to have a rough idea about the size of the between group variance and then set the scale of the Half-t prior distribution according to it. We then visit the Reference prior proposed by Berger and Bernardo (1992b). The disadvantage of this prior is that it requires the users to have some prior knowledge to order the parameters according to their inference importance. If all the parameters are at the same level of importance, then the Reference prior coincides with the Jeffreys prior. We then went to the famous Jeffreys prior. The Jeffreys prior for the one-way random effect model depends on the experimental design and this kind of dependency has been seriously disputed. More popular objective priors designed for this model will be discussed in Chapter 2.

The principle which we believe is reasonable to derive a prior is that when there is no prior knowledge available, all information that distinguishes one point from another in parameter space should come from their correspondences with probability models (Jermyn, 2005). We should spread the prior mass out in some sense equally over all the different models. How much prior weight a point in parameter space receives should depend on how much its corresponding model differs from other models represented by other points. In contrast with the Jeffreys' prior that uses local distance behaviour, we propose to use the global distance to measure the model differences. Moreover, rather than considering a pair of points by the global distance, we use the global distance structure of all points to derive a prior. We call it global distance structure prior. This will be introduced in Chapter 3.

Regarding to our initial goal of an objective prior for the one-way random effect model, the development of a global distance structure prior for such a model is discussed in Chapter 4. This and other priors mentioned in Chapter 2 are all evaluated by a simulation study in the last section of Chapter 4.

1.2 Computation Methods for Posterior Distributions

The topics addressed in part II of this thesis can be classified into two aspects. One aspect focuses on a Markov Chain Monte Carlo (MCMC) algorithm itself. More

specifically, we explore how to improve the performance of the Hamiltonian Monte Carlo sampler. The other aspect is about improving the sampling results for a sophisticated hierarchical model (developed by Craig (2013) for eco-toxicology data analysis) which has difficulty in sampling from its posterior distribution.

A MCMC Sampler

Markov Chain Monte Carlo (MCMC) methods have become one of the standard tools for Bayesian computation. MCMC methods are a class of algorithms concerning sampling from a probability distribution by constructing a Markov chain that takes the target probability distribution as its stationary equilibrium distribution. The Metropolis algorithm, as one of MCMC methods, was first developed by Metropolis et al. (1953) and became popular in statistics after the paper by Hastings in 1970. It is used widely across many sciences to sample from a probability distribution that is usually difficult to sample from directly. However, in many situations, especially Bayesian statistics, target distributions have complicated forms, highly correlated parameters and large dimensional size. The ordinary Metropolis algorithm might have slow exploration of state spaces and low acceptance rates caused by both random-walk behaviour of the traditional Metropolis methods and the complex nature of target distributions. Therefore, there is a need for the development of more efficient MCMC methods.

Hamiltonian Monte Carlo (HMC), first introduced by Duane et al. (1987), has great potential to provide efficient sampling results. It takes advantage of Hamiltonian dynamics by adding an auxiliary variable considered as a 'momentum' variable and thus transforms the problem of simulating target distributions to the problem of approximating Hamiltonian dynamics. Although HMC has good potential to give high quality simulation results, the ability to do so is limited by three hand-tuning parameters: the variance matrix M for the augmented 'momentum' variables, the number of leap-frog steps l and the step-size ε for each step of leap-frog integrator used to numerically approximate the Hamiltonian dynamics. In recent years, there has been growing interest in improving performance of HMC. Girolami and Calderhead (2011) proposed Riemann Manifold HMC (RMHMC) which exploits

the local information by setting the variance matrix M as the expected second-derivative of the log-density function and thereby improves the performance to a large degree. This expected second-derivative of the log-density function can be considered as a local metric defined in Riemann geometry. Proper tuning of l is investigated by Hoffman and Gelman (2011). They introduced the No-U-Turn Sampler (NUTS) which automatically adapts path lengths to guarantee the benefit of HMC. Generally, NUTS is an extension on HMC which tries to avoid 'double back' of the simulated path by a doubling procedure to search candidates which give 'long enough' simulated paths. Compared with basic HMC, RMHMC and NUTS adapt M and l respectively throughout whole simulations instead of using a global value. To the best of our knowledge, how to select step-size values has not been explored adequately. In Chapter 5, we will study the problem of selecting the step-size for HMC and propose an HMC variant to automatically tune the step-size throughout the whole simulation according to the local curvature information. We call it HMC with stochastic step-size. A meta-algorithm, which is realised through the development of HMC with stochastic step-size algorithm, will be given in Chapter 5. We call this meta-algorithm 'generalised Metropolis-Hastings with Dynamics'. It includes HMC, HMC with stochastic step-size and the ordinary Metropolis-Hastings as a special case.

A Real Hierarchical Model

Sampling methods, which are efficient theoretically, might lost their power when dealing with some real situations. A hierarchical model developed by Craig (2013) for eco-toxicology data has some difficulties in sampling from its posterior distributions not only because of its high dimensionality but also because it has a complex structure used to represent the taxonomical structure of species. A 'stuck' Markov chain is obtained when directly sending this model to Stan, which is a software implementing HMC or NUTS. By using the modified MCMCglmm suggested by Craig (2013), the resulting posterior samples have extremely high auto-correlations. The background and the computational problem associated with this model are discussed in Chapter 6.

Having seen the high auto-correlations in the posterior samples given by the computational method in Chapter 6, we explore how to improve simulations for such a model in Chapter 8. In this chapter, we present ways to obtain more informative posterior samples by embedding the marginalized approach and advanced samplers into the entire Gibbs structure of the modified MCMCglmm algorithm. The advanced samplers includes HMC, RMHMC, NUTS and HMC with stochastic step-size. Particularly, NUTS and RMHMC are detailed in Chapter 7 as preliminary materials.

Part I

Topics of Objective Priors

Chapter 2

Popular Objective Prior Choices

An objective prior is one that asserts no information available for parameters before data is collected. The construction and selection of a good objective prior have attracted considerable interest. Usually, a procedure for constructing objective priors depends on some external principles or assumptions since there is not a unique precise definition of ignorance. Different external principles may lead to different prior distributions. In this section, we briefly review some well-known objective priors and their underlying principles.

Laplace's rule, or the principle of insufficient reason, states that equal probability should be assigned to every point in the parameter space if we are ignorant about model parameters. The prior obeying Laplace's rule might be the one that makes the least extra assumptions in expressing ignorance. Although its simplicity is appealing, its potential usefulness has been disputed. Kass and Wasserman (1996) discussed problems caused by following Laplace's rule that implicitly suggests a uniform prior. One obvious drawback is that such a prior is not invariant to one-to-one re-parametrizations. For example, a uniform prior for the normal scale parameter would not lead to a uniform prior for the normal variance parameter.

Jeffreys (1946) proposed his famous prior—Jeffreys prior based on the connection to the local behaviour of Kullback-Leibler divergence or Hellinger distance. This prior is justified by its invariance to parameter transformations. The external assumption detailed to express the ignorance might be that two persons with different parametrizations but identical amount of prior knowledge should end up

with a same prior. Briefly speaking, Jefferys prior is proportional to the square root of the determinant of the Fisher Information matrix. Inspired by its local distance connections, George and McCulloch (1993) investigated various priors derived from other probability distances and provided a general form stating that prior is proportional to the square root of the determinant of a probability distance's differential form. Kass (1989, 1996) further elaborated Jeffreys prior from Riemannian geometry background.

About the invariance argument, the following discussion of rules for using invariance principles to assist the choice of prior distributions is based on Dawid (1983). A statistical model is a parameterized family of probability distributions with a specified domain for the parameters. In the context of a rule for assigning a prior distribution to a statistical model in the absence of prior knowledge, 1) the parameter invariance (PI) principle is that prior measures proposed for two different parameterizations of the same statistical model should respect the reparameterization; 2) the data invariance (DI) principle is that the prior measures proposed should be the same for two statistical models which differ only via a one-to-one transformation of the data; 3) the context invariance (CI) principle is that if the same statistical model is to be used in different contexts, the prior measures proposed should be the same. Jeffreys prior is an example of a rule which satisfies PI, DI and CI. Hartigan (1964) proposed that rather than assigning exact the same prior measure if a particular invariance is satisfied, equivalent prior measure should be assigned since the posterior distribution is the main issue. This results in relative invariance criteria RPI, RDI and RCI. A particular way of arriving at two versions of the same statistical model to which the (R)PI, (R)DI and (R)CI principles might be applied is via an equivariant recoding. Consider a statistical model $\mathbf{y} \sim f_{\boldsymbol{\theta}}$ and a transformation $g(\mathbf{y})$. The transformation $\bar{g}(\boldsymbol{\theta})$ is an induced recoding of $\boldsymbol{\theta}$ if $g(\mathbf{y}) \sim f_{\bar{g}(\boldsymbol{\theta})}$. This recoding g ($\bar{g}$) is called an equivariant recoding of $\mathbf{y}$ ($\boldsymbol{\theta}$). The collection of all these equivariant recodings of $\mathbf{y}$ ($\boldsymbol{\theta}$) forms a transformation group $\mathcal{G}$ ($\bar{\mathcal{G}}$). A prior which satifies RPI, RDI and RCI with respect to all equivariant recodings is called a relative invariant prior.

Box and Tiao (2011) introduced a choice of non-informative prior from the point

of data-translated likelihoods. This prior is elicited from the idea that little is known relative to the information provided by data and conveyed by the likelihood function.

The reference prior, first proposed by Bernardo in 1979 and further developed by Berger and others (Berger et al., 1988; Berger and Bernardo, 1992a; Berger et al., 2009), is constructed through the idea of maximizing the divergence between prior and posterior distribution so that the data could have maximum influence on the posterior inference.

Another category of prior distribution is conjugate prior distributions that are in the same distribution family with the corresponding posterior distribution. Due to its computational simplicity, they are quite popular in real data analysis. Usually, they do not target on representing ignorance. Non-informativeness, however, is approximately expressed by specifying the distributional parameters of conjugate priors so that the priors are flat to some degree.

The preceding priors could be easily derived if the statistical problems under consideration are trivial. However, they might be hard to derive or even not exist for a non-trivial statistical model. Here, we take the one-way random effects model as a concrete example and investigate problems with determining an objective prior for it. The selection of an objective prior for this model has attracted many researchers' attention not only because the importance of this model but also the notorious difficulties in determining a good non-informative prior for it. In the following part of this chapter, we will review some existing work on objective priors for the one-way random effects model. Apart from the above mentioned priors, two additional priors are designed especially for this model. One is the so-called uniform shrinkage prior suggested by Daniels (1999) from the point of view of assigning uniform probability on the shrinkage factor. The other is a folded-t prior distribution suggested by Gelman et al. (2006). It is an implicit conditionally-conjugate prior for variance parameters of random effects in hierarchical models. Gelman suggested that it could be used to represent weak non-informativeness by setting its distributional parameter to a large value. Both the uniform shrinkage prior and Gelman 's folded-t prior concentrate on the variance parameters of random effects in hierarchical models. Details are provided in section 2.1.

2.1 Popular non-informative priors for the one-way random effects model

The balanced one-way random effects model is expressed as follows,

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$
$$\alpha_i \sim N(0, \sigma_\alpha^2)$$
$$\varepsilon_{ij} \sim N(0, \sigma^2)$$
$$i = 1, \ldots, m; j = 1, \ldots, N$$

where i indexes groups and j indexes observations within a group; σ_α^2 is the variance of group means and σ^2 is the within-group variances.

Jeffreys Prior

There are two versions of Jeffreys prior. The first one is usually called Jeffreys general prior which is derived from the Fisher Information matrix. Mathematically, the prior determined by Jeffreys general rule is

$$\pi(\boldsymbol{\theta}) \propto |\mathbf{I}(\boldsymbol{\theta})|^{1/2}, \tag{2.1.1}$$

where $\mathbf{I}(\cdot)$ is the Fisher Information matrix of all parameters $\boldsymbol{\theta}$. The argument for this prior is that it is invariant under re-parametrizations. Intuitively, two different people with different parametrizations should end up with a same prior if their prior knowledge is on the same level. The geometric origin of the invariance is that the Kullback-Leibler discrepancy behaves locally like the square of a distance function determined by a Riemannian metric and the natural volume element of this metric is $|\mathbf{I}(\boldsymbol{\theta})|^{1/2}$ which is automatically invariant to re-parametrization (Jeffreys, 1946; Kass and Wasserman, 1996). For the one-way random effects model, the prior determined by the general rule illustrated in Equation (2.1.1) is derived as

$$\pi(\mu, \sigma, \sigma_\alpha) \propto |\mathbf{I}(\mu, \sigma, \sigma_\alpha)|^{1/2} \propto \frac{\sigma_\alpha}{\sigma(N\sigma_\alpha^2 + \sigma^2)^{3/2}}. \tag{2.1.2}$$

The modified version of Jeffreys prior concerns problems involving location parameters and other parameters. He suggested that location parameters should be

considered separately. To be specific, the modified Jeffreys prior is

$$\pi(\boldsymbol{\mu}, \boldsymbol{\theta}^{\star}) \propto |\mathbf{I}(\boldsymbol{\theta}^{\star})|^{1/2}, \tag{2.1.3}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\theta}^{\star}\}$; $\boldsymbol{\mu}$ and $\boldsymbol{\theta}^{\star}$ denote the location parameters and additional parameters respectively. $\mathbf{I}(\boldsymbol{\theta}^{\star})$ is the Fisher Information matrix derived by fixing the location parameters. The justification for this modified prior is not so clear. According to the modified rule illustrated in Equation (2.1.3), the prior for the one-way random effects model could be derived as

$$\pi(\mu, \sigma, \sigma_{\alpha}) \propto |\mathbf{I}(\sigma, \sigma_{\alpha})|^{1/2} \propto \frac{\sigma_{\alpha}}{\sigma(N\sigma_{\alpha}^2 + \sigma^2)}.$$

Data-translated Likelihood Prior

The data-translated likelihood prior, proposed by Box and Tiao (2011), attempts to express the idea that little information is available about model parameters $\boldsymbol{\theta}$ relative to the information provided by the data. Box and Tiao argued that what the data would be able to tell us is all included in the likelihood function. When the likelihood function could be expressed in terms of some particular parametrization $\boldsymbol{\phi}(\boldsymbol{\theta})$ so that different sets of data only translate the likelihood curve on the $\boldsymbol{\phi}(\boldsymbol{\theta})$ axis and maintain others unaffected, then a uniform prior would be assigned to $\boldsymbol{\phi}(\boldsymbol{\theta})$. In other words, the data-translated likelihood prior focuses on seeking parametrization such that the likelihood function is data-translated. Mathematically, the likelihood function is called data-translated if it can be expressed in the following form

$$l(\boldsymbol{\theta}|\mathbf{y}) = t_1\Big(\boldsymbol{\phi}(\boldsymbol{\theta}) - t_2(\mathbf{y})\Big),$$

where t_1 is a known function independent of $\mathbf{y}$; $\boldsymbol{\phi}(\cdot)$ is a one-to-one transformation of $\boldsymbol{\theta}$; and t_2 is a known function of $\mathbf{y}$. The data-translated prior is

$$\pi(\boldsymbol{\phi}) \propto 1.$$

The prior for $\boldsymbol{\theta}$ could be thus obtained according to a change of variables by the Jacobian factor. As might be expected, such a parametrization $\boldsymbol{\phi}$ might not exist especially for a model having a complicated likelihood function. In order to deal

with this kind of situation, Box and Tiao further proposed the approximate data-translated likelihood prior. To be specific, Box and Tiao (2011) made use of the fact that the likelihood function of $\boldsymbol{\theta}$ is approximately normal and remains approximately normal under mild one-to-one transformation if the sample size is large enough. Therefore, the log-likelihood function could be approximately expressed as

$$\begin{aligned} L(\boldsymbol{\theta}|\mathbf{y}) = \log l(\boldsymbol{\theta}|\mathbf{y}) &\approx L(\hat{\boldsymbol{\theta}}|\mathbf{y}) - \frac{n}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T V_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) \\ &\approx \text{const} - \frac{n}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T V_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}), \end{aligned}$$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimation (MLE) of $\boldsymbol{\theta}$ and $V_{\hat{\boldsymbol{\theta}}}$ is

$$V_{\hat{\boldsymbol{\theta}}} = \frac{1}{n}\mathbf{E}\Big(\frac{\partial^2 L}{\partial\theta_i \partial\theta_j}\Big)\Big|_{\hat{\boldsymbol{\theta}}} = \frac{1}{n}\mathbf{I}(\hat{\boldsymbol{\theta}}).$$

This indicates that the scale of the likelihood curve could be approximately determined by $(V_{\hat{\boldsymbol{\theta}}})^{-1/2}$. Consider a parametrization $\boldsymbol{\phi}(\boldsymbol{\theta})$. The above equation, under the new parametrization, becomes

$$V_{\hat{\boldsymbol{\phi}}} = \mathcal{J} V_{\hat{\boldsymbol{\theta}}} \mathcal{J}^T, \tag{2.1.4}$$

where $\mathcal{J} = \frac{d\boldsymbol{\theta}}{d\boldsymbol{\phi}}$. By choosing $\mathcal{J}$ so that

$$\mathcal{J} \propto \Big[\mathbf{E}\Big(\frac{\partial^2 L}{\partial\theta_i \partial\theta_j}\Big)\Big]^{-1/2},$$

then $V_{\boldsymbol{\phi}}$ expressed in Equation (2.1.4) would be independent of $\hat{\boldsymbol{\phi}}$ and thus be independent of data. Therefore, the likelihood curve under parametrization $\boldsymbol{\phi}$ would be independent of data except for the location $\hat{\boldsymbol{\phi}}$. Since a uniform prior is assigned to $\boldsymbol{\phi}$, the prior for $\boldsymbol{\theta}$ changes to be

$$\pi(\boldsymbol{\theta}) \propto \Big|\mathbf{E}\Big(\frac{\partial^2 L}{\partial\theta_i \partial\theta_j}\Big)\Big|^{1/2} = |\mathbf{I}(\boldsymbol{\theta})|^{1/2}.$$

The prior derived above from the point of using approximated data-translated likelihood changes to be the Jeffreys' prior determined by the general rule.

Returning to the one-way random effects model, its likelihood function (Box and Tiao, 2011) is

$$\begin{aligned} &l(\mu, \sigma_\alpha, \sigma|\mathbf{y}) \\ &\quad\propto \frac{1}{\sigma^{m(N-1)}} \frac{1}{(\sigma^2 + N\sigma_\alpha^2)^{m/2}} \exp\Big\{ -\frac{1}{2}\Big(\frac{mN(y_{\cdot\cdot}-\mu)^2}{\sigma^2 + N\sigma_\alpha^2} + \frac{S_2}{\sigma^2 + N\sigma_\alpha^2} + \frac{S_1}{\sigma^2}\Big)\Big\}, \end{aligned} \tag{2.1.5}$$

where $S_1 = \sum_i \sum_j (y_{ij} - y_{i\cdot})^2$ and $S_2 = N \sum_i (y_{i\cdot} - y_{\cdot\cdot})^2$; $y_{i\cdot}$ is sample mean for group i and $y_{\cdot\cdot}$ is over-all sample mean. It is not feasible to separate data and parameters in the above likelihood function so that the likelihood curve could be independent of data other than through location. Therefore, no data-translated likelihood prior for the one-way random effects model. By resorting to the approximated data-translated likelihood, the prior for this model should be the same as the one shown in Equation (2.1.2).

Relative Invariant Prior

In order to find a relative invariant prior for the one-way random effect model, we need to specify a group of equivariant recodings.

The one-way random effect model can be expressed as $\mathbf{y}_i \overset{iid}{\sim} \mathbf{N}(\mu \mathbf{1}_N, A_{N,N})$, $i = 1, \ldots, m$, where $\mathbf{1}_N$ is a N-dimensional column vector of all terms to be one; $A_{N,N} = \alpha I_{N,N} + \beta J_{N,N}$ with $J_{N,N}$ is a N-dimensional square matrix with all terms to be one and $I_{N,N}$ is a N-dimensional identity matrix (see section 4.1 for details). In this way, the model is parametrized by $\boldsymbol{\theta} = \{\mu, \alpha, \beta\}$ and the parameter space is denoted by $S_{\boldsymbol{\theta}}$.

Consider a recoding of $\mathbf{y}_i$

$$\mathbf{z}_i = g(\mathbf{y}_i) = c\,\mathbf{1}_N + B\mathbf{y}_i, \tag{2.1.6}$$

where c is a real value and B is a non-singular $N \times N$ dimensional matrix. In particular, suppose that B satisfies $B = (aI_{N,N} + bJ_{N,N})\mathbf{O}$, where a, b are some real values; $\mathbf{O}$ is an orthogonal matrix and has the property $\mathbf{O}\mathbf{1}_N = \mathbf{1}_N$. The corresponding induced equivariant recoding of $\boldsymbol{\theta} = \{\mu, \alpha, \beta\}$, which we denote by $\bar{g}$, is (see A.2.1 in appendix for proof):

$$\begin{aligned}\boldsymbol{\Phi} = \bar{g}(\boldsymbol{\theta}) &= \bar{g}(\{\mu, \alpha, \beta\}) \\ &= \{(a + Nb)\mu + c, a^2\alpha, \alpha(2a + Nb)b + \beta(a + Nb)^2\}.\end{aligned}$$

By requiring $a \neq 0$ and $a + Nb \neq 0$, the collection of these recodings forms a group $\bar{\mathcal{G}} = \{\bar{g}_{a,b,c}; \forall c \in \mathbb{R}, a \neq 0, a + Nb \neq 0\}$ (see A.2.2 in Appendix for proof). A relative invariant prior measure is the one satisfying $\Omega\big(\bar{g}(\mathcal{A})\big) \propto \Omega\big(\mathcal{A}\big)$, $\forall \mathcal{A} \subset S_{\boldsymbol{\theta}}$ for all $\bar{g} \in \bar{\mathcal{G}}$. It is not clear which prior measures satisfy this requirement.

Reference Prior

Here, we briefly review the reference prior. The main idea of the reference prior is to maximize the information presented by data in the asymptotic approach (Bernardo, 1979). The maximization of information in the data is considered as the maximization of the expected Kullback-Leibler divergence of the prior from the posterior distribution. Consider $p(\mathbf{y}|\theta)$ as the statistical model with parameter $\theta \in \mathbb{R}$. Bernardo (1979) proposed that the reference prior is the one that maximizes the expected Kullback-Leibler divergence of $\pi(\theta)$ from it corresponding posterior $p(\theta|\mathbf{y})$

$$\int p(\mathbf{y}) \Big(\underbrace{\int \log \frac{p(\theta|\mathbf{y})}{\pi(\theta)} p(\theta|\mathbf{y}) \, \mathrm{d}\theta}_{\text{K-L divergence}} \Big) \mathrm{d}\, \mathbf{y}. \tag{2.1.7}$$

The expectation of the Kullback-Leibler divergence is taken with respect to the marginal density $p(\mathbf{y}) = \int p(\mathbf{y}|\theta)\pi(\theta)\mathrm{d}\theta$. The prior that maximizes this expected divergence turns out to be the Jeffreys prior (Berger et al., 2009).

For the models that have more than one parameter, the procedure of deriving the reference prior is started from ordering and grouping parameters according to the inferential importance. For simplicity, we assume that parameter space $\boldsymbol{\theta}$ has only two elements θ_1, θ_2. When there is only one group of parameters $\boldsymbol{\theta} = \{(\theta_1, \theta_2)\}$, that is all parameters are considered to have the same inferential importance, then the reference prior coincides with the Jeffreys' general prior. When the ordering is $\boldsymbol{\theta} = \{(\theta_1), (\theta_2)\}$ with θ_1 is considered to be more important than θ_2, the reference prior is specified as (Berger and Bernardo, 1992a; Ghosh et al., 2007)

$$\pi(\theta_1, \theta_2) \propto \underbrace{|\mathbf{I}_{22}|^{1/2}}_{\pi(\theta_2|\theta_1)} \cdot \underbrace{\exp\Big\{ \int |\mathbf{I}_{22}|^{1/2} \log \Big| \frac{|\mathbf{I}|}{\mathbf{I}_{22}} \Big|^{1/2} \Big\} \, \mathrm{d}\theta_2}_{\pi(\theta_1)}, \tag{2.1.8}$$

where $\mathbf{I}$ is the Fisher information matrix; $\mathbf{I}_{22}$ stands for the lower right corner of $\mathbf{I}$ corresponding to θ_2. The function $\pi(\theta_2|\theta_1)$ is actually the general Jeffreys prior for θ_2 with θ_1 fixed. In Equation (2.1.8), the expression of $\pi(\theta_1)$, the marginal prior of θ_1, is specifically chosen so that the expected Kullback-Leibler divergence of $\pi(\theta_1)$ from its corresponding posterior is maximized in the asymptotic sense. Particularly, the expectation is taken with respect to the marginal density $p(\mathbf{y}) = \int p(\mathbf{y}|\theta_1)\pi(\theta_1) \, \mathrm{d}\theta_1$.

Mathematically, the expression of $\pi(\theta_1)$ shown in Equation (2.1.8) is

$$\begin{aligned}
&\arg\max_{\pi(\theta_1)} \lim_{n\to\infty} \mathbb{E}\Big\{ \int \log \frac{p(\theta_1|\mathbf{y})}{\pi(\theta_1)} p(\theta_1|\mathbf{y})\, \mathrm{d}\theta_1 \Big\} \\
=&\arg\max_{\pi(\theta_1)} \lim_{n\to\infty} \int \Bigg\{ \underbrace{\Big[\int \log \frac{p(\theta_1|\mathbf{y})}{\pi(\theta_1)} p(\theta_1|\mathbf{y})\, \mathrm{d}\theta_1 \Big]}_{\text{K-L divergence}} \cdot \underbrace{\int p(\mathbf{y}|\theta_1)\pi(\theta_1)\, \mathrm{d}\theta_1}_{p(\mathbf{y})} \Bigg\} \mathrm{d}\mathbf{y} \\
=&\arg\max_{\pi(\theta_1)} \lim_{n\to\infty} \int \Bigg\{ \int \Big[\int \log \frac{p(\theta_1|\mathbf{y})}{\pi(\theta_1)} p(\theta_1|\mathbf{y})\, \mathrm{d}\theta_1 \Big] p(\mathbf{y}|\theta_1)\, \mathrm{d}\mathbf{y} \Bigg\} \pi(\theta_1)\, \mathrm{d}\theta_1.
\end{aligned} \tag{2.1.9}$$

The integration part in Equation (2.1.9) is called Lindley-Bernardo functional (Ghosh et al., 2007).

Returning to the one-way random effect model, Berger and Bernardo (1992) provided a table of reference priors for the one-way random effect model corresponding to different orderings and groupings. As they pointed out, all the reference priors have the following general form

$$\pi(\mu, \sigma, \sigma_\alpha) \propto \sigma^{-a} \sigma_\alpha^{-b} (N\sigma_\alpha^2 + \sigma^2)^{-c} \psi(\frac{\sigma_\alpha^2}{\sigma^2}), \tag{2.1.10}$$

where a, b, c are some constants that are different for different orderings and groupings; $\psi(\frac{\sigma_\alpha^2}{\sigma^2})$ could be either 1 or $\Big((N-1)+(1+N\frac{\sigma_\alpha^2}{\sigma^2})^{-2}\Big)^{1/2}$. Particularly, $\{a=1, b=-1, c=\frac{3}{2}, \psi=1\}$ corresponds to the reference prior for $\{(\mu, \sigma, \sigma_\alpha)\}$ that takes all parameters as one group. Also, it turns out to have the same form as the Jeffreys' general prior. Parameters ordered as $\{\mu, (\sigma, \sigma_\alpha)\}, \{(\sigma, \sigma_\alpha), \mu\}, \{\mu, \sigma, \sigma_\alpha\}, \{\sigma, \mu, \sigma_\alpha\}, \{\sigma, \sigma_\alpha, \mu\}$ take values $\{a=1, b=-1, c=1, \psi=1\}$ and this prior coincides with the modified Jeffreys' prior that considers μ fixed.

The procedure of calculating the reference prior is closely related to the grouping and ordering of parameters by their inferential importance since different groupings and different orderings lead to different reference priors. Berger and Bernardo (1992b) suggested that it is better to consider all parameters separately and order them according to the importance but they didn't specify the reason for doing so. The reasons for grouping and ordering parameters are not clear. We should not have any preference for any grouping and ordering since we assert that no prior information is available at hand. It is natural to assign equal importance to all

three hyper-parameters and consider them as a whole if we really have no subjective information. And in such a situation, the reference prior turns out to be the general Jefferys' prior.

Uniform Shrinkage Prior

The uniform shrinkage prior, proposed by Daniels (1999), only concentrates on σ_α in the random-effect model. The posterior mean of α_i is

$$\mathrm{E}(\alpha_i|\mu,\sigma_\alpha,\sigma,\mathbf{y}) = \frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma^2/N}y_{i\cdot} + (1-\frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma^2/N})\mu,$$

where $y_{i\cdot}$ stands for the sample mean of group i. The shrinkage factor of the posterior mean for α_i is $S = \frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma^2/N}$ and a uniform prior is specified for this factor.

Conditionally-conjugate Prior

Gelman et al. (2006) commented that the parameters $\{\mu,\sigma_\alpha,\sigma\}$ of one-way random effect model do not have a simple family of conjugate prior due to the complex structure of its likelihood as illustrated in Equation (2.1.5). However, the conditionally-conjugate prior could be easily recognised. Specifically, if the conditional prior of σ_α^2 is the inverse-gamma distribution $\mathrm{InvG}(a,a)$, then the conditional posterior distribution $p(\sigma_\alpha^2|\mu,\sigma^2,\boldsymbol{\alpha},\mathbf{y})$ is also the inverse-gamma distribution

$$\sigma_\alpha^2|\mu,\sigma^2,\boldsymbol{\alpha},\mathbf{y} \sim \mathrm{InvG}(a+\frac{m}{2}, a+\frac{1}{2}\sum_{i=1}^{m}\alpha_i^2). \tag{2.1.11}$$

Note that $p(\sigma_\alpha^2|\mu,\sigma^2,\boldsymbol{\alpha},\mathbf{y})$ belongs to the inverse-gamma family while $p(\sigma_\alpha^2|\mu,\sigma^2,\mathbf{y})$ does not. The inverse-gamma prior with small value for a such as 0.01 or 0.001 is usually considered as a non-informative prior to some degree for variance parameters in the conjugate prior category. This prior is appealing in terms of its computational convenience as the posterior samples can be obtained by directly implementing Gibbs sampler that iteratively updates the full conditional distributions $p(\sigma_\alpha^2|\mu,\sigma^2,\boldsymbol{\alpha},\mathbf{y})$, $p(\boldsymbol{\alpha}|\mu,\sigma^2,\sigma_\alpha^2,\mathbf{y})$, $p(\sigma^2|\mu,\boldsymbol{\alpha},\sigma_\alpha^2,\mathbf{y})$ and $p(\mu|\sigma^2,\boldsymbol{\alpha},\sigma_\alpha^2,\mathbf{y})$. Gelman et al. (2006) pointed out two problems in the use of this prior: 1) as a approach to 0, the prior would lead to an improper posterior distribution and thus reasonable values of a should be decided; 2) the value of a is very influential for the posterior distribution and the original non-informative intention is thus violated.

Gelman's Half-t Prior

Gelman et al. (2006) suggested that one approach to deal with prior problems for the group variance parameter in a hierarchical model is to give it a parametric model with hyper-parameters. More precisely, a folded-t prior distribution is proposed for σ_α in the one-way random effect model by using an augmented model displayed as follows (Gelman et al., 2006)

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$
$$\alpha_i = \xi\eta_i,$$
$$\eta_i \sim N(0, \sigma_\eta^2).$$

Clearly, we have $\sigma_\alpha = |\xi|\sigma_\eta$ according to the formula for the random effect α_i. If prior distributions for ξ and σ_η^2 are specified as standard Normal distribution and inverse-gamma respectively, then the implicit prior for σ_α turns out to be a folded-t distribution with the scale parameter A and degree of freedom v. The prior for σ_α could be expressed as

$$\pi(\sigma_\alpha) \propto \Big(1 + \frac{1}{v}\big(\frac{\sigma_\alpha}{A}\big)^2\Big)^{-(v+1)/2}.$$

If $v = 1$, the above prior changes to be a half-Cauchy distribution. And $A \to \infty$ would lead to a uniform prior for σ_α. In order to use this prior, the value of A needs to be specified. Gelman et al. (2006) suggested to set a large but finite value for A to obtain a weakly informative prior. Particularly, the value that is a bit higher than the expected standard deviation of the underlying α_i is used in his paper. They mentioned that such a prior provides more reliable posterior distributions than that provided by the uniform prior on σ_α^2 when the number of groups m is small. Because data could only provide little information about σ_α if m is small, a uniform prior on σ_α^2 would lead to improper posterior ($m < 3$) or proper but unrealistic broad posterior distributions.

Although we see some benefits of using this half-t prior for σ_α in the one-way random effect model, this prior indeed has problems. Firstly, the principle of expanding the model as above and constructing this prior is vague. Secondly, the choice of A is unclear especially when little is known about how the underlying α_i spread.

Chapter 3

Global Distance Structure Prior

Here, we introduce the idea of the global distance structure principle. Development of priors satisfying this principle for some simple problems are discussed in this chapter.

3.1 The General Situation

Firstly, we consider the global distance structure prior in the general situation. Denote a statistical model by $f_{\boldsymbol{\theta}}(\mathbf{x};\boldsymbol{\theta})$, where $\mathbf{x} \in \mathbb{R}^n$ and $\boldsymbol{\theta} \in S_{\boldsymbol{\theta}} \subseteq \mathbb{R}^p$. The statistical model is a mapping that maps a parameter space $S_{\boldsymbol{\theta}}$ to the space of probability distributions on $\mathbb{R}^n$, that is

$$f_{\boldsymbol{\theta}} : S_{\boldsymbol{\theta}} \mapsto \mathscr{F}(\mathbb{R}^n), \tag{3.1.1}$$

where $\mathscr{F}$ is the space of all distributions on $\mathbb{R}^n$.

A distance function, denoted by d, is likewise a mapping that takes two probability distributions on the same sample space and delivers a non-negative real number, that is

$$d : \mathscr{F}(\mathbb{R}^n) \times \mathscr{F}(\mathbb{R}^n) \mapsto \mathbb{R}^+, \tag{3.1.2}$$

where $\mathbb{R}^+ = \{\forall v \in \mathbb{R}^+; v \geq 0\}$.

Together with a statistical model, the distance function could induce a new mapping $d_{\boldsymbol{\theta}}$ that maps $S_{\boldsymbol{\theta}}$, the parameter space of $\boldsymbol{\theta}$, to the non-negative real space,

i.e.

$$d_{\boldsymbol{\theta}} : S_{\boldsymbol{\theta}} \times S_{\boldsymbol{\theta}} \mapsto \mathbb{R}^{+}. \tag{3.1.3}$$

Consider a re-parametrization η that bijectively maps the above mentioned parameter space $S_{\boldsymbol{\theta}}$ to the other parameter space $S_{\boldsymbol{\varphi}} = \{\boldsymbol{\varphi}; \boldsymbol{\varphi} = \eta(\boldsymbol{\theta}), \boldsymbol{\theta} \in S_{\boldsymbol{\theta}}\}$, i.e.

$$\eta : S_{\boldsymbol{\theta}} \mapsto S_{\boldsymbol{\varphi}}. \tag{3.1.4}$$

Since the re-parametrization is a bijective mapping, the function $\eta^{-1} : S_{\boldsymbol{\varphi}} \mapsto S_{\boldsymbol{\theta}}$ is well defined. The re-parametrization induces a mapping (a statistical model) $f_{\boldsymbol{\varphi}}$ which takes the new parameter space $S_{\boldsymbol{\varphi}}$ to the space of all probability distributions on $\mathbb{R}^n$. The statistical model $f_{\boldsymbol{\varphi}}$ could be expressed as

$$f_{\boldsymbol{\varphi}} = f_{\boldsymbol{\varphi}}(\mathbf{x}; \boldsymbol{\varphi}) = f_{\boldsymbol{\theta}}(\mathbf{x}; \eta^{-1}(\boldsymbol{\varphi})). \tag{3.1.5}$$

Being similar to Equation (3.1.3), a mapping $d_{\boldsymbol{\varphi}} : S_{\boldsymbol{\varphi}} \times S_{\boldsymbol{\varphi}} \mapsto \mathbb{R}^n$ could be induced by combing the statistical model $f_{\boldsymbol{\varphi}}$ with the distance function d. And it can be expressed as

$$d_{\boldsymbol{\varphi}}(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2) = d_{\boldsymbol{\theta}}(\eta^{-1}(\boldsymbol{\varphi}_1), \eta^{-1}(\boldsymbol{\varphi}_2)), \quad \forall\, \boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2 \in S_{\boldsymbol{\varphi}}. \tag{3.1.6}$$

We want to be able to compare the function $d_{\boldsymbol{\theta}}$ and $d_{\boldsymbol{\varphi}}$ and ask if they are effectively the same. The way we do this is to require first that the two spaces $S_{\boldsymbol{\theta}}$ and $S_{\boldsymbol{\varphi}}$ are the same, i.e.

$$S_{\boldsymbol{\theta}} = S_{\boldsymbol{\varphi}} = S. \tag{3.1.7}$$

With this requirement, we only need consider the re-parametrization of the form $\eta : S \mapsto S$ so that both mappings $d_{\boldsymbol{\theta}}$ and $d_{\boldsymbol{\varphi}}$ take $S \times S$ to R^{+}. Then we could check if the functions $d_{\boldsymbol{\theta}}$ and $d_{\boldsymbol{\varphi}}$ are the same function. The definition that the distance functions are the same is as follows.

Definition 3.1.1 *Two distance functions $d_{\boldsymbol{\theta}}$ and $d_{\boldsymbol{\varphi}}$ are the same if they satisfy*

$$d_{\boldsymbol{\theta}}(s_1, s_2) = d_{\boldsymbol{\varphi}}(s_2, s_2), \quad \forall s_1, s_2 \in S. \tag{3.1.8}$$

Principle:

If the re-parametrization η could make the two mappings $d_{\boldsymbol{\theta}}$ and $d_{\boldsymbol{\varphi}}$ satisfy Equation (3.1.8), we state that the global distance structure is invariant to the re-parametrization η and thus the prior measure should also be invariant to η. Rather than requiring the invariance in parametrized family, a prior is judged with respect to the invariance in global distance structure as illustrated in Equation (3.1.8). Note that the invariance in global distance structure is used to verify the objectivity of an existing prior rather than to design a new prior.

In the following parts of this chapter, priors satisfying this global distance structure principle are discussed. In section 3.2 and 3.3, the derivations of these priors are considered in two contexts respectively: firstly, finite discrete model space and, secondly, continuous model space. In section 3.4, the these priors for the location family, scale family and location-scale family are provided.

3.2 Finite Discrete Model Space

In the case of a finite collection of models (or parameter values), the principle of insufficient reason has been used as grounds for applying a uniform prior. Kass and Wasserman (1996) discussed this and the possible issues of the partitioning paradox of this principle. They used the example provided by Shafer et al. (1976) to elaborate the paradox. Let $\Lambda = \{\lambda_1, \lambda_2\}$, where λ_1 represent the event that there is life on orbit about the star Sirius and λ_2 denotes the event there is not. According to the principle of insufficient reason, the prior weight would be $\pi(\lambda_1) = \pi(\lambda_2) = \frac{1}{2}$. But now let $\Gamma = \{\gamma_1, \gamma_2, \gamma_3\}$, where γ_1 denotes the event that there is life around the Sirius star, γ_2 denotes the event that there are planets but no life, and γ_3 denotes the event that there are no planets. The principle of insufficient reason assigns the prior weight as $\pi(\gamma_1) = \pi(\gamma_2) = \pi(\gamma_3) = \frac{1}{3}$. We shall see that global distance structure might offer some possibilities for refining the argument.

Considering that the parameter space is a finite collection of discrete points, S in Equation (3.1.7) changes to be

$$S = \{\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_n\} \subseteq \mathbb{R}^n. \tag{3.2.9}$$

A distribution on the parameter space S can be represented by a probability vector. Thus, the prior that we would like to derive here is a probability vector. Since the parameter space has finite discrete elements, the distance functions $d_{\boldsymbol{\theta}}$ and $d_{\boldsymbol{\varphi}}$ are actually matrices. If the distance is chosen to be a symmetric function, we could obtain a symmetric matrix. Also the bijection η in Equation (3.1.4) turns out to be a permutation for the elements of the parameter space and thus the matrix $d_{\boldsymbol{\varphi}}$ is a permutation of rows and columns of the matrix $d_{\boldsymbol{\theta}}$.

The principle stated in section 3.1 is: if a re-parametrization η, which makes the global distance structure invariant as defined in Equation (3.1.8), could be recognized, the prior distributions should be also invariant to η, i.e. $\pi_\theta = \pi_\varphi$. In other words, the prior under consideration is acceptable with respect to the global distance structure invariance if it is invariant to η. And, a unique such prior does not always exist. The following two examples illustrate this facts.

Example 3.2.1 *Consider a situation with 5 simple models* $\{A, B, C, D, E\}$ *corresponding to 5 parameter values* $\{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5\}$ *where, for the chosen distance, the models can be represented in the plane in one of the two following ways:*

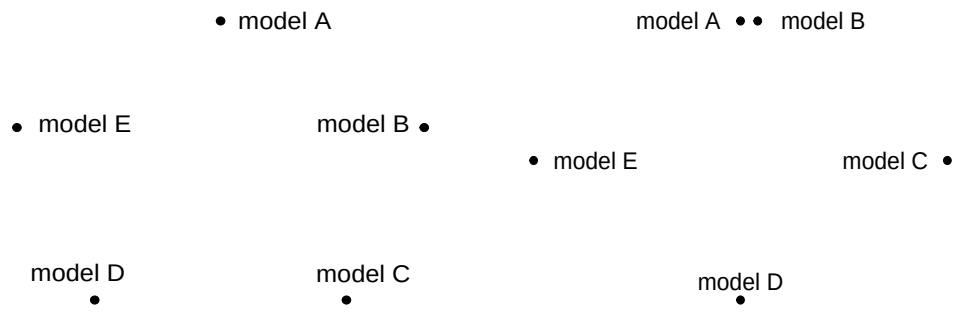


Figure 3.1: Two different five-model spaces.

For the situation on the left of Figure 3.1, the model space is represented by a regular pentagon with all sides of equal length a_1 and all diagonals of equal length a_2. The distance structure is unaffected by a rotation as illustrated in Figure 3.2.

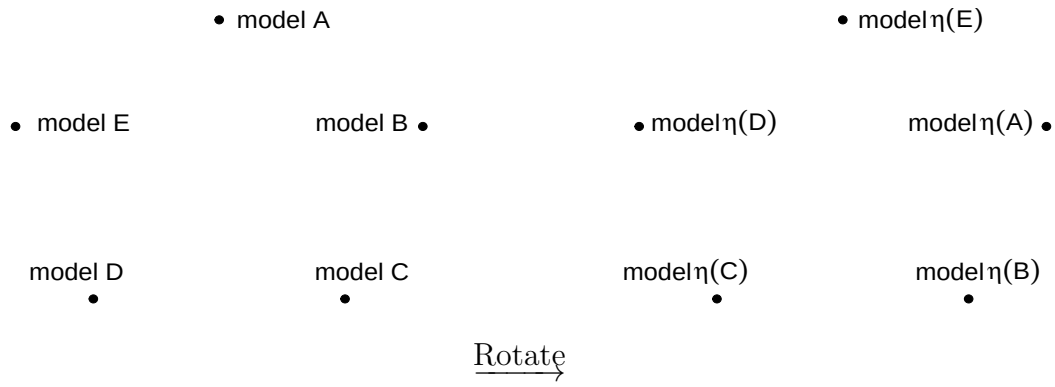


Figure 3.2: Left plot: the original model space; Right plot: model space after rotation. η denotes the rotation

The distance structure preservation could be seen by looking at the distance matrices for the original model space (left) and the rotated model space (right) displayed as follows

$$d_\theta = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 0 & a_1 & a_2 & a_2 & a_1 \\ a_1 & 0 & a_1 & a_2 & a_2 \\ a_2 & a_1 & 0 & a_1 & a_2 \\ a_2 & a_2 & a_1 & 0 & a_1 \\ a_1 & a_2 & a_2 & a_1 & 0 \end{pmatrix} \end{array},$$

$$d_\varphi = \begin{array}{c} \\ \eta(A) \\ \eta(B) \\ \eta(C) \\ \eta(D) \\ \eta(E) \end{array} \begin{array}{c} \begin{array}{ccccc} \eta(A) & \eta(B) & \eta(C) & \eta(D) & \eta(E) \end{array} \\ \begin{pmatrix} 0 & a_1 & a_2 & a_2 & a_1 \\ a_1 & 0 & a_1 & a_2 & a_2 \\ a_2 & a_1 & 0 & a_1 & a_2 \\ a_2 & a_2 & a_1 & 0 & a_1 \\ a_1 & a_2 & a_2 & a_1 & 0 \end{pmatrix} \end{array},$$

where d_θ denotes the distance matrix for the original space; d_φ is the distance matrix for the rotated model space; η, the rotation, is a re-parametrization. Clearly, these two distance matrices are exactly the same. Therefore, the prior distribution should be invariant to the rotation. That is,

$$\begin{aligned} & d_\theta = d_\varphi \\ \Longrightarrow\ & \pi_\theta = \pi_\varphi = \pi. \end{aligned} \tag{3.2.10}$$

Since both Jacobian factor and prior's propriety need not to be considered in finite discrete situations, we could have the following prior relationship:

$$\left.\begin{array}{l} \text{changing variable without Jacobian factor involved: } \pi_\theta(\theta_1) = \pi_\varphi(\eta(\theta_1)) \\ \text{Equation (3.2.10) tells that: } \pi_\varphi(\eta(\theta_1)) = \pi_\theta(\eta(\theta_1)) \\ \text{the rotation shown in Figure 3.2 tells } \theta_2 = \eta(\theta_1) \text{ and thus: } \pi_\theta(\eta(\theta_1)) = \pi_\theta(\theta_2) \end{array}\right\}$$

$$\Longrightarrow\ \pi_\theta(\theta_1) = \pi_\theta(\theta_2).$$

Likewise, $\pi(\theta_1) = \pi(\theta_2) = \pi(\theta_3) = \pi(\theta_4) = \pi(\theta_5)$ and it is a uniform prior which is

implied by the global distance structure invariance for the rotation.

The prior corresponding to the invariant distance structure, however, is not always unique. For the situation on the right of Figure 3.1, the distance structure is invariant to the flipping permutation as illustrated in Figure 3.3.

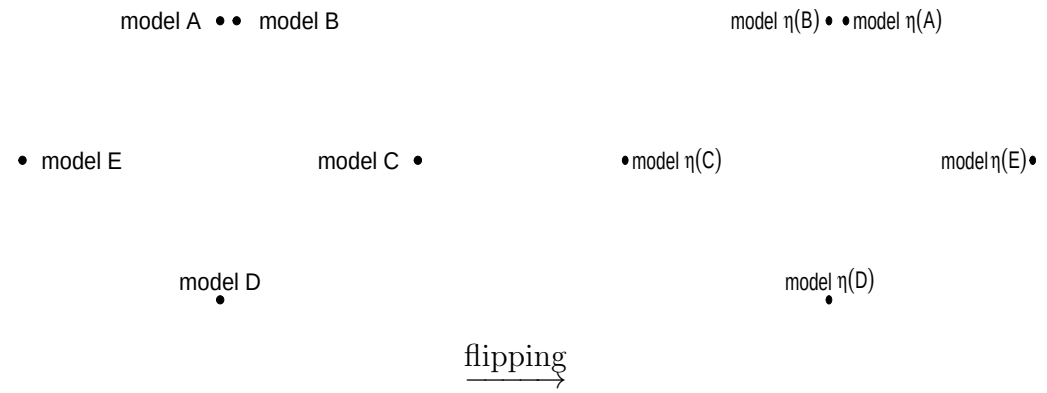


Figure 3.3: Left plot: Original plot; Right plot: Relabelling by flipping

Suppose the distance in the original model space are

$$
\begin{aligned}
d_\theta(\text{model } A, \text{model } B) &= a_1, \\
d_\theta(\text{model } A, \text{model } E) = d_\theta(\text{model } B, \text{model } C) &= a_2, \\
d_\theta(\text{model } A, \text{model } C) = d_\theta(\text{model } B, \text{model } E) &= a_3, \\
d_\theta(\text{model } A, \text{model } D) = d_\theta(\text{model } B, \text{model } D) &= a_4, \\
d_\theta(\text{model } C, \text{model } E) &= a_5, \\
d_\theta(\text{model } E, \text{model } D) = d_\theta(\text{model } C, \text{model } D) &= a_6.
\end{aligned}
$$

The distance structure preservation could be seen by looking at the distance matrices d_θ for the original space (left plot in Figure 3.3) and d_φ for the flipped model space (right plot in Figure 3.3)

$$d_\theta = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 0 & a_1 & a_3 & a_4 & a_2 \\ a_1 & 0 & a_2 & a_4 & a_3 \\ a_3 & a_2 & 0 & a_6 & a_5 \\ a_4 & a_4 & a_6 & 0 & a_6 \\ a_2 & a_3 & a_5 & a_6 & 0 \end{pmatrix} \end{array},$$

$$d_\varphi = \begin{array}{c} \\ \eta(A) \\ \eta(B) \\ \eta(C) \\ \eta(D) \\ \eta(E) \end{array} \begin{array}{c} \begin{array}{ccccc} \eta(A) & \eta(B) & \eta(C) & \eta(D) & \eta(E) \end{array} \\ \begin{pmatrix} 0 & a_1 & a_3 & a_4 & a_2 \\ a_1 & 0 & a_2 & a_4 & a_3 \\ a_3 & a_2 & 0 & a_6 & a_5 \\ a_4 & a_4 & a_6 & 0 & a_6 \\ a_2 & a_3 & a_5 & a_6 & 0 \end{pmatrix} \end{array}.$$

Therefore, we have $\pi_\theta = \pi_\varphi = \pi$. However, we could not determine a unique prior distribution. All we could say with certainty is that model C and model E should be assigned the same prior probability and that model A and model B should be assigned the same prior, that is

$$\pi(A) = \pi(B), \quad \pi(C) = \pi(E). \tag{3.2.11}$$

3.3 Continuous Model Space

Here, the derivations of priors based on global distance structure are discussed for the continuous model space (i.e. continuous parameter space). Compared with the situation of finite discrete model space in the above section, there are several differences. Firstly, the parameter space S is not a finite collection of discrete points and the distance function is not a matrix. In addition, the prior could not be expressed in a probability vector. Both proper and improper prior measures should be considered. Before the discussion of prior satisfying the global distance structure in the continuous model space, we look at the following two concepts.

Definition 3.3.1 *Given two measurable spaces $(S_{\boldsymbol{\theta}}, \mathcal{B}_{\boldsymbol{\theta}})$, $(S_{\boldsymbol{\phi}}, \mathcal{B}_{\boldsymbol{\phi}})$, a measure $\Omega_{\boldsymbol{\theta}}$ on $(S_{\boldsymbol{\theta}}, \mathcal{B}_{\boldsymbol{\theta}})$ and a measurable map $\eta : S_{\boldsymbol{\theta}} \to S_{\boldsymbol{\phi}}$, the induced measure $\Omega_{\boldsymbol{\varphi}}$ on $(S_{\boldsymbol{\phi}}, \mathcal{B}_{\boldsymbol{\phi}})$ is defined by*

$$\Omega_{\boldsymbol{\varphi}}(A) = \Omega_{\boldsymbol{\theta}}\big(\eta^{-1}(A)\big), \tag{3.3.12}$$

where $A \in \mathcal{B}_{\boldsymbol{\phi}}$.

Definition 3.3.2 *Two prior measures Ω_1 and Ω_2 are equivalent if they satisfy*

$$\Omega_1(A) = \textit{const.} \cdot \Omega_2(A), \quad \forall \textit{ measurable } A. \tag{3.3.13}$$

In other words, these two prior measures are indeed the same since their according posterior distribution are effectively the same. All the prior measures in such a equivalence class, denoted by $\mathscr{O}$, are different up to a constant.
Let us now turn to the prior satisfying the global distance structure in continuous model space.

Definition 3.3.3 *Suppose that we have*

- *a 1-1 re-parametrization $\eta : \boldsymbol{\theta} \to \boldsymbol{\varphi}$ such that the global distance structure is invariant as illustrated in Equation* (3.1.8)*. As illustrated in Equation* (3.1.7)*, we require the space of $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ to be the same in order to compare the distance function.*

- *a suggested prior measure $\Omega_{\boldsymbol{\theta}}$ to which $\pi_{\boldsymbol{\theta}}$, the prior density under consideration, corresponds*

We state that the prior measure $\Omega_{\boldsymbol{\theta}}$ is accepted with respect to the global distance structure invariance if the corresponding induced prior measure

$$\Omega_{\boldsymbol{\varphi}}(A) = \Omega_{\boldsymbol{\theta}}\big(\eta^{-1}(A)\big), \quad \forall A \subseteq S \tag{3.3.14}$$

is in the equivalent class of $\Omega_{\boldsymbol{\theta}}$, i.e.

$$\Omega_{\boldsymbol{\varphi}}(A) = const. \cdot \Omega_{\boldsymbol{\theta}}(A), \quad \forall \text{ measurable } A. \tag{3.3.15}$$

We state that the prior density $\pi_{\boldsymbol{\theta}}$, that corresponds to such a prior measure $\Omega_{\boldsymbol{\theta}}$, is accepted with respect to the global distance structure invariance.

Hartigan (1964) also considered this equivalence in prior measures. However, he concluded this equivalence from the invariance in parametrized family not from the invariance in global distance structure.

Here, we only take into account global distance structure prior measures which have finite positive measurements for bounded sets. The reason of not considering measures assigning 0 or ∞ measurements for bounded sets is as follows. Suppose there exists a bounded set A such that a prior measure Ω, which is accepted with respect to the global distance structure invariance, assigns 0 measurement to it, i.e.

$$\Omega(A) = 0.$$

Then according to the proposition 3.3.1, we can obtain

$$\Omega(B) = \Omega(A)\frac{\Omega\big(\eta(B)\big)}{\Omega\big(\eta(A)\big)} = 0,$$

where B is any bounded set. Therefore, we end up with a measure assigning 0 measurements for all bounded sets. Similarly, $\Omega(A) = \infty$ leads to a measure assigning ∞ measurements for all bounded sets. These two kinds of prior measure would not correspond to the concept of 'prior distributions'.

Proposition 3.3.1 *If the prior $\Omega_{\boldsymbol{\theta}}$ is accepted with respect to the global distance structure invariance, then $\exists\, \eta : \boldsymbol{\theta} \to \boldsymbol{\varphi}$, a 1-1 re-parametrization, such that*

$$\frac{\Omega_{\boldsymbol{\theta}}(A)}{\Omega_{\boldsymbol{\theta}}(B)} = \frac{\Omega_{\boldsymbol{\theta}}\big(\eta(A)\big)}{\Omega_{\boldsymbol{\theta}}\big(\eta(B)\big)}, \quad \forall A, B \subseteq S. \tag{3.3.16}$$

Proof : Let $A' = \eta(A), B' = \eta(B)$. As the space of $\boldsymbol{\theta}$ and the space of $\boldsymbol{\varphi}$ are forced to be the same, we have $A', B' \subseteq S$. According to Equation (3.3.15), we could have

$$\frac{\Omega_{\boldsymbol{\varphi}}(A')}{\Omega_{\boldsymbol{\varphi}}(B')} = \frac{\Omega_{\boldsymbol{\theta}}(A')}{\Omega_{\boldsymbol{\theta}}(B')}, \tag{3.3.17}$$

where $\Omega_{\boldsymbol{\varphi}}$ is the induced prior measure defined according to Equation (3.3.14). Therefore,

$$\frac{\Omega_{\boldsymbol{\varphi}}(A')}{\Omega_{\boldsymbol{\varphi}}(B')} = \frac{\Omega_{\boldsymbol{\theta}}\big(\eta^{-1}(A')\big)}{\Omega_{\boldsymbol{\theta}}\big(\eta^{-1}(B')\big)}. \tag{3.3.18}$$

By combining Equation (3.3.17) and (3.3.18), we have

$$\frac{\Omega_{\boldsymbol{\theta}}\big(\eta^{-1}(A')\big)}{\Omega_{\boldsymbol{\theta}}\big(\eta^{-1}(B')\big)} = \frac{\Omega_{\boldsymbol{\theta}}(A')}{\Omega_{\boldsymbol{\theta}}(B')}.$$

Since $A' = \eta(A), B' = \eta(B)$, the above equation changes to

$$\frac{\Omega_{\boldsymbol{\theta}}\big(A\big)}{\Omega_{\boldsymbol{\theta}}\big(B\big)} = \frac{\Omega_{\boldsymbol{\theta}}(\eta(A))}{\Omega_{\boldsymbol{\theta}}(\eta(B))}.$$

. □

The next proposition shows the prior density derived from a prior measure satisfying the global distance structure principle when the re-parametrization, to which the global distance structure is invariant, is a translation.

Proposition 3.3.2 *Suppose that a prior measure is accepted with respect to the global distance structure invariance and the 1-1 re-parametrization satisfying the global distance structure invariance is any translation*

$$\varphi = \eta(\theta) = \theta + c, \qquad S_\theta = \mathbb{R}, S_\varphi = \mathbb{R}, \tag{3.3.19}$$

then the corresponding prior density is

$$\pi(\theta) \propto \exp(\alpha\,\theta), \tag{3.3.20}$$

where α is some real value.

Proof : Denote the prior measure according to parametrization θ by Ω. According to proposition 3.3.1, we could have

$$\frac{\Omega(A)}{\Omega(B)} = \frac{\Omega(A+c)}{\Omega(B+c)},$$

where A and B are two arbitrary bounded intervals. By re-arrangement the above equation, we obtain

$$\frac{\Omega(A)}{\Omega(A+c)} = \frac{\Omega(B)}{\Omega(B+c)}.$$

Since the above equation holds for any bounded interval and any real value c, we could have the following result by setting B fixed

$$\begin{aligned} \frac{\Omega(A+c)}{\Omega(A)} &= k(c), \quad \forall \text{ bounded interval } A \subset \mathbb{R}, \forall c \in \mathbb{R} \\ &\implies \Omega(A+c) = k(c) \cdot \Omega(A), \end{aligned} \tag{3.3.21}$$

where $k(\cdot)$ is a function of c.

For bounded interval $I = (0, 1]$, we state the following two facts,

- $$\begin{aligned} &\Omega(nI) = \Omega(I) + \Omega(1+I) + \Omega(2+I) + \cdots + \Omega(n-1+I) \\ &= \Omega(I) + k(1)\Omega(I) + \big(k(1)\big)^2\Omega(I) + \cdots + \big(k(1)\big)^{n-1}\Omega(I) \qquad \text{by Equation (3.3.21)} \\ &= \Big[1 + k(1) + \big(k(1)\big)^2 + \cdots + \big(k(1)\big)^{n-1}\Big] \cdot \Omega(I), \end{aligned} \tag{3.3.22}$$

  where nI is a bounded interval $(0, n]$ and n is an integer.

- $$\begin{aligned} &\Omega(I) = \Omega(\frac{I}{m}) + \Omega(\frac{1}{m} + \frac{I}{m}) + \Omega(\frac{2}{m} + \frac{I}{m}) + \cdots + \Omega(\frac{m-1}{m} + \frac{I}{m}) \\ &= \Omega(\frac{I}{m}) + k(\frac{1}{m})\Omega(\frac{I}{m}) + \big(k(\frac{1}{m})\big)^2\Omega(\frac{I}{m}) + \cdots + \big(k(\frac{1}{m})\big)^{n-1}\Omega(\frac{I}{m}) \\ &\qquad\qquad \text{by Equation (3.3.21)} \\ &= \Big[1 + k(\frac{1}{m}) + \big(k(\frac{1}{m})\big)^2 + \cdots + \big(k(\frac{1}{m})\big)^{m-1}\Big] \cdot \Omega(\frac{I}{m}), \end{aligned} \tag{3.3.23}$$

  where $\frac{I}{m}$ is a bounded interval $(0, \frac{1}{m}]$ and m is an integer.

From the derivation of Equation (3.3.23), we have

$$\Omega(2I) = \Big[1 + k(\frac{1}{m}) + \big(k(\frac{1}{m})\big)^2 + \cdots + \big(k(\frac{1}{m})\big)^{2m-1}\Big] \cdot \Omega(\frac{I}{m}). \tag{3.3.24}$$

From the derivation of Equation (3.3.22), we have

$$\begin{aligned}\Omega(2I) &= \big(1+k(1)\big)\cdot\Omega(I)\\ &= \big(1+k(1)\big)\Big[1+k(\frac{1}{m})+\big(k(\frac{1}{m})\big)^2+\cdots+\big(k(\frac{1}{m})\big)^{m-1}\Big]\cdot\Omega(\frac{I}{m}) \qquad (3.3.25)\end{aligned}$$

by substituting Equation (3.3.23).

By comparing Equation (3.3.24) and (3.3.25), we have

$$\begin{aligned}1+k(\frac{1}{m})+\big(k(\frac{1}{m})\big)^2+\cdots+\big(k(\frac{1}{m})\big)^{2m-1}&\\ = \big(1+k(1)\big)\Big[1+k(\frac{1}{m})+\big(k(\frac{1}{m})\big)^2+\cdots+\big(k(\frac{1}{m})\big)^{m-1}\Big].&\qquad (3.3.26)\end{aligned}$$

Therefore, we have

1. If $k(\frac{1}{m})\neq 1$, then Equation (3.3.26) implies

$$\begin{aligned}\frac{1-\big(k(1/m)\big)^{2m}}{1-k(1/m)} &= \big(1+k(1)\big)\frac{1-\big(k(1/m)\big)^{m}}{1-k(1/m)}\\ \Longrightarrow \qquad k(\frac{1}{m}) &= \big(k(1)\big)^{1/m}. \qquad (3.3.27)\end{aligned}$$

2. If $k(\frac{1}{m})=1$, then Equation (3.3.26) implies

$$k(1)=1. \qquad (3.3.28)$$

Also, the above result could be written in the same form as Equation (3.3.27).

Up to now, we have some knowledge about $k(x)$ where $0<x\leq 1$ as shown in Equation (3.3.27) and (3.3.28). In order to get some information about $k(x)$ where $x\in\mathbb{R}$, we do the following job.

Firstly, we look at $\Omega(\frac{n}{m}I)$, where m and n are positive integers, as follows,

$$\begin{aligned}\Omega(\frac{n}{m}I) &= \Omega(n\cdot\frac{1}{m}I)\\ &= \Omega(\frac{1}{m}I)+\Omega(\frac{1}{m}+\frac{1}{m}I)+\cdots+\Omega\big((n-1)\frac{1}{m}+\frac{1}{m}I\big)\\ &= \Omega(\frac{1}{m}I)+k(\frac{1}{m})\Omega(\frac{1}{m}I)+\cdots+\big(k(\frac{1}{m})\big)^{n-1}\Omega(\frac{1}{m}I) \text{ by Equation (3.3.21)}\\ &= \big[1+k(\frac{1}{m})+\cdots+\big(k(\frac{1}{m})\big)^{n-1}\big]\Omega(\frac{1}{m}I). \qquad (3.3.29)\end{aligned}$$

By a simple re-arrangements to Equation (3.3.23), we obtain

$$\Omega(\frac{1}{m}I) = \big[1 + k(\frac{1}{m}) + \cdots + \big(k(\frac{1}{m})\big)^{m-1}\big]^{-1}\Omega(I). \tag{3.3.30}$$

By substituting Equation (3.3.30) to Equation (3.3.29), we have

$$\begin{aligned}
\Omega(\frac{n}{m}I) &= \frac{\big[1 + k(\frac{1}{m}) + \cdots + \big(k(\frac{1}{m})\big)^{n-1}\big]}{\big[1 + k(\frac{1}{m}) + \cdots + \big(k(\frac{1}{m})\big)^{m-1}\big]}\Omega(I) \\
&= \begin{cases} \frac{n}{m}\Omega(I) & \text{if } k(\frac{1}{m}) = 1 \\ \frac{1-\big(k(\frac{1}{m})\big)^{n}}{1-\big(k(\frac{1}{m})\big)^{m}}\Omega(I) & \text{if } k(\frac{1}{m}) \neq 1 \end{cases} \\
&= \begin{cases} \frac{n}{m}\Omega(I) & \text{if } k(\frac{1}{m}) = 1 \\ \frac{1-\big(k(1)\big)^{\frac{n}{m}}}{1-k(1)}\Omega(I) \quad \text{By Equation (3.3.27)} & \text{if } k(\frac{1}{m}) \neq 1 \end{cases}.
\end{aligned} \tag{3.3.31}$$

Both m and n are positive integers and $\{\frac{m}{n}; m, n \in \mathbb{Z}^+\}$ forms the positive rational number set $\mathbb{Q}^+$. Because the rational numbers are dense in $\mathbb{R}$, we have the following $\Omega(zI)$ with $z \in \mathbb{R}^+$,

$$\Omega(zI) = \begin{cases} z\Omega(I) & \text{if } k(\frac{1}{m}) = 1 \\ \frac{1-\big(k(1)\big)^{z}}{1-k(1)}\Omega(I) & \text{if } k(\frac{1}{m}) \neq 1 \end{cases}. \tag{3.3.32}$$

Secondly, we look at $\Omega(2zI)$, where $z \in \mathbb{R}^+$. From Equation (3.3.32), we obtain

$$\Omega(2zI) = \begin{cases} 2z\Omega(I) & \text{if } k(\frac{1}{m}) = 1 \\ \frac{1-\big(k(1)\big)^{2z}}{1-k(1)}\Omega(I) & \text{if } k(\frac{1}{m}) \neq 1 \end{cases}. \tag{3.3.33}$$

But,

$$\begin{aligned}
\Omega(2zI) &= \Omega(zI) + \Omega(z + zI) \\
&= \Omega(zI) + k(z)\Omega(zI) \quad \text{By Equation (3.3.21)} \\
&= \big(1 + k(z)\big)\Omega(zI) \\
&= \begin{cases} \big(1 + k(z)\big)z\Omega(I) & \text{if } k(\frac{1}{m}) = 1 \\ \big(1 + k(z)\big)\frac{1-\big(k(1)\big)^{z}}{1-k(1)}\Omega(I) & \text{if } k(\frac{1}{m}) \neq 1 \end{cases} \quad \text{by Equation (3.3.32).}
\end{aligned}$$

By comparing the above results with Equation (3.3.33), we could obtain $k(z)$, where $z \in \mathbb{R}^+$ in two situations:

1. If $k(\frac{1}{m}) = 1$, then

$$k(z) = 1. \tag{3.3.34}$$

2. If $k(\frac{1}{m}) \neq 1$, then

$$\frac{1 - \big(k(1)\big)^{2z}}{1 - k(1)}\Omega(I) = \big(1 + k(z)\big)\frac{1 - \big(k(1)\big)^{z}}{1 - k(1)}\Omega(I)$$

$$\implies \qquad k(z) = \big(k(1)\big)^{z}. \tag{3.3.35}$$

The result shown in Equation (3.3.34) could also be expressed in the same form of (3.3.35).

For any bounded interval A and any value $z \in \mathbb{R}^+$,

$$\begin{aligned}\Omega(A) &= \Omega(z - z + A) \\ &= k(z)k(-z)\Omega(A) \qquad \text{by Equation (3.3.21).}\end{aligned}$$

Therefore,

$$\begin{aligned}k(-z) &= \big(k(z)\big)^{-1} \\ &= \big(k(1)\big)^{-z}.\end{aligned}$$

Together with Equation (3.3.35), we obtain

$$k(z) = \big(k(1)\big)^{z}, \qquad \forall z \in \mathbb{R}. \tag{3.3.36}$$

Suppose that the bounded interval is $A = (\theta, \rho]$, then

$$\begin{aligned}\Omega(A) &= \Omega\big(\,(\theta, \rho]\,\big) = \Omega\big(\,\theta + (0, \rho - \theta]\,\big) \\ &= k(\theta) \cdot \Omega\big(\,(\rho - \theta)I\,\big) \\ &= \big(k(1)\big)^{\theta} \cdot \Omega\big(\,(\rho - \theta)I\,\big) \\ &= \begin{cases} (\rho - \theta)\Omega(I) & \text{if } k(\frac{1}{m}) = 1 \\ \frac{\big(k(1)\big)^{\theta} - \big(k(1)\big)^{\rho}}{1 - k(1)}\Omega(I) & \text{if } k(\frac{1}{m}) \neq 1 \end{cases} \qquad \text{by Equation (3.3.32).}\end{aligned}$$

By expressing ρ as $\rho = \theta + \Delta\theta$, the above result could be re-written as

$$\Omega(A) = \begin{cases} \Delta\theta \cdot \Omega(I) & \text{if } k(\frac{1}{m}) = 1 \\ \frac{\big(k(1)\big)^{\theta} - \big(k(1)\big)^{\theta + \Delta\theta}}{1 - k(1)}\Omega(I) & \text{if } k(\frac{1}{m}) \neq 1 \end{cases}.$$

Therefore,

$$\frac{\Omega(A)}{\Delta\theta} = \begin{cases} \Omega(I) & \text{if } k(\frac{1}{m}) = 1 \\ \frac{\big(k(1)\big)^{\theta}\big[1-\big(k(1)\big)^{\Delta\theta}\big]}{[1-k(1)]\cdot\Delta\theta}\Omega(I) & \text{if } k(\frac{1}{m}) \neq 1 \end{cases}.$$

As $\Delta\theta$ approaches to 0,

$$\lim_{\Delta\theta\to 0}\frac{\Omega(A)}{\Delta\theta} = \begin{cases} \Omega(I) & \text{if } k(\frac{1}{m}) = 1 \\ \frac{\big(k(1)\big)^{\theta}}{1-k(1)}\cdot\Big(-\log\big(k(1)\big)\Big)\cdot\Omega(I) & \text{if } k(\frac{1}{m}) \neq 1 \end{cases}.$$

Therefore, the density corresponding to the prior measure Ω exists, i.e.

$$\begin{aligned}\pi(\theta) &= \begin{cases} \Omega(I) & \text{if } k(\frac{1}{m}) = 1 \\ \frac{\big(k(1)\big)^{\theta}}{1-k(1)}\cdot\Big(-\ln\big(k(1)\big)\Big)\cdot\Omega(I) & \text{if } k(\frac{1}{m}) \neq 1 \end{cases} \\ &\propto \begin{cases} 1 & \text{if } k(\frac{1}{m}) = 1 \\ \big(k(1)\big)^{\theta} & \text{if } k(\frac{1}{m}) \neq 1 \end{cases}.\end{aligned}$$

Let $\alpha = \ln\big(k(1)\big)$, then $k(1) = e^{\alpha}$. With $k(\frac{1}{m}) = 1$, we have $k(1) = 1$ and thus $\alpha = 0$. Therefore, the above prior density could be re-written as

$$\pi(\theta) \propto e^{\alpha\theta}. \tag{3.3.37}$$

. □

Theorem 3.3.3 *Suppose that the prior measure Ω is accepted with respect to the global distance structure invariance and a translation illustrated as Equation* (3.3.19) *preserves the global distance structure. If a symmetrical global distance d is chosen to measure the differences between models, then the corresponding prior density is*

$$\pi(\theta) \propto 1 \tag{3.3.38}$$

Proof : Since a translation illustrated as Equation (3.3.19) preserves the global distance structure invariance, the prior density that is accepted with respect to the global distance structure invariance, according to Proposition 3.3.2, is

$$\pi(\theta) \propto e^{\alpha\theta}. \tag{3.3.39}$$

According to Equation (3.1.6), distance $d_\varphi(x, x')$ where $\forall x, x' \in S$ could be expressed as

$$\begin{aligned} d_\varphi(x, x') &= d_\theta(x - c, x' - c) \\ &= d_\theta(-x', -x) \qquad \text{by setting } c = x + x' \\ &= d_\theta(-x, -x') \qquad \text{by the symmetrical distance.} \end{aligned} \tag{3.3.40}$$

Considering the following re-parametrization

$$\phi = -\theta.$$

The corresponding distance function $d_\phi(x, x')$ where $\forall x, x' \in S$, according to Equation (3.1.6), could be expressed as

$$\begin{aligned} d_\phi(x, x') &= d_\theta(-x, -x') \\ &= d_\varphi(x, x') \qquad \text{according to Equation (3.3.40)} \\ &= d_\theta(x, x') \qquad \text{by the distance invariance to the translation } \varphi. \end{aligned}$$

From definition 3.1.1, the above result indicates that the distance structure is also invariant to a negative re-parametrization. According to proposition 3.3.1, we have

$$\frac{\Omega(A)}{\Omega(B)} = \frac{\Omega(-A)}{\Omega(-B)}, \quad \forall \text{ sets } A, B.$$

By a simple re-arrangement, the above equation changes to be

$$\frac{\Omega(A)}{\Omega(-A)} = \frac{\Omega(B)}{\Omega(-B)}.$$

The above equation holds for any choice of sets A and B. By fixing B, we have

$$\frac{\Omega(A)}{\Omega(-A)} = \text{const.} \tag{3.3.41}$$

Let $A = (\theta, \theta + \mathrm{d}\theta]$. According to Equation (3.3.39), the measure of A and $-A$ could be expressed as

$$\Omega(A) = e^{\alpha\theta}\mathrm{d}\theta, \quad \Omega(-A) = e^{-\alpha\theta}\mathrm{d}(-\theta). \tag{3.3.42}$$

By substituting the above results into Equation (3.3.41), we have

$$-e^{2\alpha\theta} = \text{const.} \tag{3.3.43}$$

Therefore, $\alpha = 0$ and the prior density in Equation (3.3.39) changes to be

$$\pi(\theta) \propto 1.$$

. □

Note that what we propose here is to take into account global distance structure rather than the global distance to derive a prior. And any divergence function, that is suitable to measure the difference between two probability distributions, can be used to derive a prior that is accepted with respect to preserving the global structure of the chosen divergence. In statistics, f-divergence, firstly introduced by Csiszar in 1963, is frequently used. Many popular divergences, such as the Kullback-Leibler divergence and Hellinger distance, are special cases of f-divergence. Let $d_\theta(\theta_1, \theta_2)$ denote the f-divergence between two statistical models under parametrization $\{\theta\}$. It could be defined as follows (Liese and Vajda, 2006)

$$d_\theta(\theta_1, \theta_2) = \int \varpi\Big(\frac{f(x;\theta_1)}{f(x;\theta_2)}\Big) f(x;\theta_2)\mathrm{d}x$$

where ϖ is a convex function such that $\varpi(1) = 0$. By denoting $w\Big(f(x;\theta_1), f(x;\theta_2)\Big) = \varpi\Big(\frac{f(x;\theta_1)}{f(x;\theta_2)}\Big) f(x;\theta_2)$, the above f-divergence can be rewritten as

$$d_\theta(\theta_1, \theta_2) = \int w\Big(f(x;\theta_1), f(x;\theta_2)\Big)\mathrm{d}x, \tag{3.3.44}$$

In particular, $w(\cdot,\cdot)$ changes along with the divergence function. For example, if Kullback-Leibler divergence is chosen,

$$w\Big(f(x;\theta_1), f(x;\theta_2)\Big) = \ln\Big(\frac{f(x;\theta_1)}{f(x;\theta_2)}\Big) f(x;\theta_1). \tag{3.3.45}$$

Since the Kullback-Leibler divergence is not symmetric, the following symmetrization is usually adopted,

$$w\Big(f(x;\theta_1), f(x;\theta_2)\Big) = \ln\Big(\frac{f(x;\theta_1)}{f(x;\theta_2)}\Big) f(x;\theta_1) + \ln\Big(\frac{f(x;\theta_2)}{f(x;\theta_1)}\Big) f(x;\theta_2). \tag{3.3.46}$$

The above formula symmetrizes Kullback-Leibler divergence. And thus the corresponding $d(\theta_1, \theta_2)$ satisfies the symmetry condition and could be considered as a metric measuring the distance between probability distributions.

If Hellinger distance is chosen,

$$w\Big(f(x;\theta_1), f(x;\theta_2)\Big) = \frac{1}{2}\Big(\sqrt{f(x;\theta_1)} - \sqrt{f(x;\theta_2)}\Big)^2.$$

If total variation distance is selected,

$$w\Big(f(x;\theta_1), f(x;\theta_2)\Big) = |f(x;\theta_1) - f(x;\theta_2)|.$$

3.4 Derivations for Simple Situations

In this section, we provide non-informative priors that are accepted with respect to the global distance structure invariance for the location family, scale family and location-scale family. In addition, the normal mean problem and the normal scale problem are considered as examples for the location family and scale family respectively. The one-way random effect model, that could be considered as an example of the location-scale family in a special situation, is discussed in Chapter 4.

3.4.1 Location Family

Let $f(y;\mu)$ denote a class of probability distributions that is parametrized by a scalar parameter μ which controls the 'location' of distribution. Mathematically, a location family must be expressible in the form

$$f(y;\mu) = h(y-\mu), \tag{3.4.47}$$

where $h(\cdot)$ is a function related to the probability density function.

Theorem 3.4.1 *Suppose that $f(\cdot)$ is a location family as defined in Equation* (3.4.47) *and a symmetric distance of the form* (3.3.44) *is chosen to measure the difference between probability distributions. The non-informative prior for the location parameter μ*

$$\pi(\mu) \propto 1 \tag{3.4.48}$$

is accepted with respect to the global distance structure invariance.

Proof : According to the distance function defined in Equation (3.3.44), the distance under parametrization $\{\mu\}$ is

$$d_\mu(\mu_1,\mu_2) = \int w\big(h(y-\mu_1), h(y-\mu_2)\big)\mathrm{d}y. \tag{3.4.49}$$

A new parametrization $\{\varphi\}$ is defined by the following translation

$$\varphi = \mu + c,$$

where c is a arbitrary constant. Under the parametrization $\{\varphi\}$, the probability distribution could be expressed as

$$f(y;\varphi) = h\big(y - (\varphi - c)\big) = h(y + c - \varphi). \tag{3.4.50}$$

According to the distance definition in Equation (3.3.44), the corresponding distance function under parametrization $\{\varphi\}$ is

$$d_\varphi(\varphi_1, \varphi_2) = \int w\Big(h(y + c - \varphi_1), h(y + c - \varphi_2)\Big)\mathrm{d}y.$$

By changing the variable $z = y + c$, the above distance function could be rewritten as

$$d_\varphi(\varphi_1, \varphi_2) = \int w\Big(h(z - \varphi_1), h(z - \varphi_2)\Big)\mathrm{d}z. \tag{3.4.51}$$

By comparing d_φ illustrated in the above line with d_μ expressed in Equation (3.4.49), these two distance functions clearly obtain the identical structure, that is

$$d_\varphi(\cdot, \cdot) = d_\mu(\cdot, \cdot). \tag{3.4.52}$$

Then, from the Theorem 3.3.3, a uniform is assigned for the location parameter

$$\pi(\mu) \propto 1.$$

. □

The fact stated by theorem 3.4.1 could be generalized to the context with vector-valued location parameters. Let $f(\mathbf{y};\boldsymbol{\mu})$ denote a class of probability distributions that is parametrized by a vector-valued parameter $\boldsymbol{\mu}$ which controls the 'location' of distribution. Mathematically, a location family must be expressible in the form

$$f(\mathbf{y};\boldsymbol{\mu}) = h(\mathbf{y} - X\boldsymbol{\mu}),$$

where $\mathbf{y} \in \mathbb{R}^n$ denotes $(n \times 1)$-dimensional observed data; $\boldsymbol{\mu} \in \mathbb{R}^p$ stands for a $(p \times 1)$-dimensional location parameter; X is a $(n \times p)$-dimensional specified matrix. Such a probability family is very common for regression models. The global distance structure prior is $\pi(\boldsymbol{\mu}) \propto 1$.

The Normal Mean

Suppose $\mathbf{y} = (y_1, \cdots, y_N)$ is a random sample from a normal distribution $N(\mu, \sigma^2)$, where σ is known. It belongs to the location family. By choosing the symmetrical Kullback-Leibler distance defined in Equation (3.3.46) to measure the corresponding distance between two models $N(\mathbf{y}; \mu_1)$ and $N(\mathbf{y}; \mu_2)$, a non-informative prior that is accepted with respect to the global distance structure invariance is uniform in μ itself, that is

$$\pi(\mu) \propto 1.$$

3.4.2 Scale Family

Let us now turn to the development of a non-informative prior distribution for scale family. Let $f(y;\sigma)$ denote a scale family. It represents a class of probability distributions which mathematically have the form

$$f(y;\sigma) = \frac{1}{\sigma}h(\frac{y}{\sigma}), \tag{3.4.53}$$

where σ is called 'scale parameter' and $h(\cdot)$ is a known function related to the probability density function. The distance function under the parametrization $\{\sigma\}$ could be expressed as

$$d_\sigma(\sigma_1,\sigma_2) = \int w\Big(\frac{1}{\sigma_1}h(\frac{y}{\sigma_1}), \frac{1}{\sigma_2}h(\frac{y}{\sigma_2})\Big)\mathrm{d}\,y, \tag{3.4.54}$$

where the form of $w(\cdot,\cdot)$ changes along with the chosen distance function.

Definition 3.4.1 *A divergence is homogeneous if it satisfies the following condition*

$$c\,w(f_1,f_2) = w(cf_1,cf_2), \tag{3.4.55}$$

where c is an arbitrary positive constant; f_1 and f_2 are two probability density functions.

Proposition 3.4.2 *Kullback-Leibler distance, Hellinger distance and total variation distance are homogeneous divergences.*

Proof : If Kullback-Leibler divergence is chosen to measure differences between probability distributions, we have

$$w(f_1,f_2) = f_1\ln\frac{f_1}{f_2}$$
$$\Longrightarrow c\,w(f_1,f_2) = (cf_1)\ln\frac{cf_1}{cf_2} = w(cf_1,cf_2).$$

That is, the Kullack-Leibler divergence is homogeneous and the symmetrical Kullback-Leibler distance automatically have this property. If Hellinger distance is chosen to measure differences between probability distributions, we have

$$w(f_1,f_2) = \frac{1}{2}(\sqrt{f_1}-\sqrt{f_2})^2$$
$$\Longrightarrow\ c\,w(f_1,f_2) = \frac{1}{2}(\sqrt{c\,f_1}-\sqrt{c\,f_2})^2 = w(cf_1,cf_2).$$

If the total variational distance is chosen to measure differences between probability distributions, we have

$$w(f_1, f_2) = |f_1 - f_2|$$
$$\Longrightarrow\ c\,w(f_1, f_2) = |cf_1 - cf_2| = w(cf_1, cf_2).$$

Therefore, Kullback-Leibler distance, Hellinger distance and total variation distance are homogeneous distances. □

Theorem 3.4.3 *Suppose that $f(\cdot)$ is a scale family as defined in Equation* (3.4.53) *and that a distance satisfying the homogeneous condition is chosen to measure the differences between probability distributions. We have*

1. *For the scale parameter σ, the non-informative prior that is accepted with respect to the global distance structure invariance is*

$$\pi(\sigma) \propto \sigma^c, \tag{3.4.56}$$

   where c is some real value.

2. *If the chosen distance is also symmetrical, then the non-informative prior that is accepted with respect to the global distance structure invariance is*

$$\pi(\sigma) \propto \frac{1}{\sigma}. \tag{3.4.57}$$

Proof : By considering the re-parametrization $\{\varphi\}$ defined as

$$\varphi = \log \sigma$$

the probability distribution could be expressed as

$$f(y; \varphi) = \frac{1}{\exp(\varphi)} h(\frac{y}{\exp(\varphi)})$$

The distance function under parametrization $\{\varphi\}$, denoted by d_φ, could be expressed as

$$d_\varphi(\varphi_1, \varphi_2) = \int w\Big(\frac{1}{\exp(\varphi_1)} h(\frac{y}{\exp(\varphi_1)}), \frac{1}{\exp(\varphi_2)} h(\frac{y}{\exp(\varphi_2)})\Big) \mathrm{d}\, y \tag{3.4.58}$$

Consider another parametrization $\{\phi\}$ constructed as

$$\phi = \varphi + c \tag{3.4.59}$$

where c is an arbitrary constant. The corresponding probability distribution under this parametrization is

$$\begin{aligned} f(y;\phi) &= \frac{1}{\exp(\phi)/\exp(c)} h\Big(\frac{y}{\exp(\phi)/\exp(c)}\Big) \\ &= \frac{1}{k\exp(\phi)} h\Big(\frac{y}{k\exp(\phi)}\Big) \end{aligned} \tag{3.4.60}$$

where $k = \exp(-c)$. The distance function d_ϕ under parametrization $\{\phi\}$ is

$$d_\phi(\phi_1,\phi_2) = \int w\Big(\frac{1}{k\exp(\phi_1)} h(\frac{y}{k\exp(\phi_1)}), \frac{1}{k\exp(\phi_2)} h(\frac{y}{k\exp(\phi_2)})\Big) \mathrm{d}\, y$$

By changing the variable $z = \frac{y}{k}$, the distance function d_ϕ in the above line changes to

$$d_\phi(\phi_1,\phi_2) = \int w\Big(\frac{1}{k\exp(\phi_1)} h(\frac{z}{\exp(\phi_1)}), \frac{1}{k\exp(\phi_2)} h(\frac{z}{\exp(\phi_2)})\Big) k\, \mathrm{d}\, z$$

Since the distance is required to have the homogeneous property, the above equation becomes

$$d_\phi(\phi_1,\phi_2) = \int w\Big(\frac{1}{\exp(\phi_1)} (\frac{z}{\exp(\phi_1)}), \frac{1}{\exp(\phi_2)} h(\frac{z}{\exp(\phi_2)})\Big) \mathrm{d}\, z \tag{3.4.61}$$

By comparing the distance function d_φ in Equation (3.4.58) and d_ϕ in Equation (3.4.61), we have

$$d_\varphi(\cdot,\cdot) = d_\phi(\cdot,\cdot) \tag{3.4.62}$$

Therefore, the global distance structure is invariant to a translation.

- To prove 1, according to proposition 3.3.2, we have

  $$\pi(\varphi) \propto \exp(\alpha\varphi)$$

  where α is some real value. By transforming back to parametrization $\{\sigma\}$ through the Jacobian factor, we have the following prior for the scale parameter

  $$\pi(\sigma) \propto \sigma^{\alpha-1} \tag{3.4.63}$$

  By letting $c = \alpha - 1$, we have

  $$\pi(\sigma) \propto \sigma^{c} \tag{3.4.64}$$

- To prove 2, if the chosen distance is also symmetrical, we have

$$\pi(\varphi) \propto 1 \tag{3.4.65}$$

according to Theorem 3.3.3. By transforming back to parametrization $\{\sigma\}$ through the Jacobian factor, we have

$$\pi(\sigma) \propto \frac{1}{\sigma}. \tag{3.4.66}$$

. □

The Normal Scale

As an example of the scale family, consider a Normal distribution for which the mean is supposed to be known. Suppose $\mathbf{y} = (y_1, \cdots, y_N)$ is a random sample from a normal distribution $N(\mu, \sigma^2)$, where μ is known. Suppose that the symmetrical Kullback-Leibler distance is chosen to measure the differences between probability distributions. According to Theorem 3.4.3, the prior, that is accepted with respect to the global distance structure invariance, is

$$\pi(\sigma) \propto \frac{1}{\sigma}.$$

3.4.3 Location-Scale family

We now turn to the development of a prior satisfying the global distance structure principle in the context of location-scale family. Let $f(y;\mu,\sigma)$ denote a location-scale family. It represents a family of probability distribution parametrized by a location parameter μ and a non-negative scale parameter σ. Mathematically, it has the form

$$f(y;\mu,\sigma) = \frac{1}{\sigma}h(\frac{y-\mu}{\sigma}), \tag{3.4.67}$$

where $g(\cdot)$ is a known function related to the probability density function.

Proposition 3.4.4 *Suppose that $f(\cdot)$ is a location-scale family as defined in Equation* (3.4.67)*. If a distance that satisfies the homogeneous condition is chosen to measure the difference between probability distributions, then*

$$d_{\{\mu,\varphi\}}(\cdot,\cdot) = d_{\{\theta,\phi\}}(\cdot,\cdot), \tag{3.4.68}$$

where $\{\mu,\varphi\}$ is a parametrization defined by

$$\mu = \mu, \qquad \varphi = \log\sigma \tag{3.4.69}$$

and $\{\theta,\phi\}$ is another parametrization defined by

$$\theta = a + c\mu, \qquad \phi = \varphi + \log c, \tag{3.4.70}$$

where a is an arbitrary value and c is an arbitrary positive value.

Proof : The probability distribution under the parametrization $\{\mu,\varphi\}$ could be expressed as

$$f(y;\mu,\varphi) = \frac{1}{\exp(\varphi)}h\Big(\frac{y-\mu}{\exp(\varphi)}\Big).$$

The distance function $d_{\{\mu,\varphi\}}$ under parametrization $\{\mu,\varphi\}$ is

$$\begin{aligned} &d_{\{\mu,\varphi\}}\Big(\{\mu_1,\varphi_1\},\{\mu_2,\varphi_2\}\Big) \\ &\qquad = \int w\bigg(\frac{1}{\exp(\varphi_1)}h\Big(\frac{y-\mu_1}{\exp(\varphi_1)}\Big), \frac{1}{\exp(\varphi_2)}h\Big(\frac{y-\mu_2}{\exp(\varphi_2)}\Big)\bigg)\mathrm{d}\,y. \end{aligned} \tag{3.4.71}$$

The probability distribution under the parametrization $\{\theta, \phi\}$ could be expressed as

$$f(y;\theta,\phi) = \frac{c}{\exp(\phi)} h\Big(\frac{cy + a - \theta}{\exp(\phi)}\Big).$$

The distance function $d_{\theta,\phi}$ under parametrization $\{\theta, \phi\}$ is

$$\begin{aligned} & d_{\{\theta,\phi\}}\Big(\{\theta_1,\phi_1\},\{\theta_2,\phi_2\}\Big) \\ & \qquad = \int w\bigg(\frac{c}{\exp(\phi_1)} h\Big(\frac{cy + a - \theta_1}{\exp(\phi_1)}\Big), \frac{c}{\exp(\phi_2)} h\Big(\frac{cy + a - \theta_2}{\exp(\phi_2)}\Big)\bigg) \mathrm{d}\, y. \end{aligned}$$

By changing the variable $z = cy + a$, the distance function $d_{\{\theta,\phi\}}$ in the above equation changes to

$$\begin{aligned} & d_{\{\theta,\phi\}}\Big(\{\theta_1,\phi_1\},\{\theta_2,\phi_2\}\Big) \\ & \qquad = \int w\bigg(\frac{c}{\exp(\phi_1)} h\Big(\frac{z - \theta_1}{\exp(\phi_1)}\Big), \frac{c}{\exp(\phi_2)} h\Big(\frac{z - \theta_2}{\exp(\phi_2)}\Big)\bigg) \frac{1}{c} \mathrm{d}\, z. \end{aligned}$$

Since the distance has the homogeneous property, the above formula could be rewritten as

$$\begin{aligned} & d_{\{\theta,\phi\}}\Big(\{\theta_1,\phi_1\},\{\theta_2,\phi_2\}\Big) \\ & \qquad = \int w\bigg(\frac{1}{\exp(\phi_1)} h\Big(\frac{z - \theta_1}{\exp(\phi_1)}\Big), \frac{1}{\exp(\phi_2)} h\Big(\frac{z - \theta_2}{\exp(\phi_2)}\Big)\bigg) \mathrm{d}\, z. \end{aligned} \tag{3.4.72}$$

By comparing the distance function $d_{\{\theta,\phi\}}$ in the above equation with the distance function $d_{\{\mu,\varphi\}}$ in Equation (3.4.71), we have

$$d_{\{\mu,\varphi\}}(\cdot,\cdot) = d_{\{\theta,\phi\}}(\cdot,\cdot). \tag{3.4.73}$$

That is, distance functions $d_{\{\mu,\varphi\}}$ and $d_{\{\theta,\phi\}}$ have the same structure. □

Theorem 3.4.5 *Suppose that $f(\cdot)$ is a location-scale family as defined in Equation* (3.4.67) *and that a distance satisfying the homogeneous condition is chosen to measure differences between probability distributions. For μ and σ, a non-informative prior that is accepted with respect to the global distance structure invariance is*

$$\pi(\mu,\sigma) \propto \sigma^{\gamma} \tag{3.4.74}$$

where γ is some real value.

Proof : According to the proposition 3.4.4, the distance structure is invariant between parametrizations $\{\mu, \varphi\}$ and $\{\theta, \phi\}$ as illustrated in Equation (3.4.69) and (3.4.70) respectively. Consider the following two Cartesian products

$$A = M \times S \qquad B = M' \times S',$$

where sets M, M' are bounded intervals of μ-space and sets S, S' are bounded intervals of φ-space

$$M = (\mu_1, \mu_2],\ M' = (\mu_3, \mu_4]; \quad S = (\varphi_1, \varphi_2],\ S' = (\varphi_3, \varphi_4].$$

According to the re-parametrization in Equation (3.4.70), we have

$$\eta(A) = \eta_\mu(M) \times \eta_\varphi(S) = (cM + a) \times (S + \log c), \tag{3.4.75}$$

$$\eta(B) = \eta_\mu(M') \times \eta_\varphi(S') = (cM' + a) \times (S' + \log c), \tag{3.4.76}$$

where η_μ and η_φ denotes the transformation on the space of μ and φ respectively. From the proposition 3.3.1, we have

$$\frac{\Omega\big(\eta(A)\big)}{\Omega(A)} = \frac{\Omega\big(\eta(B)\big)}{\Omega(B)}.$$

By fixing the Cartesian product B, the above equation changes to

$$\frac{\Omega\big(\eta(A)\big)}{\Omega(A)} = k(a, c), \qquad \forall A. \tag{3.4.77}$$

The above equation indicates that the ratio between $\Omega\Big(\eta(A)\Big)$ and $\Omega(A)$ does not depend on the set A and thus does not depend on $\mu_1, \mu_2, \varphi_1, \varphi_2$. By substituting Equation (3.4.75) into Equation (3.4.77), we have

$$\frac{\Omega\big(\eta(A)\big)}{\Omega(A)} = \frac{\Omega\Big((cM + a) \times (S + \log c)\Big)}{\Omega(M \times S)} = k(a, c). \tag{3.4.78}$$

By setting $c = 1$, Equation (3.4.78) changes to

$$\frac{\Omega\Big((M + a) \times S\Big)}{\Omega(M \times S)} = k(a, 1). \tag{3.4.79}$$

And by fixing the set S, the measure Ω could induce a new measure $\Omega^\star_S(M)$ illustrated as follows,

$$\Omega^\star_S(M) = \Omega(M \times S), \qquad \forall M. \tag{3.4.80}$$

The above equation indicates two points: 1) the new measure $\Omega_S^\star$ corresponds to the set S; 2) the measure $\Omega_S^\star$ is a measure only on the space of μ. By using the new measure $\Omega_S^\star$, Equation (3.4.79) could be re-expressed as follows

$$\frac{\Omega_S^\star(M+a)}{\Omega_S^\star(M)} = k(a,1).$$

The above result is the same with that in Equation (3.3.21). According to the proof in proposition 3.3.2, we could obtain the density on the space of μ

$$\pi_S^\star(x) = \beta(S)\cdot\exp\Big(\alpha(S)\cdot x\Big), \tag{3.4.81}$$

where $\alpha(S)$ and $\beta(S)$ are some values changing with the bounded interval $S = (\varphi_1,\varphi_2]$. And thus they could also be expressed as

$$\alpha(S) = \alpha(\varphi_1,\varphi_2), \qquad \beta(S) = \beta(\varphi_1,\varphi_2).$$

According to Equation (3.4.80) and (3.4.81), the measure $\Omega(M\times S)$ could be expressed as

$$\begin{aligned}\Omega(M\times S) = \Omega_S^\star(M) &= \int_M \pi_S^\star(x)\,\mathrm{d}x \\ &= \int_{(\mu_1,\mu_2]} \beta(\varphi_1,\varphi_2)\cdot\exp\Big(\alpha(\varphi_1,\varphi_2)\cdot x\Big)\,\mathrm{d}x \qquad (3.4.82)\\ &= \frac{\beta(\varphi_1,\varphi_2)}{\alpha(\varphi_1,\varphi_2)}\Big[\exp\big(\alpha(\varphi_1,\varphi_2)\cdot\mu_2\big) - \exp\big(\alpha(\varphi_1,\varphi_2)\cdot\mu_1\big)\Big]. \qquad (3.4.83)\end{aligned}$$

According to Equation (3.4.83), the measure of the transformed set, $\Omega\Big((cM+a)\times(S+\log c)\Big)$, could be further expressed as

$$\begin{aligned}&\Omega\Big((cM+a)\times(S+\log c)\Big) = \\ &\quad\frac{\beta(\varphi_1+\log c,\varphi_2+\log c)}{\alpha(\varphi_1+\log c,\varphi_2+\log c)}\cdot\Big[\exp\big(\alpha(\varphi_1+\log c,\varphi_2+\log c)\cdot(c\mu_2+a)\big) \\ &\qquad\qquad - \exp\big(\alpha(\varphi_1+\log c,\varphi_2+\log c)\cdot(c\mu_1+a)\big)\Big].\end{aligned} \tag{3.4.84}$$

By substituting the results of Equation (3.4.83) and (3.4.84) into Equation (3.4.78), we have

$$\begin{aligned}&\frac{\Omega\Big((cM+a)\times(S+\log c)\Big)}{\Omega(M\times S)} = \\ &\frac{\beta(\varphi_1+\log c,\varphi_2+\log c)}{\alpha(\varphi_1+\log c,\varphi_2+\log c)}\frac{\alpha(\varphi_1,\varphi_2)}{\beta(\varphi_1,\varphi_2)}\cdot\exp\big(a\cdot\alpha(\varphi_1+\log c,\varphi_2+\log c)\big)\cdot\Lambda_1\cdot\Lambda_2 = k(a,c),\end{aligned} \tag{3.4.85}$$

where

$$\Lambda_1 = \exp\Big[\big(c\cdot\alpha(\varphi_1+\log c,\varphi_2+\log c)-\alpha(\varphi_1,\varphi_2)\big)\cdot\mu_1\Big],$$
$$\Lambda_2 = \frac{\exp\big[c\cdot\alpha(\varphi_1+\log c,\varphi_2+\log c)\cdot\Delta\mu\big]-1}{\exp\big[\alpha(\varphi_1,\varphi_2)\cdot\Delta\mu\big]-1},$$

and $\Delta\mu = \mu_2 - \mu_1$. Because of the fact that $\frac{\Omega\big((cM+a)\times(S+\log c)\big)}{\Omega(M\times S)}$ does not depend on the sets M and S, the term Λ_1 that involves μ_1 must be a constant, i.e.

$$\alpha(\varphi_1+\log c,\varphi_2+\log c) = \frac{1}{c}\alpha(\varphi_1,\varphi_2). \tag{3.4.86}$$

And the term Λ_2 becomes 1 once the above equation holds. Therefore, Equation (3.4.85) changes to

$$\frac{\Omega\big((cM+a)\times(S+\log c)\big)}{\Omega(M\times S)} = \frac{c\cdot\beta(\varphi_1+\log c,\varphi_2+\log c)}{\beta(\varphi_1,\varphi_2)}\cdot\exp\big(a\cdot\alpha(\varphi_1+\log c,\varphi_2+\log c)\big) = k(a,c). \tag{3.4.87}$$

By setting $a=0$ in the above equation, we obtain the following relationship for the function $\beta(\varphi_1,\varphi_2)$

$$\frac{c\cdot\beta(\varphi_1+\log c,\varphi_2+\log c)}{\beta(\varphi_1,\varphi_2)} = k(0,c). \tag{3.4.88}$$

By substituting the above relationship into Equation (3.4.87), we could obtain

$$\frac{\Omega\big((cM+a)\times(S+\log c)\big)}{\Omega(M\times S)} = k(0,c)\cdot\exp\big(a\cdot\alpha(\varphi_1+\log c,\varphi_2+\log c)\big). \tag{3.4.89}$$

Again, by using the fact that the above ratio does not depend on values φ_1 and φ_2, we could have that $\alpha(\varphi_1+\log c,\varphi_2+\log c)$ is a constant. Together with the fact illustrated in Equation (3.4.86), we could obtain

$$\alpha(\varphi_1,\varphi_2) = 0. \tag{3.4.90}$$

According to the above result and Equation (3.4.82), we could have

$$\Omega(M\times S) = \beta(\varphi_1,\varphi_2)|M|, \tag{3.4.91}$$

where $|M|$ is the size of the set M. The above result indicates that $\beta(\varphi_1,\varphi_2)$ could also be considered as a measure. Specifically, it is the measure of the set S on the space of φ. By a simple re-arrangement of Equation (3.4.91), we could obtain

$$\beta(S) = \beta(\varphi_1,\varphi_2) = \frac{\Omega(M\times S)}{|M|}. \tag{3.4.92}$$

By substituting the above result into Equation (3.4.88), we could obtain

$$\frac{\beta(S+\log c)}{\beta(S)} = \frac{k(0,c)}{c}.$$

The above result indicates that the ratio of measures between the transformed set $S+\log c$ and the original set S does not depend on the set itself. This result is in line with that stated in Equation (3.3.21). Therefore, according to proposition 3.3.2, we could conclude that the measure $\beta(S)$ has the density $\pi_\beta(\varphi) \propto \exp(\zeta\varphi)$, where ζ is some real value. Because of the relationship between the measure $\Omega(M \times S)$ and the measure $\beta(S)$ as illustrated in Equation (3.4.91), the measure $\Omega(M \times S)$ has the same density as that of the measure $\beta(S)$, i.e.

$$\pi(\mu, \varphi) \propto \exp(\zeta\varphi).$$

By transforming back to the parametrization $\{\mu, \sigma\}$ through the Jacobian factor, we have the density

$$\pi(\mu, \sigma) \propto \sigma^\gamma, \tag{3.4.93}$$

where γ, satisfying $\gamma = \zeta - 1$, is some real value. □

Unlike the situations for the location family and the scale family in the previous sections, we have no constraint available for the power of σ in Equation (3.4.93). The main reason is that in the context of both the location family and the scale family, the distances could also be invariant to the negative re-parametrization by adding a symmetrical assumption for the distance. This extra invariance, however, does not remain for the location-scale family. There might exist some other re-parametrizations that can make the global distance structure invariant and thus can specify the value of γ in the density function illustrated by Equation (3.4.93). But at this moment, with the invariance presented in proposition 3.4.4, we could only have the density $\pi(\mu, \sigma) \propto \sigma^\gamma$, where γ is unspecified. In other words, the prior σ^γ with any power γ is accepted with respect to the global distance structure invariance. For a Normal distribution $N(\mu, \sigma^2)$ with unknown μ, σ, both Jeffreys general prior $\pi(\mu, \sigma) \propto \sigma^2$ and its modified version $\pi(\mu, \sigma) \propto \sigma$ are accepted with respect to the global distance structure invariance.

There are some connections among the context invariant prior, relative invariant prior and the global distance structure prior considered here.

- Context invariance states that if the same statistical model is used in two different contexts, then exact same prior measure should be assigned. Jeffreys general prior satisfies this condition.
- Relative invariance states that if the same statistical model is used in two different contexts, then equivalent prior measure should be assigned.
- The principle considered here is that if two statistical models have the same global distance structure, then equivalent prior measures should be assigned.

In the following table, prior for the location family, the scale family and the location-scale family according to the above three invariances are reported.

	Jeffreys	Relative	Here
location family	1	1	1
scale family	$\frac{1}{\sigma}$	σ^{γ}	$\frac{1}{\sigma}$
location-scale family	$\frac{1}{\sigma^2}$	σ^{γ}	σ^{γ}

Chapter 4

One-way Random Effect Model

In this chapter, the focus is on the development of non-informative priors that are accepted with respect to the global distance structure invariance for the one-way random effect model that has lots of difficulties in assigning a non-informative prior for its parameters. In section 4.1, the model and its parametrization are introduced. Then section 4.2 presents priors for this model from the perspective of the global distance structure invariance. In section 4.3, simulation studies are provided to analyse the performances of different prior distributions.

4.1 Model and Parametrization

Recall the one-way random effect model illustrated in section 2.1,

$$\begin{aligned} y_{ij} &= \mu + \alpha_i + \varepsilon_{ij}, \\ \alpha_i &\sim N(0, \sigma_\alpha^2), \\ \varepsilon_{ij} &\sim N(0, \sigma^2), \\ i = 1, \ldots, m; j &= 1, \ldots, N. \end{aligned}$$

The above model is parametrized by

$$\{\mu, \sigma, \sigma_\alpha\}. \tag{4.1.1}$$

This model could also be expressed in the following form,

$$\mathbf{y}_i \overset{iid}{\sim} \mathbf{N}(\mu\mathbf{1}_N, A_{N,N}),$$
$$i = 1, \ldots, m$$

where $\mathbf{1}_N$ is a N-dimensional column vector of all ones and

$$A_{N,N} = \begin{pmatrix} \vartheta^2 & \vartheta^2\rho^2 & \cdots & \vartheta^2\rho^2 \\ \vartheta^2\rho^2 & \vartheta^2 & \cdots & \vartheta^2\rho^2 \\ \vdots & \vdots & \ddots & \vdots \\ \vartheta^2\rho^2 & \vartheta^2\rho^2 & \cdots & \vartheta^2 \end{pmatrix},$$
$$\vartheta = \sqrt{\sigma^2 + \sigma_\alpha^2}, \qquad \rho = \frac{\sigma_\alpha}{\sqrt{\sigma^2 + \sigma_\alpha^2}}.$$

Particularly, $A_{N,N}$ could be written as

$$A_{N,N} = \vartheta^2\rho^2 J_{N,N} + (1-\rho^2)\vartheta^2 I_{N,N}, \tag{4.1.2}$$

where $J_{N,N}$ is a N-dimensional square matrix with all terms to be one and $I_{N,N}$ is a N-dimensional identity matrix. Therefore, the parametrization of the one-way random effect model changes to be

$$\{\mu, \vartheta, \rho\}. \tag{4.1.3}$$

This parametrization is specially chosen since μ, ϑ play the role of location parameter and scale parameter respectively with ρ fixed. This could be easily seen by looking at the likelihood function

$$\begin{aligned} L = p(\mathbf{y}; \boldsymbol{\mu}, A_{N,N}) &= \prod_{i=1}^{m} p(\mathbf{y}_i; \mu\mathbf{1}_N, A_{N,N}) \\ &= \prod_{i=1}^{m} \frac{1}{\sqrt{(2\pi)^N |A_{N,N}|}} \exp\Big(-\frac{1}{2}(\mathbf{y}_i - \mu\mathbf{1}_N)^T A_{N,N}^{-1}(\mathbf{y}_i - \mu\mathbf{1}_N)\Big), \end{aligned} \tag{4.1.4}$$

Since the determinant and the inverse operation of $A_{N,N}$ have the following results (see Appendix A.1 for detailed calculations)

$$|A_{N,N}| = (\vartheta^2)^N (1-\rho^2)^{N-1}\Big((N-1)\rho^2 + 1\Big), \tag{4.1.5}$$

$$A_{N,N}^{-1} = \frac{1}{\vartheta^2(1-\rho^2)}\Big(I_{N,N} - \frac{\rho^2}{1+(N-1)\rho^2} J_{N,N}\Big), \tag{4.1.6}$$

the likelihood in Equation (4.1.4) could be rewritten as follows,

$$L = \prod_{i=1}^{m} \frac{1}{\sqrt{(2\pi)^N(\vartheta^2)^N(1-\rho^2)^{N-1}\big((N-1)\rho^2+1\big)}} \\ \times \exp\bigg(-\frac{1}{2}(\mathbf{y}_i - \mu\mathbf{1}_N)^T \frac{1}{\vartheta^2(1-\rho^2)}\Big(I_{N,N} - \frac{\rho^2}{1+(N-1)\rho^2}J_{N,N}\Big)(\mathbf{y}_i - \mu\mathbf{1}_N)\bigg).$$

With ρ fixed, the above likelihood could be further written as

$$L = \prod_{i=1}^{m} K_1 \frac{1}{\sqrt{(2\pi)^N(\vartheta^2)^N}} \exp\bigg(-\frac{K_2}{2\vartheta^2}(\mathbf{y}_i - \mu\mathbf{1}_N)^T(\mathbf{y}_i - \mu\mathbf{1}_N)\bigg), \tag{4.1.7}$$

where

$$K_1 = \frac{1}{\sqrt{(1-\rho^2)^{N-1}\big((N-1)\rho^2+1\big)}}, \qquad K_2 = \frac{1}{(1-\rho^2)}\Big(I_{N,N} - \frac{\rho^2}{1+(N-1)\rho^2}J_{N,N}\Big).$$

Obviously, with the parameter ρ fixed as a constant, the likelihood shown in Equation (4.1.7) has the form of the location-scale family illustrated in Equation (3.4.67). Particularly, μ is the location parameter and ϑ is the scale parameter.

4.2 Non-informative Priors

In this section, the priors for the one-way random effect model parametrized by $\{\mu, \vartheta, \rho\}$ are derived from the perspective of the global distance structure invariance. Particularly, the symmetrical Kullback-Leibler distance is used to measure differences between models. Due to the independence among m groups, the Kullback-Leibler divergence of $p(\mathbf{y}|\mu_1, \vartheta_1, \rho_1)$ from $p(\mathbf{y}|\mu_2, \vartheta_2, \rho_2)$ is simply the sum of all the divergence of each group, that is

$$KL(p(\mathbf{y}|\mu_2, \vartheta_2, \rho_2)||p(\mathbf{y}|\mu_1, \vartheta_1, \rho_1)) = \sum_{i=1}^{m} KL(p(\mathbf{y}_i|\mu_2, \vartheta_2, \rho_2)||p(\mathbf{y}_i|\mu_1, \vartheta_1, \rho_1)).$$

Therefore, we could use the divergence of a single group $KL(p(\mathbf{y}_i|\mu_2, \vartheta_2, \rho_2)||p(\mathbf{y}_i|\mu_1, \vartheta_1, \rho_1))$ for simplicity.

Due to the structure of covariance matrix $A_{N\times N}$ in Equation 4.1.2 (i.e. correlated data), the divergence of each group $KL(p(\mathbf{y}_i|\mu_2, \vartheta_2, \rho_2)||p(\mathbf{y}_i|\mu_1, \vartheta_1, \rho_1))$ depends on N (the number of observations in each group) and thus the derived prior based on such a distance would be affected by N. In other words, the experiment design might have an influence on the derived prior. In our opinion, the influence of experiment design should be removed from the derived priors. Bernardo (2011) pointed out that statistical analysis is hardly to be completely objective because both experimental design and assumed models have strong subjective inputs. However, the reason that frequentist procedures are considered as 'objective' is that the frequentist inferences are only based on the assumed model and observed data. In the Bayesian framework, data is not collected at the stage of prior selection. In order to develop a prior with as least subjective input as possible, we need try to remove the influence of the experiment design on the global distance structure invariance so that the derived prior only depends on assumed models. It is, however, not always easy to remove. Therefore, we consider following two situations:

1. For some situations, the experiment design does not have influence on the global distance structure invariance and we thus can use the invariance from the global distance structure directly to derive a prior. A simple situation is

that the distance could be expressed as

$$d_{\boldsymbol{\theta}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = t(N) * d^{\star}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \tag{4.2.8}$$

where N denotes data size or experimental design; $t(N)$ is a function of N and is thus related to the experiment design; $d^{\star}$ is a function of parameters of interest and is independent of N. If the distance could be written to have the form as illustrated by Equation (4.2.8), then deriving priors by using $d_{\boldsymbol{\theta}}$ or $d^{\star}$ based on the global distance structure invariance are the same. Simple problems discussed in Chapter 3, such as the normal mean and the normal scale problem, are examples of this situation.

2. For situations where the global distance structure invariance is not clear due to the influence of experiment design, we make an attempt to investigate the structure invariance by using the averaged distances in the asymptotic sense, that is

   - $D_1 = \lim\limits_{N\to\infty} \bar{d}_{\theta}$, where $\bar{d}_{\theta} = \frac{1}{N} d_{\theta}$.
   - $D_2 = \lim\limits_{N\to\infty} (d_{\theta} - N \cdot D_1)$

   The 'taking limit' trick of D_1 provides us with two benefits: 1) the experimental design could be removed and the resulting distance describes how the model changes for the population rather than for the observed data; 2) some clear structure invariances might show up in the function D_1. Despite these attractive features, special care should be taken if D_1 is used as a distance to derive a prior since such a distance might result in some information loss. This fact is detailed in the following derivations of global distance structure priors for the one-way random effect model. The information lost is stored in the term D_2. The function D_2 might or might not provide some extra invariances for us to derive a prior. If no further global distance structure invariance can be recognised in D_2, we just consider the prior derived based on D_1 as a prior that is accepted with respect to the global distance structure invariance.

In the following part of this section, we discuss the derivations of priors from the perspective of the global distance structure invariance for the one-way random effect

model under parametrization $\{\mu, \vartheta, \rho\}$. We will look at the situation with all three parameters unknown, followed by the two-parameter situations and single-parameter situations.

4.2.1 With μ, ϑ, ρ unknown

Suppose that we are interested in all the parameters, the Kullback-Leibler divergence could be written as

$$KL(p(\mathbf{y}_i|\mu_2, \vartheta_2, \rho_2)||p(\mathbf{y}_i|\mu_2, \vartheta_1, \rho_1)) = \frac{1}{2}\Big(\mathrm{tr}(A_1^{-1}A_2) + (\mu_1 - \mu_2)^2 \mathbb{1}^T A_1^{-1}\mathbb{1} - N - \ln\frac{|A_2|}{|A_1|}\Big),$$

where

$$\mathrm{tr}(A_1^{-1}A_2) = N\frac{\vartheta_2^2}{\vartheta_1^2}\frac{1 + (N-2)\rho_1^2 - (N-1)\rho_1^2\rho_2^2}{(1-\rho_1^2)\big(1 + (N-1)\rho_1^2\big)}. \tag{4.2.9}$$

The symmetrical Kullback-Leibler distance could be expressed as

$$\begin{aligned} &d(\{\mu_1, \vartheta_1, \rho_1\}, \{\mu_2, \vartheta_2, \rho_2\}) = \\ &\frac{N}{2}\Bigg((\vartheta_2^2 - \vartheta_1^2)\Big(\frac{(N-2)\rho_1^2 + 1}{\vartheta_1^2(1-\rho_1^2)\big(1 + (N-1)\rho_1^2\big)} - \frac{(N-2)\rho_2^2 + 1}{\vartheta_2^2(1-\rho_2^2)\big(1 + (N-1)\rho_2^2\big)}\Big) \\ &\quad + (\vartheta_1^2\rho_1^2 - \vartheta_2^2\rho_2^2)\Big(\frac{(N-1)\rho_1^2}{\vartheta_1^2(1-\rho_1^2)\big(1 + (N-1)\rho_1^2\big)} - \frac{(N-1)\rho_2^2}{\vartheta_2^2(1-\rho_2^2)\big(1 + (N-1)\rho_2^2\big)}\Big) \\ &\quad + (\mu_1 - \mu_2)^2\Big(\frac{1}{\vartheta_1^2\big(1 + (N-1)\rho_1^2\big)} + \frac{1}{\vartheta_2^2\big(1 + (N-1)\rho_2^2\big)}\Big)\Bigg). \end{aligned}$$

This distance does not show clear structural invariance. Therefore, we attempt to find some structure invariances and remove N by using the limit technique in the following two situations:

- Suppose that $N \to \infty$, the averaged distance becomes

$$D_1 = \bar{d}(\{\mu_1, \vartheta_1, \rho_1\}, \{\mu_2, \vartheta_2, \rho_2\}) = \frac{1}{2}\Big(\frac{\vartheta_2^2(1-\rho_2^2)}{\vartheta_1^2(1-\rho_1^2)} + \frac{\vartheta_1^2(1-\rho_1^2)}{\vartheta_2^2(1-\rho_2^2)} - 2\Big). \tag{4.2.10}$$

  The parameter μ does not appear in the above formula and thus the structure of D_1 in the above formula would be invariant to any re-parametrization of μ. A global distance structure invariant prior $\pi(\mu, \vartheta, \rho)$ cannot be derived from this D_1 since there does not exist two prior measures that are effectively

equivalent as defined in Equation (3.3.13) with any re-parametrization. We will see later in Equation (4.2.14) that this D_1 is exactly the same as that in the context of only ϑ, ρ unknown. In other words, D_1 here only indicates some structure invariances conditional on known μ.

- In order to explore the information lost by the above D_1, we look at

$$D_2 = \lim_{N\to\infty}(d - N \cdot D_1) = \\ \frac{1}{2}\Big[\frac{\vartheta_2^2}{\vartheta_1^2}\frac{1}{\rho_1^2} + \frac{\vartheta_1^2}{\vartheta_2^2}\frac{1}{\rho_2^2} + \frac{(\mu_1-\mu_2)^2}{\vartheta_1^2}\frac{1}{\rho_1^2} + \frac{(\mu_1-\mu_2)^2}{\vartheta_2^2}\frac{1}{\rho_2^2} \\ -\big(\frac{\vartheta_2^2(1-\rho_2^2)}{\vartheta_1^2(1-\rho_1^2)}\frac{1}{\rho_1^2} + \frac{\vartheta_1^2(1-\rho_1^2)}{\vartheta_2^2(1-\rho_2^2)}\frac{1}{\rho_2^2}\big)\Big].$$

  Under the original parametrization $\{\mu, \sigma, \sigma_\alpha\}$, the above D_2 can be rewritten as

$$D_2 = \frac{1}{2}\Big[\Big(\frac{\sigma_{\alpha_2}^2}{\sigma_{\alpha_1}^2} + \frac{\sigma_{\alpha_1}^2}{\sigma_{\alpha_2}^2}\Big) + \Big(\frac{(\mu_1-\mu_2)^2}{\sigma_{\alpha_1}^2} + \frac{(\mu_1-\mu_2)^2}{\sigma_{\alpha_2}^2}\Big) - \Big(\frac{\sigma_2^2}{\sigma_1^2} + \frac{\sigma_1^2}{\sigma_2^2}\Big)\Big]$$

  This D_2 is invariant to the transformation

$$\{\mu, \sigma_\alpha, \sigma\} \to \{a + b\mu, b\sigma_\alpha, c\sigma\}$$

  where a ia an arbitrary value; b, c are arbitrary positive values. According to Theorem 3.4.3 and 3.4.5, a prior that is accepted with respect to the global distance structure invariance based on the above D_2 has the following form

$$\pi(\mu, \sigma, \sigma_\alpha) \propto \frac{\sigma_\alpha^\gamma}{\sigma} \tag{4.2.11}$$

  where γ is some real value.

In summary, we cannot conclude a prior that is accepted with respect to the global distance structure invariance based on $d(\{\mu_1, \vartheta_1, \rho_1\}, \{\mu_2, \vartheta_2, \rho_2\})$ and D_1. But, based on D_2, a prior that is accepted with respected to the global distance structure invariance can be derived and it has the form of Equation (4.2.11).

4.2.2 With ϑ, ρ unknown

Here we investigate the situation that ϑ, ρ are the parameters of interest. With μ known, we have the symmetrical Kullback-Leibler distance

$$\begin{aligned} d\Big(\{\vartheta_1,\rho_1\},\{\vartheta_2,\rho_2\}\Big) &= \frac{1}{2}\Big(\mathrm{tr}(A_1^{-1}A_2)+\mathrm{tr}(A_2^{-1}A_1)-2N\Big) \\ &= \frac{N}{2}\Big(\frac{1}{1+(N-1)\rho_1^2}\frac{\vartheta_2^2}{\vartheta_1^2}+\frac{(N-1)\rho_1^2}{1+(N-1)\rho_1^2}\frac{\vartheta_2^2(1-\rho_2^2)}{\vartheta_1^2(1-\rho_1^2)} \\ &\qquad +\frac{1}{1+(N-1)\rho_2^2}\frac{\vartheta_1^2}{\vartheta_2^2}+\frac{(N-1)\rho_2^2}{1+(N-1)\rho_2^2}\frac{\vartheta_1^2(1-\rho_1^2)}{\vartheta_2^2(1-\rho_2^2)}-2\Big). \end{aligned} \tag{4.2.12}$$

In particular, $\mathrm{tr}(A_1^{-1}A_2)$ in the above $d\Big(\{\vartheta_1,\rho_1\},\{\vartheta_2,\rho_2\}\Big)$ is the same with that in Equation (4.2.9) but can be rewritten as a weighted sum of these two terms $\frac{\vartheta_2^2}{\vartheta_1^2}=\frac{\sigma_2^2+\sigma_{\alpha\,2}^2}{\sigma_1^2+\sigma_{\alpha\,1}^2}$ and $\frac{\vartheta_2^2(1-\rho_2^2)}{\vartheta_1^2(1-\rho_1^2)}=\frac{\sigma_2^2}{\sigma_1^2}$, i.e.

$$\mathrm{tr}(A_1^{-1}A_2)=N\Big(\frac{1}{1+(N-1)\rho_1^2}\frac{\vartheta_2^2}{\vartheta_1^2}+\frac{(N-1)\rho_1^2}{1+(N-1)\rho_1^2}\frac{\vartheta_2^2(1-\rho_2^2)}{\vartheta_1^2(1-\rho_1^2)}\Big). \tag{4.2.13}$$

This distance $d\Big(\{\vartheta_1,\rho_1\},\{\vartheta_2,\rho_2\}\Big)$ in Equation (4.2.12) does not show clear structural invariance due to the influence of N. Therefore, we now look at the following two situations:

- The averaged distance with $N\to\infty$,

  $$D_1=\lim_{N\to\infty}\bar{d}\Big(\{\vartheta_1,\rho_1\},\{\vartheta_2,\rho_2\}\Big)=\frac{1}{2}\Big(\frac{\vartheta_2^2(1-\rho_2^2)}{\vartheta_1^2(1-\rho_1^2)}+\frac{\vartheta_1^2(1-\rho_1^2)}{\vartheta_2^2(1-\rho_2^2)}-2\Big) \tag{4.2.14}$$

  By the following transformation

  $$\varphi=\log\vartheta,\quad \phi=\log(1-\rho^2),$$

  D_1 changes to

  $$D_1=\frac{1}{2}\Big[\exp\big(2(\varphi_2-\varphi_1)+(\phi_2-\phi_1)\big)+\exp\big(2(\varphi_1-\varphi_2)+(\phi_1-\phi_2)\big)-2\Big]. \tag{4.2.15}$$

  If the above term D_1 is considered as a distance to measure the differences between models, a re-parametrization that such a distance structure is invariant to is

  $$\eta=\varphi+a_1,\quad \xi=\phi+a_2, \tag{4.2.16}$$

where a_1, a_2 are arbitrary values. However, this re-parametrization cannot be used to derive the global distance structure prior since the identity of Equation (3.1.7) is violated. Particularly, S_ξ, the space of ξ, is not the same with S_ϕ, the space of ϕ since

$$S_\xi = (-\infty, a_2), \quad S_\phi = (-\infty, 0). \tag{4.2.17}$$

Therefore, the re-parametrization illustrated in Equation (4.2.16) cannot be used to derive a strict global distance structure invariant prior. The re-parametrization that the distance D_1 in Equation (4.2.15) is invariant to is

$$\eta = \varphi + a, \quad \phi = \phi, \tag{4.2.18}$$

where a is an arbitrary value. That is,

$$D_1\Big(\{\varphi_1, \phi_1\}, \{\varphi_2, \phi_2\}\Big) = D_1\Big(\{\eta_1, \phi_1\}, \{\eta_2, \phi_2\}\Big).$$

The above formula indicates the conditional invariance in the structure of D_1. Denote the parameter space for parameters φ, ϕ by S_φ, S_ϕ respectively. For any $R \subset S_\phi$ and its prior measure Ω_R, we have

$$\frac{\Omega_R\big(a + M\big)}{\Omega_R(M)} = k(a), \quad \forall M \subset S_\varphi \tag{4.2.19}$$

The above equation indicates that the ratio between $\Omega_R\big(a + M\big)$ and $\Omega_R(M)$ does not depend on the sets M. This is in line with that shown in Equation (3.3.21). Note that Ω_R corresponds to the set R and that Ω_R is a measure on the space S_φ. According to the derivation in Theorem 3.3.3, we can conclude that Ω_R has the density $\pi_R \propto \frac{1}{\vartheta}$. In other words, this is a conditional prior

$$\pi(\vartheta|\rho) \propto \frac{1}{\vartheta}$$

Therefore, the global distance structure invariant prior for the unknown parameters $\{\vartheta, \rho\}$ has the following form

$$\pi(\vartheta, \rho) \propto \pi(\rho) \cdot \frac{1}{\vartheta}. \tag{4.2.20}$$

The prior for ρ cannot be specified from the invariance provided by D_1.

- In order to consider the information lost by D_1, we look at D_2 in terms of the parametrization $\{\varphi, \phi\}$

$$\begin{aligned} D_2 &= \lim_{N\to\infty}(d - N \cdot D_1) \\ &= \frac{1}{2}\Big\{\frac{\exp[2(\varphi_2-\varphi_1)]}{1-\exp(\phi_1)} + \frac{\exp[2(\varphi_1-\varphi_2)]}{1-\exp(\phi_2)} \\ &\qquad - \frac{\exp[2(\varphi_2-\varphi_1)+(\phi_2-\phi_1)]}{1-\exp(\phi_1)} - \frac{\exp[2(\varphi_1-\varphi_2)+(\phi_1-\phi_2)]}{1-\exp(\phi_2)}\Big\}. \end{aligned}$$

  The re-parametrization, that the above D_2 is invariant to, is same with that shown in Equation (4.2.18). Therefore, the global distance structure invariant prior based on D_2 has the same form with that shown in Equation (4.2.20).

In summary, both D_1 and D_2 agree on the same re-parametrization that they are invariant to and thereby have the same global distance structure invariant prior $\pi(\vartheta, \rho) \propto \pi(\rho) \cdot \frac{1}{\vartheta}$.

4.2.3 With μ, ϑ unknown

Here, we suppose that μ, ϑ are the parameters of interest. With ρ known, the symmetrical Kullback-Leibler distance is

$$\begin{aligned} &d\Big(\{\mu_1,\vartheta_1\},\{\mu_2,\vartheta_2\}\Big) = \\ &\qquad \frac{1}{2}\Big(N\frac{\vartheta_2^2}{\vartheta_1^2} + N\frac{\vartheta_1^2}{\vartheta_2^2} + \frac{N(\mu_1-\mu_2)^2}{\vartheta_1^2\big(1+(N-1)\rho^2\big)} + \frac{N(\mu_1-\mu_2)^2}{\vartheta_2^2\big(1+(N-1)\rho^2\big)} - 2N\Big). \end{aligned} \tag{4.2.21}$$

By taking the transformation $\mu = \mu, \varphi = \log\vartheta$, the above Kullback-Leibler distance could be rewritten as

$$\begin{aligned} &d\Big(\{\mu_1,\varphi_1\},\{\mu_2,\varphi_2\}\Big) = \\ &\qquad \frac{1}{2}\Big(N\exp\big(2(\varphi_2-\varphi_1)\big) + N\exp\big(2(\varphi_1-\varphi_2)\big) \\ &\qquad\qquad + \frac{N}{1+(N-1)\rho^2}\Big(\frac{(\mu_1-\mu_2)^2}{\exp(2\varphi_1)} + \frac{(\mu_1-\mu_2)^2}{\exp(2\varphi_2)}\Big) - 2N\Big). \end{aligned} \tag{4.2.22}$$

It is easy to see that

$$d\Big(\{\mu_1,\varphi_1\},\{\mu_2,\varphi_2\}\Big) = d\Big(\{a+c\mu_1,\varphi_1+\log c\},\{a+c\mu_2,\varphi_2+\log c\}\Big). \tag{4.2.23}$$

This result is in line with the distance structure invariance proved for the location-scale family in proposition 3.4.4. This coincidence is rational since the model belongs to the location-scale family with the parametrization $\{\mu, \vartheta, \rho\}$ and known ρ. Therefore, according to Theorem 3.4.5, the non-informative prior according to the global distance structure invariant principle is

$$\pi(\mu, \vartheta) \propto \vartheta^{\gamma},$$

where γ is some real values.

4.2.4 With μ, ρ unknown

Suppose that μ, ρ are the parameters of interest. With ϑ known, the symmetrical Kullback-Leibler distance could be written as

$$\begin{aligned} d\Big(\{\mu_1,\rho_1\},\{\mu_2,\rho_2\}\Big) &= \frac{N}{2}\Big(\frac{(N-1)\rho_1^2(\rho_1^2-\rho_2^2)}{(1-\rho_1^2)\big(1+(N-1)\rho_1^2\big)}+\frac{(\mu_2-\mu_1)^2}{\vartheta^2\big(1+(N-1)\rho_1^2\big)} \\ &\qquad +\frac{(N-1)\rho_2^2(\rho_2^2-\rho_1^2)}{(1-\rho_2^2)\big(1+(N-1)\rho_2^2\big)}+\frac{(\mu_2-\mu_1)^2}{\vartheta^2\big(1+(N-1)\rho_2^2\big)}\Big) \\ &= \frac{N}{2}\Big(\frac{(N-1)(\rho_1^2-\rho_2^2)^2\big((N-1)\rho_1^2\rho_2^2+1\big)}{(1-\rho_1^2)(1-\rho_2^2)\big(1+(N-1)\rho_1^2\big)\big(1+(N-1)\rho_2^2\big)} \\ &\qquad +\frac{(\mu_1-\mu_2)^2}{\vartheta^2}\big(\frac{1}{1+(N-1)\rho_1^2}+\frac{1}{1+(N-1)\rho_2^2}\big)\Big). \end{aligned} \tag{4.2.24}$$

Obviously, the above distance does not show clear structural invariance due to the influence of N. We attempt to consider the following two situations:

- Suppose that $N \to \infty$, the averaged distance becomes

  $$D_1 = \lim_{N\to\infty} \bar{d}\Big(\{\mu_1,\rho_1\},\{\mu_2,\rho_2\}\Big) = -\frac{1}{2}\Big(\frac{1-\rho_2^2}{1-\rho_1^2}-1\Big)\Big(\frac{1-\rho_1^2}{1-\rho_2^2}-1\Big).$$

  Again, the above D_1 could not tell anything about μ and thus its structure is invariant to any re-parametrization of μ. This D_1 only provides structure invariance conditional on known μ. From D_1, we cannot obtain a global distance structure invariant prior $\pi(\mu, \rho)$.

- In order to explore the information lost by D_1, we look as

$$\begin{aligned} D_2 &= \lim_{N\to\infty}(d - N\cdot D_1) \\ &= \frac{1}{2}\Big[-\frac{1}{2}\Big(\frac{1-\rho_2^2}{1-\rho_1^2}-1\Big)\Big(\frac{1-\rho_1^2}{1-\rho_2^2}-1\Big)\frac{1-(\rho_1^2+\rho_2^2)}{\rho_1^2\rho_2^2} + \frac{(\mu_1-\mu_2)^2}{\vartheta^2}\frac{\rho_1^2+\rho_2^2}{\rho_1^2\rho_2^2}\Big]. \end{aligned}$$

  The re-parametrization, that the above D_2 is invariant to, is

$$\eta = \mu + a, \quad \rho = \rho, \tag{4.2.25}$$

  where a is an arbitrary real value. Being similar with the argument for D_1 in section 4.2.2, we can obtain a conditional prior $\pi(\mu|\rho)$. According to Theorem 3.3.3, this conditional prior has the following form

$$\pi(\mu|\rho) \propto 1$$

  Therefore, the global distance structure invariant prior for the unknown parameters $\{\mu, \rho\}$ has the following form

$$\pi(\mu, \rho) \propto \pi(\rho) \tag{4.2.26}$$

  The prior for ρ cannot be specified from the invariance provided by this D_2.

In summary, the global distance structure invariant prior in the context of having $\{\mu, \rho\}$ unknown is derived based on D_2 and has the form $\pi(\mu, \rho) \propto \pi(\rho)$.

4.2.5 With only μ unknown

Suppose that the location parameter μ is the only parameter of interest. With ϑ, ρ known, the symmetrical Kullback-Leibler distance is

$$d(\mu_1, \mu_2) = \frac{(\mu_1-\mu_2)^2}{\vartheta^2(1/N + \rho^2(N-1)/N)}. \tag{4.2.27}$$

According to Equation (4.2.8), this distance d shows dependency only on the two parameter values via the difference and is obviously invariant to a translation no matther what values N takes. According to theorem 3.3.3, a non-informative prior based on the global distance structure invariance is the uniform distribution, that is

$$\pi(\mu) \propto 1.$$

4.2.6 With only ϑ unknown

Suppose that the scale parameter ϑ is the only parameter of interest. With μ, ρ known, the symmetrical Kullback-Leibler distance is

$$d(\vartheta_1, \vartheta_2) = \frac{N}{2}\Big(\frac{\vartheta_2^2}{\vartheta_1^2} + \frac{\vartheta_1^2}{\vartheta_2^2} - 2\Big). \tag{4.2.28}$$

It only depends on the two scale parameter values via the ratio. By taking the logarithm transformation of the scale parameter ϑ, i.e.

$$\varphi = \log \vartheta,$$

the distance could be invariant to any translation. Therefore, a non-informative prior that is accepted with respect to the global distance structure invariance, according to theorem 3.3.3, is $\pi(\varphi) \propto 1$. By transforming back to the original parametrization, the non-informative prior changes to

$$\pi(\vartheta) \propto \frac{1}{\vartheta}. \tag{4.2.29}$$

4.2.7 With only ρ unknown

Suppose that we are interested in ρ. With μ, ϑ unknown, the symmetrical Kullback-Leibler distance is

$$\begin{aligned} d(\rho_1, \rho_2) &= KL(p(\mathbf{y}_i|\mu, \vartheta, \rho_2)||p(\mathbf{y}_i|\mu, \vartheta, \rho_1)) + KL(p(\mathbf{y}_i|\mu, \vartheta, \rho_1)||p(\mathbf{y}_i|\mu, \vartheta, \rho_2)) \\ &= \frac{1}{2}\Big(\frac{N(N-1)\rho_1^2(\rho_1^2 - \rho_2^2)}{(1-\rho_1^2)\big(1+(N-1)\rho_1^2\big)} + \frac{N(N-1)\rho_2^2(\rho_2^2 - \rho_1^2)}{(1-\rho_2^2)\big(1+(N-1)\rho_2^2\big)}\Big). \end{aligned} \tag{4.2.30}$$

This distance does not show clear structural invariance. Therefore, we attempt to find some structure invariances and remove N by using the limit technique. As $N \to \infty$, we have

$$\begin{aligned} D_1 = \lim_{N\to\infty} \frac{1}{N} d(\rho_1, \rho_2) &= \frac{1}{2}(\rho_1^2 - \rho_2^2)(\frac{1}{1-\rho_1^2} - \frac{1}{1-\rho_2^2}) = \frac{1}{2}\frac{\big(1-\rho_2^2-(1-\rho_1^2)\big)(\rho_1^2-\rho_2^2)}{(1-\rho_1^2)(1-\rho_2^2)} \\ &= \frac{1}{2}\frac{\Big(1-\rho_2^2-(1-\rho_1^2)\Big)\Big((1-\rho_2^2)-(1-\rho_1^2)\Big)}{(1-\rho_1^2)(1-\rho_2^2)} \\ &= -\frac{1}{2}\Big(\frac{1-\rho_2^2}{1-\rho_1^2} - 1\Big)\Big(\frac{1-\rho_1^2}{1-\rho_2^2} - 1\Big). \end{aligned} \tag{4.2.31}$$

By the following transformation,

$$\varphi = \log(1-\rho^2),$$

Equation ((4.2.31)) could be rewritten as

$$D_1 = -\frac{1}{2}\Big(\exp(\varphi_2-\varphi_1)-1\Big)\Big(\exp(\varphi_1-\varphi_2)-1\Big).$$

Suppose that this D_1 is considered as the distance to measure the differences between models. Under the parametrization φ, the above result is invariant to any translation, i.e.

$$\xi = \varphi + a$$

where a is an arbitrary value. Because of the similar reason illustrated in Equation (4.2.17), this re-parametrization cannot be used to derive a strict global distance structure prior. Let us put this problem aside and still consider to use this re-parametrization to derive a prior. The resulting prior is $\pi(\varphi) \propto 1$. And by transforming back to parametrization $\{\rho\}$, we have

$$\pi(\rho) \propto \frac{\rho}{1-\rho^2}. \tag{4.2.32}$$

4.2.8 Summary

Here, the priors derived in all different contexts from the global distance structure invariant principle are summarized in the following table. The first row specified the unknown parameters and the second row states the corresponding priors. $\{\cdot\}^\star$ denote that the corresponding prior is derived according to a re-parametrization that violates the identity of parameter spaces as illustrated in Equation (3.1.7).

Parameters	$\{\mu\}$	$\{\vartheta\}$	$\{\rho\}^\star$	$\{\mu,\vartheta\}$	$\{\mu,\rho\}$	$\{\vartheta,\rho\}$	$\{\mu,\vartheta,\rho\}$
Prior	1	$\frac{1}{\vartheta}$	$\frac{\rho}{1-\rho^2}$	ϑ^γ	$\pi(\rho)$	$\pi(\rho)\cdot\frac{1}{\vartheta}$	$\frac{\vartheta^\gamma\rho^\gamma}{1-\rho^2}$

Table 4.1: Non-informative priors from global distance structure invariant principle for the one-way random effect model

Particularly, the prior reported in the last column is obtained by transforming the prior $\pi(\mu,\sigma,\sigma_\alpha) \propto \frac{\sigma_\alpha^\gamma}{\sigma}$ to that under the parametrization $\{\mu,\vartheta,\rho\}$. In addition, by

looking at the above table, we see that the prior of the following form

$$\frac{1}{\vartheta}\frac{\rho}{1-\rho^2} \tag{4.2.33}$$

respects all the forms of the above priors reported in the rest columns of the above table. In the next section, we will test the performance of two priors:

$$\pi(\mu,\vartheta,\rho) \propto \frac{\vartheta^{\gamma}\rho^{\gamma}}{1-\rho^2}; \quad \pi(\mu,\vartheta,\rho) \propto \frac{1}{\vartheta}\frac{\rho}{1-\rho^2}$$

by simulation studies to see whether their corresponding posterior distributions have unreasonable performances. Note that γ in the prior $\pi(\mu,\vartheta,\rho) \propto \frac{\vartheta^{\gamma}\rho^{\gamma}}{1-\rho^2}$ is unspecified. Note that by choosing $\gamma \in (-1,0)$, the posterior distribution can be guaranteed to be proper even with $m = 3$. We, therefore, arbitrarily set γ as $-\frac{1}{2}$ and test $\pi(\mu,\vartheta,\rho) \propto \frac{1}{\sqrt{\vartheta\rho}(1-\rho^2)}$ in the next section.

4.3 Prior Evaluation

In this section, frequentist performance of different priors are investigated for the one-way random effect model. We firstly introduce the priors to be tested, followed by the simulated data. And then, the performance of different priors is presented.

Priors

The tested priors are listed as follows.

- Global distance structure prior (GDSP for short):

$$\pi(\mu, \sigma, \sigma_\alpha) \propto \frac{1}{\sigma\sqrt{\sigma_\alpha}}$$

  This is the prior obtained by setting $\gamma = -\frac{1}{2}$ in the prior $\pi(\mu, \sigma, \sigma_\alpha) \propto \frac{\sigma_\alpha^\gamma}{\sigma}$.

- Conditional Global distance structure prior (CGDSP for short):

$$\pi(\mu, \sigma, \sigma_\alpha) \propto \frac{\sigma_\alpha}{\sigma(\sigma_\alpha^2 + \sigma^2)}$$

  This prior is obtained by transforming the prior $\pi(\mu, \vartheta, \rho) \propto \frac{1}{\vartheta}\frac{\rho}{1-\rho^2}$ under the parametrization $\{\mu, \vartheta, \rho\}$ to the original parametrization $\{\mu, \sigma, \sigma_\alpha\}$. Apart from the prior derived in the context with all three parameters unknown, it respects forms of all the priors derived for the one-way random effect model in the previous section.

- Jeffreys prior (JP for short):

$$\pi(\mu, \sigma, \sigma_\alpha) \propto \frac{\sigma_\alpha}{\sigma(N\sigma_\alpha^2 + \sigma^2)^{3/2}}$$

  This prior is same as the reference prior when setting $\{\mu, \sigma, \sigma_\alpha\}$ in one group with same importance.

- Jeffreys prior with location fixed (JPLF for short):

$$\pi(\mu, \sigma, \sigma_\alpha) \propto \frac{\sigma_\alpha}{\sigma(N\sigma_\alpha^2 + \sigma^2)}$$

- Half Cauchy prior suggested by Gelman Gelman et al. (2006):

$$\pi(\mu, \sigma, \sigma_\alpha) \propto \frac{1}{\sigma(\sigma_\alpha^2 + A^2)}$$

Following the suggestion by Gelman, we set A as a large value to obtain a weakly informative prior. Particularly, 25 and 50 are chosen. Therefore, the tested priors denoted by 'Gelman25' and 'Gelman50' respectively are $\pi(\mu, \sigma, \sigma_\alpha) \propto \frac{1}{\sigma(\sigma_\alpha^2+25^2)}$ and $\pi(\mu, \sigma, \sigma_\alpha) \propto \frac{1}{\sigma(\sigma_\alpha^2+50^2)}$.

- Uniform shrinkage prior (USP for short):

$$\pi(\mu, \sigma, \sigma_\alpha) \propto \frac{\sigma\sigma_\alpha}{(N\sigma_\alpha^2 + \sigma^2)^2}$$

The above prior is obtained by setting the prior $\pi(\mu, \sigma, S) \propto \frac{1}{\sigma}$ with $S = \frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma^2/N}$ and transforming back to parametrization $\{\mu, \sigma, \sigma_\alpha\}$.

Simulated Data

The data used to explore the performance of priors are simulated by setting the parameter values and number of observations as follows:

Parameter Values				Experimental Design	
σ_α	σ	μ	$\frac{\sigma_\alpha}{\sigma}$	m	N
2	2	5	1	3	3
				3	100
				10	3
				10	100
2	0.2	5	10	3	3
				3	100
				10	3
				10	100
2	20	5	0.1	3	3
				3	100
				10	3
				10	100
20	2	5	10	3	3
				3	100
				10	3
				10	100

Table 4.2: Parameter values and experimental designs for simulating data

In total, there would be 16 scenarios with different parameter values and experimental designs. The value of μ is fixed as 5 for all scenarios. Four sets of parameter values for σ_α, σ are chosen varying the ratio $\frac{\sigma_\alpha}{\sigma}$. Particularly, the last set of parameter values $\{\sigma_\alpha = 20, \sigma = 2\}$ are specified to have the same ratio as that for the second set $\{\sigma_\alpha = 2, \sigma_\alpha = 0.2\}$. The reason for considering data simulated using parameter values $\{\sigma_\alpha = 20, \sigma = 2\}$ in the study is to detect whether the actual values for σ_α, σ themselves have some influences on performances of priors. As for the experimental design, 4 different settings are specified for each set of parameter values so that we could explore the impacts of the number of groups and number of observations within each group.

Simulation Results

As illustrated in table 4.2, there are 16 scenarios. For each scenario, 1000 data sets are firstly generated according to the one-way random effect model and then each prior is repeatedly tested on the simulated 1000 data sets. The performance of a prior is analysed by examining the mean, median, 95% credible interval and 95% HPD (Highest Posterior Density) interval of the corresponding posterior samples obtained by the MCMC method. Particularly, Stan (Stan Development Team, 2014a) is used to obtain the Markov chain with 50000 posterior MCMC samples. The chain is thinned by 3 and has 2000 burn-in iterations. Since 1000 data sets are generated for each scenario, averaged posterior mean, averaged posterior median, the percentage of true values falling in 95% credible interval and the percentage of true values falling in 95% HPD interval are reported in the following plots. Results of different priors are distinguished by colours: GDSP is represented by light-blue dot; CGDSP is represented by blue dot; green and purple dots stand for JP and JPLF respectively; red and yellow dots denote Gelman25 prior and Gelman50 prior respectively; an orange dot is for USP. In the following plots, X-axes records the data scenarios determined by the true parameter values and experimental designs. In particular, each data scenario is expressed by σ_α_σ_m_N in the x axis. For example, 2_2_3_100 stands for the data sets generated by setting $\sigma_\alpha = 2, \sigma = 2, m = 3, N = 100$.

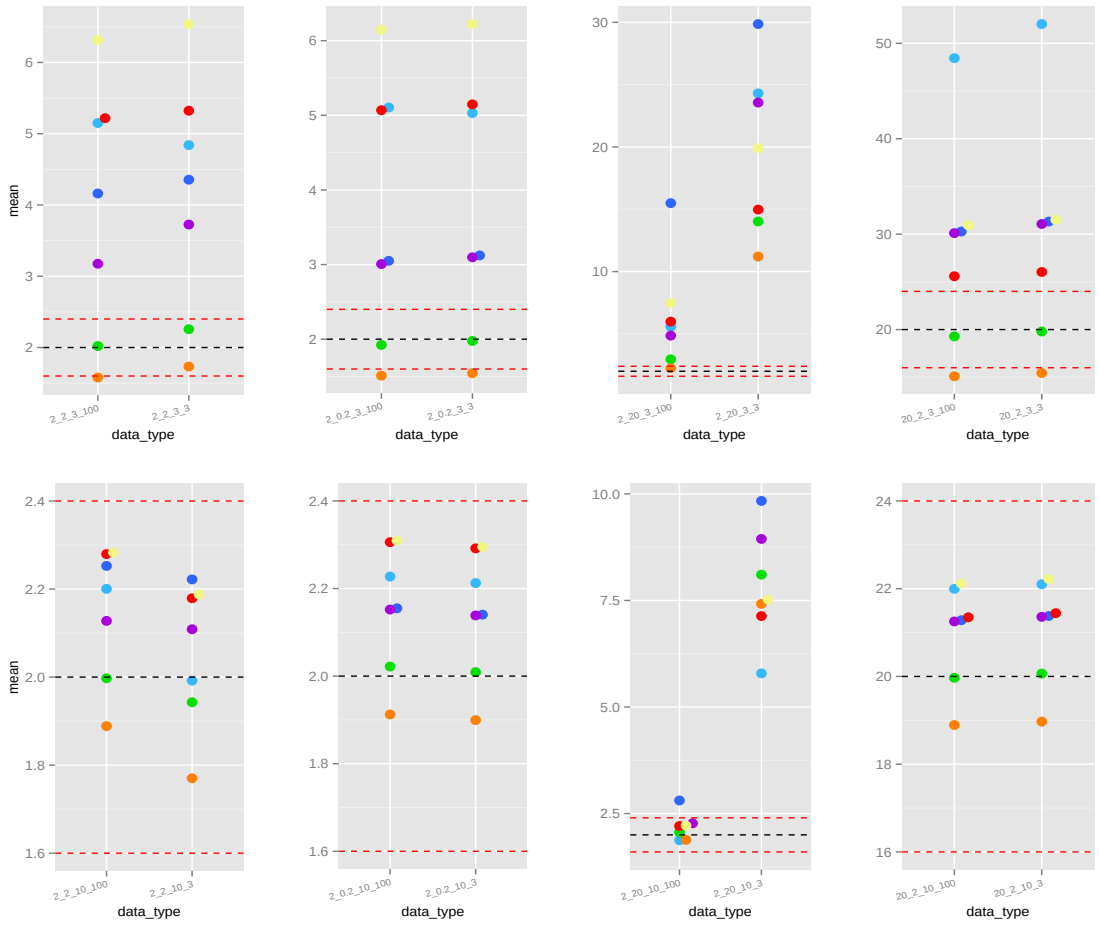


Figure 4.1: Averaged posterior mean of σ_α across 1000 data sets for each data type. The horizontal dotted line shows the true value of σ_α

The black dashed lines mark the true values of σ_α used to generate data sets. The red dashed lines mark the 20% error bounds. JP (green dot) is the prior that provides averaged posterior mean nearest to the true values for most scenarios. USP (orange dot) always gives the smallest mean values among the test priors. The Gelman25 prior (red dot) and the Gelman50 prior (yellow dot) obtain relatively large values compared with other priors in most situations. Particularly, the averaged mean values obtained by the Gelman50 are always larger than that offered by the Gelman25. This is consistent with the fact that the Gelman50 prior has larger scale and thus could be more diffuse than the Gelman25 prior. GDSP (light-blue dot) also provides relatively large values for the scenarios with σ_α larger than or equal to σ (i.e. $\{\sigma_\alpha = 2, \sigma = 2\}, \{\sigma_\alpha = 2, \sigma = 0.2\}, \{\sigma_\alpha = 20, \sigma = 2\}$) but gives results close to the

true values in the scenario with σ_α smaller than σ (i.e. $\{\sigma_\alpha = 2, \sigma = 20\}$). JPLF (purple dot) and GDSP (blue dot) have similar performances. For the scenarios with ratio $\frac{\sigma_\alpha}{\sigma} = 10$, JPLF and GDSP almost overlap with each other. Let us now turn to the influences of experimental designs and true parameter values. Firstly, the second row of Figure 4.1 illustrates prior performances for the data sets with $m = 10$ while the first row presents that for the data sets with $m = 3$. Obviously, great improvements are displayed in the second row compared to the performance of the first row since the all the results are shrinked towards the true values. Let us now turn to each individual plot containing two scenarios with only N (number of observations in each group) different. Increasing N from 3 to 100 provides big differences only for the scenarios with σ_α smaller than σ (i.e. $\{\sigma_\alpha = 2, \sigma = 20\}$). For the rest scenarios, changing N does not make obvious differences in the results. Generally speaking, the Jeffrey prior could be the best one among the test priors by simply looking at the posterior mean.

It is worthwhile exploring the median of the posterior samples due to the skewness of the posterior distribution for σ_α. Let us now look at Figure 4.2 that illustrates the averaged median of 1000 simulated data sets for each data scenario. The influences of true parameter values and experimental designs on the averaged median are similar to that exhibited in Figure 4.1 for the averaged mean. The performances of priors, however, have great differences in terms of median values. Particularly, JPLF (purple dot) replaces JP (green dot) to provide estimations that are nearest to the true values in most situations. And, thus, JPLF becomes the best choice from the perspective of posterior median. CGDSP (blue dot) exhibits similar performances with that of JPLF for the scenarios with σ_α larger than σ (i.e. $\{\sigma_\alpha = 2, \sigma = 0.2\}, \{\sigma_\alpha = 20, \sigma = 2\}$). GDSP (light-blue dot) exhibits good performances for the scenarios with $\{\sigma_\alpha = 2, \sigma = 2\}, \{\sigma_\alpha = 2, \sigma = 20\}$. In other words, CGDSP (blue dot) has good performances when the ratio $\frac{\sigma_\alpha}{\sigma}$ is large while GDSP (light-blue dot) has good performances when the ratio $\frac{\sigma_\alpha}{\sigma}$ starts to decrease. This phenomenon also exists in the Figure 4.1 reporting the averaged posterior mean.

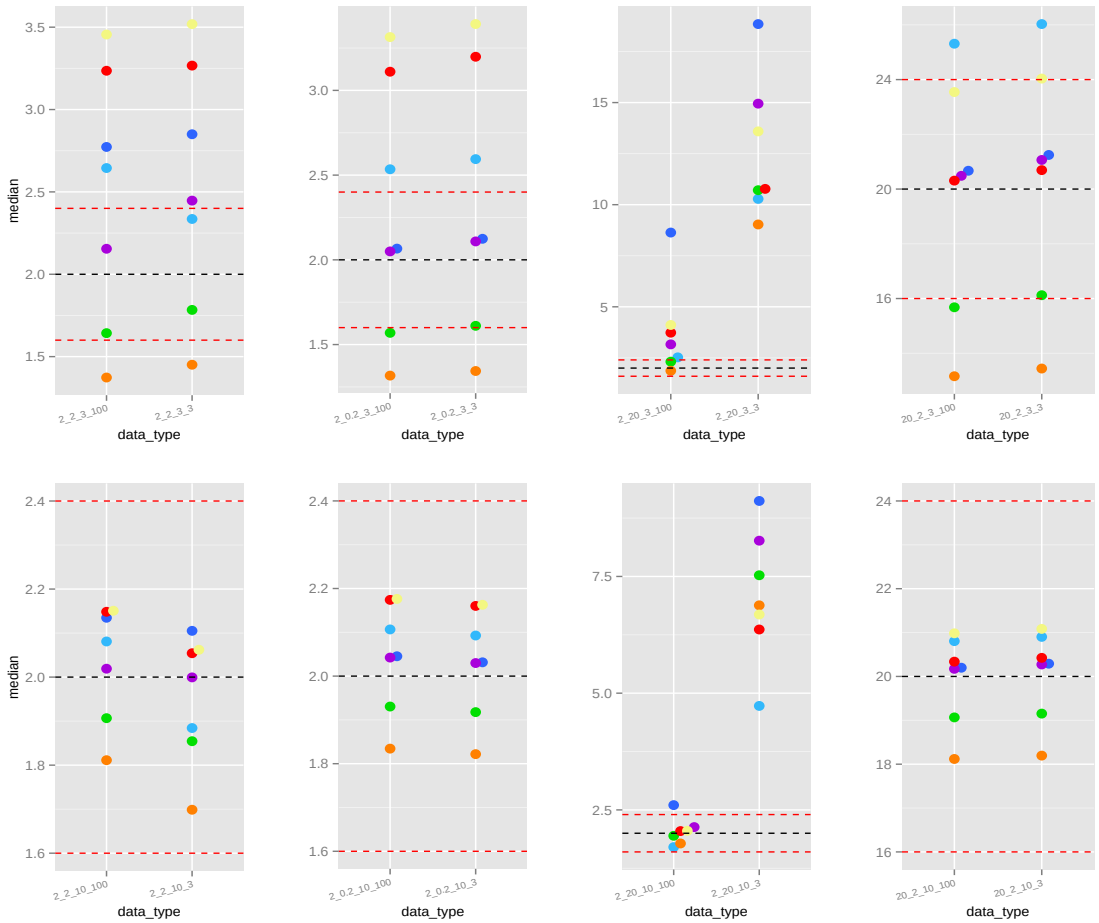


Figure 4.2: Averaged posterior median of σ_α across 1000 data sets for each data type. The horizontal dotted line shows the true value of σ_α

Let us now turn to look at Figure 4.3 showing the percentage of 1000 data sets for each data scenario that the true values of σ_α is included in the 95% credible intervals. Such a percentage is expected to be close to 95% marked by the dotted line. The texts in Figure 4.3 report the priors whose percentages of including the true values in their 95% credible interval are too low to draw on the plots. It can be easily seen that USP has the worst performances amongst the investigated priors. The JP (green dot), which is the best choice when simply looking at the averaged posterior mean in Figure 4.1, does not provide satisfactory results here because it always leads to low percentages. For scenarios with $\{\sigma_\alpha = 2, \sigma = 20\}$, the results are not as satisfactory as those for other data scenarios since most priors fall out of range. For these scenarios, GDSP (light-blue dot) provides the best performances

among all the tested priors. This figure also points out that the larger values m and N take, the better performances the priors could obtain.

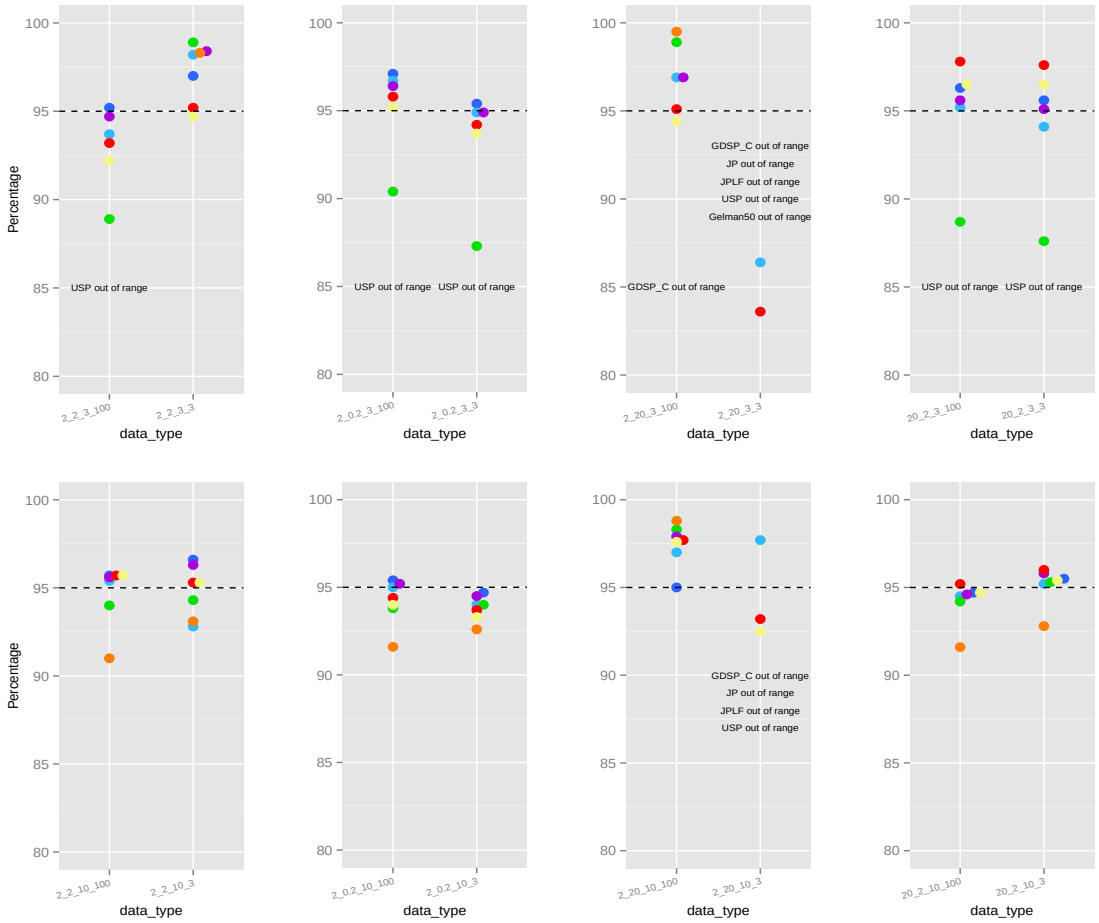


Figure 4.3: Percentage of 1000 data sets for each data type that the true value of σ_α lies in its 95% credible interval.

Again, a 95% HPD interval is explored due to the skewness of posterior distribution for σ_α. Figure 4.4 shows the percentage of 1000 data sets for each data scenario that the true value of σ_α is included in the 95% HPD intervals. This figure shows that JPLF is the best choice since the purple dots are the closest to the dotted line in most scenarios. And the CGDSP (blue dot) and JPLF (purple dot) are close or almost overlap with each other in all the scenarios apart from the ones with $\sigma_\alpha = 2, \sigma = 20$.

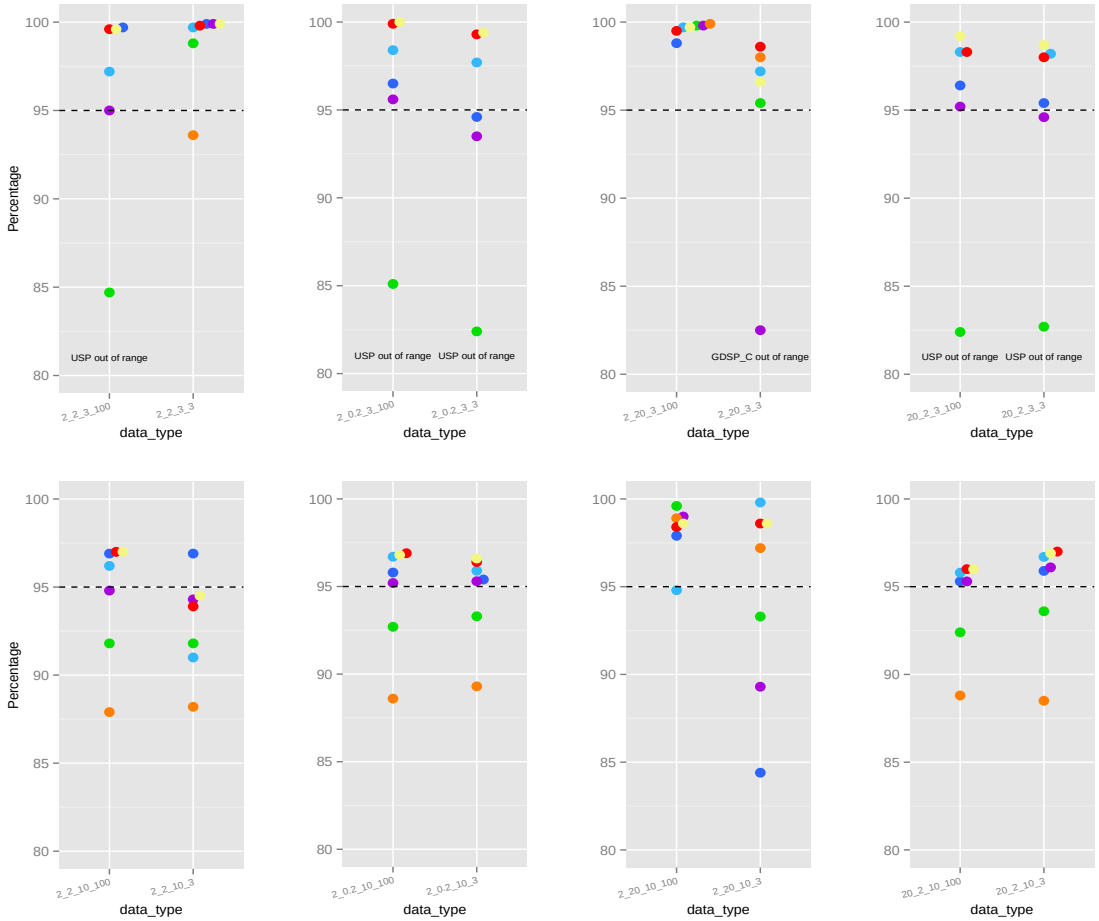


Figure 4.4: Percentage of 1000 data sets for each data type that the true value of σ_α lies in its 95% HPD interval.

As for the prior performances for σ, μ, all tested priors provide reasonable results as illustrated in Figure B.1, B.2, B.3 and B.4. And the differences amongst priors are not significant. One point to note is that both the averaged posterior mean and median for σ shrink quickly towards the true values as N increases from 3 to 100.

Taken altogether, the CGDSP offers acceptable results in scenarios when the ratio $\frac{\sigma_\alpha}{\sigma}$ is large and it displays similar performances with JPLF in such scenarios. The JPLF is the best choice when the ratio $\frac{\sigma_\alpha}{\sigma}$ is around one. For scenarios with small $\frac{\sigma_\alpha}{\sigma}$, most priors do not provide satisfactory behaviour whereas the GDSP has relatively good performances in such scenarios. Although the CGDSP has some inadequacies (violation of the identity of parameter spaces) as stated in the previous section, it does not give absurd results. It is hard to decide which prior always

performes better than others when we claim that there is no prior knowledge at all. In such a situation, it is better to test several priors and compare their performance. In particular, we suggest to test at least the JPLF, CGDSP and GDSP and compare their performances.

Part II

Topics of Bayesian Computations

Chapter 5

Generalised Metropolis-Hastings with Dynamics

Here, we propose a meta-algorithm which we call it 'generalised Metropolis-Hastings with dynamics' to construct Markov chains converging to the desired distributions. It is a class of algorithms that make the transitions by using augmented variables and dynamics. We illustrate that the Markov chains constructed according to this scheme converge to the desired distribution as long as the dynamics are volume-preserving involutions. The ordinary Metropolis-Hastings algorithm can be considered to belong to this class (see section 5.1 for details). With proper designs, the dynamics have the ability to suppress the random-walk behaviour inherent in the ordinary Metropolis-Hastings algorithm to some degree and thereby improve the algorithm efficiency. In particular, Hamiltonian Monte Carlo (HMC) can be considered as a generalised Metropolis-Hastings with dynamics algorithm which tries to avoid the random-walk behaviour and mitigates problems of highly correlated samples by defining the dynamics according to the gradient information of target distributions. This makes HMC easier to have remote proposals and converge quicker than the ordinary Metropolis-Hastings method.

In the simulation of Hamiltonian dynamics, the leap-frog integrator, a numerical method, is used to approximate Hamiltonian trajectories if the exact dynamics cannot be obtained. HMC with both exact dynamics and approximated dynamics can be considered as the generalised Metropolis-Hastings with dynamics algorithms.

To approximately solve the Hamiltonian dynamics, the user must specify two parameters: the number of leap-frog steps l and the step-size value of the leap-frog integrator ε. Neal (2011) discussed the theoretical and practical aspects of HMC.

Selecting a proper step-size value is important as ε controls the approximation errors caused by the leap-frog integrator for the calculated dynamics and thus is influential for the acceptance rate and auto-correlation of simulated samples. The issue of how to tune ε has been attracting much interest in recent years. The frequently used methods usually adapt step-size values based on optimizing an chosen objective measure (such as acceptance probability or first-order autocorrelation) which describes the behavior of an MCMC chain. Therefore, the step-size tuning problem transfers to be an optimization problem with respect to the objective measure. In particular, Hoffman and Gelman (2011) proposed to use the stochastic optimization method; Wang et al. (2013) suggested to use the Bayesian optimization method. Both methods have two disadvantages. Firstly, they require many extra parameters to be set before the start of HMC algorithm in order to control the optimization method and thus are contrary to our original aim of realizing an automatic HMC. Especially, the Bayesian optimization method is achieved through fitting Gaussian process and thus require lots of extra efforts to choose and tune an appropriate Gaussian kernel function. Secondly, these methods are both vanishing adaptions, that is the adaptive power would die out eventually and the step-size would be almost fixed after some point. Therefore, the chain behavior after these points cannot be considered. These vanishing adaptation solves many problems. They do not, however, take into account the situation in which different regions have their own requirements for step-size. In fact, the proper step-size value varies since the stable bound of dynamics, dictated by the local geometric structure, changes throughout the state space. It is not suitable to choose a global step-size value in this situation. In addition, it is usually impossible for us to get access to the information about whether the stable bound of Hamiltonian dynamics is fixed or not when the target distribution is unknown and complicated. Therefore, it is risky to use a fixed step-size value that is tuned only in burn-in iterations based on the acceptance rate.

Based on the step-size mentioned above, we propose an algorithm which exploits

the geometric structure of the log-density for a statistical model to generate step-size stochastically and thus the step-size will automatically adapt to the local structure at each MCMC iteration according to the location of the parameter. The resulting algorithm, that retains the advantages of HMC without the need to set or tune the step-size value, is also a generalised Metropolis-Hastings with dynamics method. We call this algorithm as 'HMC with stochastic step-size' in later chapters.

This chapter is divided into 8 sections. The first section details how to construct the generalised Metropolis-Hastings with dynamics algorithms. In section 5.2, we provide mathematical proof for the generalised Metropolis-Hastings with dynamics algorithms. Sections 5.3 and 5.4 describe the reason that HMC with both exact dynamics and approximated dynamics can be considered as special cases of the generalised Metropolis-Hastings with dynamics method. Section 5.5 and 5.6 contain issues concerning step-size problems of HMC with approximated dynamics and conditions to locally stabilize the approximated dynamical trajectories. In section 5.7, HMC with stochastic step-size is introduced and its performance is provided by an illustrative example. In the final section, some conclusions are drawn.

5.1 General Construction

Here, we firstly give a brief description of the ordinary Metropolis-Hastings algorithm and then introduce the generalised Metropolis-Hastings with dynamics algorithm.

The ordinary Metropolis-Hastings Algorithm

Consider a situation in which the model parameters of interest $\boldsymbol{\theta} \in \mathbb{R}^D$ have probability density function $p(\boldsymbol{\theta})$. The usual approach of the ordinary Metropolis-Hastings algorithm is to start with specifying a probability density function $f(\boldsymbol{\theta}'|\boldsymbol{\theta}^c)$ to draw the proposal sample $\boldsymbol{\theta}'$ conditional on the current state $\boldsymbol{\theta}^c$. The probability of accepting this proposal, usually called as acceptance probability, is given by

$$\min\{1, \frac{p(\boldsymbol{\theta}')f(\boldsymbol{\theta}^c|\boldsymbol{\theta}')}{p(\boldsymbol{\theta}^c)f(\boldsymbol{\theta}'|\boldsymbol{\theta}^c)}\},$$

in order to satisfy the reversibility and thus guarantee the right equilibrium distribution. The ordinary Metropolis-Hastings algorithm is now shown in the following algorithmic form.

Algorithm 1 Ordinary Metropolis-Hastings

1: Given an initial value $\boldsymbol{\theta}^1$;
2: **for** $j = 1, 2, \cdots, n$ **do**
3: Sample $\boldsymbol{\theta}' \sim f(\cdot|\boldsymbol{\theta}^j)$;
4: Let

$$\boldsymbol{\theta}^{j+1} = \begin{cases} \boldsymbol{\theta}', & \text{If Uniform}(0,\,1) \leq \min\{1, \frac{p(\boldsymbol{\theta}')f(\boldsymbol{\theta}^j|\boldsymbol{\theta}')}{p(\boldsymbol{\theta}^j)f(\boldsymbol{\theta}'|\boldsymbol{\theta}^j)}\} \\ \boldsymbol{\theta}^j, & \text{otherwise} \end{cases}.$$

5: **end for**

The Generalised Metropolis-Hastings with Dynamics Algorithm

The approach of the generalised Metropolis-Hastings with dynamics algorithm is to start with introducing auxiliary variables $\boldsymbol{\varphi} \sim g(\cdot|\boldsymbol{\theta}^c)$ conditional on the current state of the Markov chain, where $\boldsymbol{\varphi} \in \mathbb{R}^d$. Note that d, the dimension of $\boldsymbol{\varphi}$, is not necessarily the same as the dimension of the parameters of interest. The joint probability density function, composed by the parameters of interest and the augmented

variables, is

$$p(\boldsymbol{\theta}, \boldsymbol{\varphi}) = p(\boldsymbol{\theta}) g(\boldsymbol{\varphi}|\boldsymbol{\theta}). \tag{5.1.1}$$

By using a dynamic evolution U that satisfies the following two conditions:

- U is volume-preserving;
- U is an involution,

the state $\{\boldsymbol{\theta}^c, \boldsymbol{\varphi}\}$ is transfered to state $\{\boldsymbol{\theta}', \boldsymbol{\varphi}'\}$, i.e.

$$\{\boldsymbol{\theta}', \boldsymbol{\varphi}'\} = U(\{\boldsymbol{\theta}^c, \boldsymbol{\varphi}\}).$$

With probability

$$\min\{1, \frac{p(\boldsymbol{\theta}') g(\boldsymbol{\varphi}'|\boldsymbol{\theta}')}{p(\boldsymbol{\theta}^c) g(\boldsymbol{\varphi}|\boldsymbol{\theta}^c)}\},$$

state $\boldsymbol{\theta}'$ is accepted. It can be seen that simulations for the parameters of interest could be obtained by firstly sampling the joint density in Equation (5.1.1) and then simply ignoring the auxiliary variable $\boldsymbol{\varphi}$. This is because the marginal density of the joint density $p(\boldsymbol{\theta}, \boldsymbol{\varphi})$ is our desired distribution $p(\boldsymbol{\theta})$, i.e.

$$\int p(\boldsymbol{\theta}, \boldsymbol{\varphi}) \mathrm{d}\, \boldsymbol{\varphi} = \int p(\boldsymbol{\theta}) g(\boldsymbol{\varphi}|\boldsymbol{\theta}) \mathrm{d}\, \boldsymbol{\varphi} = p(\boldsymbol{\theta}). \tag{5.1.2}$$

The process of this algorithm is summarized in the following algorithmic form.

Algorithm 2 Generalised Metropolis-Hastings with Dynamics Algorithm

1: Given an initial value $\boldsymbol{\theta}^1$ and a dynamics U that is a volume-preserving involution;
2: **for** $j = 1, 2, \cdots, n$ **do**
3: Generate $\boldsymbol{\varphi} \sim g(\cdot|\boldsymbol{\theta}^j)$;
4: Obtain $\{\boldsymbol{\theta}', \boldsymbol{\varphi}'\} = U(\{\boldsymbol{\theta}^j, \boldsymbol{\varphi}\})$;
5: Let

$$\boldsymbol{\theta}^{j+1} = \begin{cases} \boldsymbol{\theta}', & \text{If Uniform}(0,\, 1) \leq \min\{1, \frac{p(\boldsymbol{\theta}') g(\boldsymbol{\varphi}'|\boldsymbol{\theta}')}{p(\boldsymbol{\theta}^j) g(\boldsymbol{\varphi}|\boldsymbol{\theta}^j)}\} \\ \boldsymbol{\theta}^j, & \text{otherwise} \end{cases}.$$

6: **end for**

Note that generating augmented variables $\boldsymbol{\varphi}$ in the third step of the above algorithm plays the role of introducing the randomness to the transitions as the dynamics of step 4 is fixed. Any dynamics that is a volume-preserving involution can be used in the above algorithm. In particular, the ordinary Metropolis-Hastings algorithm can be considered as a special case of the generalised Metropolis-Hastings with dynamics as summarized in Algorithm 2. This can be seen by considering the proposal $\boldsymbol{\theta}'$ in the third step of Algorithm 1 as the augmented variables $\boldsymbol{\varphi}$ in the third step of Algorithm 2. The dynamic which transfers the state $\{\boldsymbol{\theta}^c, \boldsymbol{\varphi}\}$ to $\{\boldsymbol{\theta}', \boldsymbol{\varphi}'\}$ in step 4 of Algorighm 2 is a swap between $\boldsymbol{\theta}^c$ and $\boldsymbol{\varphi}$, i.e.

$$\{\boldsymbol{\theta}', \boldsymbol{\varphi}'\} = U(\{\boldsymbol{\theta}^c, \boldsymbol{\varphi}\}) = \{\boldsymbol{\varphi}, \boldsymbol{\theta}^c\}.$$

This swapping dynamics reproduces the ordinary Metropolis-Hastings algorithm. And it satisfies the volume-preserving requirement since the Jacobian factor is given by

$$|\det(J)| = \left|\det \begin{pmatrix} \frac{d\boldsymbol{\theta}'}{d\boldsymbol{\theta}^c} & \frac{d\boldsymbol{\theta}'}{d\boldsymbol{\varphi}} \\ \frac{d\boldsymbol{\varphi}'}{d\boldsymbol{\theta}^c} & \frac{d\boldsymbol{\varphi}'}{d\boldsymbol{\varphi}} \end{pmatrix}\right| = \left|\det \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\right| = 1.$$

Also, U is clearly a involution, i.e.

$$U\big(U(\{\boldsymbol{\theta}^c, \boldsymbol{\varphi}\})\big) = \{\boldsymbol{\theta}^c, \boldsymbol{\varphi}\}.$$

5.2 Mathematical Proof

In this section, we give the mathematical proof for the generalised Metropolis-Hastings with dynamics method summarized in Algorithm 2. As indicated by Equation (5.1.2), the joint density of the parameters of interest and the augmented variables takes the desired statistical density $p(\boldsymbol{\theta})$ as its marginal density. Thus, this method can be justified by showing that it constructs a Markov chain converging to the joint probability $p(\boldsymbol{\theta}, \boldsymbol{\varphi})$. Tierney (1998) proposed general Metropolis-Hastings kernels which consider transition kernels with deterministic proposals as a special case. The following is a more detailed restatement proving that the transition kernels with dynamic method summarized in Algorithm 2 converge to the desired distribution.

Proposition 5.2.1 *Suppose that X has density $\pi(\cdot)$ on $\mathcal{X} \subseteq \mathbb{R}^p$ and that U is a continuously differentiable bijection almost everywhere on $\mathcal{X}$. Denote U^{-1} by T. If the transition scheme is*

$$Y|X = x \ = \begin{cases} U(x) & \text{with probability } \alpha(x) \\ x & \text{with probability } 1 - \alpha(x) \end{cases},$$

then we have

$$\pi_Y(y) = \big(1 - \alpha(y)\big)\pi(y) + \alpha\big(T(y)\big)\pi\big(T(y)\big)|T'|. \tag{5.2.3}$$

Proof : The transition scheme can be rewritten as

$$Y = IU(X) + (1 - I)X,$$

where $I|X = x$ follows a Bernoulli distribution with probability $\alpha(x)$. Let $B_r(y)$ denote the ball with radius r centered at y. The probability the variable Y is in $B_r(y)$ becomes

$$\begin{aligned} P(Y \in B_r(y)) &= P\Big(I = 1 \cap U(X) \in B_r(y)\Big) + P\Big(I = 0 \cap X \in B_r(y)\Big) \\ &= P\Big(I = 1 \cap X \in T\big(B_r(y)\big)\Big) + P\Big(I = 0 \cap X \in B_r(y)\Big) \\ &= \int_{T\left(B_r(y)\right)} \alpha(x)\pi(x)\,\mathrm{d}x + \int_{B_r(y)} \big(1 - \alpha(x)\big)\pi(x)\,\mathrm{d}x. \end{aligned} \tag{5.2.4}$$

After a change of variable by the bijection, Equation (5.2.4) becomes

$$P(Y \in B_r(y)) = \int_{B_r(y)} \alpha\big(T(z)\big)\pi\big(T(z)\big)|T'|\,\mathrm{d}z + \int_{B_r(y)} \big(1-\alpha(x)\big)\pi(x)\,\mathrm{d}x.$$

As $r \to 0$, we obtain

$$\begin{aligned}\pi_Y(y) &= \lim_{r\to 0} \frac{P(Y \in B_r(y))}{|B_r(y)|} \\ &= \alpha\big(T(y)\big)\pi\big(T(y)\big)|T'| + \big(1-\alpha(y)\big)\pi(y). \end{aligned} \tag{5.2.5}$$

Proposition 5.2.2 *The transition scheme given in proposition 5.2.1 conserves probability density function, i.e.*

$$\pi_Y(y) = \pi(y), \quad \forall y,$$

if the following conditions are satisfied:

1. *The bijection U is volume-preserving, i.e.*

$$|U'| = 1. \tag{5.2.6}$$

2. *The bijection U is an involution, i.e.*

$$U\big(U(x)\big) = x, \qquad \forall x \in X. \tag{5.2.7}$$

3. *The acceptance probability $\alpha(\cdot)$ is set to*

$$\alpha\big(x\big) = \min\Big\{1, \frac{\pi\big(U(x)\big)}{\pi(x)}\Big\}. \tag{5.2.8}$$

Proof : Since U is an involution as defined in Equation (5.2.7), we obtain

$$U(x) = U^{-1}(x) = T(x).$$

The above indicates that T is also a volume-preserving involution. Since the bijection T preserves volume, Equation (5.2.3) becomes

$$\pi_Y(y) = \big(1-\alpha(y)\big)\pi(y) + \alpha\big(T(y)\big)\pi\big(T(y)\big). \tag{5.2.9}$$

According to the involution condition, the acceptance probability defined in Equation (5.2.8) can be expressed as

$$\alpha(x) = \min\left\{1, \frac{\pi\big(U(x)\big)}{\pi(x)}\right\} = \min\left\{1, \frac{\pi\big(T(x)\big)}{\pi(x)}\right\}. \tag{5.2.10}$$

If $\pi(y) \leq \pi\big(T(y)\big)$, then according to Equation (5.2.10) we have

$$\alpha(y) = 1, \qquad \alpha\big(T(y)\big) = \frac{\pi(y)}{\pi\big(T(y)\big)}.$$

Hence, $\pi_Y(y)$ displayed in Equation (5.2.9) is simplified to

$$\pi_Y(y) = \pi(y).$$

Similarly, if $\pi(y) > \pi\big(T(y)\big)$, then

$$\alpha(y) = \frac{\pi\big(T(y)\big)}{\pi(y)}, \qquad \alpha\big(T(y)\big) = 1.$$

Hence,

$$\pi_Y(y) = \pi(y).$$

. □

The process of Algorithm 2 is justified by the following theorem.

Theorem 5.2.3 *Suppose that $X = (\boldsymbol{\theta}, \boldsymbol{\varphi})$ and its density $\pi(x) = p(\boldsymbol{\theta}, \boldsymbol{\varphi}) = p(\boldsymbol{\theta})g(\boldsymbol{\varphi}|\boldsymbol{\theta})$. If the transition from $x^c = (\boldsymbol{\theta}^c, \boldsymbol{\varphi}^c)$ to y is given by*

$$x^c = (\boldsymbol{\theta}^c, \boldsymbol{\varphi}^c) \xrightarrow[\text{fix } \boldsymbol{\theta}]{\text{generate } \boldsymbol{\varphi}} x = (\boldsymbol{\theta}^c, \boldsymbol{\varphi}) \xrightarrow{\text{by the dynamics } U} y = (\boldsymbol{\theta}', \boldsymbol{\varphi}'), \tag{5.2.11}$$

where U is a volume-preserving involution and y is accepted according to the following rule

$$Y|X = x^c = \begin{cases} (\boldsymbol{\theta}', \boldsymbol{\varphi}') & \text{with probability } \alpha(x) \\ (\boldsymbol{\theta}^c, \boldsymbol{\varphi}) & \text{with probability } 1 - \alpha(x) \end{cases},$$

with

$$\alpha\big(x\big) = \min\left\{1, \frac{\pi\big(U(x)\big)}{\pi(x)}\right\} = \min\{1, \frac{p(\boldsymbol{\theta}')g(\boldsymbol{\varphi}'|\boldsymbol{\theta}')}{p(\boldsymbol{\theta}^c)g(\boldsymbol{\varphi}|\boldsymbol{\theta}^c)}\}, \tag{5.2.12}$$

then the joint density $p(\boldsymbol{\theta}, \boldsymbol{\varphi})$ is conserved, i.e.

$$\pi_Y(y) = \pi(y),$$

and $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$.

Proof : From Equation (5.2.11), the proposal-generating process is made of two steps. Firstly, $x = (\boldsymbol{\theta}^c, \boldsymbol{\varphi})$ is generated. The second step generates the proposal $y = (\boldsymbol{\theta}', \boldsymbol{\varphi}')$ by using a volume-preserving involution U. We have the fact: if $\boldsymbol{\theta}^c$ follows $p(\boldsymbol{\theta})$, then $(\boldsymbol{\theta}^c, \boldsymbol{\varphi})$ follows $p(\boldsymbol{\theta}, \boldsymbol{\varphi})$ by sampling $\boldsymbol{\varphi}$ from $g(\boldsymbol{\varphi}|\boldsymbol{\theta}^c)$. Therefore, the first step of the transition procedure automatically conserves the density function $p(\boldsymbol{\theta}, \boldsymbol{\varphi})$. In order to justify the entire transition procedure in Equation (5.2.11), we only need to show that the second step of transition procedure conserves $p(\boldsymbol{\theta}, \boldsymbol{\varphi})$.

The second step generates the proposal $y = (\boldsymbol{\theta}', \boldsymbol{\varphi}')$ by using a volume-preserving involution U. With probability α in Equation (5.2.12), this proposal is accepted. If it is rejected, the chain stays at state $x = (\boldsymbol{\theta}^c, \boldsymbol{\varphi})$. According to proposition 5.2.2, the second transition with probability α in Equation (5.2.12) also conserves $\pi(\cdot)$.

Therefore, the combination of these two generating steps illustrated in Equation (5.2.11) with acceptance probability α in Equation (5.2.12) conserves the density $\pi(\cdot)$. That is, $p(\boldsymbol{\theta}, \boldsymbol{\varphi})$ is conserved. Therefore, $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$ since the joint density function $p(\boldsymbol{\theta}, \boldsymbol{\varphi})$ takes the desired statistical density $p(\boldsymbol{\theta})$ as its marginal density.

From Theorem 5.2.3, we can see that the generalised Metropolis-Hastings with dynamics method summarized in Algorithm 2 can provide us with a Markov chain having the desired equilibrium distribution. Therefore, the issue changes to seeking appropriate augmented variables and appropriate dynamics to provide efficient simulation results. In the following sections 5.3, 5.4 and 5.7, we will show the generalised Metropolis-Hastings method with three suitable dynamics.

5.3 Exact Hamiltonian Dynamics

Here, we illustrate that the Hamiltonian dynamics can be used as the dynamics in Algorithm 2. We begin by introducing the design of the Hamiltonian system, followed by its properties that are desired in the generalised Metropolis-Hastings with dynamics method.

Consider a situation in which the model parameters of interest $\boldsymbol{\theta} \in \mathbb{R}^D$ have probability density function $p(\boldsymbol{\theta})$. In order to build a Hamiltonian system, an auxiliary variable p_i is introduced for each such model parameter $\theta_i, 1 \leq i \leq D$. These auxiliary variables, called 'momentum' variables, are usually generated from a multivariate Gaussian distribution $N(\mathbf{p}|0, M)$. The joint probability density function composed by the parameters of interest and the 'momentum' variables is $p(\boldsymbol{\theta}, \mathbf{p}) = p(\boldsymbol{\theta})N(\mathbf{p}|0, M)$. In fact, the 'momentum' variables $\mathbf{p}$ act as the augmented variables $\boldsymbol{\varphi}$ in Algorithm 2, i.e.

$$\boldsymbol{\varphi} = \mathbf{p}.$$

So far, the variable augmentation required in the generalised Metropolis-Hastings with dynamics method is achieved. It can be seen that simulations for the parameters of interest could be obtained by firstly sampling the joint density $p(\boldsymbol{\theta}, \mathbf{p})$ and then simply ignoring the auxiliary variable $\mathbf{p}$. The parameters of interest and the augmented 'momentum' variables jointly compose a Hamiltonian system with its energy defined via the negative logarithm of the joint density function

$$H(\boldsymbol{\theta}, \mathbf{p}) = -\log p(\boldsymbol{\theta}, \mathbf{p}) = -L(\boldsymbol{\theta}) + \frac{1}{2}\log\{(2\pi)^D|M|\} + \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p}, \qquad (5.3.13)$$

where $L(\boldsymbol{\theta})$ is the log-density function of the target distribution $p(\boldsymbol{\theta})$. In physics, $\boldsymbol{\theta}$ and $-L(\boldsymbol{\theta})$ are interpreted as 'position' variable and potential energy respectively; $\mathbf{p}$ and $\frac{1}{2}\log\{(2\pi)^D|M|\} + \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p}$ are considered as 'momentum' variables and 'kinetic' energy respectively.

Here, we briefly illustrate how to construct a Markov chain converging to the right joint density function $p(\boldsymbol{\theta}, \mathbf{p})$ according to the Hamiltonian dynamics. We denote the current state of the Markov chain by $\{\boldsymbol{\theta}^c, \mathbf{p}^c\}$. According to the construction of Algorithm 2, 'momentum' variables $\mathbf{p}$ are firstly generated from $N(0, M)$ to form $\{\boldsymbol{\theta}^c, \mathbf{p}\}$. And then, the dynamical transition achieved by the Hamiltonian system is

designed by evolving the Hamiltonian dynamics with respect to dynamical time τ according to Hamiltonian equations,

$$\frac{d\boldsymbol{\theta}}{d\tau} = \frac{\partial H}{\partial \mathbf{p}} = M^{-1}\mathbf{p}, \tag{5.3.14}$$

$$\frac{d\mathbf{p}}{d\tau} = -\frac{\partial H}{\partial \boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}). \tag{5.3.15}$$

We denote the solution for the above differential equations by

$$(\boldsymbol{\theta}(\tau), \mathbf{p}(\tau)) = \Phi_\tau(\boldsymbol{\theta}(0), \mathbf{p}(0)),$$

where the starting point $\{\boldsymbol{\theta}(0), \mathbf{p}(0)\}$ of the dynamic trajectory is set as $\{\boldsymbol{\theta}^c, \mathbf{p}\}$. Note that Φ_τ plays the role of U in Algorithm 2. Therefore, the generalised Metropolis-Hastings with the Hamiltonian dynamics achieves a transition process illustrated as follows,

$$\{\boldsymbol{\theta}^c, \mathbf{p}^c\} \xrightarrow{\mathbf{p}\sim N(0,M)} \{\boldsymbol{\theta}^c, \mathbf{p}\} \xrightarrow[\text{flow}]{\text{Hamiltonian}} \{\boldsymbol{\theta}(\tau), \mathbf{p}(\tau)\}. \tag{5.3.16}$$

The transition of the Markov chain from current state $\{\boldsymbol{\theta}^c, \mathbf{p}^c\}$ to the new state $\{\boldsymbol{\theta}(\tau), \mathbf{p}(\tau)\}$ is achieved by firstly generating the augmented 'momentum' variables and then moving along the dynamic trajectory according to the Hamiltonian differential equations. As previously commented, generating 'momentum' variables in the first step plays the role of introducing the randomness of the transition as the Hamiltonian flow in the second step is determined if the starting point $\{\boldsymbol{\theta}(0), \mathbf{p}(0)\}$ and dynamical time τ are fixed. In the following part of this section, we illustrate that such a flow has appealing properties to satisfy not only the conditions required by the generalised Metropolis-Hastings with dynamics method but also guarantee the acceptance probability to be exactly one.

5.3.1 Energy Preservation

A dynamic flow satisfying Hamiltonian differential equations preserves the total energy of Hamiltonian system, i.e.

$$H\{\boldsymbol{\theta}(\tau), \mathbf{p}(\tau)\} = H\{\boldsymbol{\theta}(0), \mathbf{p}(0)\}.$$

This conservation can be demonstrated by following facts. The change of total energy with respect to dynamical time τ is

$$\frac{\mathrm{d}H}{\mathrm{d}\tau} = \sum_{i=1}^{D} \Big\{ \frac{\partial H}{\partial \theta_i} \frac{d\theta_i}{d\tau} + \frac{\partial H}{\partial p_i} \frac{dp_i}{d\tau} \Big\}.$$

Since the dynamic flow satisfies Equation (5.3.14) and (5.3.15), the above line can be rewritten as

$$\frac{\mathrm{d}H}{\mathrm{d}\tau} = \sum_{i=1}^{D} \Big\{ \frac{\partial H}{\partial \theta_i} \frac{\partial H}{\partial p_i} + \frac{\partial H}{\partial p_i} \Big(- \frac{\partial H}{\partial \theta_i} \Big) \Big\} = 0.$$

Thus, the total energy would be exactly same as long as the dynamic flow adheres to the Hamiltonian differential equations. Since the total energy and the joint probability density are in a one-to-one relationship as shown in Equation (5.3.13), the conserving energy ensures conservation of the probability density, i.e.

$$p\{\boldsymbol{\theta}(\tau), \mathbf{p}(\tau)\} = p\{\boldsymbol{\theta}(0), \mathbf{p}(0)\}.$$

This exact energy preservation ensures the acceptance probability illustrated in Equation (5.2.12) to be exactly one.

5.3.2 Volume Preservation

The dynamic flow actually makes a coordinate transformation from time 0 to time τ. Denote $\big(\theta_i, p_i\big)$ by x_i for each dimension i. The transformation can be rewritten as

$$\Phi_\tau : \Big(x_1(0), \cdots x_D(0) \Big) \to \Big(x_1(\tau), \cdots x_D(\tau) \Big).$$

The Jacobian matrix of such a transformation is

$$J\Big(x(\tau); x(0) \Big) = \frac{\partial \Big(x_1(\tau), \cdots, x_D(\tau) \Big)}{\partial \Big(x_1(0), \cdots, x_D(0) \Big)}, \tag{5.3.17}$$

with its elements denoted by

$$J_{ij} = \frac{\partial x_i(\tau)}{\partial x_j(0)}. \tag{5.3.18}$$

The Jacobian factor is

$$\det(J) = \exp\Big(\mathrm{tr}\big(\log J \big) \Big).$$

Taking the derivative of the Jacobian factor with respect to time τ, we obtain

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tau}\det(J) &= \exp\Big(\mathrm{tr}\big(\log J\big)\Big)\cdot \mathrm{tr}\big(J^{-1}\frac{\mathrm{d}J}{\mathrm{d}\tau}\big)\\ &= \det(J)\sum_{i=1}^{D}\sum_{j=1}^{D}\Big\{J_{ij}^{-1}\frac{\mathrm{d}J_{ji}}{\mathrm{d}\tau}\Big\}.\end{aligned} \tag{5.3.19}$$

According to Equations (5.3.17) and (5.3.18),

$$J_{ij}^{-1} = \frac{\partial x_i(0)}{\partial x_j(\tau)}, \tag{5.3.20}$$

$$\frac{\mathrm{d}J_{ji}}{\mathrm{d}\tau} = \frac{\partial \dot{x}_j(\tau)}{\partial x_i(0)}, \tag{5.3.21}$$

where $\dot{x}$ stands for the first derivative of the state with respect to time τ, i.e.

$$\dot{x}(\tau) = \Big(\dot{x}_1(\tau),\cdots,\dot{x}_i(\tau),\cdots,\dot{x}_D(\tau)\Big),$$

with

$$\dot{x}_i(\tau) = \Big(\frac{\mathrm{d}\theta_i(\tau)}{\mathrm{d}\tau},\frac{\mathrm{d}p_i(\tau)}{\mathrm{d}\tau}\Big) = \Big(\frac{\partial H}{\partial p_i(\tau)},-\frac{\partial H}{\partial \theta_i(\tau)}\Big). \tag{5.3.22}$$

Substituting Equation (5.3.20) and (5.3.21) into (5.3.19),

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tau}\det(J) &= \det(J)\sum_{i=1}^{D}\sum_{j=1}^{D}\Big\{\frac{\partial x_i(0)}{\partial x_j(\tau)}\frac{\partial \dot{x}_j(\tau)}{\partial x_i(0)}\Big\}\\ &= \det(J)\sum_{i=1}^{D}\sum_{j=1}^{D}\sum_{k=1}^{D}\Big\{\frac{\partial x_i(0)}{\partial x_j(\tau)}\frac{\partial \dot{x}_j(\tau)}{\partial x_k(\tau)}\frac{\partial x_k(\tau)}{\partial x_i(0)}\Big\} \qquad (5.3.23)\\ &= \det(J)\sum_{j=1}^{D}\sum_{k=1}^{D}\Big\{\frac{\partial \dot{x}_j(\tau)}{\partial x_k(\tau)}\sum_{i=1}^{D}\Big(\frac{\partial x_i(0)}{\partial x_j(\tau)}\frac{\partial x_k(\tau)}{\partial x_i(0)}\Big)\Big\}\\ &= D\cdot\det(J)\sum_{j=1}^{D}\sum_{k=1}^{D}\Big\{\frac{\partial \dot{x}_j(\tau)}{\partial x_k(\tau)}\delta_{kj}\Big\}\\ &= D\cdot\det(J)\sum_{j=1}^{D}\Big\{\frac{\partial \dot{x}_j(\tau)}{\partial x_j(\tau)}\Big\} \qquad (5.3.24)\end{aligned}$$

where the chain rule is inserted into the line (5.3.23); and δ_{kj} stands for the delta function, i.e.

$$\delta_{kj} = \begin{cases} 1 & \text{if } k=j \\ 0 & \text{otherwise}\end{cases}.$$

By substituting Equation (5.3.22) into (5.3.24), we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tau}\det(J) &= D\cdot\det(J)\sum_{j=1}^{D}\left\{\frac{\partial}{\partial\theta_j(\tau)}\Big(\frac{\mathrm{d}\theta_j(\tau)}{\mathrm{d}\tau}\Big)+\frac{\partial}{\partial p_j(\tau)}\Big(\frac{\mathrm{d}p_j(\tau)}{\mathrm{d}\tau}\Big)\right\}\\ &= D\cdot\det(J)\sum_{j=1}^{D}\left\{\frac{\partial}{\partial\theta_j(\tau)}\Big(\frac{\partial H}{\partial p_j(\tau)}\Big)+\frac{\partial}{\partial p_j(\tau)}\Big(-\frac{\partial H}{\partial\theta_i(\tau)}\Big)\right\}=0.\end{aligned} \tag{5.3.25}$$

The above equation indicates the fact that the Jacobian factor does not change along the Hamiltonian dynamic flow. In addition, the initial value of the Jacobian factor is

$$\det\Big(J\big(x(0);x(0)\big)\Big)=\left|\frac{\partial\Big(x_1(0),\cdots,x_D(0)\Big)}{\partial\Big(x_1(0),\cdots,x_D(0)\Big)}\right|=1.$$

Combining the above equation with the assertion in Equation (5.3.25), we have

$$\det\Big(J\big(x(\tau);x(0)\big)\Big)=1.$$

Therefore, along the dynamic flow, the volume element is preserved.

5.3.3 Involution

Recall that the Hamiltonian dynamics in Equation (5.3.14) and (5.3.15) is given by the evolution operator Φ_τ,

$$\Phi_\tau:\{\boldsymbol{\theta}(0),\mathbf{p}(0)\}\xrightarrow[\text{flow}]{\text{Hamiltonian}}\{\boldsymbol{\theta}(\tau),\mathbf{p}(\tau)\}.$$

In order to illustrate the involution property, we could add an extra step—-changing sign of 'momentum' variables——to the Hamiltonian dynamics. More specifically, the dynamics U can be considered as evolving the Hamiltonian dynamics with the sign changed 'momentum' variables. Let us denote

$$\{\boldsymbol{\theta}^\star,\mathbf{p}^\star\}=R(\{\boldsymbol{\theta},\mathbf{p}\})=\{\boldsymbol{\theta},-\mathbf{p}\},$$

where R denotes the transformation of changing sign of 'momentum' variables. The 'momentum' variables are generated from Gaussian distribution $N(0,M)$ which is a symmetrical distribution about 0 and thus changing the sign of $\mathbf{p}$ would not bring changes to the total energy H. The sign change of $\mathbf{p}$ also does not have any influence

on the volume-preserving property. According to the Hamiltonian equations (5.3.14) and (5.3.15) for $\{\boldsymbol{\theta}, \mathbf{p}\}$, we have the following differential equations for $\{\boldsymbol{\theta}^\star, \mathbf{p}^\star\}$,

$$\begin{cases} \frac{d\boldsymbol{\theta}^\star}{d\tau} = \frac{d\boldsymbol{\theta}}{d\tau} = \frac{\partial H}{\partial \mathbf{p}} = -\frac{\partial H}{\partial \mathbf{p}^\star} \\ \frac{d\mathbf{p}^\star}{d\tau} = -\frac{d\mathbf{p}}{d\tau} = \frac{\partial H}{\partial \boldsymbol{\theta}} = \frac{\partial H}{\partial \boldsymbol{\theta}^\star} \end{cases} \Longrightarrow \begin{cases} \frac{d\boldsymbol{\theta}^\star}{d\tau} = -\frac{\partial H}{\partial \mathbf{p}^\star} \\ \frac{d\mathbf{p}^\star}{d\tau} = \frac{\partial H}{\partial \boldsymbol{\theta}^\star} \end{cases}$$

That is, the dynamics U is defined as the above differential equation for $\{\boldsymbol{\theta}^\star, \mathbf{p}^\star\}$. In terms of evolution operator, this dynamics implies (Lamb and Roberts, 1998),

$$U = \Phi_\tau^{-1} \circ R = R \circ \Phi_\tau$$

where $\circ$ denotes function composition. Therefore, U is an involution since

$$\begin{aligned} U \circ U &= (R \circ \Phi_\tau) \circ (\Phi_\tau^{-1} \circ R) \\ &= R \circ \Phi_\tau \circ \Phi_\tau^{-1} \circ R \\ &= R \circ R = I. \end{aligned}$$

where I represents the identity function.

According to theorem 5.2.3, we can conclude that the exact Hamiltonian dynamics can be used to construct a Markov chain converging to the desired statistical distribution.

5.4 Approximated Hamiltonian Dynamics

Usually, the Hamiltonian equations in Equations (5.3.14) and (5.3.15) cannot be solved analytically and thus we cannot use the exact Hamiltonian dynamics as the dynamics required in Algorithm 2 to construct a Markov chain. Therefore, a suitable numerical method is required to approximate the dynamical flows. As long as the approximated dynamics provided by the chosen numerical method can satisfy the conditions of volume preservation and involution, it can be used in Algorithm 2 to construct a Markov chain converging the desired distribution. The leap-frog integrator, which is often successfully used to approximate Hamiltonian trajectories, is reviewed as follows,

$$\mathbf{p}(\tau + \frac{\varepsilon}{2}) = \mathbf{p}(\tau) + (\frac{\varepsilon}{2})\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)}, \tag{5.4.26}$$

$$\boldsymbol{\theta}(\tau + \varepsilon) = \boldsymbol{\theta}(\tau) + \varepsilon M^{-1}\mathbf{p}(\tau + \frac{\varepsilon}{2}), \tag{5.4.27}$$

$$\mathbf{p}(\tau + \varepsilon) = \mathbf{p}(\tau + \frac{\varepsilon}{2}) + (\frac{\varepsilon}{2})\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau+\varepsilon)}, \tag{5.4.28}$$

where $\big(\boldsymbol{\theta}(\tau), \mathbf{p}(\tau)\big)$ is the current state of the Hamiltonian trajectory, $\big(\boldsymbol{\theta}(\tau+\varepsilon), \mathbf{p}(\tau+\varepsilon)\big)$ is the next state of the trajectory given by the leap-frog integrator and ε is the step-size. Consecutively applying the leap-frog integrator provides us with approximate trajectory paths. In fact, this is the so called Hamiltonian Monte Carlo that is firstly introduced by Duane et al. (1987) and popularized by Neal (2011) in statistics field.

In the following part, we will illustrate that the approximated Hamiltonian dynamics given by the leap-frog integrator is a volume-preserving involution. In addition, the approximation errors of the leap-frog integrator is investigated since the approximated dynamics, unlike the exact Hamiltonian dynamics, introduce errors when calculating the total energies.

5.4.1 Volume Preservation

It is straightforward to verify the volume preservation property since the transformation carried out by the leap-frog integrator can be considered as a composite of

three shear mappings as illustrated in Equations (5.4.26) to (5.4.28), i.e.

$$\begin{aligned}
&\text{by Equation (5.4.26) ,} && \{\boldsymbol{\theta}(\tau), \mathbf{p}(\tau)\} \to \{\boldsymbol{\theta}(\tau), \mathbf{p}(\tau + \frac{\varepsilon}{2})\} \\
&\text{by Equation (5.4.27) ,} && \{\boldsymbol{\theta}(\tau), \mathbf{p}(\tau + \frac{\varepsilon}{2})\} \to \{\boldsymbol{\theta}(\tau + \varepsilon), \mathbf{p}(\tau + \frac{\varepsilon}{2})\} \\
&\text{by Equation (5.4.28) ,} && \{\boldsymbol{\theta}(\tau + \varepsilon), \mathbf{p}(\tau + \frac{\varepsilon}{2})\} \to \{\boldsymbol{\theta}(\tau + \varepsilon), \mathbf{p}(\tau + \varepsilon)\}
\end{aligned}$$

The above three transformations are shear mappings according to the definition of shear mapping as follows

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} x + g(y) \\ y \end{pmatrix}$$

with Jacobian matrix

$$\begin{pmatrix} 1 & g'(y) \\ 0 & 1 \end{pmatrix}.$$

Since the Jacobian factor of a shear mapping is 1, the Jacobian factor of the transformation from $\{\boldsymbol{\theta}(\tau), \mathbf{p}(\tau)\} \to \{\boldsymbol{\theta}(\tau + \varepsilon), \mathbf{p}(\tau + \varepsilon)\}$ is also 1 as it is the product of the Jacobian factors of three shear mappings. Therefore, the approximated dynamics provided by the leap-frog integrator is volume-preserving.

5.4.2 Involution

We denote the mapping constructed by the leap-frog integrator as LF_ε. In order to illustrate that the involution requirement is satisfied, we add an extra step—-changing the sign of 'momentum' variables—-to the leap-frog dynamics. That is, the dynamics U required in Algorithm 2 is given by

$$U : \{\boldsymbol{\theta}(\boldsymbol{\tau}), \mathbf{p}(\boldsymbol{\tau})\} \xrightarrow[\text{sign}]{\text{change}} \{\boldsymbol{\theta}(\boldsymbol{\tau}), -\mathbf{p}(\boldsymbol{\tau})\} \xrightarrow{LF_\varepsilon} \{\boldsymbol{\theta}(\boldsymbol{\tau} + \boldsymbol{\varepsilon}), \mathbf{p}(\boldsymbol{\tau} + \boldsymbol{\varepsilon})\}. \quad (5.4.29)$$

Note that the total energy H is not influenced by the sign changes since the 'momentum' variables are generated from a Gaussian distribution that is symmetrical about 0. The sign change also preserves the volume and thereby U is also volume-preserving. According to Equations (5.4.26) to (5.4.28), $\boldsymbol{\theta}(\tau + \varepsilon)$ and $\mathbf{p}(\tau + \varepsilon)$ can

be expressed as

$$\boldsymbol{\theta}(\tau+\varepsilon) = \boldsymbol{\theta}(\tau) + \varepsilon M^{-1}\Big(-\mathbf{p}(\tau) + \frac{1}{2}\varepsilon\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)}\Big), \tag{5.4.30}$$

$$\mathbf{p}(\tau+\varepsilon) = -\mathbf{p}(\tau) + \frac{1}{2}\varepsilon\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} + \frac{1}{2}\varepsilon\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau+\varepsilon)}. \tag{5.4.31}$$

Suppose that the trajectory is now started from $\{\boldsymbol{\theta}(\boldsymbol{\tau}+\boldsymbol{\varepsilon}), \mathbf{p}(\boldsymbol{\tau}+\boldsymbol{\varepsilon})\}$. By applying the transformation U, that is composed of a sign change for 'momentum' variables and the leap-frog mapping LF_ε, to $\{\boldsymbol{\theta}(\boldsymbol{\tau}+\boldsymbol{\varepsilon}), \mathbf{p}(\boldsymbol{\tau}+\boldsymbol{\varepsilon})\}$, we obtain

$$\{\boldsymbol{\theta}(\boldsymbol{\tau}+\boldsymbol{\varepsilon}), \mathbf{p}(\boldsymbol{\tau}+\boldsymbol{\varepsilon})\} \xrightarrow[\text{sign}]{\text{change}} \{\boldsymbol{\theta}(\boldsymbol{\tau}+\boldsymbol{\varepsilon}), -\mathbf{p}(\boldsymbol{\tau}+\boldsymbol{\varepsilon})\} \xrightarrow{LF_\varepsilon} \{A, B\},$$

where

$$A = \boldsymbol{\theta}(\tau+\varepsilon) + \varepsilon M^{-1}\Big(-\mathbf{p}(\tau+\varepsilon) + \frac{1}{2}\varepsilon\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau+\varepsilon)}\Big),$$
$$B = -\mathbf{p}(\tau+\varepsilon) + \frac{1}{2}\varepsilon\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau+\varepsilon)} + \frac{1}{2}\varepsilon\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=A}.$$

By substituting Equation (5.4.30) and (5.4.31) into A and B,

$$\begin{aligned} A &= \boldsymbol{\theta}(\tau) + \varepsilon M^{-1}\Big(\mathbf{p}(\tau) + \frac{1}{2}\varepsilon\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)}\Big) \\ &\quad + \varepsilon M^{-1}\Big(-\mathbf{p}(\tau) - \frac{1}{2}\varepsilon\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} - \frac{1}{2}\varepsilon\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau+\varepsilon)} + \frac{1}{2}\varepsilon\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau+\varepsilon)}\Big) \\ &= \boldsymbol{\theta}(\tau), \end{aligned} \tag{5.4.32}$$

$$\begin{aligned} B &= \mathbf{p}(\tau) - \frac{1}{2}\varepsilon\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} - \frac{1}{2}\varepsilon\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau+\varepsilon)} + \frac{1}{2}\varepsilon\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau+\varepsilon)} + \frac{1}{2}\varepsilon\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=A} \\ &= \mathbf{p}(\tau) - \frac{1}{2}\varepsilon\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} + \frac{1}{2}\varepsilon\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} \qquad \text{By Equation (5.4.32)} \\ &= \mathbf{p}(\tau). \end{aligned}$$

That is, the dynamics U is an involution. Therefore, the approximated Hamiltonian dynamics provided by the leap-frog integrator can be used to construct a Markov chain converging to the desired distribution according to Theorem 5.2.3. According to Equation (5.2.12), the acceptance probability is not exactly one and is given by

$$\min\{1, \frac{\exp(-H(\boldsymbol{\theta}(\tau), \mathbf{p}(\tau)))}{\exp(-H(\boldsymbol{\theta}^c, \mathbf{p}))}\}.$$

Clearly, the smaller the approximation error is, the higher the acceptance probability is. As the generalised Metropolis-Hastings with the approximated Hamiltoninan dynamics is the usually used HMC, we will refer it as HMC in the following parts. It is summarized in the following algorithmic form.

Algorithm 3 Hamiltonian Monte Carlo

1: Given an initial value $\boldsymbol{\theta}^1$ and values for ε, l;
2: **for** $j = 1, 2, \cdots, n$ **do**
3: Sample $\mathbf{p} \sim N(0, M)$
4: Set $\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta}^j, \mathbf{p}' \leftarrow \mathbf{p}$
5: **for** $i = 1$ **to** l **do**
6: Set $\boldsymbol{\theta}', \mathbf{p}' \leftarrow \text{Leapfrog}(\boldsymbol{\theta}', \mathbf{p}', \varepsilon)$
7: **end for**
8: Let

$$\boldsymbol{\theta}^{j+1} = \begin{cases} \boldsymbol{\theta}', & \text{If Uniform}(0,\ 1) \leq \min\{1, \frac{\exp(-H(\boldsymbol{\theta}', \mathbf{p}'))}{\exp(-H(\boldsymbol{\theta}^j, \mathbf{p}))}\} \\ \boldsymbol{\theta}^j, & \text{otherwise} \end{cases}.$$

9: **end for**
10: **Function** Leapfrog $\{\boldsymbol{\theta}, \mathbf{p}, \varepsilon\}$
11: Set $\mathbf{p}' \leftarrow \mathbf{p} + (\frac{\varepsilon}{2})\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$
12: Set $\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} + \varepsilon M^{-1}\mathbf{p}$
13: Set $\mathbf{p}' \leftarrow \mathbf{p}' + (\frac{\varepsilon}{2})\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}')$
14: **Return** $\boldsymbol{\theta}', \mathbf{p}'$

5.4.3 Approximately Conserving Energy

Any numerical method will introduce approximation errors and thus the energy could not be conserved exactly by the leap-frog integrator. The approximation errors of a numerical method are the differences between numerical solutions and exact solutions. Here, the local error of the leapfrog integrator, which turns out to be $\mathcal{O}(\varepsilon^3)$, is illustrated. We denote the exact solutions at time $\tau + \varepsilon$ by $\boldsymbol{\theta}(\tau + \varepsilon)$ and $\mathbf{p}(\tau + \varepsilon)$; and we denote the numerical solutions by $\tilde{\boldsymbol{\theta}}(\tau + \varepsilon)$ and $\tilde{\mathbf{p}}(\tau + \varepsilon)$. For the sake of simplicity, the dimensionality of both $\boldsymbol{\theta}$ and $\mathbf{p}$ are set to one. The exact solution of Hamiltonian equations at time $\tau + \varepsilon$ is expressed by applying a Taylor

expansion as follows,

$$\boldsymbol{\theta}(\tau+\varepsilon) = \boldsymbol{\theta}(\tau) + \varepsilon M^{-1}\mathbf{p}(\tau) + \frac{1}{2}\varepsilon^2 M^{-1}\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} + \frac{1}{3!}\varepsilon^3 M^{-1}\frac{\partial^2 L}{\partial \boldsymbol{\theta}^2}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)}\frac{\partial H}{\partial \mathbf{p}}|_{\mathbf{p}=\mathbf{p}(\tau)} + \mathcal{O}(\varepsilon^4), \tag{5.4.33}$$

$$\mathbf{p}(\tau+\varepsilon) = \mathbf{p}(\tau) + \varepsilon\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} + \frac{1}{2}\varepsilon^2 M^{-1}\mathbf{p}(\tau)\frac{\partial^2 L}{\partial \boldsymbol{\theta}^2}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} + \frac{1}{3!}\varepsilon^3\left(M^{-1}\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)}\frac{\partial^2 L}{\partial \boldsymbol{\theta}^2}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} + \Big(M^{-1}\mathbf{p}(\tau)\Big)^2\frac{\partial^3 L}{\partial \boldsymbol{\theta}^3}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)}\right) + \mathcal{O}(\varepsilon^4). \tag{5.4.34}$$

Recall the leap-frog integrator expressed in Equation (5.4.30) and (5.4.31), the numerical solutions provided by the leap-frog integrator could be written as

$$\tilde{\boldsymbol{\theta}}(\tau+\varepsilon) = \boldsymbol{\theta}(\tau) + \varepsilon M^{-1}\mathbf{p}(\tau) + \frac{1}{2}\varepsilon^2 M^{-1}\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)}, \tag{5.4.35}$$

$$\tilde{\mathbf{p}}(\tau+\varepsilon) = \mathbf{p}(\tau) + \frac{1}{2}\varepsilon\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} + \frac{1}{2}\varepsilon\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}(\tau+\varepsilon)}. \tag{5.4.36}$$

Expand $\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}(\tau+\varepsilon)}$ at $\boldsymbol{\theta}(\tau)$ by Taylor expansion,

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}(\tau+\varepsilon)} &= \frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} + \Big(\tilde{\boldsymbol{\theta}}(\tau+\varepsilon) - \boldsymbol{\theta}(\tau)\Big)\frac{\partial^2 L}{\partial \boldsymbol{\theta}^2}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} \\
&\quad + \frac{1}{2}\Big(\tilde{\boldsymbol{\theta}}(\tau+\varepsilon) - \boldsymbol{\theta}(\tau)\Big)^2\frac{\partial^3 L}{\partial \boldsymbol{\theta}^3}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} + \mathcal{O}(\varepsilon^3) \\
&= \frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} + \Big(\varepsilon M^{-1}\mathbf{p}(\tau) + \frac{1}{2}\varepsilon^2 M^{-1}\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)}\Big)\frac{\partial^2 L}{\partial \boldsymbol{\theta}^2}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} \\
&\quad + \frac{1}{2}\Big(\varepsilon M^{-1}\mathbf{p}(\tau) + \frac{1}{2}\varepsilon^2 M^{-1}\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)}\Big)^2\frac{\partial^3 L}{\partial \boldsymbol{\theta}^3}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} + \mathcal{O}(\varepsilon^3) \\
&= \frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} + \varepsilon M^{-1}\mathbf{p}(\tau)\frac{\partial^2 L}{\partial \boldsymbol{\theta}^2}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} \\
&\quad + \frac{1}{2}\varepsilon^2\left(M^{-1}\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)}\frac{\partial^2 L}{\partial \boldsymbol{\theta}^2}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} + \big(M^{-1}\mathbf{p}(\tau)\big)^2\frac{\partial^3 L}{\partial \boldsymbol{\theta}^3}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)}\right) + \mathcal{O}(\varepsilon^3).
\end{aligned} \tag{5.4.37}$$

By substituting Equation (5.4.37) into (5.4.36), we have

$$\tilde{\mathbf{p}}(\tau+\varepsilon) = \mathbf{p}(\tau) + \varepsilon\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} + \frac{1}{2}\varepsilon^2 M^{-1}\mathbf{p}(\tau)\frac{\partial^2 L}{\partial \boldsymbol{\theta}^2}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} + \frac{1}{4}\varepsilon^3\left(M^{-1}\frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)}\frac{\partial^2 L}{\partial \boldsymbol{\theta}^2}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} + \big(M^{-1}\mathbf{p}(\tau)\big)^2\frac{\partial^3 L}{\partial \boldsymbol{\theta}^3}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)}\right) + \mathcal{O}(\varepsilon^4). \tag{5.4.38}$$

Denote the error for $\boldsymbol{\theta}$ by $\text{Err}(\boldsymbol{\theta})$. By comparing Equation (5.4.35) with (5.4.33), we have

$$\begin{aligned}\text{Err}(\boldsymbol{\theta}) &= \boldsymbol{\theta}(\tau+\varepsilon) - \tilde{\boldsymbol{\theta}}(\tau+\varepsilon) \\ &= \frac{1}{3!}\varepsilon^3 (M^{-1})^2 \boldsymbol{p}(\tau) \frac{\partial^2 L}{\partial \boldsymbol{\theta}^2}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} + \mathcal{O}(\varepsilon^4). \end{aligned} \tag{5.4.39}$$

This shows that for small ε, the error for $\boldsymbol{\theta}$ is approximately proportional to ε^3 and is controlled by term $(M^{-1})^2 \boldsymbol{p}(\tau) \frac{\partial^2 L}{\partial \boldsymbol{\theta}^2}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)}$ that is related to the curvature information of the desired density. It is important to ensure $\text{Err}(\boldsymbol{\theta})$ to be moderately small so that the Markov chain would not go to extreme regions. Similarly, the error for $\mathbf{p}$ could be written as

$$\begin{aligned}\text{Err}(\mathbf{p}) &= \mathbf{p}(\tau+\varepsilon) - \tilde{\mathbf{p}}(\tau+\varepsilon) \\ &= -\frac{1}{12}\varepsilon^3 \bigg(M^{-1} \frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} \frac{\partial^2 L}{\partial \boldsymbol{\theta}^2}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} + \big(M^{-1}\mathbf{p}(\tau)\big)^2 \frac{\partial^3 L}{\partial \boldsymbol{\theta}^3}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} \bigg) + \mathcal{O}(\varepsilon^4). \end{aligned}$$

The above equation indicates that the accuracy of $\mathbf{p}$ is related to the first derivative, the second derivative and even the third derivative of the target log-density function at the current state of the approximated dynamical trajectory. This approximation error would have an influence on the accuracy of the total energy of the Hamiltonian system. We now turn to the corresponding error in the total energy of the Hamiltonian system caused by the leap-frog integrator. By using a Taylor expansion, we could express $H\Big(\tilde{\boldsymbol{\theta}}(\tau+\varepsilon), \tilde{\mathbf{p}}(\tau+\varepsilon)\Big)$ at the point $\Big(\boldsymbol{\theta}(\tau+\varepsilon), \mathbf{p}(\tau+\varepsilon)\Big)$ as follows,

$$\begin{aligned} & H\Big(\tilde{\boldsymbol{\theta}}(\tau+\varepsilon), \tilde{\mathbf{p}}(\tau+\varepsilon)\Big) \\ &= H\Big(\boldsymbol{\theta}(\tau+\varepsilon), \mathbf{p}(\tau+\varepsilon)\Big) \\ &\quad + \varepsilon^3 \Big(\frac{1}{4}(M^{-1})^2 \mathbf{p}(\tau) \big(\frac{\partial^2 L}{\partial \boldsymbol{\theta}^2} \frac{\partial L}{\partial \boldsymbol{\theta}} \big)|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} + \frac{1}{12} \big(M^{-1}\mathbf{p}(\tau)\big)^3 \frac{\partial^3 L}{\partial \boldsymbol{\theta}^3}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} \Big) + \mathcal{O}(\varepsilon^4). \end{aligned}$$

We denote the approximation error for the energy of the Hamiltonian system by $\text{Err}(H)$ and obtain

$$\begin{aligned}\text{Err}(H) &= H\Big(\tilde{\boldsymbol{\theta}}(\tau+\varepsilon), \tilde{\mathbf{p}}(\tau+\varepsilon)\Big) - H\Big(\boldsymbol{\theta}(\tau+\varepsilon), \mathbf{p}(\tau+\varepsilon)\Big) \\ &= \varepsilon^3 \Big(\frac{1}{4}(M^{-1})^2 \mathbf{p}(\tau) \big(\frac{\partial^2 L}{\partial \boldsymbol{\theta}^2} \frac{\partial L}{\partial \boldsymbol{\theta}} \big)|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} + \frac{1}{12} \big(M^{-1}\mathbf{p}(\tau)\big)^3 \frac{\partial^3 L}{\partial \boldsymbol{\theta}^3}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)} \Big) + \mathcal{O}(\varepsilon^4). \end{aligned} \tag{5.4.40}$$

The above equation illustrates that the local error for $H(\boldsymbol{\theta}, \mathbf{p})$ has order ε^3. And this error influenced by $\text{Err}(\mathbf{p})$ is also related to the first derivative, the second derivative and the third derivative of the target log-density function at the current state of the approximated dynamical trajectory. Clearly, the closer to zero the error $\text{Err}(H)$ is, the higher the acceptance probability is. However, moderate size of this local error in Hamiltonian energy would be acceptable.

5.4.4 Example

Here, we use a toy example to illustrate the performance of HMC compared with the ordinary Metropolis algorithm. Considering the following 'banana' example

$$y_i \overset{\text{i.i.d.}}{\sim} N\left(\theta_1 + \theta_2^2, \sigma_y^2\right) \qquad i = 1, \ldots, N$$

with prior distribution for θ_1 and θ_2 chosen as

$$\boldsymbol{\theta} \sim N\left(0, \sigma_\theta^2 I\right)$$

where σ_θ and σ_y are fixed as 1 and 2 respectively. The data $\{y_i; i = 1, \cdots, 100\}$ are simulated from the above model with specified parameter values. The mean and the standard deviation of the simulated data were 1.26 and 2.16 respectively. The corresponding posterior density contour is displayed in Figure 5.1.

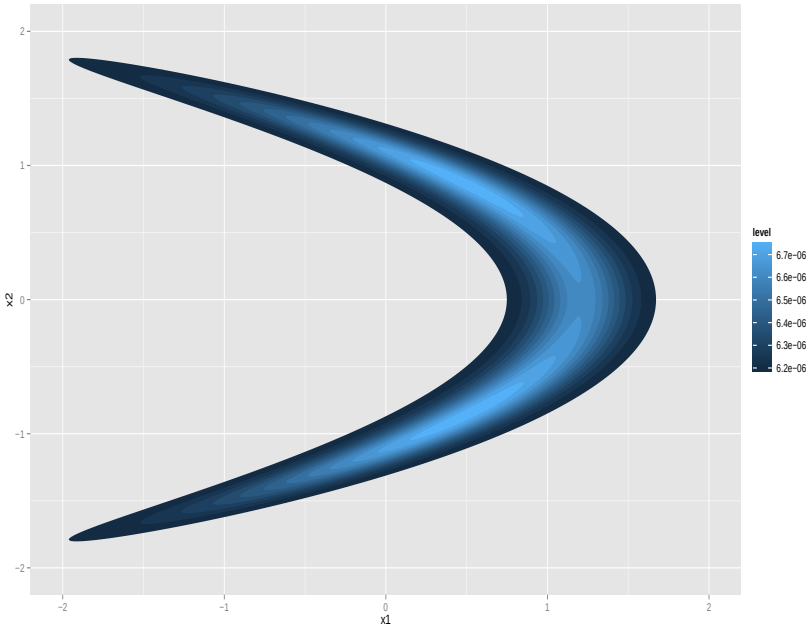


Figure 5.1: Target density contour of 'Banana example'

Both HMC and RWMH (random walk Metropolis-Hastings algorithm) are used to sample the posterior distribution for this model. Particularly, algorithm parameters required by HMC is specified as follows,

$$\{\varepsilon = 0.1,\ l = 4,\ M = I\}$$

For the RWMH method, Laplace approximation is often used to initiate the variance matrix of proposal density. Moreover, a parameter that is used to scale this variance matrix is tuned in the burn-in iterations according to the acceptance rate. Figure 5.2 shows the proposal densities tuned by Laplace approximation with two different initial guess points. Different initial guess points lead to completely different proposal densities due to two local maximum states and the special shape of the target density ('banana' shape). Neither of these two proposal densities could recognize the shape of target density well and give rise to distant proposals. In a real simulation problem, the actual target density is unknown. Therefore, traditional methods possessing the random-walk behaviour could not usually provide us with efficient sampling results.

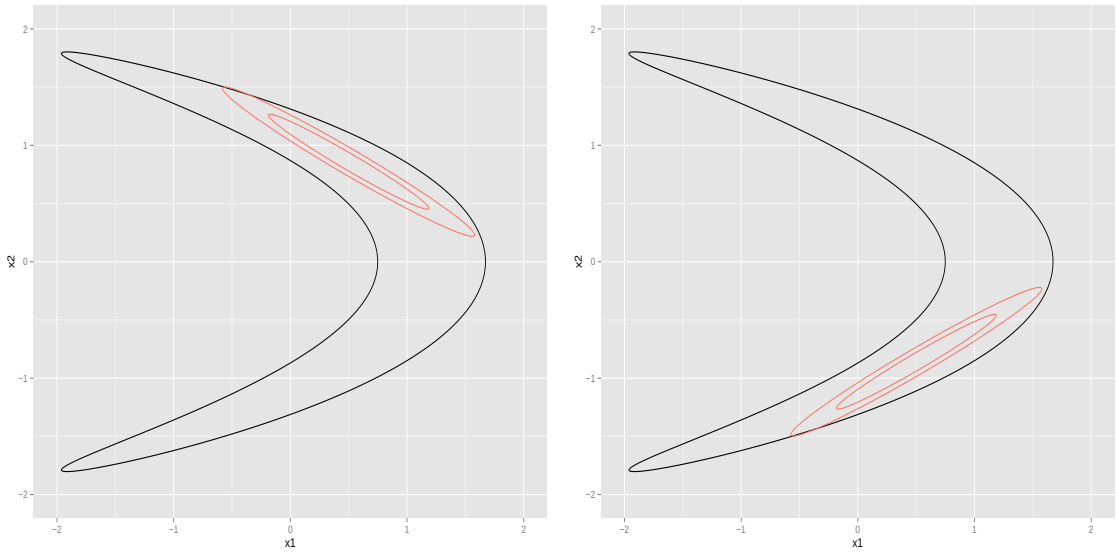


Figure 5.2: Proposal density provided by Laplace approximation with different initial guess points for RWMH sampler. Left plot starts from $(-1, 1.4)$; right plot starts from $(-1, -1.4)$. The black contour represents the target density; red contour lines stand for the tuned proposal densities.

We started both HMC and RWMH from point $(-1, 1.4)$ and implemented them to obtain 20000 posterior samples without thinning. For the sake of clarity, the first 600 simulated samples are displayed in Figure 5.3. HMC illustrated in the left plot traversed the state space quickly compared with the traditional RWMH sampler reported in the right plot. Moreover, posterior samples provided by the

RWMH sampler displayed a similar shape with the proposal density of sampler as illustrated in the left plot of Figure 5.2.

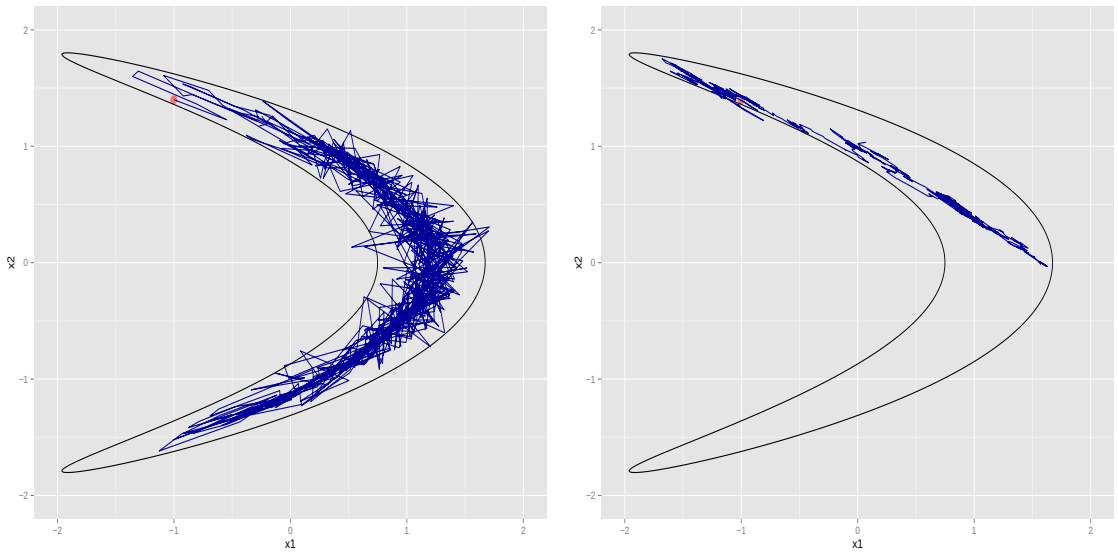


Figure 5.3: 600 posterior samples provided by HMC sampler (left plot) and RWMH sampler (right plot). Red points illustrate starting points.

Therefore, HMC has potentials to provide distant proposals as it exploits the Hamiltonian dynamics to avoid the random-walk behaviour and guide the proposals.

5.5 Step-Size Problems

As highlighted in the previous section, HMC is a powerful sampling method in providing distant proposals. Its performance, however, is very sensitive to its own algorithm parameters: M, l and ε. RMHMC, proposed by Girolami and Calderhead (2011), and NUTS, proposed by Hoffman and Gelman (2011), are two HMC variants designed to deal with problems of M and l respectively; See Chapter 7 for more details. In this section, we will discuss the influences of step-size to the performances of HMC sampler from two aspects. Firstly, the result of inappropriate step-size is illustrated in section 5.5.1. Secondly, section 5.5.2 shows that a fixed global step-size is not suitable.

5.5.1 Inappropriate Step-Size

Obviously, when the step-size is too small, the energy of a Hamiltonian system is well conserved by the leap-frog integrator to some degree and the acceptance rate is high. However, the problem is that the performance of HMC is just like a random walk Metropolis-Hastings MCMC which has high auto-correlations, low effective sample size and thus slow explorations of the state space which wastes much computation time. When the step-size is too large, the leap-frog integrator could not conserve the energy of Hamiltonian system well enough and thus lots of proposals would be rejected. Recall the acceptance probability,

$$\alpha = \min\{1, \frac{\exp\big(-H[\boldsymbol{\theta}(\tau), \mathbf{p}(\tau)]\big)}{\exp\big(-H[\boldsymbol{\theta}(0), \mathbf{p}(0)]\big)}\},$$

where $\{\boldsymbol{\theta}(0), \mathbf{p}(0)\}$ is the starting point of the dynamical trajectory and is identical to $\{\boldsymbol{\theta}^c, \mathbf{p}\}$; $\{\boldsymbol{\theta}(\tau), \mathbf{p}(\tau)\}$ is the end point of the approximated trajectory provided by the leap-frog integrator. If the numerical integrator used could conserve the energy exactly, then the acceptance rate would always be one. It is, however, unrealistic since numerical integrators always introduce errors. The leap-frog integrator has local error of order ε^3 and global error of order ε^2 with a fixed length of the trajectory. The acceptance probability is determined by the difference between the Hamiltonian energies at the starting point and ending point of the approximated trajectory.

A rejected trajectory is caused by large difference between these two values. To highlight this fact, we reuse the 'banana' example stated in section 5.4.4. The dynamic trajectory is initialized from point $(-3, 5, 3.5)$ and approximated by the leap-frog integrator with step-size value 0.08. Table 5.1 reports how the Hamiltonian energy changes for 4 leap-frog steps in approximating such a trajectory. The total energy underwent a big change even after one leapfrog step and became extremely large after 4 steps.

	step 0	step 1	step 2	step 3	step 4
energy	9.37e+02	2.24e+03	4.28e+04	8.37e+09	1.55e+26

Table 5.1: Energy changes when leapfrog starts from $(-3.5, 3.5)$ with step-size 0.08

Such a trajectory would definitely be rejected. Therefore, an inappropriately large step-size would lead to many unstable trajectories like the case highlighted in Table 5.1 and thus give rise to low acceptance rate and a stuck MCMC chain, i.e. no new proposals are accepted. For such an unstable trajectory caused by an inappropriate step-size, as illustrated by Table 5.1, two points are worthy to conclude:

- The approximation error in energy increases quickly as the trajectory grows. Neal (2011) pointed out that the approximation error in simulated Hamiltonian trajectories is independent of l as long as the step-size value is small enough to make the Hamiltonian dynamic stable. Therefore, a stable boundary is required to be satisfied by the step-size to prevent the local approximation error from accumulating as the number of leap-frog integrator increases.

- The approximation error after one leap-frog step (the local approximation error) might be large enough to result in a low acceptance probability. Therefore, by controlling the step-size, the local approximation error should be managed in a moderate magnitude to provide reasonable acceptance probability.

5.5.2 Changing Step-Size

Depending upon the problems caused by inappropriate step-sizes, the issue is detecting the boundary which could guarantee the leap-frog integrator to give stable trajectory approximations with reasonable local approximation errors. As Neal commented, a constant boundary is not dangerous since the step-size problem can be fixed by preliminary runs. In these preliminary runs, HMC could be started with a big step-size and then we could reduce the step-size gradually until a satisfactory acceptance rate is reached. The real problem arises when there is no fixed boundary that has the ability to make the leap-frog integrator provide stable trajectories globally. If the stable boundaries vary based on different regions of state space, then a step-size which is stable for one region might turn out to be: 1) too large for other regions and thus the chain might not visit other regions; 2) too small for other regions so that close proposals are obtained. Therefore, in situations with changing stable boundaries, usual adaptive methods are not suitable any more. In addition, if local stable boundaries become smaller and smaller as $\boldsymbol{\theta}$ changes, then there might not exist a single step-size which is appropriate for the HMC to run throughout the whole state space. Here, we will illustrate two points:

- The optimal or sub-optimal step-size value might change as current state changes.
- Even starting points could have an influence on the choice of the step-size value.

To illustrate the step-size problem due to changing stable boundaries, we carried out two experiments by running HMC for the 'banana' example with different initial points and step-size values. There are 500 iterations in each MCMC chain generated by HMC algorithm with 4 Leapfrog steps per iteration; the variance matrix for 'momentum' variables is set to the identity matrix for simplicity. If the proposal point is accepted, the trajectory path is displayed by the blue line, otherwise the red line.

Experiment 1:

The first experiment proceeded by implementing HMC with the same step-size value but two different starting points. To be specific, these two Markov chains constructed by HMC initialized from points $(-3, 2.8)$ and $(-3, 3)$ respectively. Also the step-size value is set to be 0.1 for both of them.

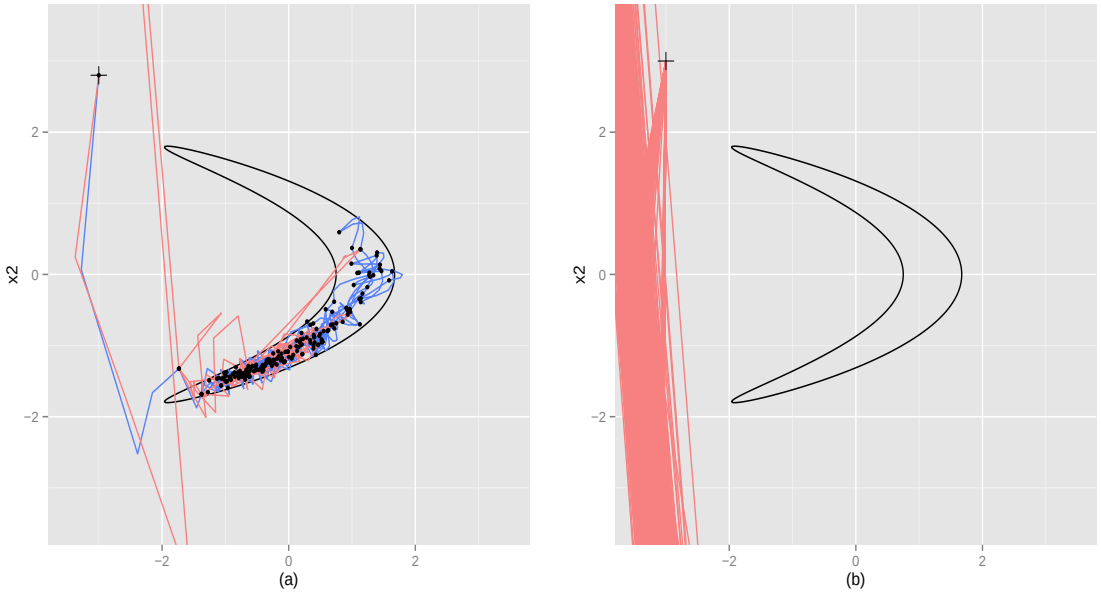


Figure 5.4: 200 posterior samples provided by HMC sampler with step-size 0.1 and starting point $(-3, 2.8)$ (plot a) and $(-3, 3)$ (plot b). Red lines represent rejected paths; blue lines mean accepted paths; '+' illustrates initial point; '·' means an accepted state.

The sampler performances are illustrated in Figure 5.4. It can be seen that small differences in starting points lead to completely different results: the one in plot (a) started from point $(-3, 2.8)$ could traverse the state space although there were several rejected iterations at the beginning; the other one in plot (b) which is initialized from point $(-3, 3)$ always fails to be accepted. This phenomenon indicates that step-size 0.1 is stable for $(-3, 2.8)$ or its small neighbourhood but not suitable for the vicinity of $(-3, 3)$.

Experiment 2: The second experiment set up bears a close resemblance to the previous experiment. We only altered the step-size value from 0.1 to 0.08 and specified starting points as $(-3, 3)$ and $(-3, 3.5)$. Figure 5.5 reports 200 posterior samples generated by HMC sampler. Obviously, reducing the step-size value from

0.1 to 0.08 fixes the problem shown in Experiment 1 for point $(-3, 3)$. However, step-size 0.08 cannot satisfy the stability requirement as long as the initial point changes from $(-3, 3)$ to $(-3, 3.5)$ (plot (b)). In order for the chain to move from point $(-3, 3.5)$, the step-size value needs to be decreased again. This circulating phenomenon, that would occur again and again as initial point goes further, implies that the stability boundary is changing locally.

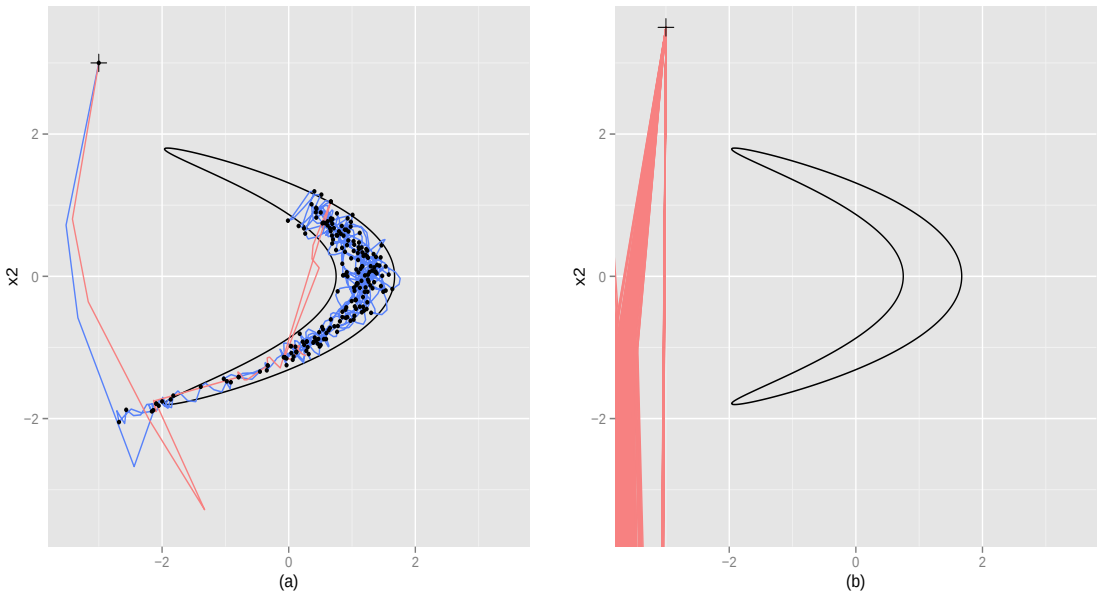


Figure 5.5: 200 posterior samples provided by HMC sampler with step-size 0.08 and starting point $(-3, 3)$ (plot a) and $(-3, 3.5)$ (plot b). Red lines represent rejected paths; blue lines mean accepted paths; '+' illustrates initial point; '·' means an accepted state.

As shown by Experiment 1 & Experiment 2, different regions might have different stability boundaries. In addition, the more extreme position at which the chain is currently located, the smaller step-size is needed. In other words, an appropriate step-size is dependent on where the point is. In plot (b) of both Figure 5.4 & Figure 5.5, all the unstable trajectories (red lines) illustrate that it is the inappropriate step-size used for the current position point that will drive paths to extreme places which have extremely low probabilities. Therefore, to solve the step-size problems, local conditions is needed in order for the leap-frog integrator to give stable trajectories with local approximation errors of moderate size.

5.6 Step-Size Local Conditions

In this section, we will focus on the local conditions for the step-size so that the leap-frog integrator can provide stable trajectories with local approximation errors of moderate size. In general, such local conditions, especially the one for the stability of the leap-frog integrator, cannot be derived easily. We therefore approximate the local area of the target statistical distribution and explore the step-size conditions for this local approximation. And the step-size conditions explored for the local approximations are considered as the local conditions for the original target problem approximately.

Here, we illustrate how to locally approximate the target statistical distribution in the Hamiltonian system. Recall the Hamiltonian system illustrated in Equation (5.3.13)

$$H(\boldsymbol{\theta}, \mathbf{p}) = -\log p(\boldsymbol{\theta}, \mathbf{p}) = -L(\boldsymbol{\theta}) + \frac{1}{2}\log\{(2\pi)^D |M|\} + \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p}. \tag{5.6.41}$$

Denote $-L$ by $\mathcal{L}$, the above line can be re-written as

$$H(\boldsymbol{\theta}, \mathbf{p}) = \mathcal{L}(\boldsymbol{\theta}) + \frac{1}{2}\log\{(2\pi)^D |M|\} + \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p}. \tag{5.6.42}$$

The local approximation is made for $\mathcal{L}(\boldsymbol{\theta})$ through its second-order Taylor expansion around $\boldsymbol{\theta}^c$, the current state of the Markov chain iterations. Let us denote such a approximation by $\mathfrak{L}_{\boldsymbol{\theta}^c}(\boldsymbol{\theta})$, we then have the following expression

$$\mathcal{L}(\boldsymbol{\theta}) \approx \mathfrak{L}_{\boldsymbol{\theta}^c}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}^c) + \frac{d}{d\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^c} \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}^c) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^c)^T \cdot \frac{d^2}{d\boldsymbol{\theta}^2}\mathcal{L}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^c} \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}^c). \tag{5.6.43}$$

For $\mathfrak{L}_{\boldsymbol{\theta}^c}(\boldsymbol{\theta})$, it is a quadratic function and thus has the first derivative and the second derivative of the following forms

$$\frac{d}{d\boldsymbol{\theta}}\mathfrak{L}_{\boldsymbol{\theta}^c}(\boldsymbol{\theta}) = \frac{d}{d\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^c} + (\boldsymbol{\theta} - \boldsymbol{\theta}^c)^T \cdot \frac{d^2}{d\boldsymbol{\theta}^2}\mathcal{L}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^c} \tag{5.6.44}$$

$$\frac{d^2}{d\boldsymbol{\theta}^2}\mathfrak{L}_{\boldsymbol{\theta}^c}(\boldsymbol{\theta}) = \frac{d^2}{d\boldsymbol{\theta}^2}\mathcal{L}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^c} \tag{5.6.45}$$

$$\frac{d^k}{d\boldsymbol{\theta}^k}\mathfrak{L}_{\boldsymbol{\theta}^c}(\boldsymbol{\theta}) = 0, k > 2 \tag{5.6.46}$$

Therefore, by expanding the quadratic function $\mathfrak{L}_{\boldsymbol{\theta}^c}(\boldsymbol{\theta})$ around its maximum $\boldsymbol{\theta}^c_\star$ according to the Taylor expansion, $\mathfrak{L}_{\boldsymbol{\theta}^c}(\boldsymbol{\theta})$ can be re-expressed as

$$\mathfrak{L}_{\boldsymbol{\theta}^c}(\boldsymbol{\theta}) = \mathfrak{L}_{\boldsymbol{\theta}^c}(\boldsymbol{\theta}^c_\star) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^c_\star)^T \cdot \frac{d^2}{d\boldsymbol{\theta}^2}\mathcal{L}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^c} \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}^c_\star) \tag{5.6.47}$$

where $\boldsymbol{\theta}^c_\star$ is the maximum of $\mathfrak{L}_{\boldsymbol{\theta}^c}(\boldsymbol{\theta})$ and thus the first derivative in Equation (5.6.44) evaluated at this point is zero. By substituting the above expression into Equation (5.6.43), we have

$$\mathcal{L}(\boldsymbol{\theta}) \approx \mathfrak{L}_{\boldsymbol{\theta}^c}(\boldsymbol{\theta}^c_\star) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^c_\star)^T \cdot \frac{d^2}{d\boldsymbol{\theta}^2}\mathcal{L}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^c} \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}^c_\star)$$

By substituting the above approximation for $\mathcal{L}(\boldsymbol{\theta})$ into Equation (5.6.42), we obtain a local approximation $\mathcal{H}(\boldsymbol{\theta}, \mathbf{p})$ for the original Hamiltonian system $H(\boldsymbol{\theta}, \mathbf{p})$, i.e.

$$H(\boldsymbol{\theta}, \mathbf{p}) \approx \mathcal{H}(\boldsymbol{\theta}, \mathbf{p}) = \\ \mathfrak{L}_{\boldsymbol{\theta}^c}(\boldsymbol{\theta}^c_\star) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^c_\star)^T \cdot \frac{d^2}{d\boldsymbol{\theta}^2}\mathcal{L}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^c} \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}^c_\star) + \frac{1}{2}\log\{(2\pi)^D|M|\} + \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p}. \tag{5.6.48}$$

Note that $\frac{d^2}{d\boldsymbol{\theta}^2}\mathcal{L}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^c}$ represents the curvature information of the statistical model of interest around $\boldsymbol{\theta}^c$. Suppose that $\frac{d^2}{d\boldsymbol{\theta}^2}\mathcal{L}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^c}$ is a positive-definite matrix. Through such an approximation, at the start of each simulated trajectory, $\mathcal{H}(\boldsymbol{\theta}, \mathbf{p})$ is like a scenario taking a Gaussian distribution with mean $\boldsymbol{\theta}^c_\star$ and variance matrix $\left(\frac{d^2}{d\boldsymbol{\theta}^2}\mathcal{L}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^c}\right)^{-1}$ as its target distribution. In other words, the Gaussian distribution $\mathcal{N}(\boldsymbol{\theta}^c_\star, \left(\frac{d^2}{d\boldsymbol{\theta}^2}\mathcal{L}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^c}\right)^{-1})$ is used to locally approximate the area around $\boldsymbol{\theta}^c$ of the original target distribution and is thus adapted to the current state of the Markov chain. While, for each dynamical trajectory that is going to be calculated by the leap-frog integrator, the target distribution of $\mathcal{H}$ is simply a Gaussian distribution with fixed mean and fixed variance. In the following parts, we discuss the conditions controlling stability and the local approximation errors of the leap-frog integrator for $\mathcal{H}$. And these conditions are considered as local conditions for H approximately.

Local Stability Condition

For $\mathcal{H}$ shown in Equation (5.6.48), the analytical trajectory of its Hamiltonian equations can be derived. For the sake of simplicity, the terms $\mathfrak{L}_{\boldsymbol{\theta}^c}(\boldsymbol{\theta}^c_\star)$ and $\frac{1}{2}\log\{(2\pi)^D|M|\}$

can be dropped since they are both constant. By denoting $\Sigma = \left(\frac{d^2}{d\boldsymbol{\theta}^2}\mathcal{L}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^c}\right)^{-1}$ and shifting the approximated Gaussian distribution to have zero mean, i.e. $\mathcal{N}(\mathbf{0}, \Sigma)$, the Hamiltonian system $\mathcal{H}$ in Equation (5.6.48) can be written as

$$\mathcal{H}(\boldsymbol{\theta}, \mathbf{p}) = \frac{1}{2}\boldsymbol{\theta}^T\Sigma^{-1}\boldsymbol{\theta} + \frac{1}{2}\mathbf{P}^T M^{-1}\mathbf{p}. \tag{5.6.49}$$

Its corresponding Hamiltonian equations are

$$\dot{\boldsymbol{\theta}} = \frac{\partial \mathcal{H}}{\partial \mathbf{P}} = M^{-1}\mathbf{P}$$
$$\dot{\mathbf{P}} = -\frac{\partial \mathcal{H}}{\partial \boldsymbol{\theta}} = -\Sigma^{-1}\boldsymbol{\theta}$$

This Hamiltonian equations are equivalent to

$$\ddot{\boldsymbol{\theta}} + M^{-1}\Sigma^{-1}\boldsymbol{\theta} = 0$$

Its analytical solution can be expressed as (José and Saletan, 1998)

$$\boldsymbol{\theta}(\tau) = C\sum_{j=1}^{D}\cos(\omega_j\tau + \delta_j)N_j, \tag{5.6.50}$$

where ω_j are square root of the eigenvalues λ_j of the matrix $M^{-1}\Sigma^{-1}$, i.e. $\omega_j = \sqrt{\lambda_j}$; N_j are the corresponding eigenvectors of matrix $M^{-1}\Sigma^{-1}$; C and δ_j are amplitude and phases, both determined by the initial conditions. Equation 5.6.50 indicates that the analytical solution can be considered as a combination of D independent harmonic oscillators. To obtain a good numerical approximation to such a solution, we would like the numerical method to provide stable results for each of these harmonic oscillations. In other words, we should investigate the leap-frog stability problem for D trajectories

$$\mathbf{u}_j(\tau) = C\cos(\omega_j\tau + \delta_j)N_j, \tag{5.6.51}$$
$$j = 1, \cdots, D.$$

Each of these trajectories is the solution of the following differential equation

$$\ddot{\mathbf{u}}_j + \omega_j^2\mathbf{u}_j = 0, \tag{5.6.52}$$

which is equivalent to

$$\dot{\mathbf{u}}_j = m_j^{-1}\mathbf{I}\mathbf{v}_j$$
$$\dot{\mathbf{v}}_j = -k_j^{-1}\mathbf{I}\mathbf{u}_j$$

where m_j and k_j are any values satisfying $\omega_j^2 = m_j^{-1}k_j^{-1}$; $\mathbf{I}$ is a $D \times D$-dimensional identity matrix. Each of these oscillators conserves a Hamiltonian system of the form

$$\mathcal{H}_j(\mathbf{u}_j, \mathbf{v}_j) = \frac{1}{2}\mathbf{u}_j^T(k_j^{-1}\mathbf{I})\mathbf{u}_j + \frac{1}{2}\log(2\pi|m_j\mathbf{I}|) + \frac{1}{2}\mathbf{v}_j^T(m_j^{-1}\mathbf{I})\mathbf{v}_j.$$

Suppose that the amplitude of $\mathbf{u}_j, \mathbf{v}_j$ are denoted by u_j, v_j respectively. The numerical solution of a one-step leap-frog integrator applied to its differential equations yields

$$\begin{pmatrix} u_j(\tau+\varepsilon) \\ v_j(\tau+\varepsilon) \end{pmatrix} = S \begin{pmatrix} u_j(\tau) \\ v_j(\tau) \end{pmatrix},$$

where

$$\begin{aligned} S &= \begin{pmatrix} 1 & 0 \\ -\frac{\varepsilon}{2}k_j^{-1} & 1 \end{pmatrix} \begin{pmatrix} 1 & \varepsilon m_j^{-1} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\frac{\varepsilon}{2}k_j^{-1} & 0 \end{pmatrix} \\ &= \begin{pmatrix} 1 - \frac{\varepsilon^2}{2m_jk_j} & \frac{\varepsilon}{m_j} \\ -\frac{\varepsilon}{k_j} + \frac{\varepsilon^3}{4k_j^2m_j} & 1 - \frac{\varepsilon^2}{2m_jk_j} \end{pmatrix}. \end{aligned}$$

The eigenvalues of the above matrix S determine stability (the long-time behaviour of the numerical solution) and stability requires the eigenvalues to be less than or equal to one in modulus (Hairer et al., 2006). The matrix S has $2D$ eigenvalues

$$\underbrace{\xi_1, \cdots \xi_1}_{D \text{ eigenvalues}}, \underbrace{\xi_2, \cdots \xi_2}_{D \text{ eigenvalues}}$$

where

$$\xi_{1,2} = 1 - \frac{\varepsilon^2}{2m_jk_j} \pm \frac{\varepsilon}{\sqrt{m_jk_j}}\sqrt{\frac{\varepsilon^2}{4m_jk_j} - 1}.$$

Note that the leap-frog integrator consists of shear transformations and thus $|S| = (\xi_1\xi_2)^D = 1$. If the two eigenvalues ξ_1, ξ_2 are both real values, then one of them must

be larger than 1 and thereby violates the stability requirement. In order to satisfy the stability requirements, we should let

$$\begin{aligned} &\frac{\varepsilon^2}{4m_jk_j} - 1 < 0 \\ \Longrightarrow\ &\frac{\varepsilon}{\sqrt{m_jk_j}} < 2 \end{aligned} \tag{5.6.53}$$

and ξ_1, ξ_2 become complex values

$$\xi_{1,2} = 1 - \frac{\varepsilon^2}{2m_jk_j} \pm i\frac{\varepsilon}{\sqrt{m_jk_j}}\sqrt{1 - \frac{\varepsilon^2}{4m_jk_j}}$$

with $|\xi_{1,2}| = 1$. Therefore, we obtain D stability conditions of the form in Equation (5.6.53) corresponding to the D trajectories in Equation (5.6.51). Note that $\sqrt{m_j^{-1}k_j^{-1}}$ are identical to ω_j, the square root of the eigenvalues of matrix $M^{-1}\Sigma^{-1}$. The stability conditions in Equation (5.6.53) can be rewritten as

$$\varepsilon\omega_j < 2,$$

If $M = I$, ω_j become square root of the eigenvalues of the matrix Σ^{-1}. Therefore, the higher the eigenvalue of Σ^{-1} is, the smaller the step-size is required to make the corresponding oscillator stable. Since there are D such oscillators, the condition that can make all the oscillators stable is

$$\varepsilon\omega < 2, \tag{5.6.54}$$

where ω is the square root of the largest eigenvalue of the matrix Σ^{-1}, i.e. $\omega = \sqrt{\lambda}$, where $\lambda = \max\{\lambda_j; j = 1 \cdots, D\}$.

Recall the dependence of Σ^{-1} on the current state of the Markov chain,

$$\Sigma^{-1} = -\frac{\partial^2 L}{\partial \boldsymbol{\theta}^2}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^c}.$$

Therefore, the local stability condition for the original target $L(\boldsymbol{\theta})$ is approximately

$$\varepsilon\omega_{\boldsymbol{\theta}^c} < 2, \tag{5.6.55}$$

where $\omega_{\boldsymbol{\theta}^c} = \sqrt{\lambda_{\boldsymbol{\theta}^c}}$ and $\lambda_{\boldsymbol{\theta}^c}$ is the largest eigenvalue of $-\frac{\partial^2 L}{\partial \boldsymbol{\theta}^2}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^c}$, the Hessian matrix at the current state. The larger the eigenvalue is, the smaller the step-size is required

to keep the trajectory stable. Since the Hessian matrix represents the local curvature information, we can conclude that the higher the curvature is, the smaller the step-size should be. This finding is in line with the intuition visualized in Figure 5.6.

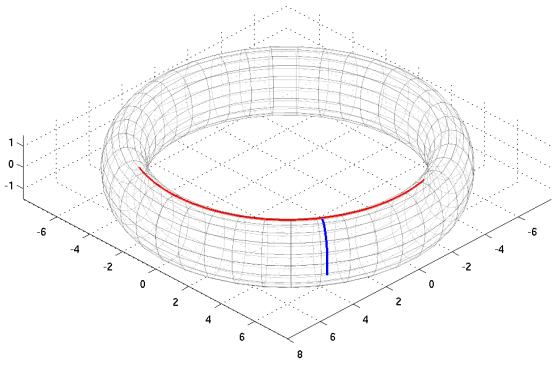


Figure 5.6: Surface of Torus

The rotation speed (curvature) along the red curve is much more gentle than that on the blue curve. Suppose that a particle is moving on the surface of the torus. The step-size that could make the particle move successfully along the red curve is larger than that for the blue curve. A step-size which satisfies the stable condition of the red curve might turn out to be too large to keep the particle on the surface due to the influence along the blue curve. Also note that the conclusion in Equation (5.6.55) is derived under the assumption that the Hessian matrix is positive-definite. This assumption, however, cannot always be guaranteed and thus it might lead to negative eigenvalues. A negative eigenvalue with large absolute value still depicts a large curvature but in an opposite direction compared to the positive one. Intuitively, a small step-size is also needed in this situation. Therefore, we change the local stability condition in Equation (5.6.55) to the following one

$$\varepsilon\omega_{\boldsymbol{\theta}^c} < 2 \tag{5.6.56}$$

where $\omega_{\boldsymbol{\theta}^c} = \sqrt{|\lambda_{\boldsymbol{\theta}^c}|}$ and $\lambda_{\boldsymbol{\theta}^c}$ is the eigenvalue of the matrix of $-\frac{\partial^2 L}{\partial \boldsymbol{\theta}^2}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^c}$ with the largest absolute value. RMHMC exploits this fact to design 'momentum' variables; see details in section 7.2.

Approximation Errors

For the Hamiltonian system shown in Equation (5.6.48), according to Equation (5.4.40), the local approximation error $\mathrm{Err}(\mathcal{H})$ caused by the leap-frog integrator changes to be

$$\mathrm{Err}(\mathcal{H}) = \varepsilon^3\Big(\frac{1}{4}\mathbf{p}^T(\tau)(M^{-1})^2\big(\frac{\partial^2 L}{\partial \boldsymbol{\theta}^2}\frac{\partial L}{\partial \boldsymbol{\theta}}\big)|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\tau)}\Big) + \mathcal{O}(\varepsilon^4)$$

The above equation indicates that not only the curvature of the log-density function but also the gradient (interpreted as the 'force' in physic) should be considered in controlling the local approximation error. There is no explicit condition for the approximation errors as that for the stability. And reasonable approximation errors are acceptable. Neal (2011) has already noted that a small step-size value is required when the gradient (the first derivative) of the log density is large. However, compared with the first derivative, the curvature controlling the stability is more influential since it makes sure that the parameters of interest $\boldsymbol{\theta}$ do not go to wild places in long time. In addition, as illustrated by Equation (5.4.39), it is the curvature information that controls the accuracy of $\boldsymbol{\theta}$ given by the leap-frog integrator.

5.7 HMC with Stochastic Step-Size

As previous discussions have shown, the stability requirement varies as the current state changes. This fact leads us to consider using variable step-sizes according to where the state is currently located. In this section, problems of the variable step-sizes method are firstly stated and then a new algorithm is proposed to achieve state-dependent step-size HMC in a stochastic way without encountering the described problems. This new algorithm can also be considered as a generalised Metropolis-Hastings with dynamics method.

5.7.1 Variable Step-Size Problems

The leap-frog integrator with a constant appropriate step-size could guarantee that the approximated dynamics are involutions as the leap-frog integration is time symmetrical. This desirable property, however, will be lost if variable step-sizes are used. Hut et al. (1995) proposed an implicit method to recover this appealing property. It calculates step-size by a given symmetry function

$$\varepsilon = \frac{1}{2}\big(h(\phi_t) + h(\phi_{(t+\varepsilon)})\big)$$

where $\phi_t = [\boldsymbol{\theta}(t), \boldsymbol{p}(t)]$ and $\phi_{(t+\varepsilon)} = [\boldsymbol{\theta}(t+\varepsilon), \boldsymbol{p}(t+\varepsilon)]$. The function $h(\cdot)$ is some criterion of choosing step-size according to where the state is. The symmetry is recovered by the symmetry function but the unwanted property of this method is that extra iterations are required to calculate step-size values because of the implicit function involved in the symmetry function. The RMHMC sampler, which selects the variance matrix for 'momentum' variables according to where the current state is, also involves implicit calculations and thus requires expensive computations especially for complex models.

5.7.2 Stochastic Step-Size

Our goal is a sampler which retains the good features of HMC without either needing the user to choose a leap-frog step-size or assuming that there is a global lower limit to the stability boundary. We propose an algorithm which automatically updates

the step-size according to the current state. The basic idea is to generate a step-size for each iteration from a distribution determined by the local curvature (local geometric information) at the current state of the Markov chain. As discussed in section 5.6, the curvature information solely determines a stable trajectory and is more influential in keeping the parameters of interest not falling in wild regions compared with the gradient. Therefore, only the local curvature is exploited to simulate step size values. This changing step-size scheme violates the involution property of HMC and we overcome this difficulty by re-defining the augmented variables $\boldsymbol{\varphi}$ in Algorithm 2 rather than using implicit symmetry functions to recover the involution. We consider the 'momentum' variables and the generated step-size altogether as the augmented variables $\boldsymbol{\varphi}$ required in Algorithm 2, i.e.

$$\boldsymbol{\varphi} = \{\mathbf{p}, \varepsilon\}$$

The dynamics used here are the same with that illustrated in Equations (5.4.26) to (5.4.28). Note that the dynamics do not not involve varying step-sizes since changing step-sizes is achieved before the start of the dynamics. These dynamics are involutions and preserve volume. The acceptance probability is given by

$$\min\{1, \frac{\exp(-H(\boldsymbol{\theta}', \boldsymbol{p}'))g(\varepsilon|\boldsymbol{\theta}')}{\exp(-H(\boldsymbol{\theta}^c, \boldsymbol{p}))g(\varepsilon|\boldsymbol{\theta}^c)}\}. \tag{5.7.57}$$

A benefit of this stochastic scheme is that there is no need for the user to specify a step-size value. In addition, this stochastic scheme allows the step-size chances to take small values to get out of 'sticky' points and large values to move to distant proposals if possible. Finally, the novelty is that the scheme exploits the local geometric information to update the step-size distribution automatically. This process is described in the following algorithmic form.

Algorithm 4 Hamiltonian Monte Carlo with Stochastic Step-size

1: Given an initial value $\boldsymbol{\theta}^1$, value for l and matrix M ;
2: **for** $j = 1, 2, \cdots, n$ **do**
3: Sample $\boldsymbol{p} \sim N(0, I)$
4: Sample $\varepsilon \sim g(\varepsilon|\boldsymbol{\theta}^j)$
5: Set $\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta}^j, \boldsymbol{p}' \leftarrow \boldsymbol{p}$
6: **for** $i = 1$ **to** l **do**
7: Set $\boldsymbol{\theta}', \boldsymbol{p}' \leftarrow$ Leap-frog$(\boldsymbol{\theta}', \boldsymbol{p}', \varepsilon)$
8: **end for**
9: Let

$$\boldsymbol{\theta}^{j+1} = \begin{cases} \boldsymbol{\theta}', & \text{If Uniform}(0,\,1) \leq \min\{1, \frac{\exp(-H(\boldsymbol{\theta}', \boldsymbol{p}'))g(\varepsilon|\boldsymbol{\theta}')}{\exp(-H(\boldsymbol{\theta}^j, \boldsymbol{p}))g(\varepsilon|\boldsymbol{\theta}^j)}\} \\ \boldsymbol{\theta}^j, & \text{otherwise} \end{cases}.$$

10: **end for**
11: **Function** Leap-frog $\{\boldsymbol{\theta}, \boldsymbol{p}, \varepsilon\}$
12: Set $\boldsymbol{p}' \leftarrow \boldsymbol{p} + (\frac{\varepsilon}{2})\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$
13: Set $\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} + \varepsilon M^{-1}\boldsymbol{p}'$
14: Set $\boldsymbol{p}' \leftarrow \boldsymbol{p}' + (\frac{\varepsilon}{2})\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}')$
15: **Return** $\boldsymbol{\theta}', \boldsymbol{p}'$

Specifically, we propose that $g(\varepsilon|\theta^c)$ is any appropriate distribution which has positive support and is scaled by $\frac{1}{\sqrt{|\lambda|}}$, where λ is the eigenvalue with the largest absolute value of the matrix $M^{-1}\frac{\partial^2(-L)}{\partial\,\boldsymbol{\theta}^2}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^c}$. The convergence of Algorithm 4 is demonstrated in the following corollary.

Corollary 5.7.1 *Suppose that $X = (\boldsymbol{\theta}, \boldsymbol{\varphi})$, where $\boldsymbol{\varphi} = \{\mathbf{p}, \varepsilon\}$, has density $\pi(\cdot)$ on $\mathcal{X} \subseteq \mathbb{R}^p$. The Markov chain described in Algorithm 4 converges to $\pi(\cdot)$.*

Proof : The joint density of the parameters of interest $\boldsymbol{\theta}$ and the augmented variables $\boldsymbol{\varphi}$ is

$$\pi(X) = p(\boldsymbol{\theta})N(\mathbf{p}|0, M)g(\varepsilon|\boldsymbol{\theta}) = \exp\big(-H(\{\boldsymbol{\theta}, \mathbf{p}\})\big)g(\varepsilon|\boldsymbol{\theta}).$$

Let $LF(\{\boldsymbol{\theta}, \mathbf{p}\}, l, \varepsilon)$ denote the leap-frog integrator which starts at $\{\boldsymbol{\theta}, \mathbf{p}\}$ with fixed step l and fixed step-size ε. Similarly as that in section 5.4, by changing the sign of

momentum variables $LF(\{\boldsymbol{\theta}, \mathbf{p}\}, L, \varepsilon)$ is also a continuously differentiable volume-preserving involution. Algorithm 4 changes the step-size before the beginning of the dynamics and keeps step-size unchanged during the dynamics. Therefore, the dynamics can be expressed as

$$U : (\{\boldsymbol{\theta}, \mathbf{p}\}, \varepsilon) \rightarrow \Big(LF(\{\boldsymbol{\theta}, \mathbf{p}\}, L, \varepsilon), \varepsilon\Big) = (\{\boldsymbol{\theta}', \mathbf{p}'\}, \varepsilon).$$

It is a continuously differentiable volume-preserving involution if we change the sign of momentum variable. With the acceptance probability defined according to Equation (5.2.12),

$$\alpha(x) = \min\Big\{1, \frac{\exp\Big(-H\big(\{\boldsymbol{\theta}', \mathbf{p}'\}\big)\Big) g(\varepsilon|\boldsymbol{\theta}')}{\exp\Big(-H\big(\{\boldsymbol{\theta}, \mathbf{p}\}\big)\Big) g(\varepsilon|\boldsymbol{\theta})}\Big\},$$

the transition achieved by Algorithm 4 conserves the joint density $\pi(\cdot)$ according to Theorem 5.2.3. □

5.7.3 Illustrative Example

In this section, the new designed algorithm 'HMC with stochastic step-size' is applied to two examples: the previously discussed 'banana' example and a multivariate t distribution.

Banana example

As stated in Equation (5.6.56), step-size are dictated by the local curvature. Thus, step-size are generated from a half-standard-normal distribution scaled by the eigenvalue with the largest absolute value of the local curvature matrix and thus adapted automatically according to the current state. We investigate the performances of the new algorithm if the simulation is started from a extreme point. The starting point is set to $(-10, 10)$ which is a very extreme starting position compared to previously mentioned starting points. The variance matrix of the 'momentum' variables are set to the identity matrix and the number of the leap-frog steps is set to 4. The trace plots and autocorrelations plots of simulation results are shown in Figure 5.7. In order to visualize the tract plots clearly, the samples drawn on the plots are obtained by thinning 10. The autocorrelations plots are for posterior samples without thinning. The simulated chain mixes quickly and converges well even with such a starting point. In fact, this algorithm works well even with a starting point $(-100, 100)$ where has extremely large gradient value. This indicates that without the gradient information involved in, adapting the step-size values only according to the local curvature information is enough to deal with the step-size problems.

The marginal simulation result is compared to the theoretical marginal density generations by carrying out a Kolmogorov-Smirnov test. In Figure 5.8, the blue line is the empirical cumulative density curve provided by the theoretical marginal density generation and the red curve is provided by the simulation result from Algorithm 4. The curves overlap well with significant small distances in both plots. The joint simulation result is tested by a Chi-squared test. Using a grid of 25 cells, the chi-squared goodness-of-fit statistic for this simulation is 26.15057 which is smaller than 36.41503, the critical value at 5% significance.

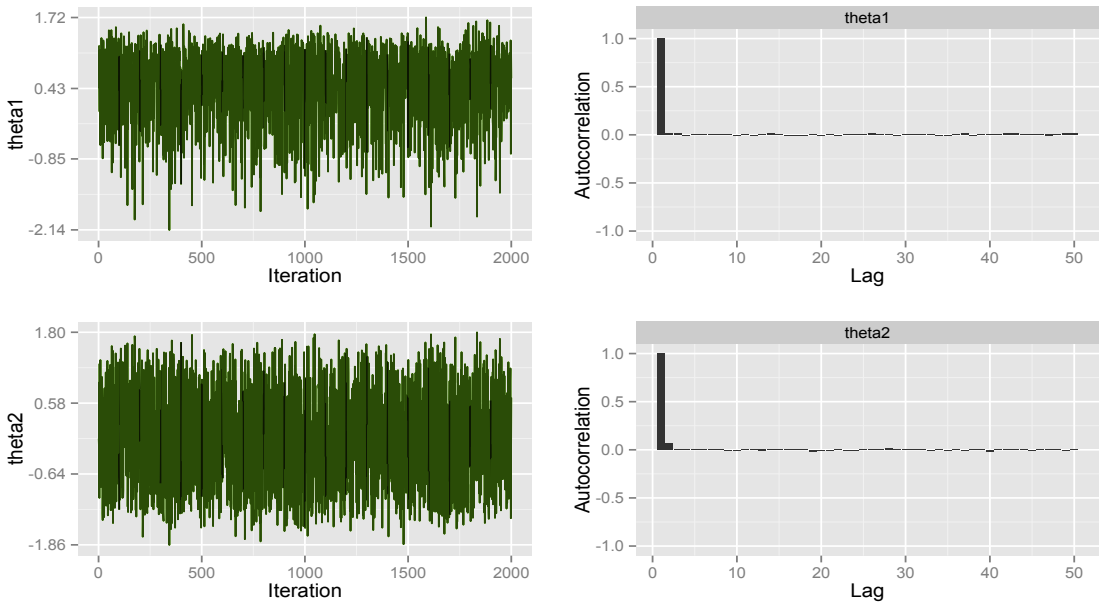


Figure 5.7: Trace plots (left column) and autocorrelations plots (right column) of simulated samples for 'banana' example.

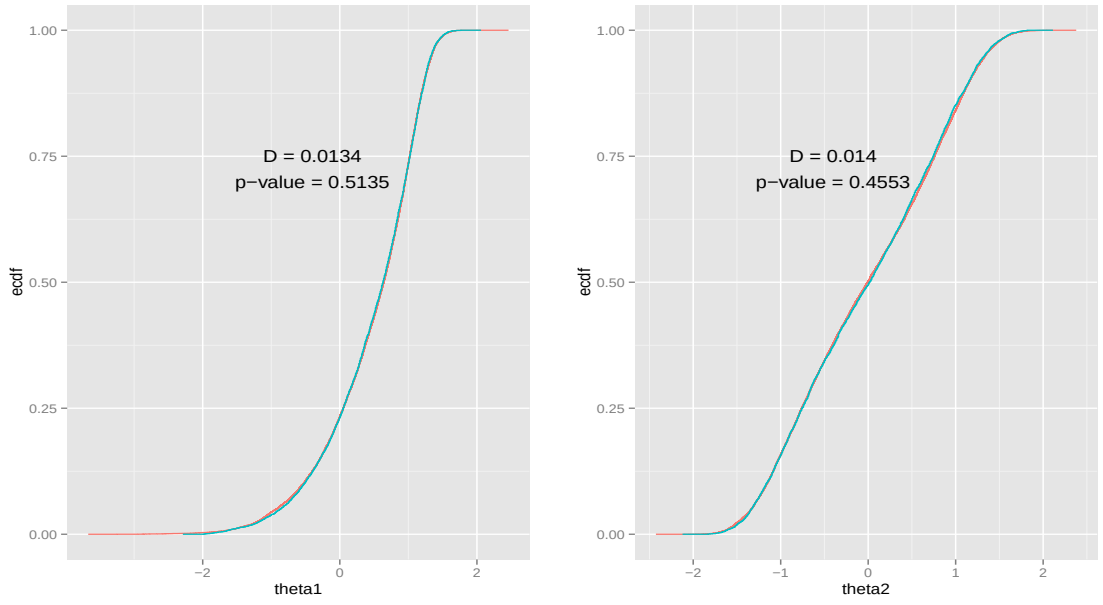
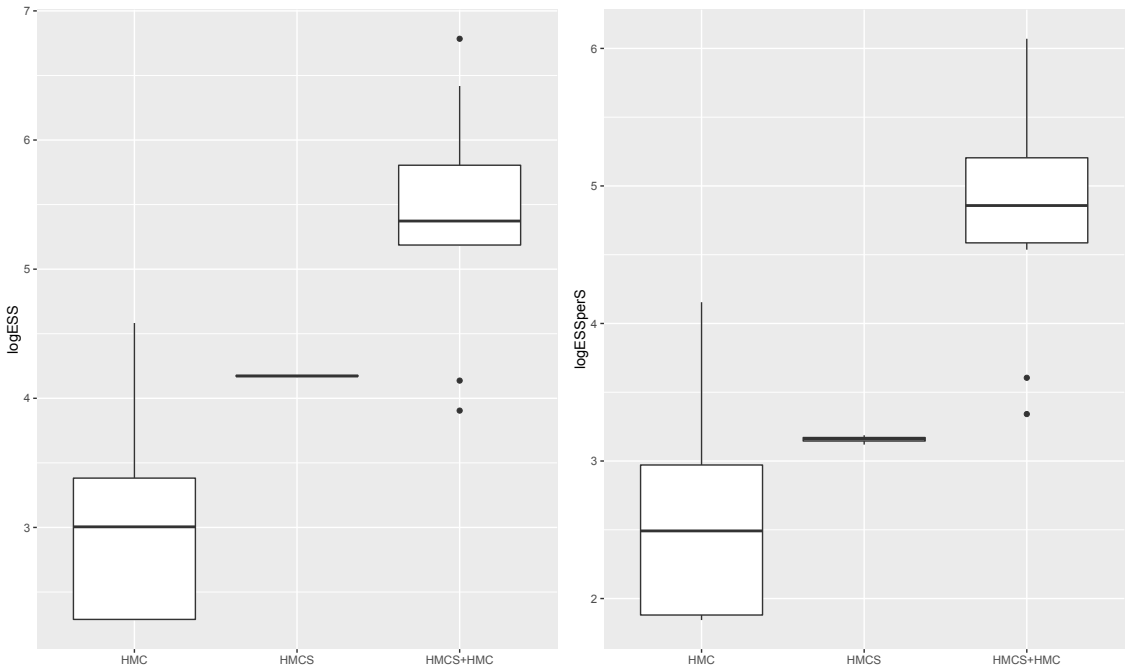


Figure 5.8: Comparison of the empirical distributions of samples generated from the theoretical marginal density and samples provided by HMC with stochastic step-size algorithm. The left plot is for θ_1 and the right plot is for θ_2.

Multivariate t distribution

Considering a 10-dimensional t distribution $t_{\nu,\boldsymbol{\mu},\Sigma}(\boldsymbol{\theta})$ with $\nu = 1 \times 10^6$, $\boldsymbol{\mu}$ is a 10-dimensional vector with all terms being zero and $\Sigma = 1 \times 10^{-5}\boldsymbol{I}$ where $\boldsymbol{I}$ is a 10×10 identity matrix. Both HMC and HMC with stochastic step-size algorithm are applied to this 10-dimensional t distribution. In order to be comparable, the covariance matrix of the momentum variables for both algorithms are set to be identity matrix and the number of leap-frog steps are both set to be 4.



(a) Box-plot of logarithm of ESS (b) Box-plot of logarithm of ESS/s

Figure 5.9: Efficiency comparisons

For the basic HMC algorithm, the step-sizes are adapted according to the acceptance rate in the burn-in iterations. Specifically, the step-size is doubled if the current acceptance rate is larger than 0.8 and is halved if the current acceptance rate is smaller than 0.6. For the HMC with stochastic step-size, there is no need to specify step-sizes. We run 10 chains each having 20000 iterations without thinning in three situations: HMC, HMC with stochastic step-size and the combination of HMC and HMC with stochastic step-size. Specifically, in the situation of using HMC and HMC with stochastic step-size together, HMC with stochastic step-size

is implemented in the burn-in iterations and HMC is used in the main iterations with fixed step-size set as the mean of step-size obtained by HMC with stochastic step-size algorithm. Figure 5.9 illustrates the box-plots of logarithm of the effective sample size (plot a) and logarithm of ESS per second (plot b) of obtained chains in all three situations. All the logarithms are taken based on 10. Using HMC solely results in unstable results. This further reflects that the performance of HMC is sensitive to the chosen step-size and thus the step-size tuning method. As for HMC with stochastic step-size algorithm, both plots indicates that it is the most stable one among all three and provides competitive performances compared with HMC. By comparing ESS with ESS/s for HMC with stochastic step-size algorithm, it is easy to see that it is more computational expensive than HMC. The reasons is that it requires curvature calculations. Clearly, the combination of HMC and HMC with stochastic step-size has generally larger effective sample size and effective sample size per second than the other two. It makes use of HMC with stochastic step-size algorithm to obtain reasonable step-size in the burn-in iterations and retains the speed of HMC in the main iterations.

5.8 Conclusions

We have presented a meta-algorithm 'generalised Metropolis-Hastings with dynamics' which includes, but not limited to, the ordinary Metropolis-Hastings algorithm, HMC with both exact and approximated dynamics and HMC with stochastic step-sizes. Any dynamics that are volume-preserving involutions can be used to design a algorithm converging to the desired distribution according to Theorem 5.2.3. Any algorithm (such as HMC) that exploits the dynamics to suppress the random-walk behaviour is worthy of investigation.

The HMC with stochastic step-size algorithm automatically adapts step-size according to the local curvature information of statistical model surfaces. This sampler eliminates the basic HMC's dependence on the chosen step-size value and is robust to extreme starting points.

Chapter 6

Background of a Complex Hierarchical Model

Hierarchical models have wide applications due to their flexibility in modelling a range of data across many sciences. Especially in Bayesian analysis, hierarchical models have become more and more prevalent after great computing power, efficient algorithms and user-friendly software have become available. The hierarchical model considered here, and further on, was firstly developed by Craig (2013) to model eco-toxicological data about variations in sensitivity of species to chemicals. It is particularly designed to characterize the non-exchangeable and taxonomic structure of species. Rather than using the frequently chosen Gaussian distributions, the Student's t-distribution is selected for the response variable since its heavy-tailed behaviour is observed in the preliminary data analysis (see Craig (2013) for details). In order to sample the posterior distribution resulting from the use of this model, two main computation tools are considered: Markov chain Monte Carlo (MCMC) methods and the MCMCglmm method.

Among a class of MCMC methods, random-walk Metropolis-Hastings algorithm (Metropolis et al., 1953) and Gibbs sampling algorithm (Geman and Geman, 1984) are traditional methods that depend on proposal distributions and conditional distributions respectively; Hamiltonian Monte Carlo (Duane et al., 1987) is a technique that exploits the gradient information through the Hamiltonian scheme. Some software tools, eg. BUGS (Spiegelhalter et al., 1996) and Stan (Stan Development Team,

2014b), have been developed to implement these MCMC samplers without onerous programming by users. These methods and software tools, however, result in poor performances when they are expected to deal with more sophisticated models involving high dimensionality and complex patterns of dependence. To be specific, the successful design of the most practical MCMC algorithms to sample from a target distribution in scenarios involving high dimensionality and complex dependence patterns relies on the appropriate choice of the proposal distribution. This holds true even for the Hamiltonian Monte Carlo sampler since the problem of tuning proposal distributions transfers to that of tuning distributions for 'momentum' variables.

As the model of choice becomes complicated, the solution is to break up the original sampling algorithm into smaller and simpler sampling problems by targeting the subcomponents of the entire parameter space. Efficient design of algorithms is often feasible in the block of such subcomponents. MCMCglmm (Hadfield et al., 2010) is particularly designed to sample posterior distributions of the generalized linear mixed models by classifying the whole parameter space into two subcomponents —- one block of linear predictors and another block of variance for the linear predictors. The R package 'MCMCglmm' is available to implement this method directly. However, it only works for models with response variables from a limited range of distributions. Unfortunately, the Student's t-distribution, that is assumed by our model, is not included in the predefined list. Craig (2013) described how to modify the MCMCglmm to calculate a model with t-distributed response variables. Although some improvements have been achieved by using the modified MCMCglmm for the targeted model, the obtained results still display high auto-correlations.

In this chapter, the background to our model of interest is introduced in section 6.1. Section 6.2 presents computation results and problems for the targeted model by using MCMC and MCMCglmm. The following three chapters deal with the computation problems and concentrate on improving simulation performances by designing different computation strategies for such a model. In chapter 7, some advanced MCMC methods, that would be used in the design of computation strategies for the chosen model, are introduced as preliminary materials. Chapter 8 details the design of computation strategies for this hierarchical model by combining and mod-

ifying advanced computation tools. In chapter 8.5, computation results obtained from different methods proposed in chapter 8 are compared.

6.1 Model Structure

The hierarchical model considered here is designed to model ecotoxicological data, especially the half maximal effective concentration (EC50). EC50 refers to the concentration of a chemical that provokes 50% of the maximal possible response after a specified exposure time (Motulsky, 1995). According to the exposure time, the test from which the data are recorded can be roughly classified as acute test or chronic test. As their names indicate, acute test is a short-term exposure test (usually hours or days) while chronic test is a long-term exposure test (weeks, months or years). The analysis of exotoxicological data mainly deals with variations in sensitivity of species to different chemicals. There is a large literature on ecotoxicological risk assessment and much of the existing work make some underlying assumptions: for example, Gaussian distributed errors and exchangeability among species. However, it has been found by examining a database of acute test results for a wide variety of chemicals and aquatic species that those assumptions might be inappropriate (Craig, 2013). Measurement errors for the same chemical-species combination display heavy-tailed behaviour. Moreover, species sensitivities are not priori exchangeable and exhibit taxonomic structure. In order to model these features, Craig (2013) proposed the hierarchical model

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \tag{6.1.1}$$

and

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \psi_{ij}, \tag{6.1.2}$$

where

- y_{ijk} is the k-th measured log-sensitivity by using the log-EC50 for chemical i tested on species j;
- μ_{ij} is the true log-sensitivity for species j exposed to chemical i;
- ε_{ijk} is measurement error modelled as Student's-t distribution by the parametrization

$$\varepsilon_{ijk} = \sigma_\varepsilon \frac{z_{ijk}}{\sqrt{\kappa_{ijk}}},$$

where z_{ijk} are iid (independently and identically distributed) $N(0,1)$ and κ_{ijk} are iid $\Gamma(\frac{1}{2}\nu_\kappa, \frac{1}{2}\nu_\kappa)$. z_{ijk} and κ_{ijk} are independent with each other.

- μ is the overall central value of log-sensitivity across all chemicals and species;
- α_i is the difference between the central value of log-sensitivity for chemical i and μ. They are modelled as random effects which are iid $N(0, \sigma_\alpha^2)$;
- β_j is the logarithm of the sensitivity-tendency for species j. In order to incorporate the taxonomical structure, β_j is modelled as

$$\beta_j = \beta_{1t_1(j)} + \cdots + \beta_{lt_l(j)} + \cdots + \beta_{Lt_L(j)},$$

where $\beta_{lt_l(j)}$ is the tendency component at taxonomical level l for species j whose classification at level l is t. Moreover, all the β_{lt}'s are exchangeable at same taxonomic level l. The species tendency components $\beta_{l\cdot}$ at each level are iid $N(0, \sigma_{\beta l}^2)$.

- ψ_{ij} is the interaction between chemical i and species j. In order to incorporate both the chemical specific variability and the taxonomically-related structure, the interaction factors are written as

$$\psi_{ij} = \phi_i \xi_{ij}$$

and

$$\xi_{ij} = \xi_{i1t_1(j)} + \cdots + \xi_{ilt_l(j)} + \cdots + \xi_{iLt_L(j)},$$

where ϕ_i scales the log-sensitivity variation for chemical i. This allows some chemicals to exhibit more variation in sensitivity than others. ξ_{ij} is constructed to introduce the taxonomical structure in a similar way as that for β_j. Moreover, all the ξ_{ilt}'s are exchangeable for fixed taxonomic level l, i.e. $\xi_{\cdot l\cdot}$ are iid $N(0, \sigma_{\xi l}^2)$. Thus, ξ_{ij} retain partial exchangeability between interactions to some degree. In this way, ξ_{ij} is directly comparable between different chemicals but ψ_{ij} is not. If we let $\lambda_i = \frac{1}{\phi_i^2}$, then the interactions could be rewritten as

$$\psi_{ij} = \frac{1}{\sqrt{\lambda_i}}\xi_{i1t_1(j)} + \cdots + \frac{1}{\sqrt{\lambda_i}}\xi_{ilt_l(j)} + \cdots + \frac{1}{\sqrt{\lambda_i}}\xi_{iLt_L(j)},$$

where all λ_i are defined to be iid $\Gamma(\frac{1}{2}\nu_\phi, \frac{1}{2}\nu_\phi)$. In the right hand-side of the above equation, each component term $\frac{1}{\sqrt{\lambda_i}}\xi_{ilt_l(j)}$ can be considered as normal distribution with spread controlled by λ_i corresponding to chemical i. Technically, such a term follows a Student's-t distribution. Since all the component terms in the right hand-side are scaled by a same λ_i, these terms are not independent. The sum of these correlated Student's-t distributed terms, ψ_{ij}, does not follow a Student's-t distribution.

- Conditional on the hyper-parameters $\sigma_\alpha, \left\{\sigma_{\beta l}\right\}_{l=1,\cdots,L}, \left\{\sigma_{\xi l}\right\}_{l=1,\cdots,L}, \nu_\phi, \sigma_\epsilon$ and ν_κ, the following blocks are independent,

$$\mu, \{\alpha_i\}, \{\beta_{1t}\}, \cdots, \{\beta_{Lt}\}, \{\phi_i\}, \{\xi_{i1t}\}, \cdots, \{\xi_{iLt}\}, \{\kappa_{ijk}\}, \{z_{ijk}\}.$$

6.2 Computations

The prior distribution for the hyper-parameters are assumed to be independent

$$p(\mu, \sigma_\alpha, \sigma_{\beta 1}, \cdots, \sigma_{\beta L}, \sigma_{\xi 1}, \cdots, \sigma_{\xi L}, \nu_\phi, \sigma_\varepsilon, \nu_\kappa) \propto$$
$$p(\mu)p(\sigma_\alpha)p(\sigma_{\beta 1}, \cdots, \sigma_{\beta L})p(\sigma_{\xi 1}, \cdots, \sigma_{\xi L})p(\sigma_\varepsilon)p(\nu_\phi)p(\nu_\kappa),$$

where

- $p(\mu)$ is a diffuse prior distribution $N(0, 100)$;
- $p(\sigma_\epsilon) \propto \frac{1}{\sigma_\epsilon}$;
- $p(\sigma_\alpha) \propto 1$;
- $p(\sigma_{\beta 1}, \cdots, \sigma_{\beta L}) \propto 1$;
- $p(\sigma_{\xi 1}, \cdots, \sigma_{\xi L}) \propto 1$;
- $p(\nu_\phi) \propto \frac{1}{\nu_\phi^2}$;
- $p(\nu_\kappa) \propto \frac{1}{\nu_\kappa^2}$.

The joint posterior probability density function is

$$\begin{aligned} &p(\mu, \{\alpha_i\}, \{\beta il\}, \{\psi_{ilt}\}, \{\lambda\}, \{\kappa_{ijk}\}, \sigma_\alpha, \sigma_{\beta 1}, \cdots, \sigma_{\beta L}, \sigma_{\xi 1}, \cdots, \sigma_{\xi L}, \nu_\phi, \sigma_\varepsilon, \nu_\kappa | \{y_{ijk}\}) \\ &\quad \propto p(\mu)p(\nu_\kappa)p(\nu_\phi)p(\sigma_\epsilon) \prod_{i \in \mathcal{I}} N(\alpha_i | 0, \sigma_\alpha) \prod_{l=1}^{L} \prod_{t \in \mathcal{L}_l} N(\beta_{lt} | 0, \sigma_{\beta l}) \\ &\qquad \times \prod_{i \in \mathcal{I}} \text{Gamma}(\lambda_i | \frac{1}{2}\nu_\phi, \frac{1}{2}\nu_\phi) \prod_{i \in \mathcal{I}} \prod_{l=1}^{L} \prod_{t \in \mathcal{L}_{il}} N(\psi_{ilt} | 0, \sigma_{\xi l}^2 / \lambda_i) \\ &\qquad \times \prod_{(i,j) \in \mathcal{IJ}} \prod_{k=1}^{K_{ij}} \text{Gamma}(\kappa_{ijk} | \frac{1}{\nu_\kappa}, \frac{1}{2}\nu_\kappa) N(y_{ijk} | \mu_{ij}, \frac{\sigma_\epsilon^2}{\kappa_{ijk}}), \end{aligned} \tag{6.2.3}$$

where $\mathcal{I}$ is the set of all chemicals i in the database; $\mathcal{IJ}$ is the set of all chemical-species combinations (i, j) in the database; $\mathcal{L}_l$ is the set of classifications at level l for species in the database and $\mathcal{L}_{il}$ is the set of classifications at level l in the database tested on chemical i. The database used here contains a wide variety of chemicals and aquatic species. It has 8997 records involving 1896 chemicals

with 'CAS' number. Each species is classified into 4 taxonomical levels: Phylum-division, Class, Order and Latin. In order to make statistical inferences, the above posterior distribution needs to be calculated. Considering its complexity structure and high dimensionality, two recently developed computation packages, 'rstan' and 'MCMCglmm', are used to draw posterior samples as they are known to deal with complicated and high-dimensional models.

6.2.1 Stan

Stan, a software which implements NUTS (No-U-Turn-Sampler), could be used directly to simulate the posterior probability density function of the constructed model; see section 7.1 for a detailed description of NUTS. The model code which is fed to the argument of the 'Stan' function provided by R package 'rstan(version:2.2.0)' is displayed in Appendix D. The performance, however, is very poor. The following rejection warning message is obtained almost in every iteration

```
"Informational Message: The current Metropolis proposal is about to be
rejected becuase of the following issue:
Error in function stan::prob::normal_log(N4stan5agrad3varE): Location
parameter[764] is -nan:0, but must be finite! If this warning occurs
sporadically, such as for highly constrained variable types like co-
variance matrices, then the sampler is fine, but if this warning oc-
curs often then your model may be either severely ill-conditioned or
misspecified."
```

Clearly, even with an advanced MCMC sampler, we still cannot obtain reliable posterior samples for such a sophisticated model involving 20316 latent parameters contained in the location component μ_{ij} and 12 hyper-parameters if we crudely apply the MCMC method to the entire parameter set.

6.2.2 Modified MCMCglmm

The MCMCglmm package (Hadfield et al., 2010) was developed to implement MCMC sampling methods for generalized linear mixed models. It classifies parameters as two components: 1)linear predictors; 2) covariance structures for fixed and random effects

in the linear predictors. Generally, computation steps are iterated between these two components by using the conditional distributions. It allows response variables to follow many distributions, e.g. Gaussian, Poisson and exponential, but Student's t-distributed response variables are not considered. Craig (2013) exploited the idea of MCMCglmm and modified it to make it suitable for the model considered here with Student's t-distributed errors. The modified algorithm, which is used to simulate the chosen model by sampling iteratively between its corresponding conditional distribution of linear predictors and the rest of the hyper-parameters, is restated here.

1. Simulate the linear predictors:

   Conditional on $\sigma_\alpha, \{\sigma_{\beta l}\}_{l=1,\cdots,L}, \{\sigma_{\xi l}\}_{l=1,\cdots,L}, \{\kappa_{ijk}\}, \{\lambda_i\}$, the model could be expressed in the following matrix form

   $$Y = X\theta + z$$

   where θ is a column vector containing all the predictors and has a prior distribution $\theta \sim N(\theta_0, \Sigma)$; X is the design matrix; z is also a column vector and has prior distribution $z \sim N(0, R)$. The posterior distribution for θ is

   $$\begin{aligned} p(\theta|Y) &\propto N(\theta|\theta_0, \Sigma) \times N(Y|X\theta, R) \\ &\propto \exp\Big(-\frac{1}{2}\big(\theta^T(\Sigma^{-1} + X^T R^{-1} X)\theta - 2(\Sigma^{-1}\theta_0 + X^T R^{-1} Y)\theta\big)\Big) \\ &= N\Big(C^{-1}(\Sigma^{-1}\theta_0 + X^T R^{-1} Y), C^{-1}\Big) \end{aligned} \tag{6.2.4}$$

   where $C = \Sigma^{-1} + X^T R^{-1} X$. Simulation from distribution displayed in Equation (6.2.4) is achieved by the following algorithm:

Algorithm 5 Sampling the linear predictor

1: Simulate $\theta^\star$ from $N(\theta_0, \Sigma)$ and $\varepsilon^\star$ from $N(0, R)$.

2: Set $Y^\star = X\theta^\star + \varepsilon^\star$.

3: Compute $\tilde{\theta} = C^{-1} X^T R^{-1}(Y - Y^\star)$.

4: Set $\theta = \tilde{\theta} + \theta^\star$.

The reason that these steps give a correct simulation is shown below:

$$\begin{aligned} C\theta &= C\tilde{\theta} + C\theta^{\star} \\ &= X^T R^{-1}(Y - Y^{\star}) + \Sigma^{-1}\theta^{\star} + X^T R^{-1} X\theta^{\star} \\ &= X^T R^{-1} Y - X^T R^{-1} X\theta^{\star} - X^T R^{-1}\varepsilon^{\star} + \Sigma^{-1}\theta^{\star} + X^T R^{-1} X\theta^{\star} \\ &= X^T R^{-1} Y + \Sigma^{-1}\theta_0 - \Sigma^{-1}\theta_0 - X^T R^{-1}\varepsilon^{\star} + \Sigma^{-1}\theta^{\star} \\ &= X^T R^{-1} Y + \Sigma^{-1}\theta_0 + \Sigma^{-1}(\theta^{\star} - \theta_0) - X^T R^{-1}\varepsilon^{\star} \end{aligned} \tag{6.2.5}$$

In the above equation, the first two terms are constant; the third and final term follow $N(0, \Sigma^{-1})$ and $N(0, X^T R^{-1} X^T)$ respectively. Obviously, $C\theta$ follows $N(\Sigma^{-1}\theta_0 + X^T R^{-1} Y,\ \Sigma^{-1} + X^T R^{-1} X)$. Therefore, Equation (6.2.4) is satisfied. Note that the sparseness of C according to its definition makes its inverse matrix solved efficiently by the sparse Cholesky decomposition provided by R package 'Matrix'. The details of calculation for C^{-1} are described in Craig's technical report.

2. Simulate other parameters:

   Conditional on the linear predictors obtained in step 1, the simulations for parameters $\sigma_\alpha, \{\sigma_{\beta l}\}_{l=1,\cdots,L}, \{\sigma_{\xi l}\}_{l=1,\cdots,L}, \{\kappa_{ijk}\}, \{\lambda_i\}, \nu_\kappa, \nu_\phi$ are provided by Gibbs sampling which iterates by sampling in succession from the conditional distribution of each parameter given current values of other parameters. The full conditional distributions for all of them could be recognised as known distribution families except that for ν_κ and ν_ϕ. Therefore, the simulations for ν_κ, ν_ϕ are done by using the random-walk Metropolis-Hastings method. According to the joint posterior shown in Equation (6.2.3), the full conditional distributions are displayed as follows,

   for $\{\kappa_{ijk}\}$ block:

$$\kappa_{ijk}|\text{others} \sim \Gamma\Big(\frac{1}{2}(\nu_\kappa + 1), \frac{1}{2}\big(\nu_\kappa + \frac{(y_{ijk} - \mu_{ij})^2}{\sigma_\varepsilon^2}\big)\Big) \tag{6.2.6}$$

   for σ_ε: After making transformation $\tau_\varepsilon = \frac{1}{\sigma_\epsilon^2}$,

$$\tau_\varepsilon|\text{others} \sim \Gamma\Big(\frac{1}{2}\sum_{(i,j)\in\mathcal{IJ}} K_{ij}, \frac{1}{2}\sum_{(ij)\in\mathcal{IJ}}\sum_{k=1}^{K_{ij}} \kappa_{ijk}(y_{ijk} - \mu_{ij})^2\Big)$$

   for σ_α: After making transformation $\tau_\alpha = \frac{1}{\sigma_\alpha^2}$,

$$\tau_\alpha|\text{others} \sim \Gamma\Big(\frac{1}{2}\big(|\mathcal{I}| - 1\big), \frac{1}{2}\sum_{i\in\mathcal{I}} \alpha_i^2\Big)$$

for $\{\sigma_{\beta l}\}$ block: After making transformation $\tau_{\beta l} = \frac{1}{\sigma^2_{\beta l}}$,

$$\tau_{\beta l} \sim \Gamma\Big(\frac{1}{2}\big(|\mathcal{L}_l| - 1\big), \frac{1}{2}\sum_{t\in\mathcal{L}_l}\beta^2_{lt}\Big),\ l = 1, \cdots, L$$

for $\{\sigma_{\xi l}\}$ block: After making transformation $\tau_{\xi l} = \frac{1}{\sigma^2_{\xi l}}$,

$$\tau_{\xi l} \sim \Gamma\Big(\frac{1}{2}\big(\sum_{i\in\mathcal{I}}|\mathcal{L}_{il}| - 1\big), \frac{1}{2}\sum_{i\in\mathcal{I}}\sum_{t\in\mathcal{L}_{il}}\frac{\psi^2_{ilt}}{\phi^2_i}\Big),\ l = 1, \cdots, L$$

for $\{\lambda_i\}$ block:

$$\lambda_i|\text{others} \sim \Gamma\Big(\frac{1}{2}\big(\nu_\phi + \sum_{l=1}^{L}|\mathcal{L}_{il}|\big), \frac{1}{2}\big(\nu_\phi + \sum_{l=1}^{L}\sum_{t\in\mathcal{L}_{il}}\frac{\psi^2_{ilt}}{\sigma^2_{\xi l}}\big)\Big) \tag{6.2.7}$$

for ν_κ:

$$p(\nu_\kappa|\text{others}) \propto \frac{1}{\nu^2_\kappa}\Big(\frac{(\nu_\kappa/2)^{\nu_\kappa/2}}{\Gamma(\nu_\kappa/2)}\Big)^{\sum_{(i,j)\in\mathcal{J}_\mathcal{I}}K_{ij}}\Big(\prod_{(i,j)\in\mathcal{J}_\mathcal{I}}\prod_{k=1}^{K_{ij}}\kappa_{ijk}\Big)^{\nu_\kappa/2}$$
$$\exp\{-\frac{1}{2}\nu_\kappa\sum_{(i,j)\in\mathcal{J}_\mathcal{I}}\sum_{k=1}^{K_{ij}}\kappa_{ijk}\}$$

for ν_ϕ:

$$p(\nu_\phi|\text{others}) \propto \frac{1}{\nu^2_\phi}\Big(\frac{(\nu_\phi/2)^{\nu_\phi/2}}{\Gamma(\nu_\phi/2)}\Big)^{\sum_{i\in\mathcal{I}}\lambda_i}\Big(\prod_{i\in\mathcal{I}}\lambda_i\Big)^{\nu_\phi/2}\exp\{-\frac{1}{2}\nu_\phi\sum_{i\in\mathcal{I}}\lambda_i\}$$

The simulations for ν_κ and ν_ϕ are obtained by using the random-walk Metropolis-Hastings method as their conditional distribution cannot be recognized to some known distribution families. In the burn-in period, the scale of proposal distribution is tuned by using the acceptance rate. Briefly, if current observed acceptance rate is lower than a given lower bound, then the scale is reduced by half; if the observed accepted rate is higher than a given upper bound, then the scale is doubled. Roberts et al. (2001) stated that Metropolis-Hastings MCMC algorithms with acceptance rate between 0.15 and 0.5 is at least 80% efficient. We therefore set the above mentioned lower bound and upper bound to be 0.15 and 0.5 respectively. Other algorithm parameters needed in the random-walk Metropolis-Hastings method and the resulting acceptance rates for ν_κ, ν_ϕ are displayed in the following grey box.

Number of Iterations: $N = 20000$;

Burn-in: burn $= 2000$;

Thin: thin $= 1$;

The scale of proposal distribution tuned during the burn-in period:

- for ν_κ: 0.025
- for ν_ϕ: 0.05

The accepted rate of the main iterations:

- for ν_κ: 0.3197
- for ν_ϕ: 0.2929

Some improvements are achieved by MCMCglmm method compared to the results given by Stan in the previous section. However, posterior samples have high auto-correlations for most parameters, especially those for $\nu_\kappa, \nu_\phi, \sigma_\varepsilon, \{\sigma_{\beta l}\}_{l=1,\cdots,L}$. The trace-plot and auto-correlation plot for ν_κ and ν_ϕ are displayed in Figure 6.1. The sticky behaviour indicates the low efficiency of the algorithm. In order to measure number of independent samples in

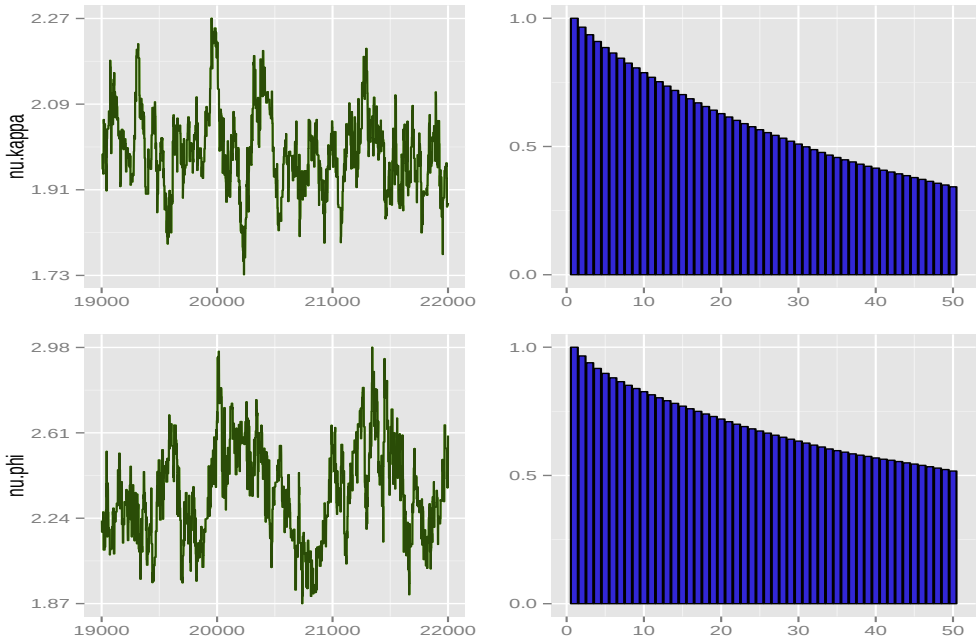


Figure 6.1: Trace-plot and Auto-correlations of ν_κ and ν_ϕ

the simulations, ESS (effective sample size) is used to show the efficiency of the algorithm. Particularly, ESS, which is closely related to auto-correlations, is defined as

$$ESS = \frac{N}{1 + 2\sum_{k=1}^{\infty} \rho_k}$$

where N is Number of iterations and ρ_k is auto-correlations at lag k. In 'coda', an R package, ESS is provided by function 'effectiveSize' which fits an autoregressive (AR) model to calculate ESS. In order to obtain a reliable value for ESS, we make some transformations to our simulation results if they are skewed heavily. The simulations of $\sigma_{\beta 1}$ and $\sigma_{\beta 2}$, which correspond to taxonomical level 'Class' and 'Phylum-division' respectively, have obvious skewness (as shown by red curves in Figure 6.2). Therefore, we take square root of simulations for them and calculate ESS after transformations.

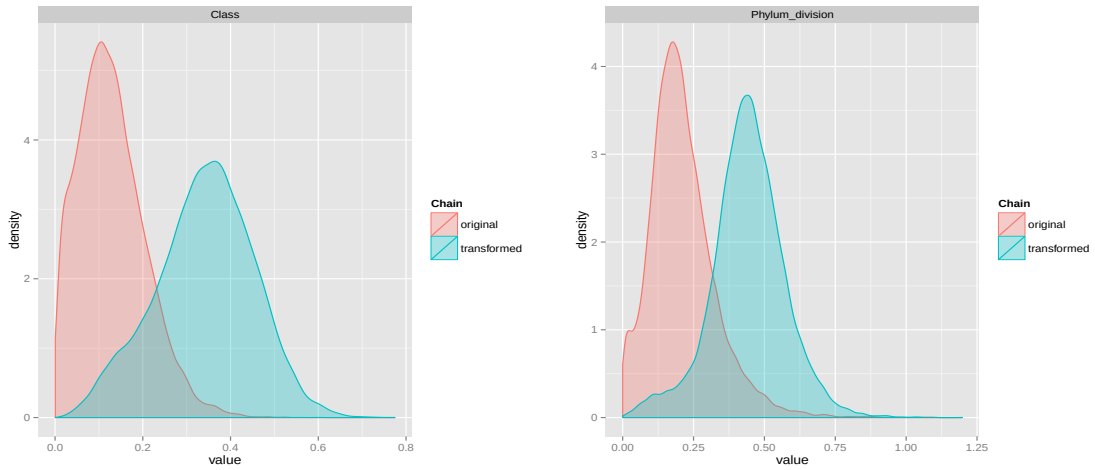


Figure 6.2: Density plots for the original and transformed samples for $\sigma_{\beta 1}$ and $\sigma_{\beta 2}$.

In the Table 6.1, auto-correlations at lag $1, 5, 10$ and effective sample size for all hyper-parameters are provided in ascending order of effective sample size value. Extreme high auto-correlations exist in the samples for $\{\sigma_{\xi l}\}_{l=1,\cdots,L}, \nu_\phi, \sigma_\varepsilon, \nu_\kappa$ as shown in this table. Therefore, we need to change our computation strategy and try other more efficient algorithms in order to obtain high quality simulations.

parameters	lag 1	lag 5	lag 10	ESS
$\sigma_{\xi 4}$ (Latin:CAS)	0.990	0.950	0.902	103
$\sigma_{\xi 3}$ (Order:CAS)	0.981	0.916	0.843	149
ν_{ϕ}	0.966	0.881	0.814	154
$\sigma_{\xi 1}$ (Phylum-division:CAS)	0.983	0.918	0.848	176
$\sigma_{\xi 2}$ (Class:CAS)	0.981	0.910	0.835	182
ν_{κ}	0.965	0.864	0.769	242
σ_{ε}	0.952	0.814	0.709	248
$\sigma_{\beta 2}$ (Class)	0.919	0.705	0.555	316
$\sigma_{\beta 4}$ (Latin)	0.881	0.567	0.356	1040
$\sigma_{\beta 3}$ (Order)	0.882	0.568	0.349	1074
σ_{α} (CAS)	0.504	0.222	0.144	1923
$\sigma_{\beta 1}$ (Phylum-division)	0.671	0.227	0.123	1934
μ	0.075	0.022	0.022	11323

Table 6.1: Auto-correlations and ESS

Chapter 7

Advanced MCMC

In this chapter, we review two advanced MCMC algorithms, NUTS (No-U-Turn Sampler) and RMHMC (Riemann Manifold Hamiltonian Monte Carlo), which are made use of later to improve simulation quality for the model described in the previous chapter. As discussed in Chapter 5, Hamiltonian Monte Carlo is a powerful MCMC tool which suppresses random walk behaviour by taking advantage of Hamiltonian dynamic system. In spite of the potential efficiency provided by this scheme of HMC, tuning of HMC algorithm parameters, ϵ (step-size), l (path length) and M (variance matrix of 'Momentum'), is still an important issue which is influential on the efficiency of the algorithm. NUTS and RHMC are two HMC variants that are designed to automatically tune algorithm parameters l and M respectively.

7.1 NUTS

Hoffman and Gelman (2011) proposed NUTS, a relatively new MCMC method which extends HMC to eliminate the need of hand-tuning l by users. The trajectories, that are used to approximate the exact dynamic flow satisfying the Hamiltonian equations, are numerically calculated by the leap-frog integrator. Obviously, to implement this sampler, one must choose an appropriate length for these trajectories to reach distant proposals efficiently. The need for such a special choice limits the routine use of HMC, and inhibits the development of software that automatically construct an HMC sampler. As commented by Hoffman and Gelman (2011), calculating the length of the simulated trajectory is not an easy task. A trajectory that is too short might result in a high auto-correlated chain

that turns out to have low efficiency. A too long trajectory might cause the chain to trace back. The trace back behaviour not only leads to a waste of computation but also results in low efficiency due to proposals which go back towards the current states. This fact is shown in Figure 7.1, in which we consider a bivariate Gaussian distribution

$$N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1.8 & 0.99\\0.99 & 1.8\end{pmatrix}\right)$$

as the target distribution. The blue curves are the simulated trajectories which start from

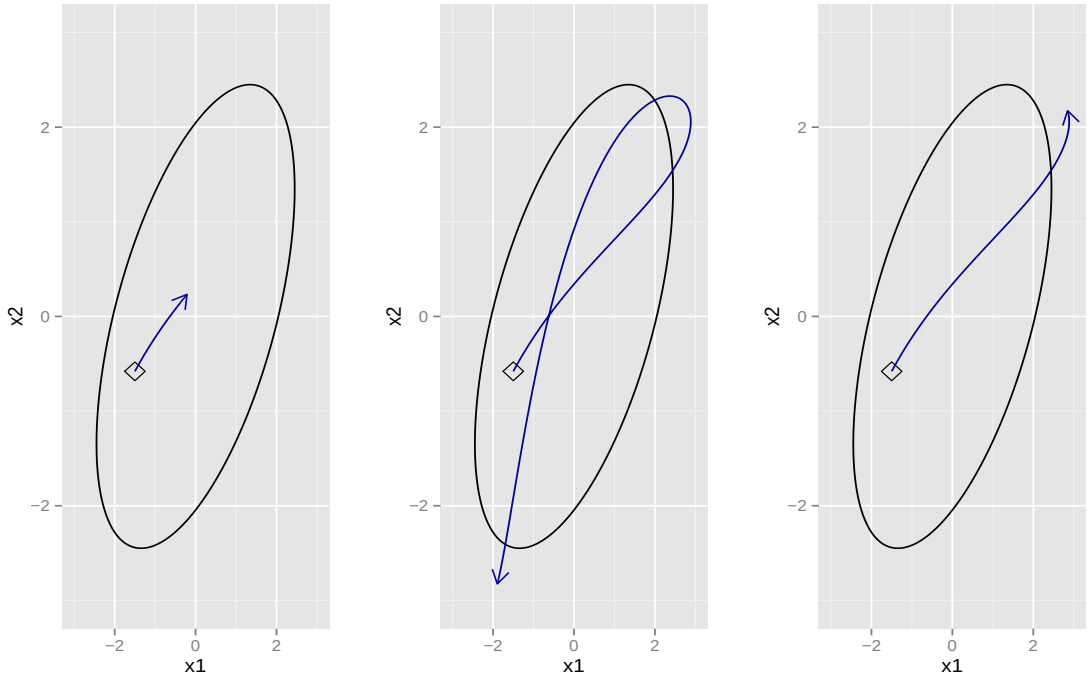


Figure 7.1: Trajectories with different lengths. The black contour is the target bivariate Gaussian distribution. Simulated trajectories are displayed by blue curves with same starting point marked by □ and ending points marked by arrows.

same point $(-1.5, -0.58)$ marked by a 'square' and ends at different points marked by arrows. The 'momentum' variable generated from standard bivariate Gaussian distribution is $(1.52, 1.22)$ and step-size is chosen as 0.08. In the leftmost plot, the trajectory, which has the ability to move to further place, is halted too early. In middle plot, the path starts to trace back to the initial point due to a too long trajectory. The rightmost plot shows the trajectory with the 'just right' length to move to the furthest place without a waste of computation.

The real difficulty is that some areas require small l while some need large l. Therefore, a fixed global l which is always right for every state to move to the most distant place along the simulated trajectory might not exist. The commonly used adaptive methods which tune l only in the burn-in period would be inadequate to maximize the potential ability of HMC. Moreover, unlike HMC with stochastic step-size (introduced in Chapter 5) and RMHMC (described in section 7.2), it is difficult to make use of some geometrical tools like the local curvature to tune the length of trajectory. In Figure 7.2, the target distribution is a simple bivariate Gaussian distribution which has constant curvature throughout the whole state space. Two blue trajectories are initialized from different points: the one in

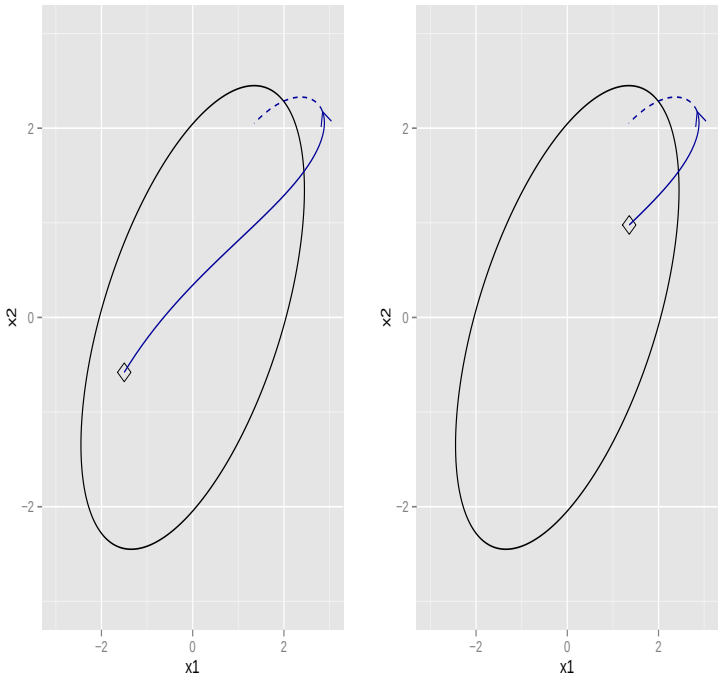


Figure 7.2: Trajectories with different starting points. The simulated trajectories are shown by solid blue curves starting from different points marked by □ and ending at same point marked by arrows. The length of trajectories are $l = 40$ and $l = 20$ respectively. The dotted lines show paths after stopping points.

the left hand-side plot is initialized from point $(-1.5, -0.58)$; the one in the right hand-side plot starts from point $(1.36, 0.98)$. The best stopping points are marked by arrows since the trajectories start to trace back as shown by the dotted lines after arrows. Conspicuously, the left hand-side plot needs a larger l while a smaller l is adequate for the right hand-side trajectory. It is noteworthy that the trace-back behaviour is due to the periodical

feature of the solutions for the Hamiltonian equations as illustrated in Equations (5.6.50). The period of the trajectory can be worked out analytically in such a trivial example while in most real computation problems it cannot be derived. In fact, the period of the Hamiltonian solution is not useful in determining the length l of the trajectory. Because it is where the trajectory starts that dominates the tuning as shown in Figure 7.2 in which two trajectories have the same period but should have different l due to different starting points.

In fact, maximizing the ability of HMC in terms of length of trajectory is identical to keeping moving trajectories until an appropriate stopping point is achieved. Therefore, a criterion that judges whether the trajectory has reached a far enough point is necessary during main iterations. Denote $(\boldsymbol{\theta}, \boldsymbol{p})$ and $(\boldsymbol{\theta}', \boldsymbol{p}')$ as states where a trajectory starts and currently reaches respectively. The criterion used in NUTS is based on the dot product between the vector $\boldsymbol{\theta}' - \boldsymbol{\theta}$ and $\boldsymbol{p}'$, i.e.

$$(\boldsymbol{\theta}' - \boldsymbol{\theta}) \cdot \boldsymbol{p}' \tag{7.1.1}$$

Once the above dot product changes the sign to negative, it indicates that the 'momentum' variables start to pull the trajectory back to its initial point and thus we should stop the trajectory.

7.1.1 Reversibility

After confirming a stopping rule that seeks an appropriate length of the simulated trajectory, a problem with this rule comes into view. Such a rule cannot retain the reversibility which is mandatory for a MCMC algorithm to converge to the desired distribution. We illustrate the irreversibility in Figure 7.3. In the left hand-side plot, the trajectory is initialized from point $(-1.5, -0.58)$ which is marked by 'square'; and it is terminated at the point $(2.88, 2.10)$ according to the stopping rule in Equation (7.1.1) since the value of that formula is -0.245. Let us consider the reverse trajectory in the right hand-side plot. The trajectory is started from point $(2.88, 2.10)$ that is the terminated point of the trajectory in the left hand-side plot. The blue curve violates the reversibility as it passes by the point $(-1.5, -0.58)$ (initial point of the trajectory in the left hand-side plot) and terminates at point $(-2.38, -2.64)$ according to the stopping rule.

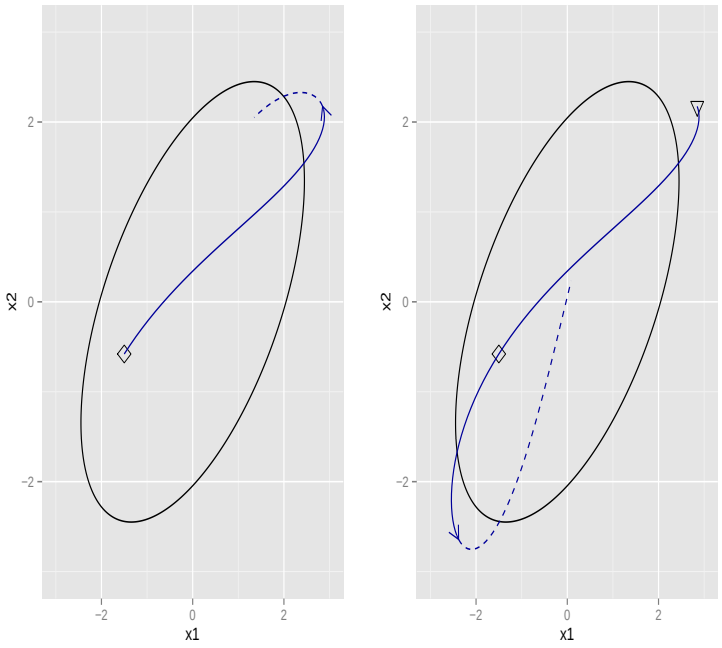


Figure 7.3: Non-reversible trajectory. Blue solid curves with arrows pointing at terminated points represent simulated trajectories. □ and ▽ represent starting points of the trajectory of the left and right plot respectively. Dotted lines show paths after terminated points.

The strategy used in NUTS to recover the reversibility is called doubling procedure. This procedure builds a tree as shown in Figure 7.4. The doubling step at each tree level j is implemented by moving the trajectory 2^{j-1} leap-frog steps after choosing a direction (backward or forward marked by the red arrows) uniformly. The development from level j to level $j+1$ is completed recursively. Suppose the tree currently has j levels. It develops the $(j+1)$-th level by recursively calculating two $(j-1)$-level sub-tree. For example, suppose that the tree has 2 levels. It grows level 3 by adding two 1-level sub-trees with nodes marked by 3. For a tree with j levels, it contains $2^j - 1$ balanced binary sub-trees marked by the blue dashed lines. At level j, the new double procedure increases the number of sub-trees by 2^{j-1}. For example, the number of sub-trees are increased from 3 to 7 after the doubling procedure at level 3. After each doubling, the stopping rule (Equation (7.1.1)) needs to be tested. More specifically, the leftmost and rightmost points of each sub-trees are examined by the stopping rule. Therefore, the stopping rule is tested by $2^j - 1$ times if the tree has j levels. Once the stopping criterion is satisfied by a pair of points, the tree evolution stops. The level that contains the points causing the stopping is

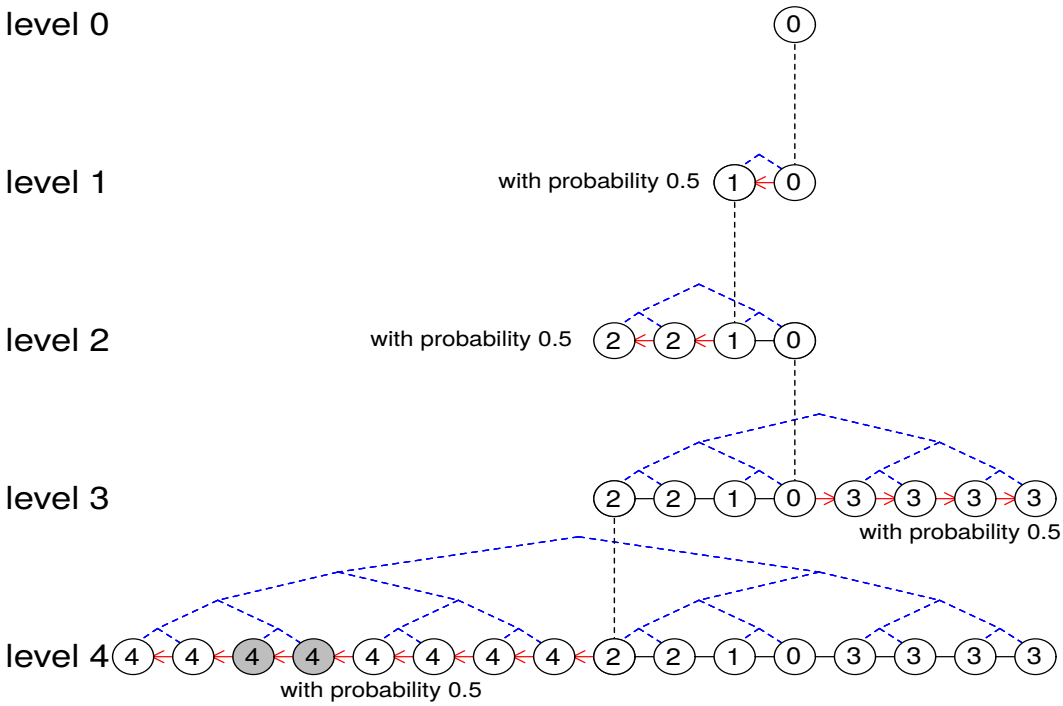


Figure 7.4: Tree evolution. This illustrated 4-level tree is constructed by 4 doubling steps. Starting from the initial point (recorded by 0) located at level 0 (the root of the tree), after randomly choosing the direction the trajectory moves 2^0 step backward from node 0 to node 1 at level 1. If the two nodes at level 1 does not satisfy the stopping rule, the tree is growing to level 2 where the trajectory moves 2^1 steps backward from node 1 to the leftmost node 2. If the nodes at level 2 do not meet the stopping criterion, the tree grows to level 3 and so on.

removed from the tree; we return to the last level and uniformly select one point as the proposal of this MCMC iteration. For example, suppose that the pair containing the two grey nodes shown in the Figure 7.4 meets the stopping criterion. The tree evolution is stopped and we then delete level 4, and uniformly select one node from level 3 (2^3 nodes available) as terminating point of the simulated trajectory.

We briefly state how the reversibility is guaranteed by such a tree constructed via the doubling procedure; see Hoffman and Gelman (2011) for strict proof. Suppose that the stopping arises at level 4 as previously assumed and the point uniformly selected from level 3 is coloured by red as shown in the top plot of Figure 7.5. In other words, the trajectory is started from node 0 (yellow node) and terminated at node 3 (red node). The

last level with gray nodes are removed from the tree due to the fact that it contains points satisfying the stopping rule. Although being removed, the last backward movement still contributes to the whole tree development (level 0 to level 3) since the final tree might be different if the direction is chosen to be forward after level 3. Therefore, the probability of transferring from node 0 to the red node 3 is

$$\underbrace{\frac{1}{2}}_{\text{backward}} \times \underbrace{\frac{1}{2}}_{\text{backward}} \times \underbrace{\frac{1}{2}}_{\text{forward}} \times \underbrace{\frac{1}{2}}_{\text{backward and removed}} \times \underbrace{\frac{1}{8}}_{\text{uniform selection}} .$$

In the bottom plot of Figure 7.5, the tree is initiated from node 3 (the terminated state of the top tree) at level 0. This tree can be built by moving forward once and backward three times in order. The structure of the sub-trees marked by the blue dashed lines in the bottom plot is identical to that in the top plot. Therefore, the pairs of nodes being tested by the stopping rule in the bottom plot are exactly the same with those in the top plot. This indicates that the tree keeps on growing until level 4. For the bottom plot, the probability of obtaining this tree and terminating trajectory at node 0 is

$$\underbrace{\frac{1}{2}}_{\text{forward}} \times \underbrace{\frac{1}{2}}_{\text{backward}} \times \underbrace{\frac{1}{2}}_{\text{backward}} \times \underbrace{\frac{1}{2}}_{\text{backward and removed}} \times \underbrace{\frac{1}{8}}_{\text{uniform selection}} .$$

That is, the probability of moving from the yellow node to the red node is the same as that of moving from the red one to the yellow one under this tree and thus the reversibility is recovered.

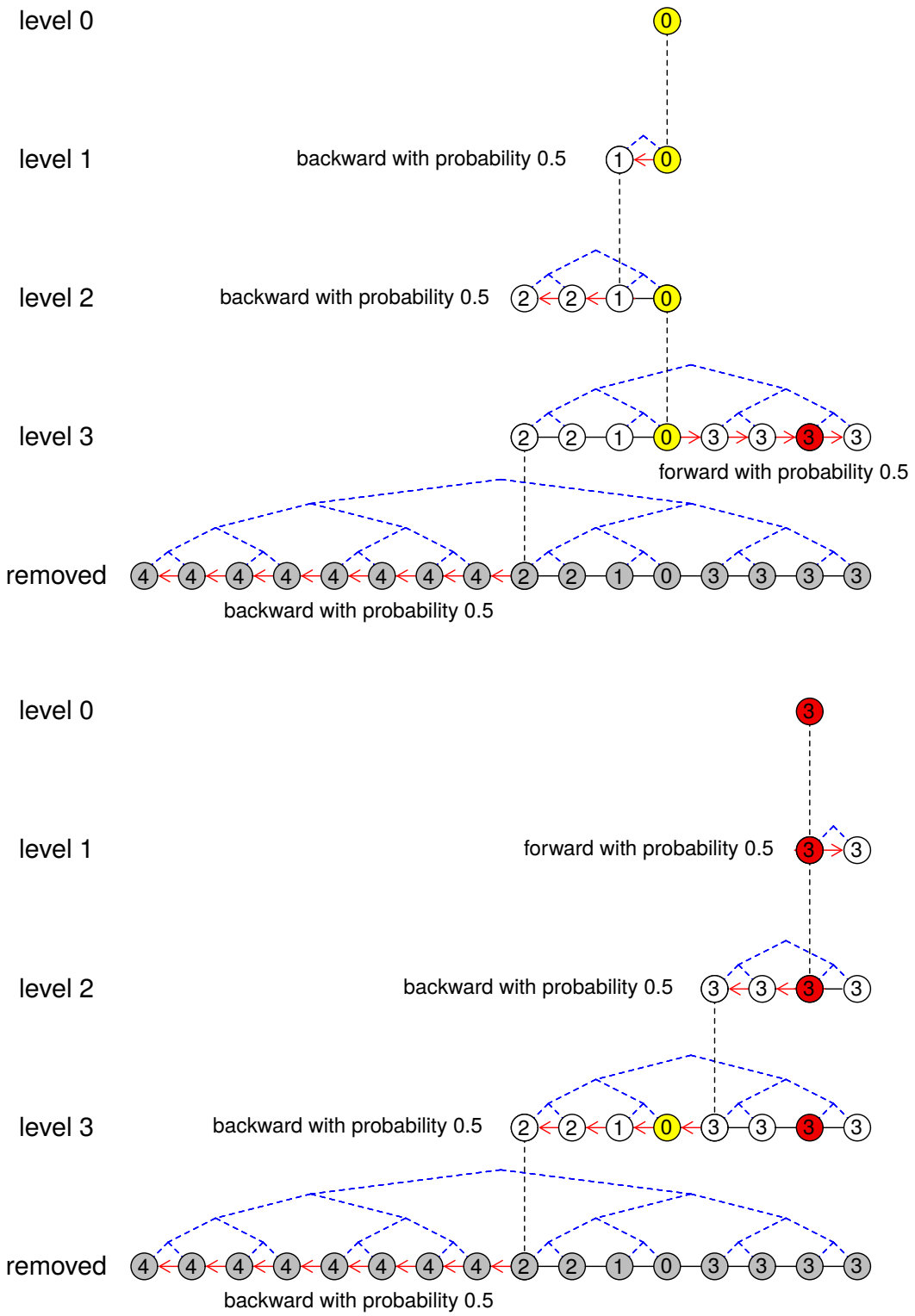


Figure 7.5: Reversibility of Tree

7.2 RMHMC

RMHMC, proposed by Girolami and Calderhead (2011), exploits the geometric information of statistical models to choose M (the variance matrix of 'momentum' variables) and thus eliminates the needs for hand-tuning of it. Moreover, the choice for M depends on $\boldsymbol{\theta}$. HMC has been demonstrated to have extraordinary potential ability to provide distant proposals owing to its special proposal strategy that takes advantages of the Hamiltonian dynamic flow by augmenting an auxiliary variable to establish the Hamiltonian system. In fact, the augmented 'momentum' variable plays the role of proposal distribution that introduces randomness in HMC. Therefore, M is influential in the performance of HMC.

7.2.1 Effect of M

By reusing the 'banana' example described in section 5.4.4, we illustrate the effect of M. For the 'banana' example, we show the performance of HMC with M chosen as identity matrix and RMHMC which chooses M as the expected Hessian matrix of the log-density function. These two sampler are implemented for one iteration with 30 leap-frog steps and step-size 0.1. The simulated trajectory paths are displayed in the following figure by using the blue lines. These two trajectories are started from the same starting point $(-1, 1.5)$ marked by $+$. The black dots are terminated points.

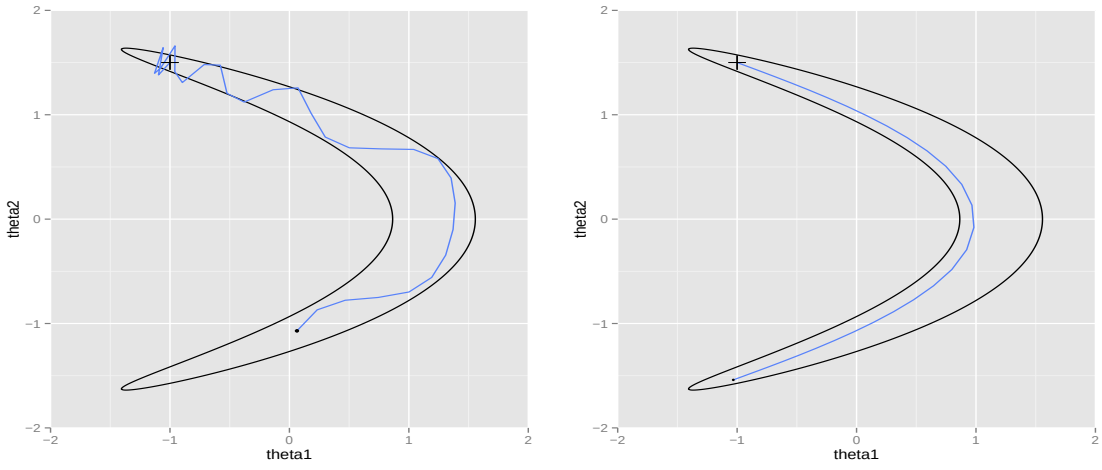


Figure 7.6: HMC and RMHMC trajectory path for one iteration. Left: HMC; Right: RMHMC.

The left-hand side plot displays the trajectory given by the HMC and the right-hand side plot displays the trajectory given by the RMHMC. The 'momentum' variables used

to construct the left-hand side trajectory are generated from $N(0, \mathbf{I})$ while the 'momentum' variables used to construct the right-hand side trajectory are generated from $N(0, \mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2}(-L)\right])$. In order to make comparisons, we use the same random number to generate 'momentum' variables. The trajectory given by the HMC has zigzag behaviour while the trajectory given by the RMHMC is much more smooth.

According to Equation 5.6.50, by locally approximating the target distribution, the trajectory can be considered as a combination of independent oscillators along directions of eigenvectors of matrix $M^{-1}\Sigma_{\boldsymbol{\theta}}^{-1}$, where $\Sigma_{\boldsymbol{\theta}}^{-1} = -\frac{\partial^2 L}{\partial \boldsymbol{\theta}^2}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^c}$. If $M = \Sigma_{\boldsymbol{\theta}}^{-1}$, then the matrix $M^{-1}\Sigma_{\boldsymbol{\theta}}^{-1}$ becomes a identity matrix. In this way, the simulation becomes easy since the target distribution is locally standardized to a standard Gaussian distribution and the curvature of the target log-density function is locally corrected to 1. Since we need to generate the 'momentum' variables from $N(0, M)$, M must be a positive-definite matrix. And this can be fixed by using the expectation of $-\frac{\partial^2 L}{\partial \boldsymbol{\theta}^2}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^c}$.

7.2.2 Implementation

As M depends on $\boldsymbol{\theta}$, we denote it by $M(\boldsymbol{\theta})$. The 'momentum' variables come from

$$\mathbf{p} \sim N(0, M(\boldsymbol{\theta}))$$

where $M(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2}(-L)\right]$. The Hamiltonian system formed by the parameters of interest and such augmented 'momentum' variables is

$$H(\boldsymbol{\theta}, \mathbf{p}) = -L(\boldsymbol{\theta}) + \frac{1}{2}\log\{|M(\boldsymbol{\theta})|\} + \frac{1}{2}\mathbf{p}^T M(\boldsymbol{\theta})^{-1}\mathbf{p} \tag{7.2.2}$$

The energy shown in Equation (7.2.2) is not separable and the corresponding Hamiltonian equations are

$$\frac{d\theta_i}{dt} = \frac{\partial H}{\partial p_i} = \{M^{-1}(\boldsymbol{\theta})\mathbf{p}\}_i$$

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial \theta_i} = \frac{\partial L}{\partial \theta_i} - \frac{1}{2}\mathrm{tr}\{M(\boldsymbol{\theta})^{-1}\frac{\partial M(\boldsymbol{\theta})}{\partial \theta_i}\} + \frac{1}{2}\mathbf{p}^T M(\boldsymbol{\theta})^{-1}\frac{\partial M(\boldsymbol{\theta})}{\partial \theta_i}M(\boldsymbol{\theta})^{-1}\mathbf{p}$$

The numerical integrator exploited to solve above differential equations is the generalized leap-frog algorithm

$$\boldsymbol{p}(t + \frac{\varepsilon}{2}) = \boldsymbol{p}(t) - \frac{\varepsilon}{2}\nabla_{\boldsymbol{\theta}} H\{\boldsymbol{\theta}(t), \mathbf{p}(t + \frac{\varepsilon}{2})\} \tag{7.2.3}$$

$$\boldsymbol{\theta}(t + \varepsilon) = \boldsymbol{\theta}(t) + \frac{\varepsilon}{2}\left[\nabla_{\mathbf{p}} H\{\boldsymbol{\theta}(t), \mathbf{p}(t + \frac{\varepsilon}{2})\} + \nabla_{\mathbf{p}} H\{\boldsymbol{\theta}(t + \varepsilon), \boldsymbol{p}(t + \frac{\varepsilon}{2})\}\right] \tag{7.2.4}$$

$$\mathbf{p}(t + \varepsilon) = p(t + \frac{\varepsilon}{2}) - \frac{\varepsilon}{2}\nabla_{\boldsymbol{\theta}} H\{\boldsymbol{\theta}(t + \varepsilon), p(t + \frac{\varepsilon}{2})\} \tag{7.2.5}$$

Equations (7.2.3) and (7.2.4) are implicit equations for $\boldsymbol{p}(t+\frac{\varepsilon}{2}), \boldsymbol{\theta}(t+\varepsilon)$ and thus extra numerical iterations are required to solve them. The method used in RMHMC to solve these implicit functions is fixed-point iterations. Suppose an implicit function $x = g(x)$. The fixed-point iteration scheme is summarized as follows.

Algorithm 6 Fix-Point Iteration

1: Given an initial guess x_0;
2: **for** $n = 0, \cdots, N$ **do**
3: $n = n + 1$;
4: $x_{n+1} = g(x_n)$;
5: **end for**

$\boldsymbol{p}(t+\frac{\varepsilon}{2}), \boldsymbol{\theta}(t+\varepsilon)$ in Equation (7.2.3) and (7.2.4) are calculated according to the above scheme. Girolami and Calderhead (2011) suggested to set N to 5 or 6 in the fix-point iteration scheme for solving the implicit function in the RMHMC algorithm.

Disadvantages of RMHMC

Admitting the perfect performance of RMHMC, it is much more computationally expensive than HMC since it requires not only significant effort in matrix calculations (such as matrix decompositions and matrix derivatives) but also in solving implicit equations. In addition, the expectation of the Hessian matrix also need efforts to derive before using this sampler. There is no good solution which could simplify RMHMC and retain its ability at the same time. In order to speed up the program which runs RMHMC, one might resort to other high-efficient programming languages such as C++ or Python.

Chapter 8

Improving Simulations for a Real Model

As discussed in section 6.2, current computational solutions to the model under consideration are unsatisfactory: Stan failed to provide us with even one simulation; modified MCMCglmm performed better than Stan but also provided very high-autocorrelated simulation results. In this chapter, we carry out a study to investigate suitable computational strategies to improve the simulation quality for the hierarchical model described in chapter 6.

As illustrated in section 6.2.1 and 6.2.2, the computation difficulties associated with such a model are due to its high-dimensional parametric space, complicated model structure and the limitations of simulation algorithms. Stan's use of both HMC and NUTS is fully justified but submitting the whole model directly to Stan causes a stuck Markov chain due to the complicated model structure (involving interactions, taxonomically related structures and partial exchangeability) and large number of parameters (31221 parameters in total). The modified MCMCglmm seems to be acceptable as it eventually provided us with a simulation result after substantial thinning (such as 100). Generally speaking, it classifies parameters according to their roles (linear predictors and variance structure parameters) in the model and then simulates them separately. The major drawback of this method when applied to the target model is the 'sticky' behaviour of simulations for variance structure parameters provided by the MCMC sampler. Particularly, MCMC simulations for the parameters $\{\nu_\kappa, \sigma_\epsilon, \nu_\phi, \sigma_{\xi 1}, \sigma_{\xi 2}, \sigma_{\xi 3}, \sigma_{\xi 4}\}$ display very high autocorrelations according to Table 6.1. With these facts in mind, our experimental set up bears

a close resemblance to the modified MCMCglmm that deals with parameters separately according to their roles. However, the following two changes are made:

1. Blocking parameters

As shown in Table 6.1, the original sampling method that simulates the variance structure parameters one at a time according to their full conditional distributions obtained highly auto-correlated posterior samples for these parameters. We therefore consider blocking these challenging variance-structural parameters

$$\{\nu_\kappa, \sigma_\epsilon, \nu_\phi, \sigma_{\xi 1}, \sigma_{\xi 2}, \sigma_{\xi 3}, \sigma_{\xi 4}\}.$$

Rather than considering all the above parameters in one single block, we further divided them into two small blocks:

$$\left\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}, \{\lambda_i\}_{i\in I}\right\}, \quad \left\{\nu_\kappa, \sigma_\varepsilon, \{\kappa_{ijk}\}_{k=1,\ldots,K_{ij};(i,j)\in\mathcal{IJ}}\right\}$$

The reason why we adopted the above partition is that the above two blocks are conditionally independent with each other. Such independence is illustrated as follows. Recall the hierarchical model illustrated in section 6.1. The structure of model is displayed in Figure 8.1. In order to obtain the conditional independence clearly, the direct acyclic graph and its corresponding moral graph are shown in Figure 8.2 and 8.3 respectively. The moral graph is obtained by connecting the nodes that have a common child. We should connect every two nodes in $\{\mu, \alpha_i, \beta_{lt_j}, \psi_{ij}, \kappa_{ijk}, \sigma_\varepsilon\}$. In order to make the moral graph clear, we use the oval with double green edge to mean that every two nodes on its edge are connected. In Figure 8.3, the variables inside the red dotted ellipse are given by Algorithm 5. By looking at the moral graph in Figure 8.3, we have the following conditional independence

$$\left\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}, \{\lambda_i\}_{i\in I}\right\} \perp \left\{\nu_\kappa, \sigma_\varepsilon, \{\kappa_{ijk}\}_{k=1,\ldots,K_{ij};(i,j)\in\mathcal{IJ}}\right\} \mid \left\{\mu, \alpha_i, \beta_{lt_j}, \psi_{ij}, y_{ijk}\right\}.$$

It is natural to group those challenging parameters into two blocks:

$$\text{Block } \left\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}, \{\lambda_i\}_{i\in I}\right\} \text{ and Block } \left\{\nu_\kappa, \sigma_\varepsilon, \{\kappa_{ijk}\}_{k=1,\ldots,K_{ij};(i,j)\in\mathcal{IJ}}\right\}.$$

The parameter ν_ϕ controls variance of λ_i and the whole block $\left\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}, \{\lambda_i\}_{i\in I}\right\}$ decides the degrees of freedom and the variance of the t-distributed random effect interaction ψ_{ij}. Similarly, ν_κ governs all of κ_{ijk} and the whole block $\left\{\nu_\kappa, \sigma_\varepsilon, \{\kappa_{ijk}\}_{k=1,\ldots,K_{ij};(i,j)\in\mathcal{J}_\mathcal{I}}\right\}$ dictates the degree of freedom and the variance of the the t-distributed measurement error ε_{ijk}.

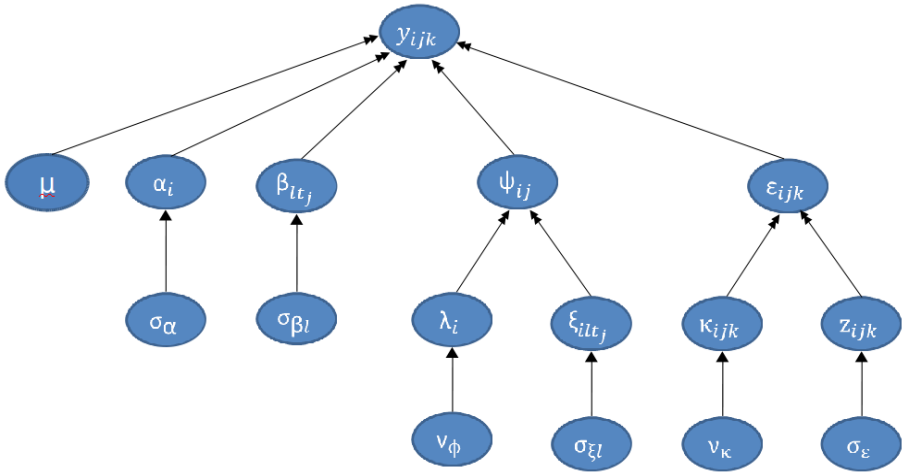


Figure 8.1: Model Structure. Double arrows represent deterministic dependencies. For example, $\varepsilon_{ijk} = \frac{z_{ijk}}{\sqrt{\kappa_{ijk}}}$.

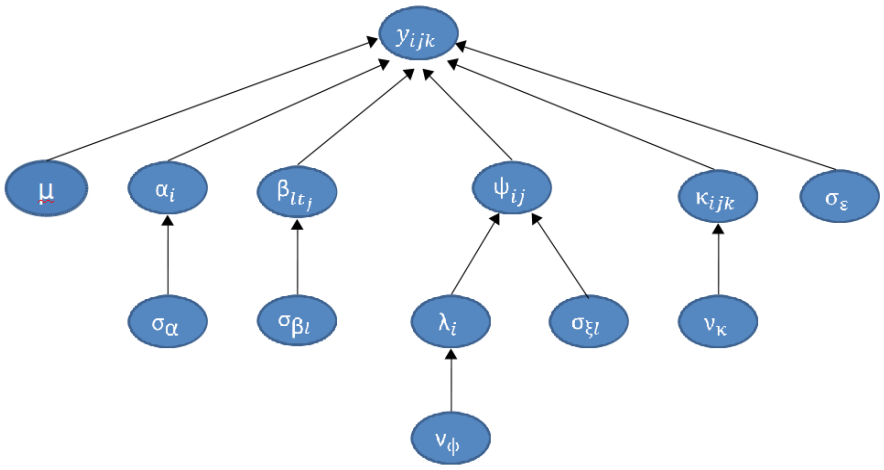


Figure 8.2: Direct Acyclic Graph

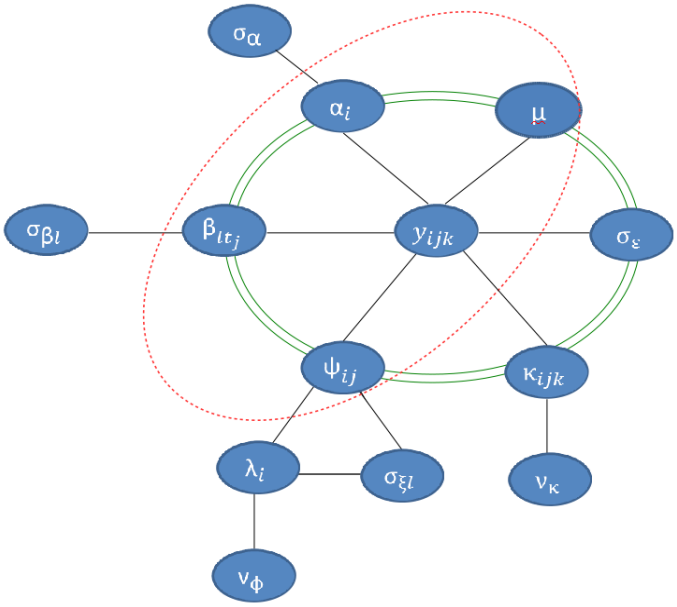


Figure 8.3: Moral Graph.

2. Sampling the marginalized conditional distributions

In order to simulate the two blocks mentioned above, we apply the marginal MCMC sampling method instead of directly using MCMC methods to the blocks. To illustrate the computation design for the full conditional distributions of these two blocks, we assume a scenario where we are interested in sampling from the distribution $p(\mathbf{x}_1, \mathbf{x}_2|\mathbf{x}_3)$, where $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ are arbitrary random variables. Consider the standard decomposition

$$p(\mathbf{x}_1, \mathbf{x}_2|\mathbf{x}_3) = p(\mathbf{x}_1|\mathbf{x}_3)p(\mathbf{x}_2|\mathbf{x}_1, \mathbf{x}_3), \tag{8.0.1}$$

where $p(\mathbf{x}_1|\mathbf{x}_3)$ is the marginalized distribution of $\mathbf{x}_1$ after integrating out $\mathbf{x}_2$ from $p(\mathbf{x}_1, \mathbf{x}_2|\mathbf{x}_3)$. We use the following sampling procedure for an update

$$T\big((\mathbf{x}_1^\star, \mathbf{x}_2^\star)|(\mathbf{x}_1, \mathbf{x}_2)\big) = T(\mathbf{x}_1^\star|\mathbf{x}_1)p(\mathbf{x}_2^\star|\mathbf{x}_1^\star, \mathbf{x}_3) \tag{8.0.2}$$

That is,

$$(\mathbf{x}_1, \mathbf{x}_2) \xrightarrow[\text{fix } \mathbf{x}_2]{\mathbf{x}_1^\star \sim p(\mathbf{x}_1|\mathbf{x}_3)} (\mathbf{x}_1^\star, \mathbf{x}_2) \xrightarrow[\text{fix } \mathbf{x}_1^\star]{\mathbf{x}_2^\star \sim p(\mathbf{x}_2|\mathbf{x}_1^\star, \mathbf{x}_3)} (\mathbf{x}_1^\star, \mathbf{x}_2^\star), \tag{8.0.3}$$

This procedure is automatically justified by the decomposition shown in Equation (8.0.1). This procedure was chosen because it is a feasible way to transform a high dimensional simulation problem into a low dimensional simulation that is much more economic to deal with. This marginal scheme is also exploited in particle MCMC proposed by Andrieu et al. (2010) for the state space model. They mentioned that, "this proposed $\mathbf{x}_2^\star$ is perfectly adapted to the proposed $x_1^\star$ and the only degree of freedom of the algorithm (which will affect its performance) is $T(\mathbf{x}_1^\star|\mathbf{x}_1)$" (Andrieu et al., 2010). And, therefore, a good sampling from the marginalized distribution for $\mathbf{x}_1$ is essential. This procedure to update $(\mathbf{x}_1, \mathbf{x}_2)$ is summarized in the following algorithmic form.

Algorithm 7 The Marginal Sampling Approach

1. Given current states $\mathbf{x}^t = \Big\{\mathbf{x}_1^t, \mathbf{x}_2^t\Big\}$, sampler's parameters Λ;
2. Set $\mathbf{x}_1^{t+1} = \text{Sampler}\Big(p(\mathbf{x}_1|\mathbf{x}_3), \boldsymbol{x}_1^t, \Lambda\Big)$;
3. Simulate $\mathbf{x}_2^{t+1} \sim p(\mathbf{x}_2|\mathbf{x}_1^{t+1}, \mathbf{x}_3)$.

where Λ denotes all parameters needed by the specific sampler implemented to simulate $\boldsymbol{x}_1$. For example, $\Lambda = \{\epsilon, \mathfrak{l}, M\}$ representing the step-size, the number of the leap-frog steps and the variance matrix of 'momentum' variables if HMC is chosen to be the sampler.

Let us now turn to our real problem: sampling from the conditional distribution of the block $\left\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}, \{\lambda_i\}_{i\in I}\right\}$ and the block $\left\{\nu_\kappa, \sigma_\varepsilon, \{\kappa_{ijk}\}_{k=1,\ldots,K_{ij};(i,j)\in\mathcal{J}_\mathcal{I}}\right\}$. We respectively denote these two conditional distributions by

$$p(\nu_\phi, \sigma_{\xi 1}, \cdots, \sigma_{\xi L}, \{\lambda\}_{i\in I}|\text{others})$$

and

$$p(\nu_\kappa, \sigma_\varepsilon, \{\kappa_{ijk}\}|\text{others})$$

where 'others' denotes the parameters not in the targeted block. For each block, we firstly sample a marginalized conditional distribution and then a conditional distribution. To be specific, for block $\left\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}, \{\lambda_i\}_{i\in I}\right\}$, the marginalized conditional distribution $p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}|\text{others})$ and the conditional distribution of $p(\{\lambda_i\}_{i\in I}|\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}, \text{others})$ are sampled in order to achieve the simulation for the conditional distribution of this block. For block $\left\{\nu_\kappa, \sigma_\varepsilon, \{\kappa_{ijk}\}_{k=1,\ldots,K_{ij};(i,j)\in\mathcal{J}_\mathcal{I}}\right\}$, the marginalized conditional distribution $p(\nu_\kappa, \sigma_\varepsilon|\text{others})$ and the conditional distribution of $p(\{\kappa_{ijk}\}_{k=1,\ldots,K_{ij};(i,j)\in\mathcal{J}_\mathcal{I}}|\nu_\kappa, \sigma_\varepsilon, \text{others})$ are sampled in order to achieve the simulation for the conditional distribution of this block. The parameters $\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}$ and parameters $\nu_\kappa, \sigma_\varepsilon$ play the role of $\mathbf{x}_1$, the parameters $\{\lambda_i\}_{i\in I}$ and $\{\kappa_{ijk}\}_{k=1,\ldots,K_{ij};(i,j)\in\mathcal{J}_\mathcal{I}}$ act as $\mathbf{x}_2$ and the parameters not in this block are $\mathbf{x}_3$ in Equation (8.0.1).

There are two reasons of adopting this marginalized sampling method. Firstly, high-dimensional simulation problems can be simplified to low-dimensional simulation problems. For example, for block $\left\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}, \{\lambda_i\}_{i\in I}\right\}$, a 1901-dimensional (1896 λ_i's, 4 $\sigma_{\xi l}$'s and 1 ν_ϕ) simulation is divided into a 5-dimensional sampling problem for $\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}$ and a sampling problem concerning 1896 λ_i's. Particularly, the simulation problem of 1896 λ_i's could be efficiently solved by sampling them all in a single line of code as the conditional distribution $p(\{\lambda_i\}_{i\in I}|\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L})$, that is the same as the one in Equation (6.2.7), belongs independently to the gamma distribution family. Secondly, $\{\lambda_i\}_{i\in I}$ and $\{\kappa_{ijk}\}_{k=1,\ldots,K_{ij};(i,j)\in\mathcal{J}_\mathcal{I}}$ play the role of latent variable; $\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}$ and $\nu_\kappa, \sigma_\varepsilon$ can be considered as hyper-parameters of these two blocks respectively conditional on 'other' parameters. Integrating out latent variables can break the correlation between hyper-parameters and latent variables and thus ease simulating difficulties of the corresponding hyper-parameters. This fact can be illustrated in the following experimental results where the posterior samples for ν_ϕ and ν_κ have been improved to a large extent (please see Figure C.6 for details).

Therefore, the main issue is to choose an appropriate sampler to sample from the marginalized conditional distribution. We apply different MCMC samplers to achieve informative simulations. Particularly, RWMH, HMC, and HMC's variants are used as the proposal methods for the desired marginalized conditional distributions.

This chapter is divided into 5 sections. In section 8.1, mathematical details that are needed in the algorithms for sampling the block $\left\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}, \{\lambda_i\}_{i\in I}\right\}$ and the block $\left\{\nu_\kappa, \sigma_\varepsilon, \{\kappa_{ijk}\}_{k=1,\ldots,K_{ij};(i,j)\in\mathcal{J}_\mathcal{I}}\right\}$ are presented. Sections 8.2 and 8.3 provide simulation results obtained by applying RWMH, basic HMC, NUTS, RMHMC and HMC with stochastic step-size sampler to the marginalized conditional distributions respectively. Section 8.4 discusses the autocorrelations left in the simulation results presented in section 8.3. In the final section, simulation results from different sampling methods are compared.

8.1 Blocking Parameters

Here, we provide the mathematical details that are required in sampling from the block $\left\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}, \{\lambda_i\}_{i\in I}\right\}$ and the block $\left\{\nu_\kappa, \sigma_\varepsilon, \{\kappa_{ijk}\}_{k=1,\ldots,K_{ij};(i,j)\in\mathcal{J}_\mathcal{I}}\right\}$ according to the marginal approach illustrated in Algorithm 7. To be specific, the marginalized distribution, its first derivatives, second derivatives and the expected Hessian matrix are provided for each block.

8.1.1 Block $\left\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}, \{\lambda_i\}_{i\in I}\right\}$

Considering the joint posterior distribution shown in Equation (6.2.3), the full conditional distribution of the parameters in block $\left\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}, \{\lambda_i\}_{i\in I}\right\}$, given parameters not in this block, can be expressed as

$$p(\nu_\phi, \sigma_{\xi 1}, \cdots, \sigma_{\xi L}, \{\lambda\}_{i\in I}|\text{others}) \propto p(\nu_\phi)\prod_{i\in\mathcal{I}}\Gamma(\lambda_i|\frac{1}{2}\nu_\phi, \frac{1}{2}\nu_\phi)\prod_{i\in\mathcal{I}}\prod_{l=1}^{L}\prod_{t\in\mathcal{L}_{il}}N(\psi_{ilt}|0, \sigma_{\xi l}^2/\lambda_i)$$

In the following section, 'other' is omitted for convenience and $p(\nu_\phi, \sigma_{\xi 1}, \cdots, \sigma_{\xi L}, \{\lambda\}_{i\in I})$ is used to denote the full conditional distribution of this block. As described in Equation (8.0.2), the strategy used to simulate the full conditional distribution $p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}, \{\lambda_i\}_{i\in I})$ is composed by two steps:

1. sampling from the marginalized conditional distribution $p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L})$ that is the result of integrating out all of $\{\lambda_i\}_{i\in I}$ from the full conditional distribution $p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}, \{\lambda_i\}_{i\in I})$;

2. given $\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}$ obtained from the previous step, sampling from the conditional distribution $p(\{\lambda_i\}_{i\in I}|\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L})$.

By using this marginal approach, a 1901-dimensional (1896 λ_i's, 4 $\sigma_{\xi l}$'s and 1 ν_ϕ) simulation is divided into a 5-dimensional sampling problem stated in step 1 and a sampling problem concerning 1896 λ_i's. Particularly, the simulation problem of 1896 λ_i's could be efficiently solved by sampling them all in a single line of code as the conditional distribution $p(\{\lambda_i\}_{i\in I}|\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L})$, that is the same as the one in Equation (6.2.7), belongs in dependently to the gamma distribution family. Therefore, the main issue is to choose an appropriate sampler to sample from the 5-dimensional marginalized conditional distribution.

Marginal Distribution

By integrating out all of λ_i (see Appendix C.1 for the integration details), we obtained the following marginalized conditional distribution $p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L})$,

$$\begin{aligned} p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}) &= \int \cdots \int p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}, \{\lambda_i\}_{i\in I})\, \mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_{|I|} \\ &= \underbrace{\frac{1}{\nu_\phi^2}}_{\text{prior}} \underbrace{\prod_{i\in I} t_{\nu_\phi}(W_i|0,\Sigma_i)}_{\text{'likelihood'}} \end{aligned} \tag{8.1.4}$$

with

$$W_i = \begin{pmatrix} W_{i1} \\ W_{i2} \\ \vdots \\ W_{il} \\ \vdots \\ W_{iL} \end{pmatrix}_{P_i,1}$$

$$\Sigma_i = \sigma^2_{\xi 1} I_{|L_{i1}|,|L_{i1}|} \oplus \sigma^2_{\xi 2} I_{|L_{i2}|,|L_{i2}|} \oplus \cdots \oplus \sigma^2_{\xi l} I_{|L_{il}|,|L_{il}|} \oplus \cdots \oplus \sigma^2_{\xi L} I_{|L_{iL}|,|L_{iL}|}$$

$$P_i = \sum_{l=1}^{L} |L_{il}|$$

where W_{il} is the column vector of length $|L_{il}|$ with entries $\{\psi_{ilt}\}_{t\in L_{il}}$; $I_{x,x}$ stands for $x \times x$-dimensional identity matrix and '$\oplus$' is direct sum. The term Σ_i represents a $P_i \times P_i$ dimensional covariance matrix. In this block, the relevant linear predictor components, which we denote by W_i, are considered as 'data' that are provided by Algorithm 5. Each $\{W_i; i \in I\}$ is independently from a P_i-dimensional t distribution with parameters Σ_i and ν_ϕ, i.e.

$$W_i \sim t_{\nu_\phi}(0, \Sigma_i)$$

Derivatives and Fisher Information Matrix

In order to sample from the above marginalized conditional distribution by using different MCMC samplers, logarithm of the marginal distribution, its first derivatives and Fisher Information matrix are required. We denote the logarithm of $p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L})$ by l_ϕ. After

some simplifying, it is given by

$$l_\phi = -2\log\nu_\phi + |I|\Big(\frac{1}{2}\nu_\phi\log\frac{1}{2}\nu_\phi - \log\Gamma(\frac{1}{2}\nu_\phi)\Big) + \sum_{i\in I}\Big\{-\frac{1}{2}\log|\Sigma_i| + \log\Gamma(\frac{1}{2}\nu_\phi+\frac{1}{2}P_i) - (\frac{1}{2}\nu_\phi+\frac{1}{2}P_i)\log\big(\frac{1}{2}\nu_\phi+\frac{1}{2}W_i^T\Sigma_i^{-1}W_i\big)\Big\} \tag{8.1.5}$$

With the above formula, it is straightforward to obtain its derivatives which are listed as follows.

- First Derivatives:

  The first derivative of l_ϕ with respect to ν_ϕ is

$$\frac{\partial}{\partial\nu_\phi}l_\phi = -\frac{2}{\nu_\phi} + \frac{1}{2}|I|\Big(1-\Psi(\frac{1}{2}\nu_\phi)\Big) + \frac{1}{2}\sum_{i\in I}\Big\{\Psi(\frac{1}{2}\nu_\phi+\frac{1}{2}P_i) - \log(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}) - \frac{\nu_\phi+P_i}{\nu_\phi}\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-1}\Big\} \tag{8.1.6}$$

  where

$$\Psi(x) = \frac{d}{dx}\log(\Gamma(x)) = \frac{\Gamma'(x)}{\Gamma(x)}$$

  Particularly, $\Psi(\cdot)$ could be calculated by calling the R function 'digamma()' directly.

  The first derivative of l_ϕ with respect to $\sigma_{\xi l}$ $(l = 1, \cdots, L)$ is

$$\frac{\partial}{\partial\sigma_{\xi l}}l_\phi \;=\; -\sum_{i\in I}\frac{|L_{il}|}{\sigma_{\xi l}} + \sum_{i\in I}\Big\{(\frac{\nu_\phi+P_i}{\sigma_{\xi l}})\Big(W_{il}^T\Sigma_{il}^{-1}W_{il}\Big)\Big(\nu_\phi + W_i^T\Sigma_i^{-1}W_i\Big)^{-1}\Big\} \tag{8.1.7}$$

  where

$$\Sigma_{il} = \sigma_{\xi l}^2 J_{|L_{il}|\times|L_{il}|}$$

$$\text{i.e. } \Sigma_i = \Sigma_{i1}\oplus\Sigma_{i2}\oplus\ldots\Sigma_{il}\oplus\ldots\Sigma_{iL}.$$

- Second Derivatives:

  According to the above first derivatives, the following second derivatives are obtained. The second derivative of l_ϕ with respect to ν_ϕ is

$$\frac{\partial^2}{\partial\nu_\phi^2}l_\phi = C_1 - \frac{1}{2}\sum_{i\in I}\Big\{\frac{\nu_\phi-P_i}{\nu_\phi^2}\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-1} + \frac{\nu_\phi+P_i}{\nu_\phi^3}(W_i^T\Sigma_i^{-1}W_i)\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-2}\Big\} \tag{8.1.8}$$

where

$$C_1 = \frac{2}{\nu_\phi^2} + \frac{1}{2}|I|\Big(\frac{1}{\nu_\phi} - \frac{1}{2}\Psi'(\frac{1}{2}\nu_\phi)\Big) + \frac{1}{4}\sum_{i\in I}\Psi'(\frac{1}{2}\nu_\phi + \frac{1}{2}P_i)$$

$$\Psi'(x) = \frac{d^2}{dx^2}\log(\Gamma(x))$$

where $\Psi'(\cdot)$ could be obtained by calling the R function 'trigamma()' directly.

The second derivative of l_ϕ with respect to ν_ϕ and $\sigma_{\xi l}$ $(l = 1, \cdots, L)$ is

$$\frac{\partial^2}{\partial\nu_\phi\partial\sigma_{\xi l}} l_\phi = \sum_{i\in I}\left\{\frac{W_{il}^T\Sigma_{il}^{-1}W_{il}}{\nu_\phi\sigma_{\xi l}}\Big(1 + \frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-1} - \frac{\nu_\phi + P_i}{\nu_\phi^2\sigma_{\xi l}}W_{il}^T\Sigma_{il}^{-1}W_{il}\Big(1 + \frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-2}\right\} \tag{8.1.9}$$

The second derivative of l_ϕ with respect to $\sigma_{\xi l}$ $(l = 1, \cdots, L)$ is

$$\frac{\partial^2}{\partial\sigma_{\xi l}^2} = \frac{\sum_{i\in I}|L_{il}|}{\sigma_{\xi l}^2} + \sum_{i\in I}\left\{\frac{-3(\nu_\phi + P_i)}{\sigma_{\xi l}^2\nu_\phi}W_{il}^T\Sigma_{il}^{-1}W_{il}\Big(1 + \frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-1} + \frac{2(\nu_\phi + P_i)}{\sigma_{\xi l}^2\nu_\phi^2}\Big(W_{il}^T\Sigma_{il}^{-1}W_{il}\Big)^2\Big(1 + \frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-2}\right\} \tag{8.1.10}$$

The second derivative of l_ϕ with respect to $\sigma_{\xi l}$ and $\sigma_{\xi j}$ with $l \neq j$ is

$$\frac{\partial^2}{\partial\sigma_{\xi l}\partial\sigma_{\xi j}}\Big|_{l\neq j} = \sum_{i\in I}\left\{\frac{2(\nu_\phi + P_i)}{\nu_\phi^2\sigma_{\xi l}\sigma_{\xi j}}\Big(W_{il}^T\Sigma_{il}^{-1}W_{il}\Big)\Big(W_{ij}^T\Sigma_{ij}^{-1}W_{ij}\Big)\Big(1 + \frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-2}\right\} \tag{8.1.11}$$

The Fisher Information matrix is essential in implementing RMHMC. The calculation of such a matrix involves some challenging integrations. The following part shows resulting Fisher Information matrix.

- Fisher Information Matrix:

  In order to achieve Fisher Information matrix, we need the expected values of the above second derivatives. According to Equation (8.1.8), the corresponding expec-

tation is

$$\begin{aligned}\mathbb{E}\Big(-\frac{\partial^2}{\partial\nu_\phi^2}l_\phi\Big) &= -\int\cdots\int\frac{\partial^2}{\partial\nu_\phi^2}l_\phi\prod_{i\in I}t_{\nu_\phi}(W_i|0,\Sigma_i)\,\mathrm{d}W_1\cdots\mathrm{d}W_{|I|}\\ &= -C_1+\frac{1}{2}\sum_{i\in I}\frac{\nu_\phi-P_i}{\nu_\phi^2}C_2+\frac{1}{2}\sum_{i\in I}\frac{\nu_\phi+P_i}{\nu_\phi^3}\,C_3\\ &= -C_1+\frac{1}{2}\sum_{i\in I}\left(\frac{\nu_\phi-P_i}{\nu_\phi(\nu_\phi+P_i)}+\frac{P_i}{\nu_\phi(\nu_\phi+P_i+2)}\right)\end{aligned}\tag{8.1.12}$$

where

$$C_2=\mathbb{E}\bigg[\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-1}\bigg]$$

$$C_3=\mathbb{E}\bigg[(W_i^T\Sigma_i^{-1}W_i)\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-2}\bigg]$$

See proposition C.2.1 and C.2.5 in Appendix C.2 for the results and derivations for C_2 and C_3.

According to Equation (8.1.9), its corresponding expectation is given by

$$\begin{aligned}\mathbb{E}\Big(-\frac{\partial^2}{\partial\nu_\phi\partial\sigma_{\xi l}}l_\phi\Big) &= -\int\cdots\int\frac{\partial^2}{\partial\nu_\phi\partial\sigma_{\xi l}}l_\phi\prod_{i\in I}t_{\nu_\phi}(W_i|0,\Sigma_i)\,\mathrm{d}W_1\cdots\mathrm{d}W_{|I|}\\ &= \sum_{i\in I}\frac{\nu_\phi+P_i}{\nu_\phi^2\sigma_{\xi l}}\,C_3-\sum_{i\in I}\frac{1}{\nu_\phi\sigma_{\xi l}}\,C_4\\ &= \sum_{i\in I}\frac{-2|L_{il}|}{\sigma_{\xi l}(\nu_\phi+P_i)(\nu_\phi+P_i+2)}\end{aligned}\tag{8.1.13}$$

where

$$C_4=\mathbb{E}\bigg[W_{il}^T\Sigma_{il}^{-1}W_{il}\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-1}\bigg]$$

See proposition C.2.6 for the result and detailed calculations for the term C_4.

According to Equation (8.1.10), its corresponding expectation is given by

$$\begin{aligned}\mathbb{E}\Big(-\frac{\partial^2}{\partial\sigma_{\xi l}^2}l_{\nu_\phi}\Big) &= -\int\cdots\int\frac{\partial^2}{\partial\sigma_{\xi l}^2}l_\phi\prod_{i\in I}t_{\nu_\phi}(W_i|0,\Sigma_i)\,\mathrm{d}W_1\cdots\mathrm{d}W_{|I|}\\ &= -\sum_{i\in I}\frac{|L_{il}|}{\sigma_{\xi l}^2}-\sum_{i\in I}\frac{3(\nu_\phi+P_i)}{\nu_\phi\sigma_{\xi l}^2}C_4-\sum_{i\in I}\frac{2(\nu_\phi+P_i)}{\nu_\phi^2\sigma_{\xi l}^2}C_5\\ &= -\sum_{i\in I}\frac{|L_{il}|}{\sigma_{\xi l}^2}-\sum_{i\in I}\frac{|L_{il}|}{\sigma_{\xi l}^2}\left\{\frac{|L_{il}|+2}{1+(\nu_\phi+P_i)/2}-3\right\}\end{aligned}\tag{8.1.14}$$

where

$$C_5=\mathbb{E}\bigg[\Big(W_{il}^T\Sigma_{il}^{-1}W_{il}\Big)^2\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-2}\bigg]$$

See proposition C.2.7 for the result and the derivation of the term C_5.

According to Equation (8.1.11), its corresponding expectation is given by

$$\begin{aligned}\mathbb{E}\Big(-\frac{\partial^2}{\partial\sigma_{\xi l}\sigma_{\xi j}}l_{\nu_\phi}\Big) &= -\int\cdots\int\frac{\partial^2}{\partial\sigma_{\xi l}\sigma_{\xi j}}l_\phi\ \prod_{i\in I}t_{\nu_\phi}(W_i|0,\Sigma_i)\,\mathrm{d}W_1\cdots\mathrm{d}W_{|I|}\\ &= -\sum_{i\in I}\frac{2(\nu_\phi+P_i)}{\nu_\phi^2\sigma_{\xi l}\sigma_{xij}}C_6\\ &= -\sum_{i\in I}\frac{|L_{il}||L_{ij}|}{\sigma_{\xi l}\sigma_{\xi j}\Big(1+(\nu_\phi+P_i)/2\Big)}\end{aligned}\tag{8.1.15}$$

where

$$C_6=\mathbb{E}\bigg[\Big(W_{il}^T\Sigma_{il}^{-1}W_{il}\Big)\Big(W_{ij}^T\Sigma_{ij}^{-1}W_{ij}\Big)\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-2}\bigg]$$

See proposition C.2.8 for the result and detailed derivation for the term C_6.

8.1.2 Block $\Big\{\nu_\kappa,\sigma_\varepsilon,\{\kappa_{ijk}\}_{k=1,\ldots,K_{ij};(i,j)\in\mathcal{J}_\mathcal{I}}\Big\}$

According to the joint distribution shown in Equation (6.2.3), the full conditional distribution of parameters $\{\nu_\kappa,\sigma_\varepsilon,\{\kappa_{ijk}\}_{k=1,\ldots,K_{ij};(i,j)\in\mathcal{J}_\mathcal{I}}\}$ given all other parameters not in this block is

$$p(\nu_\kappa,\sigma_\varepsilon,\{\kappa_{ijk}\}|\text{others})\propto$$
$$p(\nu_\kappa)p(\sigma_\epsilon)\prod_{(i,j)\in\mathcal{I}\mathcal{J}}\prod_{k=1}^{K_{ij}}\mathrm{Gamma}(\kappa_{ijk}|\frac{1}{2}\nu_\kappa,\frac{1}{2}\nu_\kappa)N(y_{ijk}|\mu_{ij},\frac{\sigma_\epsilon^2}{\kappa_{ijk}})$$

In the following section, 'others' is omitted again for the sake of simplicity and thus $p(\nu_\kappa,\sigma_\varepsilon,\{\kappa_{ijk}\})$ denotes the full conditional distribution of this block. Similar to the previous block, the update strategy used to simulate this full conditional distribution $p(\nu_\kappa,\sigma_\varepsilon,\{\kappa_{ijk}\})$ is also divided into two parts as described in Equation (8.0.2):

1. sampling from the marginalized conditional distribution $p(\nu_\kappa,\sigma_\varepsilon)$ that is the result of integrating out all of $\{\kappa_{ijk}\}$ from the full conditional distribution $p(\nu_\kappa,\sigma_\varepsilon,\{\kappa_{ijk}\})$;

2. given $\nu_\kappa,\sigma_\varepsilon$ simulated from the marginalized conditional distribution in the first step, simulating $\{\kappa_{ijk}\}$ from the conditional distribution $p(\{\kappa_{ijk}\}|\nu_\kappa,\sigma_\varepsilon)$.

By using this procedure, a 8999-dimensional (8997 κ_{ijk}'s, 1 σ_ε and 1 ν_κ) sampling problem is divided into a 2-dimensional simulation problem stated in step 1 and a 8997-dimensional simulation for κ_{ijk} in step 2. Particularly, the 8997-dimensional sampling problem for

κ_{ijk} could be solved by sampling them all in a single line of code. This is because the conditional distribution $p(\{\kappa_{ijk}\}|\nu_\kappa, \sigma_\varepsilon)$, that is the same as the one stated in Equation (6.2.6), belongs to the gamma distribution family. Therefore, our focus is the simulation from the marginalized conditional distribution $p(\nu_\kappa, \sigma_\varepsilon)$ required in the first step.

Marginal Distribution

Through integrating out all of $\{\kappa_{ijk}\}$, we get the marginalized conditional distribution shown as follows,

$$p(\nu_\kappa, \sigma_\varepsilon) \propto \underbrace{\frac{1}{\sigma_\varepsilon \nu_\kappa^2}}_{\text{prior}} \prod_{(i,j)\in\mathcal{J}_\mathcal{I}} \prod_{k=1}^{K_{ij}} \underbrace{\frac{1}{\sigma_\varepsilon} \frac{\Gamma(\frac{\nu_\kappa+1}{2})}{\Gamma(\frac{\nu_\kappa}{2})\sqrt{\nu_\kappa}} \Big(1 + \frac{(y_{ijk}-\mu_{ij})^2/\sigma_\varepsilon^2}{\nu_\kappa}\Big)^{-\frac{\nu_\kappa+1}{2}}}_{\text{density of Student's t-distribution}} \tag{8.1.16}$$

It is straightforward to verify that this marginalized conditional distribution is in compliance with the model assumption defining the t-distributed measurement error through

$$y_{ijk} \sim N(\mu_{ij}, \frac{\sigma}{\sqrt{\kappa_{ijk}}})$$

where $\kappa_{ijk} \sim \Gamma(\frac{1}{2}\nu_\kappa, \frac{1}{2}\nu_\kappa)$. After integrating out all of $\{\kappa_{ijk}\}$, y_{ijk} are independent t-distributed with mean μ_{ij}, scale σ_ε and degree of freedom ν_κ conditional on the parameters not in this block. For the reason of simplicity, we adopted the following standardization

$$T_{ijk} = \frac{y_{ijk} - \mu_{ij}}{\sigma_\varepsilon}$$

and thus

$$T_{ijk} \sim t_{\nu_\kappa}$$

Derivatives and Fisher Information Matrix

In order to simulate above marginalized conditional distribution shown in Equation (8.1.16) by using different MCMC samplers, the first derivatives, second derivatives and the Fisher information matrix need to be calculated. After some simplifying, the logarithm of $p(\nu_\kappa, \sigma_\varepsilon)$, which we shall call by l_κ, has the following form

$$\begin{aligned} l_\kappa = -2\log\nu_\kappa - (K+1)\log\sigma_\varepsilon + K\Big(\frac{1}{2}\nu_\kappa \log(\frac{1}{2}\nu_\kappa) - \log\Gamma(\frac{1}{2}\nu_\kappa) + \log\Gamma(\frac{1}{2}\nu_\kappa + \frac{1}{2})\Big) \\ - (\frac{1}{2}\nu_\kappa + \frac{1}{2}) \sum_{(i,j)\in\mathcal{J}_\mathcal{I}} \sum_{k=1}^{K_{ij}} \log\Big(\frac{1}{2}\big(\nu_\kappa + \frac{(y_{ijk}-\mu_{ij})^2}{\sigma_\varepsilon^2}\big)\Big) \end{aligned} \tag{8.1.17}$$

where

$$K = \sum_{(i,j)\in\mathcal{J}_{\mathcal{I}}} K_{ij}$$

Its corresponding first derivatives and second derivatives are provided as follows.

- First Derivatives:

  The first derivative of l_κ with respect to ν_κ is

$$\frac{\partial}{\partial\nu_\kappa} l_\kappa = -\frac{2}{\nu_\kappa} + \frac{1}{2}K\Big(\log(\frac{1}{2}\nu_\kappa) + 1 - \Psi(\frac{1}{2}\nu_\kappa) + \Psi(\frac{1}{2}\nu_\kappa + \frac{1}{2})\Big)$$
$$-\frac{1}{2}\sum_{(i,j)\in\mathcal{J}_{\mathcal{I}}}\sum_{k=1}^{K_{ij}}\log\Big(\frac{1}{2}\big(\nu_\kappa + \frac{(y_{ijk}-\mu_{ij})^2}{\sigma_\varepsilon^2}\big)\Big) - (\frac{1}{2}\nu_\kappa + \frac{1}{2})\sum_{(i,j)\in\mathcal{J}_{\mathcal{I}}}\sum_{k=1}^{K_{ij}}\Big(\nu_\kappa + \frac{(y_{ijk}-\mu_{ij})^2}{\sigma_\varepsilon^2}\Big)^{-1} \tag{8.1.18}$$

  The first derivative of l_κ with respect to σ_ε is

$$\frac{\partial}{\partial\sigma_\varepsilon} l_\kappa = -\frac{K+1}{\sigma_\varepsilon} + (\nu_\kappa + 1)\sum_{(i,j)\in\mathcal{J}_{\mathcal{I}}}\sum_{k=1}^{K_{ij}}\frac{(y_{ijk}-\mu_{ij})^2}{\nu_\kappa\sigma_\varepsilon^3 + (y_{ijk}-\mu_{ij})^2\sigma_\varepsilon} \tag{8.1.19}$$

- Second Derivatives:

  According to the above first derivatives, the second derivatives are displayed as follows. The second derivative of l_κ with respect to ν_κ is

$$\frac{\partial^2}{\partial\nu_\kappa^2} l_\kappa = C_7 - \frac{1}{\nu_\kappa}\sum_{(i,j)\in\mathcal{J}_{\mathcal{I}}}\sum_{k=1}^{K_{ij}}\Big(1 + \frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-1} + \frac{\nu_\kappa+1}{2\nu_\kappa^2}\sum_{(i,j)\in\mathcal{J}_{\mathcal{I}}}\sum_{k=1}^{K_{ij}}\Big(1 + \frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-2} \tag{8.1.20}$$

  where

$$C_7 = \frac{2}{\nu_\kappa^2} + \frac{K}{2}\Big(\frac{1}{\nu_\kappa} - \frac{1}{2}\Psi'(\frac{\nu_\kappa}{2}) + \frac{1}{2}\Psi'(\frac{\nu_\kappa+1}{2})\Big)$$

  The second derivative of l_κ with respect to ν_κ and σ_ε is

$$\frac{\partial^2}{\partial\nu_\kappa\partial\sigma_\varepsilon} = \sum_{(i,j)\in\mathcal{J}_{\mathcal{I}}}\sum_{k=1}^{K_{ij}}\left\{\frac{1}{\nu_\kappa\sigma_\varepsilon}T_{ijk}^2\Big(1 + \frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-1} - \frac{\nu_\kappa+1}{\nu_\kappa^2\sigma_\varepsilon}T_{ijk}^2\Big(1 + \frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-2}\right\} \tag{8.1.21}$$

  The second derivative of l_κ with respect to σ_ε is

$$\frac{\partial^2}{\partial\sigma_\varepsilon^2} l_\kappa = \frac{K+1}{\sigma_\varepsilon^2} - \frac{\nu_\kappa+1}{\nu_\kappa\sigma_\varepsilon^2}\sum_{(i,j)\in\mathcal{J}_{\mathcal{I}}}\sum_{k=1}^{K_{ij}}\left\{3T_{ijk}^2\Big(1 + \frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-2} + \frac{1}{\nu_\kappa}T_{ijk}^4\Big(1 + \frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-2}\right\} \tag{8.1.22}$$

The Fisher Information Matrix is obtained by taking expectations of the previously derived second derivatives. Here, each term in the Fisher Information matrix is provided.

- Fisher Information Matrix:

  According to Equation (8.1.20), the corresponding term in Fisher Information Matrix is

$$\begin{aligned}
&\mathbb{E}\Big(-\frac{\partial^2}{\partial\nu_\kappa^2}l_\kappa\Big)\\
&=-C_7+\frac{1}{\nu_\kappa}\sum_{(i,j)\in\mathcal{J}_\mathcal{I}}\sum_{k=1}^{K_{ij}}\mathbb{E}\bigg[\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-1}\bigg]-\frac{\nu_\kappa+1}{2\nu_\kappa^2}\sum_{(i,j)\in\mathcal{J}_\mathcal{I}}\sum_{k=1}^{K_{ij}}\mathbb{E}\bigg[\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-2}\bigg]\\
&=-C_7+K\bigg(\frac{1}{(\nu_\kappa+1)}-\frac{(\nu_\kappa+2)}{2\nu_\kappa(\nu_\kappa+3)}\bigg) \qquad (8.1.23)
\end{aligned}$$

  The trick related to the calculations of the expectation $\mathbb{E}\bigg[\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-1}\bigg]$ and $\mathbb{E}\bigg[\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-2}\bigg]$ is shown in proposition C.3.1.
  According to Equation (8.1.21), the corresponding expectation result becomes

$$\begin{aligned}
&\mathbb{E}\Big(-\frac{\partial^2}{\partial\nu_\kappa\sigma_\varepsilon}\Big)\\
&=-\sum_{(i,j)\in\mathcal{J}_\mathcal{I}}\sum_{k=1}^{K_{ij}}\bigg\{\frac{1}{\nu_\kappa\sigma_\varepsilon}\mathbb{E}\bigg[T_{ijk}^2\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-1}\bigg]-\frac{\nu_\kappa+1}{\nu_\kappa^2\sigma_\varepsilon}\mathbb{E}\bigg[T_{ijk}^2\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-2}\bigg]\bigg\}\\
&=-\frac{2K}{\sigma_\varepsilon(\nu_\kappa+1)(\nu_\kappa+3)} \qquad (8.1.24)
\end{aligned}$$

  According to Equation (8.1.22), the corresponding expectation is given by

$$\begin{aligned}
&\mathbb{E}\Big(-\frac{\partial^2}{\partial\sigma_\varepsilon^2}\Big)\\
&=-\frac{K+1}{\sigma_\varepsilon^2}+\frac{\nu_\kappa+1}{\nu_\kappa\sigma_\varepsilon^2}\sum_{(i,j)\in\mathcal{J}_\mathcal{I}}\sum_{k=1}^{K_{ij}}\bigg\{3\mathbb{E}\bigg[T_{ijk}^2\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-2}\bigg]+\frac{1}{\nu_\kappa}\mathbb{E}\bigg[T_{ijk}^4\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-2}\bigg]\bigg\}\\
&=-\frac{K+1}{\sigma_\varepsilon^2}+\frac{3K(\nu_\kappa+1)}{\sigma_\varepsilon^2(\nu_\kappa+3)} \qquad (8.1.25)
\end{aligned}$$

  The derivation for $\mathbb{E}\bigg[T_{ijk}^2\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-1}\bigg]$, $\mathbb{E}\bigg[T_{ijk}^2\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-2}\bigg]$ in Equation (8.1.24) and that for $\mathbb{E}\bigg[T_{ijk}^2\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-2}\bigg]$, $\mathbb{E}\bigg[T_{ijk}^4\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-2}\bigg]$ in Equation (8.1.25) are addressed in proposition C.3.2.

8.2 RWMH for Marginalized Conditional Distributions

In this section, the RWMH is chosen as the sampler mentioned in Algorithm 7 to simulate the marginalized conditional distributions $p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L})$ and $p(\nu_\kappa, \sigma_\varepsilon)$ as shown in Equation (8.1.4) and (8.1.16) respectively. To implement it, the variance matrix of the proposal distribution needs to be specified. The usual approach is to customize such a matrix by exploiting a Laplace approximation that provides us with an initial guess about the spread of the target distribution. To be specific, it approximates the target marginalized conditional distribution by using a Gaussian distribution with the mean and the variance setting as the mode and the variance matrix at the mode of the desired distribution respectively. We also need to note that the variance matrix of the proposal distribution is adapted during the whole iteration process since the random-walk Metropolis-Hastings sampling for the target marginalized conditional distribution is embedded into a Gibbs sampling structure dealing with entire model parameters. At each iteration, updates in other blocks would change the spread information of the desired marginalized conditional distribution and thus the approximation for the spread also needs to be renewed. This procedure that achieves the second step of Algorithm 7 is listed below.

Algorithm 8 RWMH for marginals within Gibbs Structure

1: Given current states $\mathbf{x}_1^t, \Lambda = \{\Delta\}$
2: Set $V = \Delta \times \text{laplace}\big(\log[p(\mathbf{x}_1^t)], \mathbf{x}_1^t\big)\$\text{var}$
3: Set $\mathbf{x}_1' = \mathbf{x}_1^t + \text{Gaussian}(0, V)$
4: With probability $\alpha = \min\{1, \frac{p(\mathbf{x}_1')}{p(\mathbf{x}_1^t)}\}$, set $\mathbf{x}_1^{t+1} = \mathbf{x}_1'$

where Δ, a scale parameter tuned according to the acceptance rate during the burn-in period, is used to modify the matrix given by the Laplace approximation. The function 'laplace()', which calculates the Laplace approximation, is provided by R package 'LearnBayes'. It returns mode and variance at the mode of the distribution placed in its first argument. The simulation information is displayed in the following grey box. The resulting Markov chain consists of 2000 burn-in iterations and 20000 main iterations without thinning. This set-up is the same as that for the Markov chain described in section 6.2.2. In addition, the algorithm in section 6.2.2 and the algorithm implemented here both use the RWMH sampler to simulate distributions that do not belong to a known family. The

difference is that the algorithm stated in section 6.2.2 uses the RWMH sampler to sample the full conditional distribution of each parameter while the algorithm proposed here uses the RWMH sampler to deal with the marginalized conditional distribution of the problematic parameters.

Number of Iterations: $N = 20000$;

Burn-in: burn $= 2000$;

Thinning: thin $= 1$;

The scale Δ:

- for $\{\nu_\kappa, \sigma_\varepsilon\}$: $\Delta = 1.6$
- for $\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\}$: $\Delta = 0.8$

The acceptance rate of the main iterations:

- for $\{\nu_\kappa, \sigma_\varepsilon\}$: 0.3713
- for $\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\}$: 0.4151

The simulation results of parameters $\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\}$ and $\{\nu_\kappa, \sigma_\varepsilon\}$ provided by the marginal approach with RWMH sampler are displayed in Figure C.2 in the appendix C.4. Particularly, the last 3000 samples of the simulated chain for the parameter ν_κ and ν_ϕ are reported in the following figure.

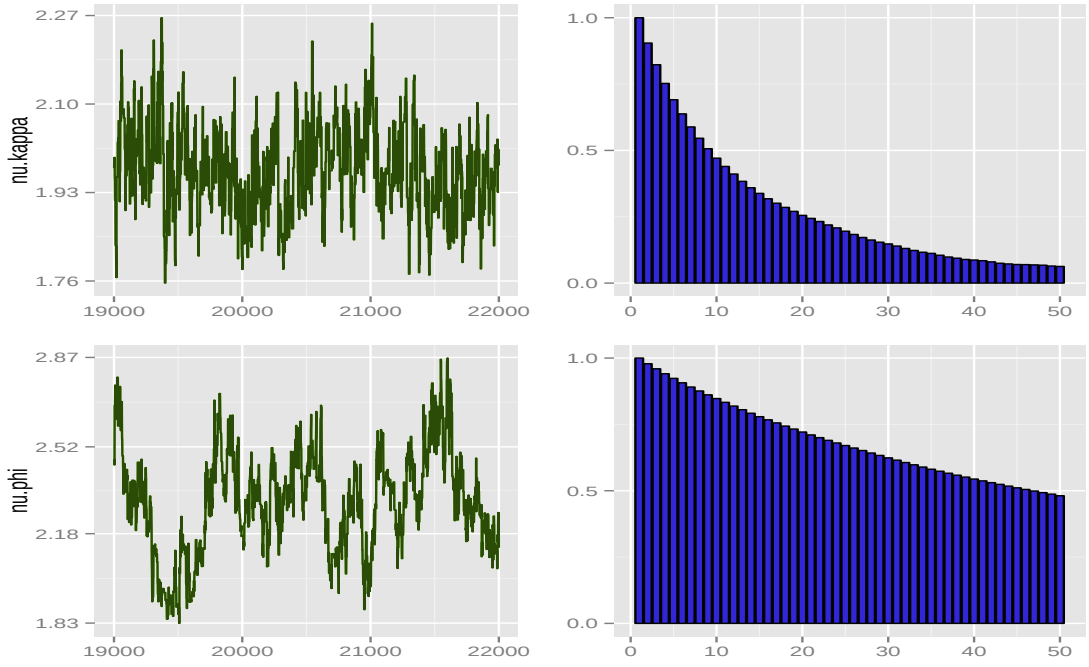


Figure 8.4: Trace plot and auto-correlations for the parameter ν_κ and ν_ϕ

Compared with that in Figure 6.1, there is clear improvement in the case of the parameter ν_κ while no significant improvement was achieved for the parameter ν_ϕ by this marginal procedure. Figure C.2 further confirms that the chain, given by using the marginal approach and the RWMH sampler together, can reduce the auto-correlation for $\{\nu_\kappa, \sigma_\varepsilon\}$ but does not works well for $\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\}$. Indeed, the distribution $p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L})$ has a more complex structure and higher dimensionality than that of $p(\{\nu_\kappa, \sigma_\varepsilon\})$. Therefore, the RWMH sampler might not be suitable in such a scenario.

8.3 Advanced Samplers for Marginal Distributions

As was reported before, the RWMH sampler has limitations in sampling from distributions which have complicated structures and high-dimensional spaces. As for the problem here, it demonstrates poor performance in sampling from the marginalized conditional distribution $p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L})$. We, therefore, consider replacing the RWMH sampler by some advanced sampler to improve the simulation results. Particularly, HMC and its variants are chosen to replace the RWMH to simulate both $p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L})$ and $p(\nu_\kappa, \sigma_\varepsilon)$ as displayed in Equation (8.1.4) and (8.1.16) respectively.

8.3.1 Basic HMC

Let us firstly turn to the basic HMC sampler. To implement the basic HMC sampler, we set the parameters Λ needed for the sampler as follows,

$$\Lambda_\kappa = \{\epsilon_\kappa = 0.1, \mathfrak{l}_\kappa = 5, M_\kappa = \begin{pmatrix} 134 & 0 \\ 0 & 19487 \end{pmatrix}\};$$

$$\Lambda_\phi = \{\epsilon_\kappa = 0.1, \mathfrak{l}_\kappa = 5, M_\phi = \begin{pmatrix} 26 & 0 & 0 & 0 & 0 \\ 0 & 1304 & 0 & 0 & 0 \\ 0 & 0 & 1883 & 0 & 0 \\ 0 & 0 & 0 & 8140 & 0 \\ 0 & 0 & 0 & 0 & 12995 \end{pmatrix}\}$$

where $\epsilon_\kappa, \epsilon_\phi$ denote the step-size values used for HMC sampling of $p(\nu_\kappa, \sigma_\varepsilon)$ and $p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L})$ respectively; $\mathfrak{l}_\kappa, \mathfrak{l}_\phi$ represent the number of leap-frog steps used for HMC sampling of $p(\nu_\kappa, \sigma_\varepsilon)$ and $p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L})$ respectively; M_κ, M_ϕ are the variance matrices for momentum variables in the HMC sampling for distribution $p(\nu_\kappa, \sigma_\varepsilon)$ and $p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L})$ respectively. Particularly, M_κ, M_ϕ are chosen according to the variance of samples illustrated in Figure C.1. Other algorithm information and resulting acceptance rates are listed in the following grey box.

Number of Iterations: $N = 20000$;

Burn-in: burn = 2000;

Thin: thin = 1;

The acceptance rate of the main iterations:

- for $\{\nu_\kappa, \sigma_\varepsilon\}$: 0.99195
- for $\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\}$: 0.9284

The simulation results for parameters $\{\nu_\kappa, \sigma_\varepsilon\}$ and $\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\}$ are displayed in Figure C.3. Compared with Figure C.1 and C.2, general improvements have been achieved by using the HMC sampler in the marginal approach since the auto-correlations for all parameters listed in Figure C.3 are reduced by different degrees. Particularly, the most significant improvements lie in the decrease of auto-correlations among the posterior samples for the parameters ν_κ and ν_ϕ as illustrated in the following figure. For the sampling problem from the marginalized conditional distribution $p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L})$, despite the corresponding auto-correlations have been reduced to some degree by using the HMC sampler, there are still conspicuous auto-correlations, especially for the parameter $\{\sigma_{\xi l}\}_{l=1:L}$.

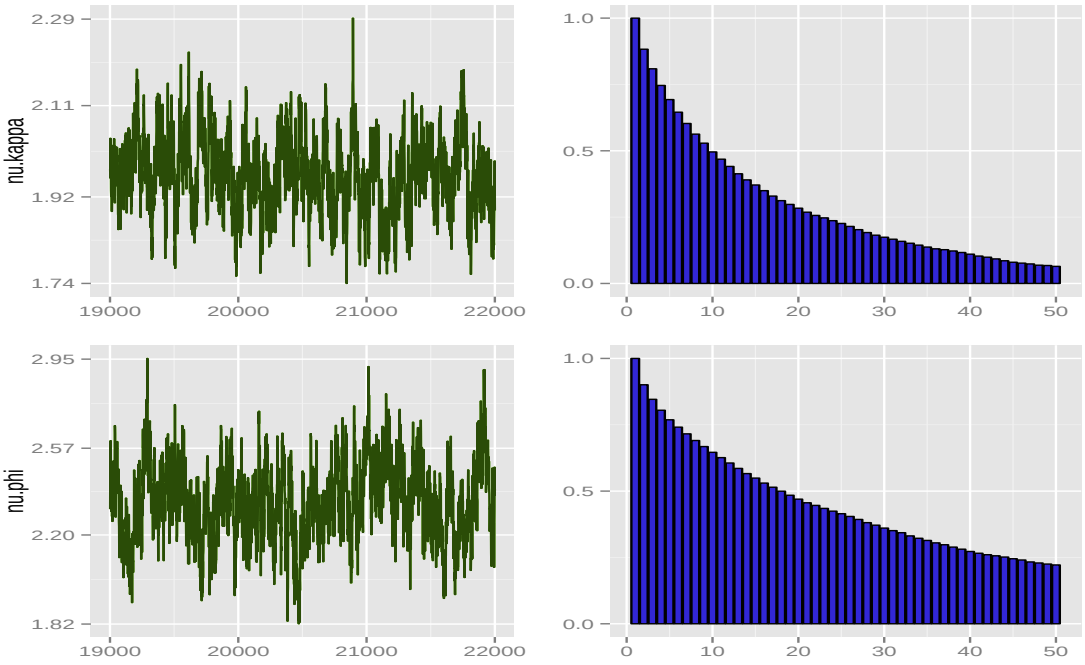


Figure 8.5: Trace plot and auto-correlations for the parameter ν_κ and ν_ϕ

To sum up, the combination of the marginalized strategy and an advanced sampler to sample from marginalized conditional distributions is demonstrated to improve the mixing behaviour of the Markov chain. Neither of them would be powerful if it were employed alone. Further work will concentrate on testing other advanced samplers for the marginalized conditional distributions to obtain more informative simulation results.

8.3.2 NUTS

Now, NUTS is selected to simulate the marginalized conditional distributions of $p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L})$ and $p(\nu_\kappa, \sigma_\varepsilon)$ displayed in Equation (8.1.4) and (8.1.16) respectively. Rather than setting the number of the leap-frog integrator $\mathfrak{l}$ to an arbitrary value, NUTS is considered here as the 'Sampler' in Algorithm 7 to apply HMC and automatically tune the number of leap-frog steps. The step-size values $\epsilon_\kappa, \epsilon_\phi$ and variance matrices M_κ, M_ϕ required by this sampler are chosen to be the same as that in section 8.3.1. Other algorithm information and resulting acceptance rates are stated in the following grey box.

Number of Iterations: $N = 20000$;

Burn-in: burn $= 2000$;

Thin: thin $= 1$;

The acceptance rate of the main iterations:

- for block $\{\nu_\kappa, \sigma_\varepsilon\}$: 1
- for block $\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\}$: 0.999

The obtained acceptance rates are close to one. One possible reason that could explain these high values is that the chosen step-size values are appropriate to make most binary trees grow at least one level. The depths of binary trees constructed by the NUTS during the main iterations are reported in Figure 8.6.

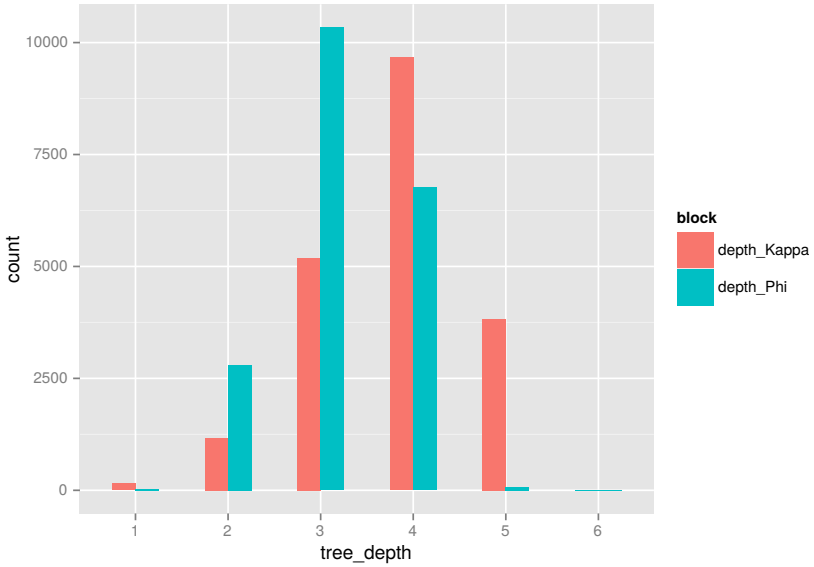


Figure 8.6: Histograms of depth of constructed binary trees. Red: block $\{\nu_\kappa, \sigma_\varepsilon\}$; Blue: block $\left\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\right\}$

From the illustrated histograms, tree depths are clustered around small values (such as 3 or 4) for both blocks. In addition, block $\{\nu_\kappa, \sigma_\varepsilon\}$ tends to have slightly deeper binary trees than the other block. The computational time is very sensitive to the tree depth. Suppose a tree is constructed to have depth 8, it is equivalent to $2^8 = 256$ leap-frog steps in a single iteration which would reduce the speed of program. Too many trees with high depth indicates that the chosen step-size is too small. Therefore, choosing a proper step-size is crucial in NUTS as well.

Due to the complex procedure and recursive nature of building a binary tree required by NUTS to recover the reversibility, it would be inefficient to implement NUTS by using R. Therefore, the code for carrying out NUTS to sample from these two marginalized conditional distributions is written in C++ and it is integrated with R code written for simulating other parameters through 'Rcpp' package. The simulation results for parameters $\{\nu_\kappa, \sigma_\varepsilon\}$ and $\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\}$ are displayed in Figure C.4. Compared to the chain given by the basic HMC, NUTS did not provide us with significant improvements. It only did slightly better for parameters $\{\nu_\kappa, \sigma_\varepsilon\}$.

8.3.3 RMHMC

Let us now use the RMHMC as the 'Sampler' in Algorithm 7 to simulate the marginalized conditional distributions $p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L})$ and $p(\nu_\kappa, \sigma_\varepsilon)$ displayed in Equation (8.1.4) and (8.1.16) respectively. Rather than using a fixed variance matrix for 'momentum' variables, the RMHMC automatically tunes the variance matrix by utilizing the local expected Hessian matrix that provides us with not only the local curvature information but also a positive definite matrix. The parameters needed by the RMHMC sampler to simulate targeted two marginalized distributions are specified as

$$\Lambda_\kappa = \{\epsilon_\kappa = 0.85, \mathfrak{l}_\kappa = 5\}; \quad \Lambda_\phi = \{\epsilon_\phi = 0.80, \mathfrak{l}_\phi = 5\} \tag{8.3.26}$$

By using the local expected Hessian matrix as the variance matrix of 'momentum' variables, the influence of the step-size problem is automatically relieved in the RMHMC sampler. As shown by ϵ_κ and ϵ_ϕ in the above equation, they are larger than those used in the previously mentioned samplers. However, the RMHMC sampler is a computational expensive algorithm since it requires not only matrix decompositions for each leap-frog step but also extra iterations to deal with the implicit functions involved in the generalised leap-frog integrator. Considering the limited speed of R, the implementation of the

RMHMC sampler is written in C++ code and calling through R by using 'Rcpp' package. Other algorithm information and acceptance rates of these two blocks are displayed in the following grey box. The reported acceptance rates are satisfactory even with such high step-size values.

Number of Iterations: $N = 20000$;

Burn-in: burn = 2000;

Thin: thin = 1;

The acceptance rate of the main iterations:

- for $\{\nu_\kappa, \sigma_\varepsilon\}$: 0.90925
- for $\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\}$: 0.87565

The simulation results for parameters $\{\nu_\kappa, \sigma_\varepsilon\}$ and $\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\}$ are illustrated in Fig. C.5. All trace plots demonstrate that the resulting Markov chain has better mixing behaviour than those provided by other methods. The most striking results are the simulations for parameters $\{\nu_\kappa, \sigma_\varepsilon\}$ because of the huge reductions of auto-correlations as illustrated in the corresponding auto-correlation plot. As for the other block, obvious improvements also emerge to different extents with the parameter ν_ϕ improving the most.

8.3.4 HMC with Stochastic Step-size

Let us now try HMC with stochastic step-size to sample the two marginalized conditional distributions $p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L})$ and $p(\nu_\kappa, \sigma_\varepsilon)$. This sampler proceeds very much in the same way as the basic HMC in section 8.3.1. The only difference is that the HMC with stochastic step-size does not need the user to specify a step-size value. It automatically adapts the step-size values according to the largest eigenvalue of the expected Hessian matrix at the current state. Note that we can also just use the Hessian matrix evaluated at current states and its corresponding eigenvalue with the largest absolute value. Compared to the basic HMC sampler, it is more computationally expensive as it needs eigen-decomposition at states where each simulated trajectory starts and terminates in order to generate step-size values and recover the reversibility. However, its computational complexity is much lower than that of the other two HMC variants: NUTS and RMHMC. The step-size value of each iteration is generated from a probability distribution with parameters determined

by the local curvature information. Particularly, we choose

$$\epsilon \sim \mathcal{TN}\Big(\text{mean} = \frac{1}{2}r, \text{sd} = \frac{1}{8}r, a = 0, b = r\Big) \quad \text{with } r = \frac{2}{\sqrt{\lambda_{\boldsymbol{\theta}}}} \tag{8.3.27}$$

where $\mathcal{TN}(\cdot)$ stands for truncated normal distribution bounded between a and b; $\lambda_{\boldsymbol{\theta}}$ denotes the largest eigenvalue of the curvature matrix evaluated at point $\boldsymbol{\theta}$. The term r is the maximum step-size allowed by the stability condition illustrated in Equation (5.6.55). The variance matrices of 'momentum' variables and the number of leap-frog steps required by this sampler are set to be the same as that for the basic HMC sampler described in section 8.3.1. Other algorithm information and resulting acceptance rates of the targeted two blocks are listed in the following grey box.

Number of Iterations: $N = 20000$;

Burn-in: burn = 2000;

Thin: thin = 1;

The acceptance rate of the main iterations:

- for $\{\nu_\kappa, \sigma_\varepsilon\}$: 0.91815
- for $\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\}$: 0.87735

The simulation results for parameters $\{\nu_\kappa, \sigma_\varepsilon\}$ and $\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\}$ are reported in Figure C.6. It also apparent that simulation results for parameters $\{\nu_\kappa, \sigma_\varepsilon\}$ demonstrate much better performances than that for parameters $\Big\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\Big\}$. Trace plots for block $\{\nu_\kappa, \sigma_\varepsilon\}$ display good mixing behaviours and their associated autocorrelation plots demonstrate just small amount of autocorrelations among posterior samples. On the other hand, trace plots for parameters $\Big\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\Big\}$ indicate that the chain does not traverse their marginal distribution as quick as that for parameters $\{\nu_\kappa, \sigma_\varepsilon\}$. Their associated autocorrelation plots still display high autocorrelations among posterior samples. By comparing the simulation results displayed in Figure C.6 with others simulation results, we draw two conclusions. Firstly, for sampling parameters $\{\nu_\kappa, \sigma_\varepsilon\}$, although this sampler is not as good as RMHMC (Figure C.5), it does much better than the basic HMC sampler (Figure C.3) and NUTS (Figure C.4). Secondly, for sampling parameters $\Big\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\Big\}$, there is no significant further decrease in auto-correlations among posterior samples achieved by this sampler. Particularly, the RMHMC sampler outperforms HMC with stochastic step-size algorithm in the simulation for $\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\}$.

The step-size values, that are simulated during the main iterations, are illustrated in Figure 8.7.

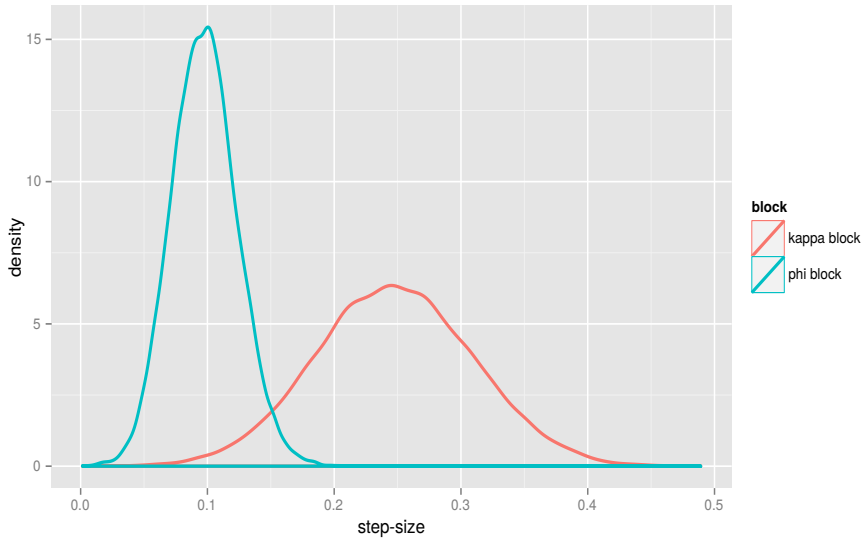


Figure 8.7: step-size used for sampling block $\{\nu_\kappa, \sigma_\varepsilon\}$ and block $\left\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\right\}$

It shows that the marginalized conditional distribution $p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L})$ (blue line) needs smaller step-size values than the distribution $p(\nu_\kappa, \sigma_\varepsilon)$ (red line). This indicates that sampling for the block $\left\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\right\}$ is more challenging that that for the block $\{\nu_\kappa, \sigma_\varepsilon\}$.

8.4 Explanation of Tenacious Autocorrelations

Despite the fact that the joint use of the marginalized approach and advanced samplers (HMC and its associated variants) have obtained great achievements in respect of decreasing auto-correlations, noticeable amounts of auto-correlation are still persistent in posterior samples especially for block $\left\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\right\}$. All HMC variants including HMC itself obtained similar simulation performances for the marginalized conditional distribution $p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L})$. For $p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L})$, although the RMHMC sampler did better than other samplers, it still cannot reduce the autocorrelations to a satisfactory level. In this section, the reason of these tenacious auto-correlations are explored from the perspective of the Gibbs structure for the entire sampling.

These apparent and tenacious auto-correlations are not due to the implemented sampler used to sample from these two marginalized conditional distributions but should be attributed to parameters not in the target block. As previously stated, simulations for these two marginalized conditional distributions are embedded into a big block Gibbs sampling structure. Therefore, sampling results for these two blocks are influenced by parameters in the remaining sampling parts of the entire Gibbs structure. To verify the fact that the tenacious autocorrelations do indeed come from the Gibbs structure, the following experiment is carried out. The experiment is to proceed by first fixing parameters not in block $\{\nu_\kappa, \sigma_\varepsilon\}$ and block $\left\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\right\}$ to eliminate the influence of those parameters and then simulating these two blocks solely by RMHMC. Figure 8.8 illustrates autocorrelations of 3000 samples obtained by this experiment. Autocorrelations for both blocks become significantly small. This finding confirms the statement that those tenacious autocorrelations displayed in Figure C.3 to Figure C.6 for HMC and its variants were caused by parameters in the remaining of sampling parts of the entire Gibbs structure.

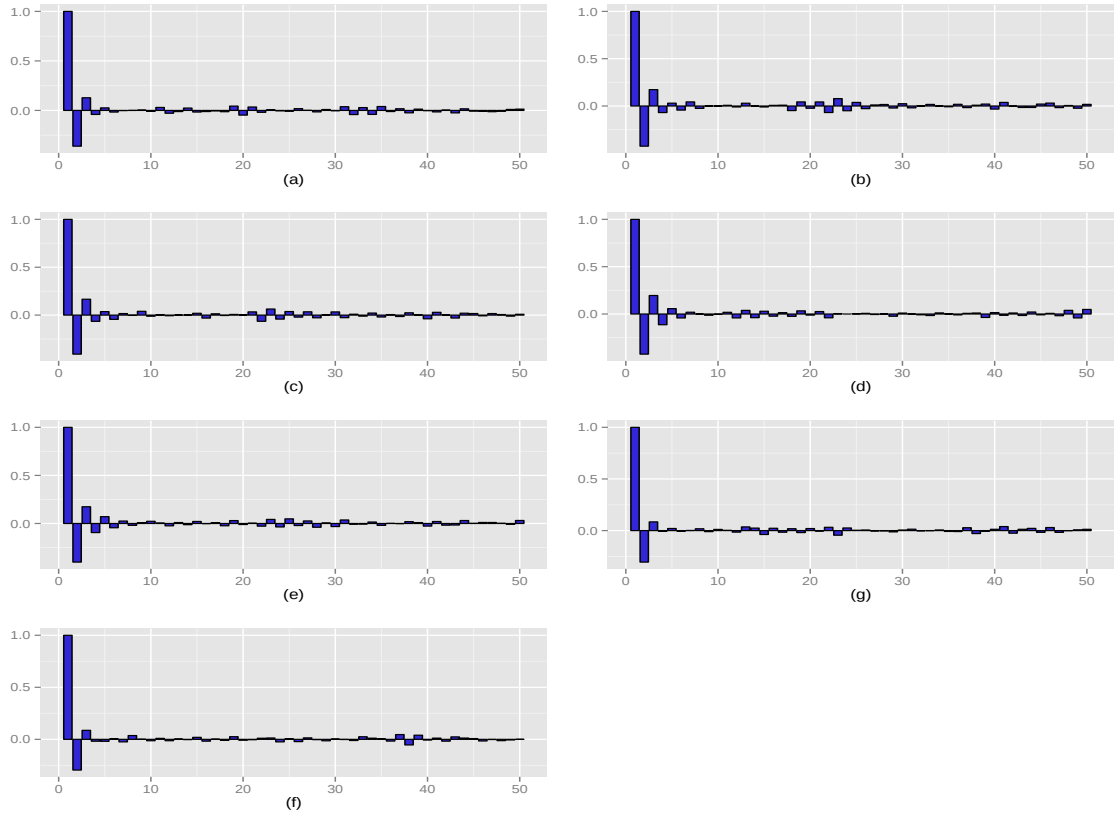


Figure 8.8: Auto-correlations of posterior samples obtained by fixing other parameters and simulating only block $\{\nu_\kappa, \sigma_\varepsilon\}$ and block $\left\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\right\}$ by RMHMC. plot (a)-(e): for $\left\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\right\}$; plot (f)-(g): for $\{\nu_\kappa, \sigma_\varepsilon\}$

8.5 Conclusion

In a real situation, an issue of true concern is which sampling method should be used to achieve the best performance within the shortest possible time. Some sampling schemes like RMHMC might give simulations with relatively low autocorrelations but they are computationally expensive in each iteration. In respect of computation time, these sampling methods might not be good choices. In this section, we compare all the previously mentioned sampling methods in terms of both the effective sample size and the computational time.

As shown in Figures C.1 to C.6, the computational method's ability of providing informative samples varies a lot. The lower autocorrelations among the posterior samples are, the more information these samples can offer. In order to measure the information of simulations provided by each sampling scheme, we use the effective sample size provided by the R function 'effectiveSize()' from 'coda' package. For the samples from the posterior provided by a particular sampling method, we obtain ESS for all the hyper-parameters. In Table C.1 of Appendix C.5, ESS values for all the hyper-parameters of the chains shown in Figures C.1 to C.6 are reported. In order to compare these ESS values given by different sampling methods, we choose the original sample method (the modified MCMCglmm without the marginalized approach detailed in section 6.2.2) as the base-line method and calculate the ratios between ESS for the chains provided by the other sampling methods (all sampling methods in section 8.2 and 8.3) and that for the chain provided by the base-line method. For clear visualization, the logarithms of obtained ratios of ESS are shown in Figure 8.9. In particular, the green line represents the basic HMC; the purple line represents NUTS; the blue line represents RMHMC; the red line represents HMC with stochastic step-size (HMC_S); the black represents RWMH. The black dashed horizontal line marks 1. The line higher than this dotted line means that the corresponding chain can provide more effective samples than the base-line method, otherwise the base-line method is better. From Figure 8.9, lines show great difference for the parameters $\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}, \nu_\kappa, \sigma_\varepsilon\}$ that are simulated by using the marginalized distributions as illustrated in section 8.1 while for the other parameters, the lines sit around 0. Compared with the base-line method, the marginalized approach with other samplers, apart from the RWMH sampler (black line), generally increases the ESS values.

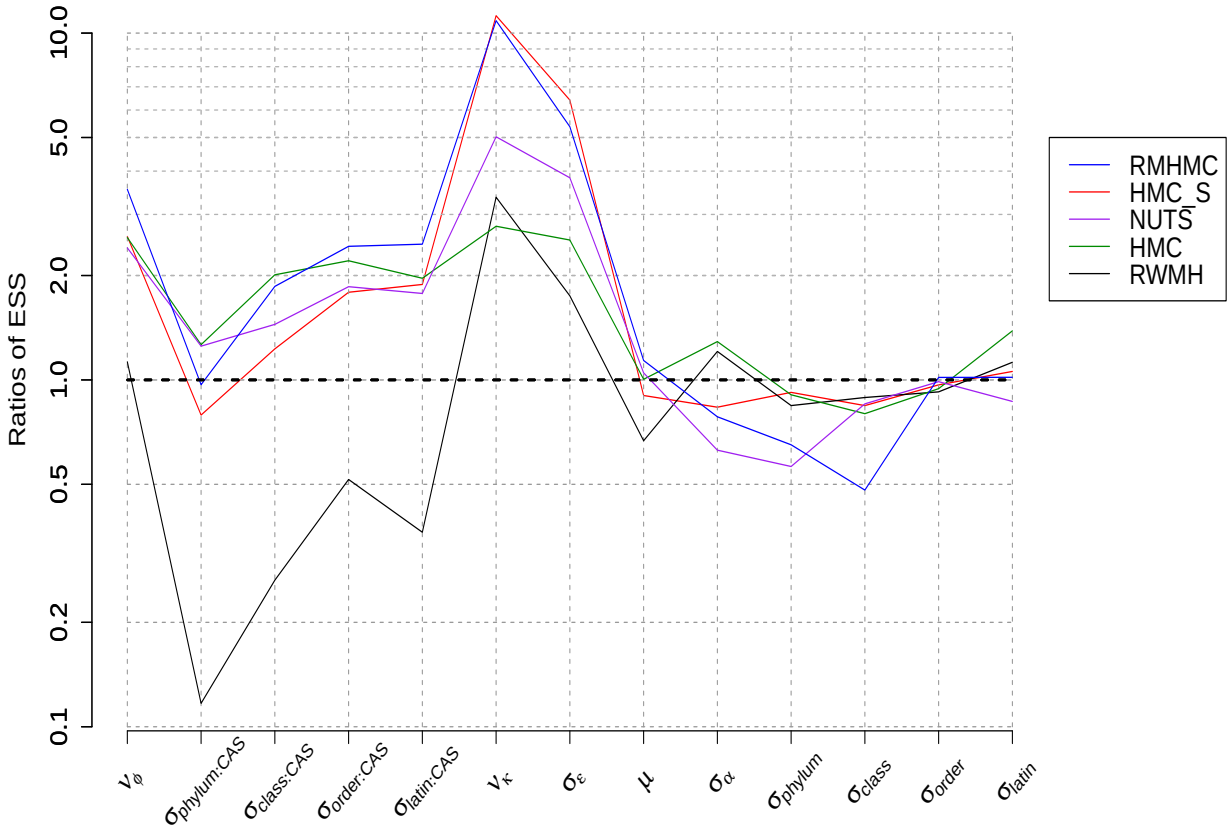


Figure 8.9: Logarithm of Ratios of ESS for all the hyper-parameters.

In spite of the increased effective sample size that is achieved by the marginalized approaches, their computations are also quite expensive. The computation time must be taken into account when measuring the efficiency of a sampling method. We therefore calculate ESS/s (ESS per second) for each sampling method by dividing the ESS values by the its corresponding computation time. Figure 8.10 displays the ratios of ESS/s in the same way as that for Figure 8.9. For the parameters that are not simulated by the marginalized distributions, all the lines are under the black dashed line since compared with the base-line method, rest of the samplers with marginalized approach give similar ESS values for these parameters but use longer computation time. For the parameters that are simulated by using the marginalized distributions, HMC with stochastic step-size (red line) provided the highest ESS/s for $\{\nu_\kappa, \sigma_\varepsilon\}$ and the basic HMC sampler (green line) provided the highest ESS/s for $\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\}$.

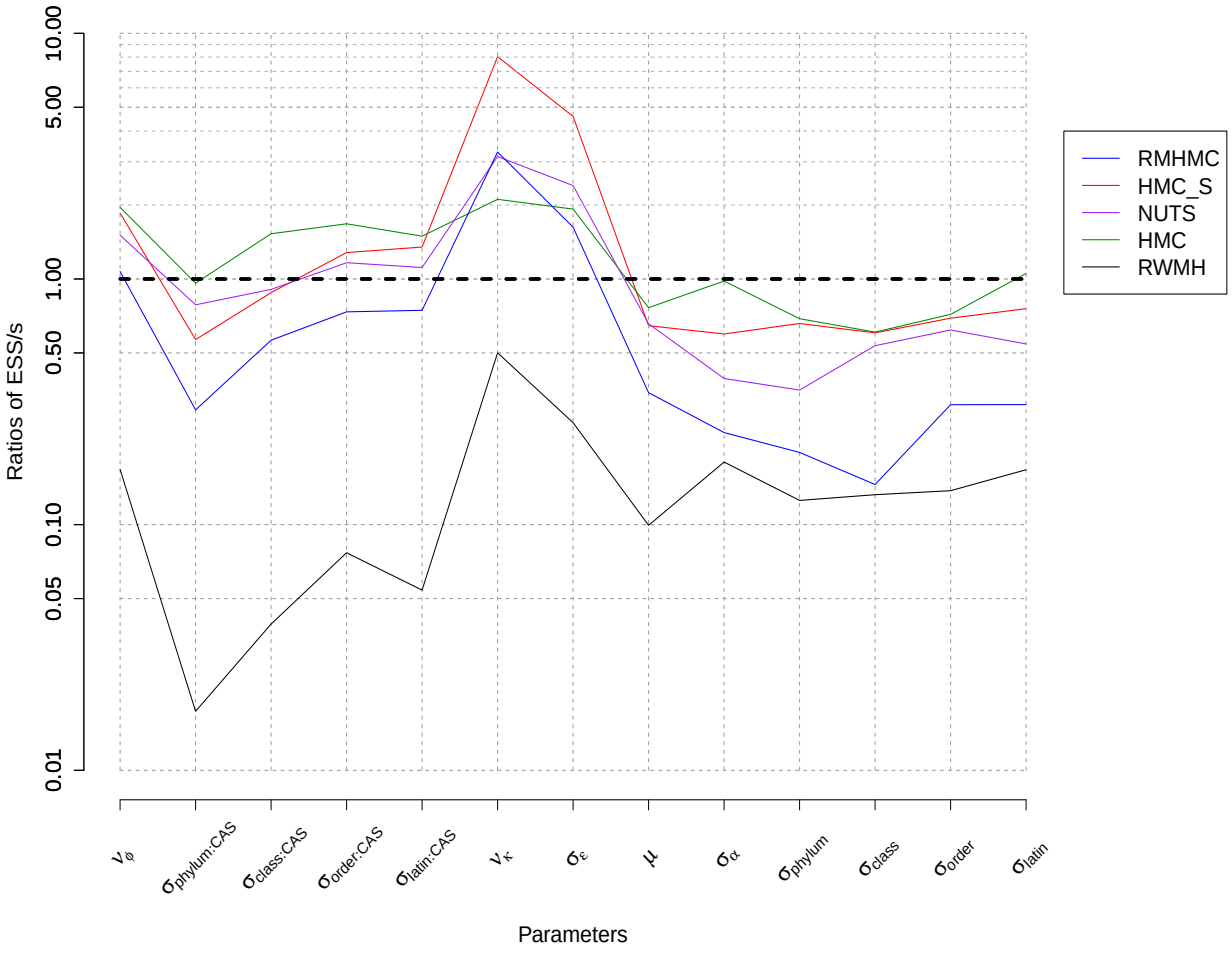


Figure 8.10: Ratios of ESS/SEC for the unblocked hyper-parameters.

To sum up, sampling the marginalized conditional distributions by the basic HMC sampler or the HMC with stochastic step-size sampler provided us with the best simulation results for this model in terms of both effective sample size and computation time.

Chapter 9

Conclusion

The study was set out to explore some topics of two important areas in Bayesian statistics: objective priors and MCMC simulations for posterior distributions.

Objective Priors

The initial goal of the research for objective priors was to develop a principle for an objective prior to satisfy in order to represent ignorance. This aim was motivated by the observation that for the one-way random effect model which is a simple model, most objective priors depend on additional knowledge about parameters and experimental designs. It was hoped that the principle discussed in this thesis might enable us to consider representing the ignorance in a different way. In particular, the principle was applied to the one-way random effect model which is simple but is notoriously difficult to specify an objective prior for it.

Our principle was introduced in Chapter 3. The main idea of the principle is that if the global distance structure is invariant to a re-parametrization, then equivalent prior measure should be assigned to these two parametrizations. This idea was motivated by the belief that when there is no prior knowledge available, all information that distinguishes one point from another should be obtained by considering how its corresponding statistical model differs from other statistical models. We used the global distance to present the differences among statistical model and, in order to avoid considering only a pair of points, the global distance structure of all points was actually used to derive a prior. Based on this global distance structure principle, we derived corresponding priors for three simple problems: location family, scale family and location-scale family.

In Chapter 4, the global distance structure was applied to the one-way random effect model. For the one-way random effect model, most objective priors, such as the Jeffreys prior, the Jeffreys prior with location parameter fixed and the uniform shrinkage prior, depend on the experimental design (i.e. the number of observations). In order to avoid such dependencies, we considered the structure of the averaged global distance by using the limit technique, i.e.

$$D_1 = \lim_{N\to\infty} \frac{1}{N} d_{\boldsymbol{\theta}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \qquad D_2 = \lim_{N\to\infty} [d_{\boldsymbol{\theta}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) - N \cdot D_1]$$

The benefits of considering these distances are the removal of the influence of the experiment design and the simplification of the distance structure. Apart from D_1, D_2 has also been taken into account since the limit technique usually leads to information loss in D_1. Based on the structure of such averaged distances, the priors were derived in different contexts with all three parameters $\{\mu, \vartheta, \rho\}$ unknown, only two of the parameters unknown and only one of the parameters unknown. Two priors resulted from these derivations: the GDSP, $\pi(\mu, \sigma, \sigma_\alpha) \propto \frac{\sigma_\alpha^\gamma}{\sigma}$, is obtained by considering all parameters as unknown; the CGDSP, $\pi(\mu, \sigma, \sigma_\alpha) \propto \frac{\sigma_\alpha}{\sigma(\sigma_\alpha^2+\sigma^2)}$, is the one that respects all the forms of the priors derived in the context with two parameters unknown and one parameter unknown. The performances of the GDSP with $\gamma = -\frac{1}{2}$, the CGDSP and other popular objective priors were evaluated by using a simulation study. The conclusion drawn from the simulation studies is that no prior could always perform better than others. When the true value of between group variance is much larger than that of within group variance, the CGDSP and the JPLF had similar performances and were the best choices. When the true value of between group variance is much smaller than that of within group variance, most priors did not show good performance while the GDSP gave relatively satisfactory performance. When the true value of between group variance is similar to that of within group variance, the JPLF is the best choice. We, therefore, suggested to use GDSP, CGDSP and JPLF together.

The limitation of the GDSP is that γ, the power of σ_α, is unspecified. In the simulation study, γ was arbitrarily chosen as $-\frac{1}{2}$. Choice of the value of γ could perhaps be an area for further exploration and might have the potential to increase the frequentist behaviour of the corresponding posterior distribution. The limitation of the CGDSP is that although it did not lead to posterior distributions with undesirable frequentist behaviour, it did not satisfy the requirement that the parameter spaces should not change after the reparametrization. Seeking other parametrizations so that other invariant structures could

display in D_1 and D_2 could perhaps be an area for further opportunity to improve the performance of priors.

Computation

The research into the Bayesian computation included two aspects: one is for the HMC algorithm itself; the other one is the simulation problem for a real complex hierarchical model developed for ecotoxicology data analysis.

The initial goal of the research for the HMC algorithm was to improve its performance from the perspective of the step-size. This aim was motivated by the observation that although the HMC has the potential to avoid the random-walk behaviour of the traditional MCMC algorithm, its ability depends largely on the choices of the step-size values. It was hoped that the method proposed in this thesis could enable us to tune the step-size values automatically.

HMC with stochastic step-size, our method proposed to automatically tune the step-size values, was introduced in Chapter 5. After an exploration of the problem of the step-size, we found out that the real difficulty of choosing a good step-size value is that a good global step-size value might not actually exist. In other words, appropriate step-size values changes as the Markov chain moves. We therefore investigated the local step-size conditions which turned out to depend on the local curvature information of the target log-density function. The main idea of the proposed method is to consider the step-size as an augmented random variable generated according to the curvature information at the current state of the Markov chain. In this way, the step-size could change long Markov chain iterations. The proposed method was applied to the 'banana' example. It displayed good performance even with an extreme starting point. In addition, through the exploration of the method of choosing the step-size values, we found a new way, called by generalised Metropolis-Hastings with dynamics, to represent a series of algorithms including the ordinary Metropolis-Hastings algorithm, the HMC algorithm, and the HMC with stochastic step-size algorithm.

The limitation of the HMC with stochastic step-size algorithm is that it only exploited the largest eigenvalue of the local curvature information matrix. Researching how to make full use of the local curvature information, without requiring as much expensive computation as the RMHMC, might be an area for further exploration to improve the algorithm performance.

As for the simulation problem for a real complex hierarchical model developed for ecotoxicology data analysis, our goal was to decrease the autocorrelation of the posterior samples. This aim was motivated by the observation that the existing method, the modified MCMCglmm, for this model, led to highly-correlated posterior samples. It was hoped that the method suggested in this thesis could help decrease the autocorrelation living in the posterior samples given by the original method.

Firstly, the background of this model and its computation difficulties were given in Chapter 6. After an attempt to apply some existing methods to this model, we found out that the procedure adopted in the modified MCMCglmm, breaking the simulation for the entire parameters into small pieces and simulating them alternatively under a Gibbs structure, is a possible way to compute this model even though its corresponding posterior samples had high autocorrelations. This led to a study of improving the simulation efficiency for small pieces inside the big Gibbs structure. Secondly, two HMC variants, NUTS and RMHMC, used to improve the simulation were reviewed in Chapter 7 as preliminary materials.

The strategy, used to improve the simulation within the big Gibbs structure, was introduced in Chapter 8. The strategy contained two aspects. One is to group parameters into blocks and then simulate the blocks by using the marginalized distributions to break the correlations among parameters within the blocks. The other is to use advanced samplers, HMC and its variants, to simulate the marginalized distributions. After the comparisons, the combination of the marginalized approach and the HMC with stochastic step-size was the best choice in terms of both effective sample size and computation time.

The limitation of this method is that it only targeted on parts of the parameters. Integrating out all the random effects analytically or approximately and simulate the resulting marginalized distribution could perhaps provide possible ways to further decrease autocorrelations of the posterior samples.

Appendix A

Computations for one-way random effect models

A.1 Covariance Matrix

Here, two facts about a matrix having the form illustrated in Equation (4.1.2) are provided.
1) If $A_{N,N} = aI_{N,N} + bJ_{N,N}$, then we have

$$|A_{N,N}| = (a + Nb)a^{N-1}$$

Proof Denote the orthogonal eigenvectors of $A_{N,N}$ are $\mathbf{1}_N, u_2, \ldots, u_N$, i.e.

$$u_j \perp \mathbf{1}_N,\ j = 2, \ldots, N$$

where $\mathbf{1}_N$ is a column vector with all terms to be one. According to the definition of eigenvalue and eigenvector,

$$A_{N,N} \cdot v = \lambda v$$

we have

$$\begin{aligned} A_{N,N} \cdot \mathbf{1}_N &= aI_{N,N}\mathbf{1}_N + bJ_{N,N}\mathbf{1}_N = a\mathbf{1}_N + Nb\mathbf{1}_N \\ &= (a + Nb)\mathbf{1}_N = \lambda_i \mathbf{1}_N \\ A_{N,N} \cdot u_j &= aI_{N,N}u_j + bJ_{N,N}u_j = aI_{N,N}u_j + b\mathbf{1}_N\mathbf{1}_N^T u_j \\ &= au_j = \lambda_j u_j \qquad \text{where } j = 2, \ldots, N \end{aligned}$$

Note that $\mathbf{1}_N^T u_j = 0$ since the orthogonal property. Therefore, the eigenvalues of $A_{N,N}$ are $a + Nb, a, \ldots, a$ and the determinant of $A_{N,N}$ is the product of these eigenvalues, i.e.

$$|A_{N,N}| = (a + Nb)a^{N-1}$$

. □

2) If $A_{N,N} = I_{N,N} + cJ_{N,N}$, then we have

$$A_{N,N}^{-1} = I_{N,N} + dJ_{N,N} \qquad \text{where } d = -\frac{c}{Nc+1}$$

Proof

$$A_{N,N} \cdot A_{N,N}^{-1} = (I_{N,N} + cJ_{N,N})(I_{N,N} + dJ_{N,N}) = I_{N,N} + (c + d + Ncd)J_{N,N} = I_{N,N}$$

Therefore,

$$c + d + Ncd = 0 \implies d = -\frac{c}{Nc+1}$$

. □

A.2 Equivariant Recodings

Consider the one-way random effect model

$$\mathbf{y} \overset{iid}{\sim} \mathbf{N}(\mu\mathbf{1}_N, A_{N,N});\ i = 1, \ldots, m \tag{A.2.1}$$

where

$$A_{N,N} = \alpha I_{N,N} + \beta J_{N,N} \tag{A.2.2}$$

with $J_{N,N}$ is a N-dimensional square matrix with all terms to be one, $\mathbf{1}_N$ is a N-dimensional column vector of all terms to be one and $I_{N,N}$ is a N-dimensional identity matrix.

Proposition A.2.1 *Suppose that a recoding of* $\mathbf{y_i}$ *has the following form*

$$\mathbf{z}_i = g(\mathbf{y}_i) = c\,\mathbf{1}_N + B\mathbf{y}_i, \tag{A.2.3}$$

where c *is a real value and* B *is a non-singular* $N \times N$ *dimensional matrix. In particular, suppose that* B *satisfies*

$$B = (aI_{N,N} + bJ_{N,N})\mathbf{O}, \tag{A.2.4}$$

where a, b are some real values; $\mathbf{O}$ *is an orthogonal matrix and has the property*

$$\mathbf{O}\mathbf{1}_N = \mathbf{1}_N. \tag{A.2.5}$$

We show here that the corresponding equivariant recoding of $\boldsymbol{\theta} = \{\mu, \alpha, \beta\}$ *induced by this* g *is*

$$\begin{aligned}\boldsymbol{\Phi} &= \bar{g}(\boldsymbol{\theta}) = \bar{g}(\{\mu, \alpha, \beta\}) \\ &= \{(a + Nb)\mu + c, a^2\alpha, \alpha(2a + Nb)b + \beta(a + Nb)^2\}.\end{aligned} \tag{A.2.6}$$

Proof : After the transformation g, the variable $\mathbf{z}_i$ has the mean

$$\begin{aligned}\mathbb{E}(\mathbf{z}_i) &= c\mathbf{1}_N + \mu(aI_{N,N} + bJ_{N,N})\mathbf{O}\mathbf{1}_N \\ &= c\mathbf{1}_N + \mu(aI_{N,N} + bJ_{N,N})\mathbf{1}_N \quad \text{by Equation (A.2.5)} \\ &= c\mathbf{1}_N + a\mu\mathbf{1}_N + b\mu J_{N,N}\mathbf{1}_N \\ &= c\mathbf{1}_N + a\mu\mathbf{1}_N + Nb\mu\mathbf{1}_N \qquad \text{by } J_{N,N}\mathbf{1}_N = N\mathbf{1}_N \\ &= (c + a\mu + Nb\mu)\mathbf{1}_N\end{aligned} \tag{A.2.7}$$

The covariance matrix of $\mathbf{z}_i$ is

$$\begin{aligned}\mathbb{C}ov(\mathbf{z}_i) &= B\,\mathbb{C}ov(\mathbf{y}_i)\,B^T = BA_{N,N}B^T \\ &= (aI_{N,N} + bJ_{N,N})(\alpha I_{N,N} + \beta J_{N,N})(aI_{N,N} + bJ_{N,N})^T \\ &= a^2\alpha\mathbf{O}\mathbf{O}^T + ab\alpha\mathbf{O}\mathbf{O}^T J_{N,N} + a^2\beta\mathbf{O}J\mathbf{O}^T + ab\beta\mathbf{O}J_{N,N}\mathbf{O}^T J_{N,N} + ba\alpha J_{N,N}\mathbf{O}\mathbf{O}^T \\ &\quad + b^2\alpha J_{N,N}\mathbf{O}\mathbf{O}^T J_{N,N} + ab\beta J_{N,N}\mathbf{O}J_{N,N}\mathbf{O}^T + b^2beta J_{N,N}\mathbf{O}J_{N,N}\mathbf{O}^T J_{N,N}\end{aligned} \tag{A.2.8}$$

Since the property shown in Equation (A.2.5), we have the following fact

$$\mathbf{O}J_{N,N} = J_{N,N} \tag{A.2.9}$$

By multiplying both sides of the above equation by a matrix $\mathbf{O}^{-1}$, we obtain that

$$J_{N,N} = \mathbf{O}^{-1}J_{N,N} \tag{A.2.10}$$

Together with the fact that $\mathbf{O}$ is an orthogonal matrix, i.e. $\mathbf{O}\mathbf{O}^T = I_{N,N}$, we have

$$\mathbf{O}^T J_{N,N} = \mathbf{O}^{-1}J_{N,N} = J_{N,N} \tag{A.2.11}$$

Because of the special structure of $J_{N,N}$, we have

$$J_{N,N}J_{N,N} = NJ_{N,N} \tag{A.2.12}$$

By substituting Equations (A.2.9), (A.2.11) and (A.2.12) into Equation (A.2.8), the covariance matrix of $\mathbf{z}_i$ changes to

$$\mathbb{C}ov(\mathbf{z}_i) = a^2\alpha I_{N,N} + [\alpha(2a+Nb)b + \beta(a+Nb)^2]J_{N,N} \tag{A.2.13}$$

After the transformation g, $\mathbf{z}_i$ still obeys the one-way random effect model. By comparing the mean and covariance matrix of $\mathbf{y}_i$ and $\mathbf{z}_i$, we obtain that $\mathbf{z}_i$ follows the one-way random effect model with parameters $\{(a+Nb)\mu + c, a^2\alpha, \alpha(2a+Nb)b + \beta(a+Nb)^2\}$. That is, the induced equivariant recoding of $\boldsymbol{\theta} = \{\mu, \alpha, \beta\}$, which we call it by $\bar{g}$, is

$$\boldsymbol{\Phi} = \bar{g}(\boldsymbol{\theta}) = \bar{g}(\{\mu, \alpha, \beta\}) = \{(a+Nb)\mu + c, a^2\alpha, \alpha(2a+Nb)b + \beta(a+Nb)^2\}.$$

. □

Proposition A.2.2 *For the equivariant recoding g shown in proposition A.2.1, the collection of all the induced equivariant recodings $\bar{g}$ shown in Equation* (A.2.6) *forms a group*

$$\bar{\mathcal{G}} = \{\bar{g}_{a,b,c}; \forall c \in \mathbb{R}, a \neq 0, a+Nb \neq 0\} \tag{A.2.14}$$

Proof : In order to qualify as a group, there are four requirements: closure, associativity, identity element and inverse element. We will check them in order.
Closure: $\forall \bar{g}_1 = \bar{g}_{a_1,b_1,c_1} \in \bar{\mathcal{G}}$ and $\forall \bar{g}_2 = \bar{g}_{a_2,b_2,c_2} \in \bar{\mathcal{G}}$, according to Equation (A.2.6), we have

$$\begin{aligned}\bar{g}_1\bar{g}_2(\{\mu,\alpha,\beta\}) &= \bar{g}_1(\{(a_2+Nb_2)\mu + c_2, a_2^2\alpha, \alpha(2a_2+Nb_2)b_2 + \beta(a_2+Nb_2)^2\}) \\ &= \{(a^\star+Nb^\star)\mu + c^\star, (a^\star)^2\alpha, \alpha(2a^\star+Nb^\star)b^\star + \beta(a^\star+Nb^\star)^2\}, \end{aligned} \tag{A.2.15}$$

where $a^\star = a_1a_2, b^\star = a_1b_2 + a_2b_1 + Nb_1b_2,$ and $c^\star = (a_1+Nb_1)c_2 + c_1$.
Therefore, $\bar{g}_1\bar{g}_2 \in \bar{\mathcal{G}}$ and the closure is satisfied.
Associativity: $\forall \bar{g}_1 = \bar{g}_{a_1,b_1,c_1} \in \bar{\mathcal{G}}, \forall \bar{g}_2 = \bar{g}_{a_2,b_2,c_2} \in \bar{\mathcal{G}}$ and $\forall \bar{g}_3 = \bar{g}_{a_3,b_3,c_3} \in \bar{\mathcal{G}}$, by applying the fact in Equation (A.2.15) to $\bar{g}_1\bar{g}_2$ and $(\bar{g}_1\bar{g}_2)\bar{g}_3$ in order, we have

$$(\bar{g}_1\bar{g}_2)\bar{g}_3 = \{(a_4+Nb_4)\mu + c_4, a_4^2\alpha, \alpha(2a_4+Nb_4)b_4 + \beta(a_4+Nb_4)^2\}, \tag{A.2.16}$$

where $a_4 = a_1a_2a_3$, $b_4 = a_1a_2b_3 + a_3(a_1b_2 + a_2b_1 + Nb_1b_2) + N(a_1b_2 + a_2b_1 + Nb_1b_2)b_3$ and $c_4 = [a_1a_2 + N(a_1b_2 + a_2b_1 + Nb_1b_2)]c_3 + (a_1 + Nb_1)c_2 + c_1$.
Similarly, by applying the fact in Equation (A.2.15) to $\bar{g}_2\bar{g}_2$ and $\bar{g}_1(\bar{g}_2\bar{g}_3)$ in order, we have

$$\bar{g}_1(\bar{g}_2\bar{g}_3) = \{(a_5+Nb_5)\mu + c_5, a_5^2\alpha, \alpha(2a_5+Nb_5)b_5 + \beta(a_5+Nb_5)^2\}, \tag{A.2.17}$$

where $a_5 = a_1a_2a_3$, $b_5 = a_1(a_2b_3+a_3b_2+Nb_2b_3)+a_2a_3b_1+Nb_1(a_2b_3+a_3b_2+Nb_2b_3)$ and $c_5 = (a_1+Nb_1)[(a_2+Nb_2)c_3+c_2]+c_1$. By some formula simplifications, we can see that a_4, b_4, c_4 from $(\bar{g}_1\bar{g}_2)\bar{g}_3$ are the same as a_5, b_5, c_5 from $\bar{g}_1(\bar{g}_2\bar{g}_3)$. Therefore, $(\bar{g}_1\bar{g}_2)\bar{g}_3 = \bar{g}_1(\bar{g}_2\bar{g}_3)$ and the associativity is satisfied.

Identity element: Consider $\bar{g}_{1,0,0}$ and $\forall \bar{g}_{a,b,c} \in \bar{\mathcal{G}}$, according to Equation (A.2.15), we have

$$\begin{aligned}
&\bar{g}_{a,b,c}\,\bar{g}_{1,0,0}(\{\mu,\alpha,\beta\}) = \bar{g}_{1,0,0}\,\bar{g}_{a,b,c}(\{\mu,\alpha,\beta\}) \\
&= \{(a^\star + Nb^\star)\mu + c^\star, (a^\star)^2\alpha, \alpha(2a^\star + Nb^\star)b^\star + \beta(a^\star + Nb^\star)^2\} \\
&= \bar{g}_{a,b,c}(\{\mu,\alpha,\beta\})
\end{aligned}$$

Therefore, $\bar{g}_{1,0,0}$ is the identity element in $\bar{\mathcal{G}}$.

Inverse element: $\forall \bar{g}_{a,b,c} \in \bar{\mathcal{G}}$, consider $\bar{g}_{a',b',c'}$, where $a' = \frac{1}{a}$, $b' = -\frac{b}{a^2+Nab}$, and $c' = -\frac{c}{a+Nb}$. By applying Equation (A.2.15) to $\bar{g}_{a,b,c}\,\bar{g}_{a',b',c'}$ and $\bar{g}_{a',b',c'}\,\bar{g}_{a,b,c}$, we have

$$\begin{aligned}
&\bar{g}_{a,b,c}\,\bar{g}_{a',b',c'}(\{\mu,\alpha,\beta\}) = \bar{g}_{a',b',c'}\,\bar{g}_{a,b,c}(\{\mu,\alpha,\beta\}) \\
&= \{(a_6 + Nb_6)\mu + c_6, a_6^2\alpha, \alpha(2a_6 + Nb_6)b_6 + \beta(a_6 + Nb_6)^2\},
\end{aligned}$$

where $a_6 = aa' = 1$, $b_6 = ab' + a'b + Nbb' = a'b + ab' + Nbb' = 0$ and $c_6 = (a+Nb)c' + c = (a' + Nb')c + c' = 0$. Therefore,

$$\bar{g}_{a,b,c}\,\bar{g}_{a',b',c'}(\{\mu,\alpha,\beta\}) = \bar{g}_{a',b',c}\,\bar{g}_{a,b,c}(\{\mu,\alpha,\beta\}) = \bar{g}_{1,0,0}(\{\mu,\alpha,\beta\}).$$

That is, the requirement for the inverse element is satisfied. □

Appendix B

Simulation Results of Priors

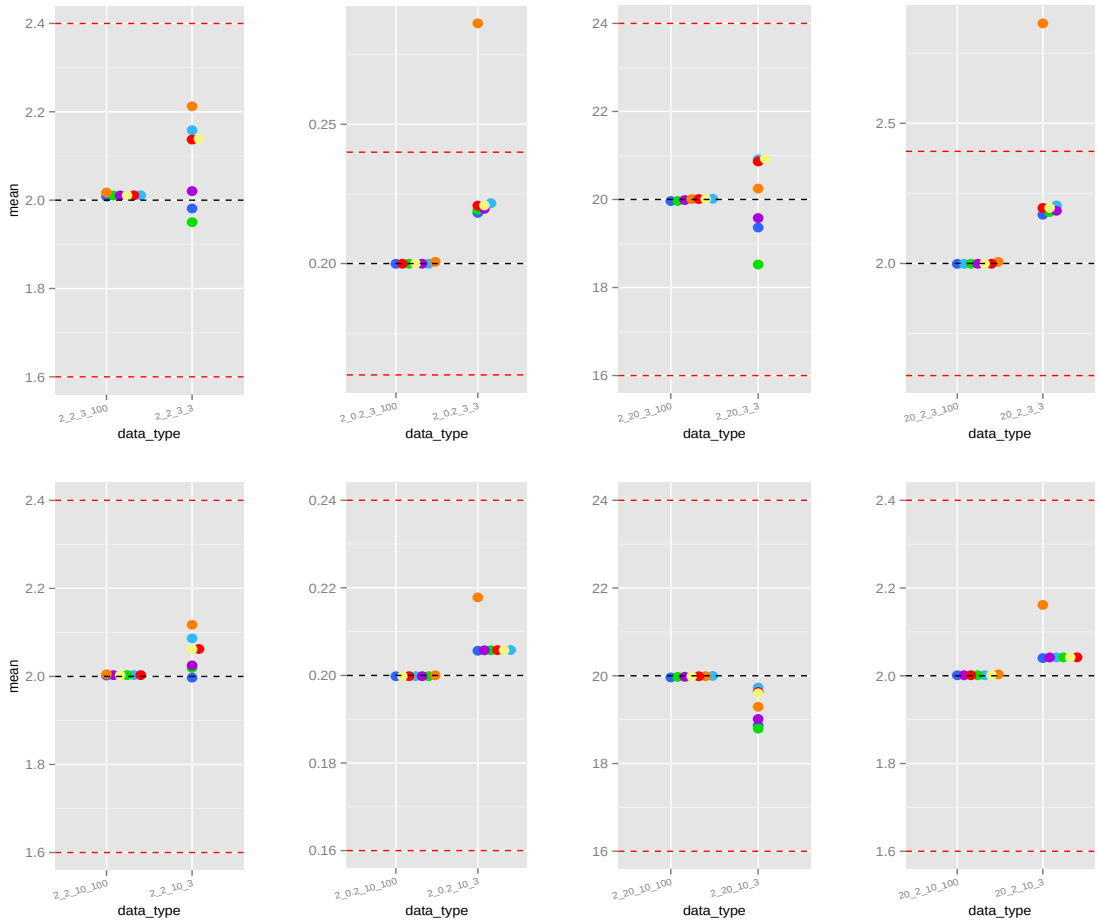


Figure B.1: Averaged posterior mean of σ across 1000 data sets for each data type.

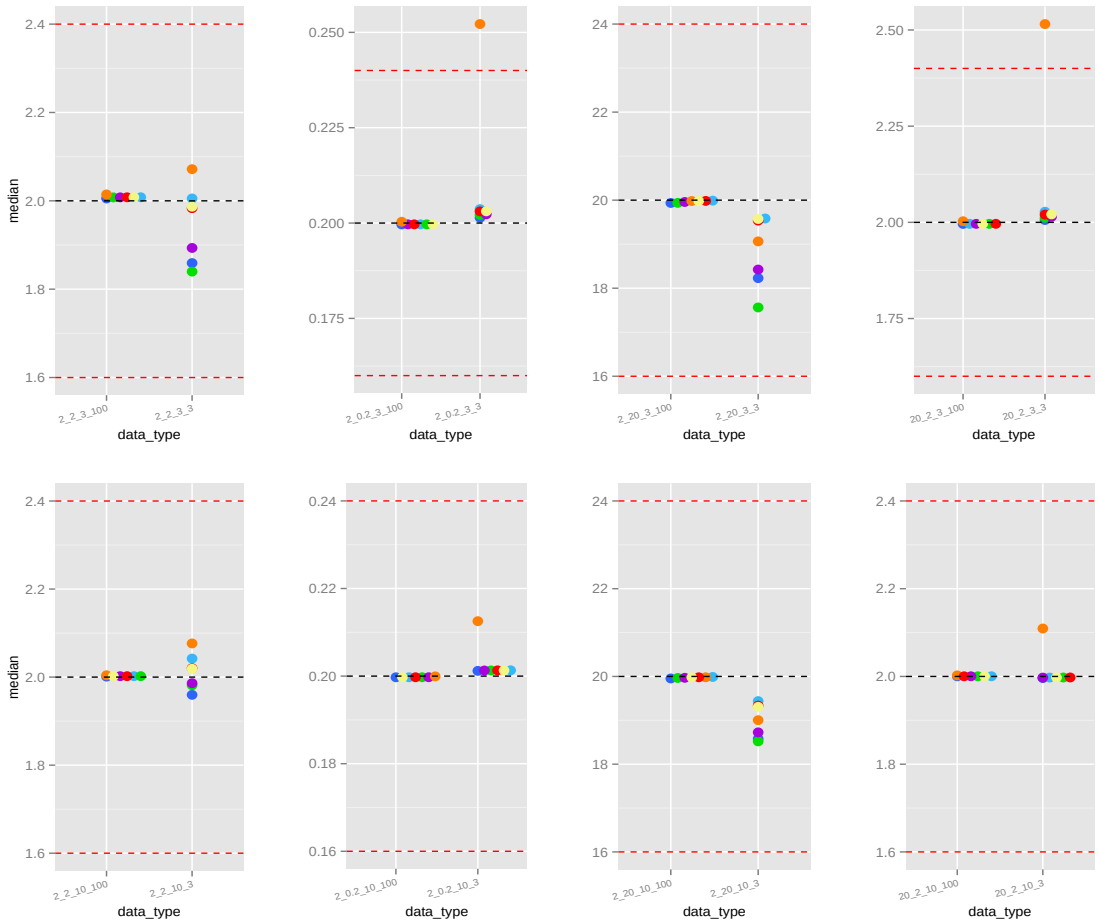


Figure B.2: Averaged posterior median of σ across 1000 data sets for each data type.

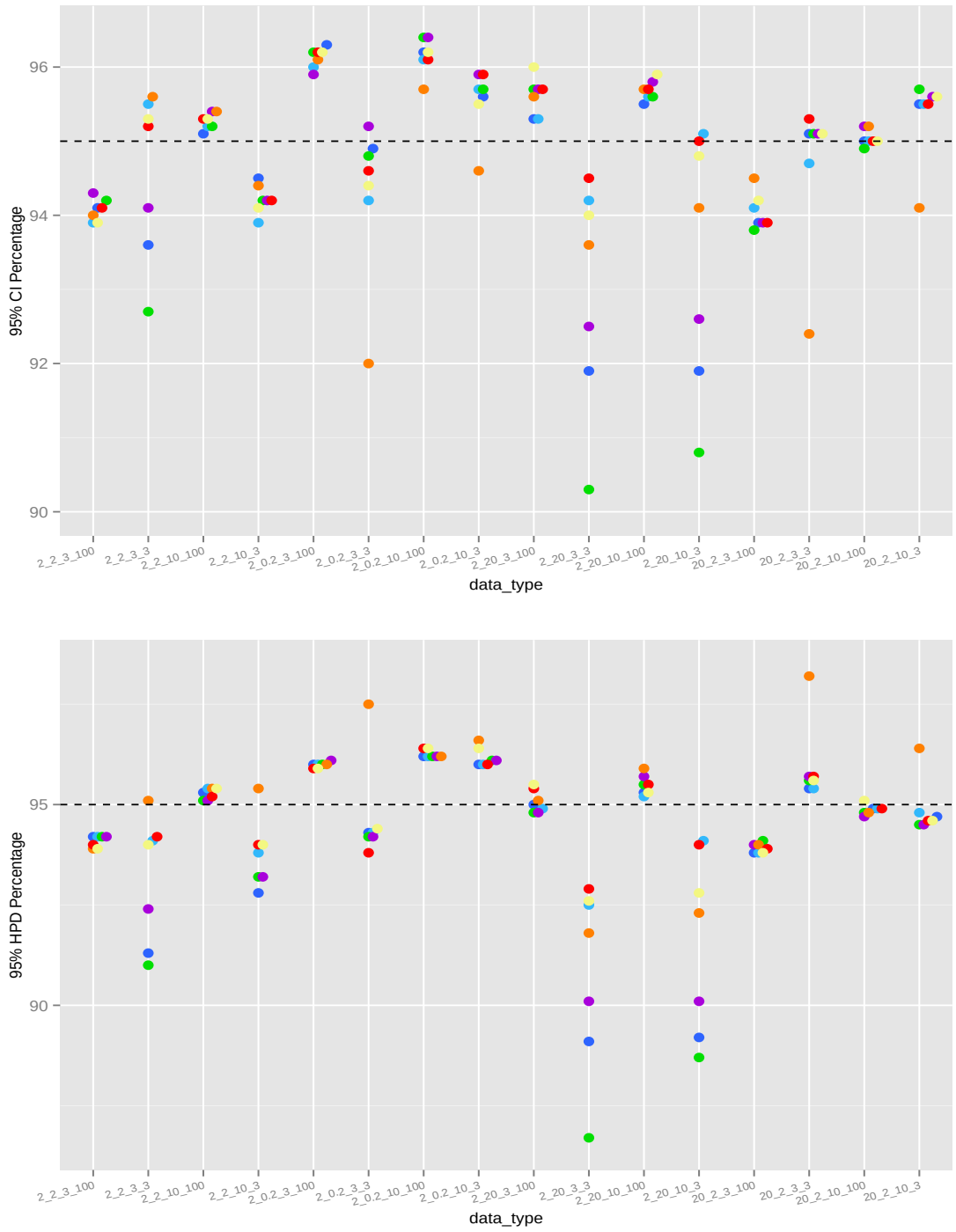


Figure B.3: Percentage for σ. Top plot: percentage of 1000 data sets for each data type that the true value lies in 95% credible interval; Bottom plot: percentage of 1000 data sets for each data type that ture value lies in 95% HPD

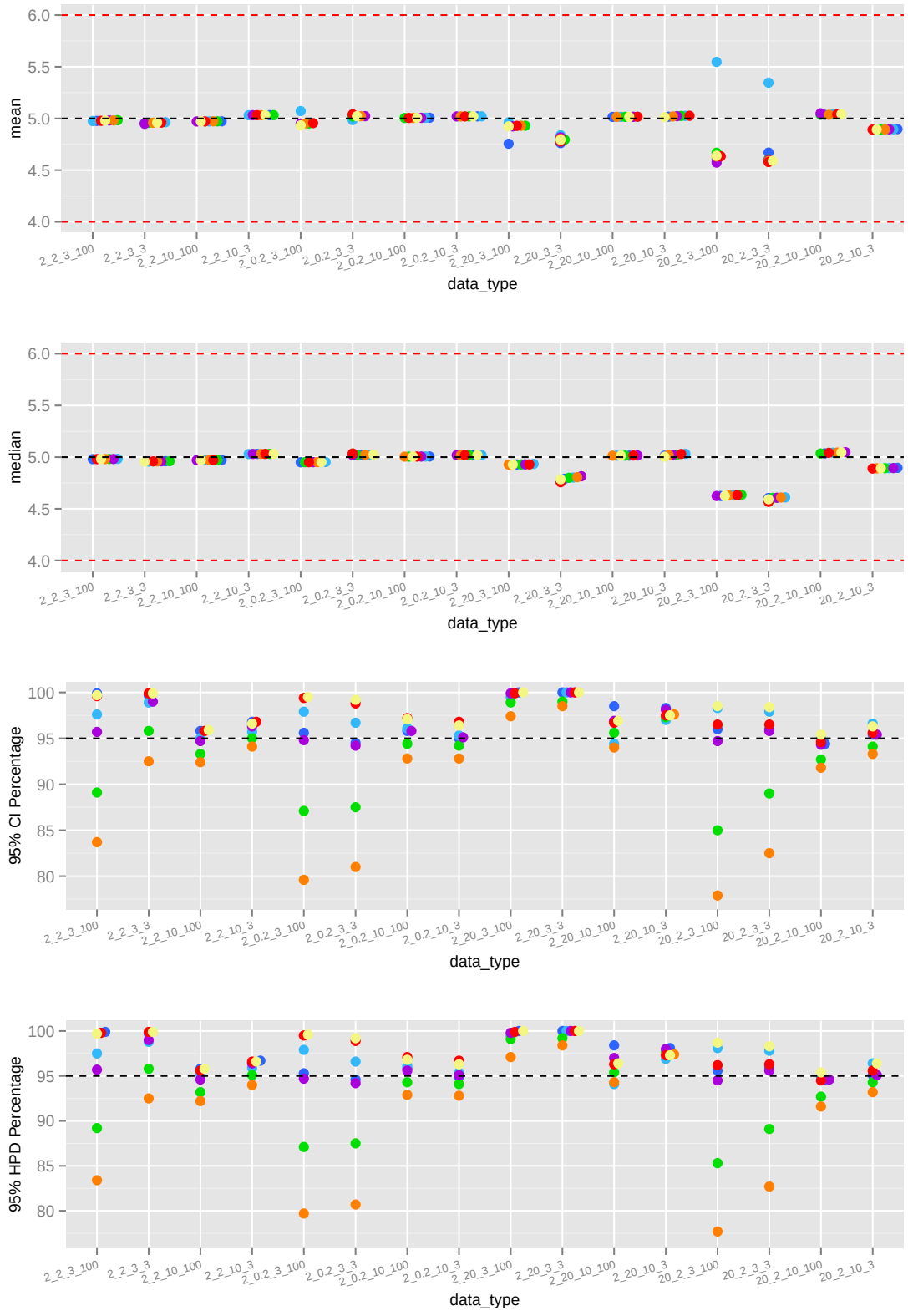


Figure B.4: For μ. The four plots focus on posterior mean, posterior median, 95% credible interval, 95 % HPD respectively

Appendix C

Detailed Calculations

C.1 Marginal Distribution $p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L})$

The marginal distribution $p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L})$ is

$$
\begin{aligned}
p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}) &= \int \cdots \int p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}, \{\lambda_i\}_{i\in I})\, \mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_{|I|} \\
&= \frac{1}{\nu_\phi^2} \int \cdots \int \prod_{i\in I} \Big\{ \frac{(\frac{1}{2}\nu_\phi)^{\frac{1}{2}\nu_\phi}}{\Gamma(\frac{1}{2}\nu_\phi)} \lambda_i^{\frac{1}{2}\nu_\phi - 1} \exp\big(-\frac{1}{2}\nu_\phi \lambda_i\big) \prod_{l=1}^{L} \prod_{t\in L_{il}} \frac{1}{\sqrt{2\pi}\sigma_{\xi l}/\sqrt{\lambda_i}} \\
&\qquad\qquad \exp\big(-\frac{\psi_{ilt}^2 \lambda_i}{2\sigma_{\xi l}^2}\big) \Big\} \mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_{|I|} \\
&= \frac{1}{\nu_\phi^2} \Big(\prod_{i\in I} \frac{(\frac{1}{2}\nu_\phi)^{\frac{1}{2}\nu_\phi}}{\Gamma(\frac{1}{2}\nu_\phi)} \Big(\prod_{l=1}^{L} \frac{1}{(\sqrt{2\pi}\sigma_{\xi l})^{|L_{il}|}} \Big) \Big) \\
&\qquad \int \cdots \int \prod_{i\in I} \Big\{ \lambda_i^{\frac{1}{2}\nu_\phi + \frac{1}{2}\sum_{l=1}^{L}|L_{il}| - 1} \exp\Big[-\lambda_i \big(\frac{1}{2}\nu_\phi + \sum_{l=1}^{L}\sum_{t\in L_{il}} \frac{\psi_{ilt}^2}{2\sigma_{\xi l}^2}\big) \Big] \Big\} \mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_{|I|}
\end{aligned}
\tag{C.1.1}
$$

It is clear that λ_i appears in the form of the pdf

$$
\Gamma\big(\frac{1}{2}\nu_\phi + \frac{1}{2}\sum_{l=1}^{L}|L_{il}|, \frac{1}{2}\nu_\phi + \sum_{l=1}^{L}\sum_{t\in L_{il}} \frac{\psi_{ilt}^2}{2\sigma_{\xi l}^2}\big)
$$

but without the normalizing constant and thus the normalizing constant could be used to do the integration in line (C.1.1) as follows

$$
\begin{aligned}
&p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}) \\
&= \frac{1}{\nu_\phi^2} \prod_{i\in I} \Bigg\{ \frac{(\frac{1}{2}\nu_\phi)^{\frac{1}{2}\nu_\phi}}{\Gamma(\frac{1}{2}\nu_\phi)} \Big(\prod_{l=1}^{L} \frac{1}{(\sqrt{2\pi}\sigma_{\xi l})^{|L_{il}|}} \Big) \frac{\Gamma(\frac{1}{2}\nu_\phi + \frac{1}{2}\sum_{l=1}^{L}|L_{il}|)}{(\frac{1}{2}\nu_\phi + \sum_{l=1}^{L}\sum_{t\in L_{il}} \frac{\psi_{ilt}^2}{2\sigma_{\xi l}^2})^{\frac{1}{2}\sum_{l=1}^{L}|L_{il}| + \frac{1}{2}\nu_\phi}} \Bigg\} \\
&= \frac{1}{\nu_\phi^2} \prod_{i\in I} \Bigg\{ (\nu_\phi \pi)^{-\frac{1}{2}\sum_{l=1}^{L}|L_{il}|} \Big(\prod_{l=1}^{L} \sigma_{\xi l}^{-|L_{il}|} \Big) \frac{\Gamma(\frac{1}{2}\nu_\phi + \frac{1}{2}\sum_{l=1}^{L}|L_{il}|)}{\Gamma(\frac{1}{2}\nu_\phi)} \\
&\qquad\qquad \Big(1 + \frac{\sum_{l=1}^{L}\sum_{t\in L_{il}} \psi_{ilt}^2/\sigma_{\xi l}^2}{\nu_\phi} \Big)^{-(\frac{1}{2}\nu_\phi + \frac{1}{2}\sum_{l=1}^{L}|L_{il}|)} \Bigg\}
\end{aligned}
$$

Denote

$$
P_i = \sum_{l=1}^{L} |L_{il}|
$$

$$
W_i = \begin{pmatrix} W_{i1} \\ W_{i2} \\ \vdots \\ W_{il} \\ \vdots \\ W_{iL} \end{pmatrix}_{P_i, 1}
$$

$$
\Sigma_i = \sigma_{\xi 1}^2 I_{|L_{i1}|,|L_{i1}|} \oplus \sigma_{\xi 2}^2 I_{|L_{i2}|,|L_{i2}|} \oplus \cdots \oplus \sigma_{\xi L}^2 I_{|L_{iL}|,|L_{iL}|}
$$

where W_{il} is the column vector of length $|L_{il}|$ with entries $\{\psi_{ilt}\}_{t\in L_{il}}$; $I_{x,x}$ is $x \times x$ dimensional identity matrix and '$\oplus$' is direct sum. The term Σ_i represents a $P_i \times P_i$ dimensional covariance matrix. Thus, the marginalized distribution can be simplified as

$$
\begin{aligned}
&p(\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}) \\
&= \frac{1}{\nu_\phi^2} \prod_{i\in I} \Bigg\{ (\nu_\phi \pi)^{-\frac{1}{2}P_i} |\Sigma_i|^{-\frac{1}{2}} \frac{\Gamma(\frac{1}{2}\nu_\phi + \frac{1}{2}P_i)}{\Gamma(\frac{1}{2}\nu_\phi)} \Big(1 + \frac{W_i^T \Sigma_i^{-1} W_i}{\nu_\phi} \Big)^{-(\frac{1}{2}\nu_\phi + \frac{1}{2}P_i)} \Bigg\} \\
&= \underbrace{\frac{1}{\nu_\phi^2}}_{\text{prior}} \underbrace{\prod_{i\in I} t_{\nu_\phi}(W_i|0, \Sigma_i)}_{\text{'likelihood'}}
\end{aligned} \tag{C.1.2}
$$

C.2 Expectations for Block $\left\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}, \{\lambda_i\}_{i\in I}\right\}$

The following calculations are under the assumption that P_i dimensional variable $W_i \sim t_{\nu_\phi}(0, \Sigma_0)$, i.e.

$$p(W_i) = C\Big(1 + \frac{1}{\nu_\phi} W_i^T \Sigma_i^{-1} W_i\Big)^{-\frac{\nu_\phi + P_i}{2}} \tag{C.2.1}$$

where

$$C = (\nu_\phi \pi)^{-\frac{P_i}{2}} \frac{\Gamma(\frac{\nu_\phi}{2} + \frac{P_i}{2})}{\Gamma(\frac{\nu_\phi}{2})|\Sigma_i|^{\frac{1}{2}}}$$

Proposition C.2.1

$$\mathbb{E}\left[\Big(1 + \frac{W_i^T \Sigma_i^{-1} W_i}{\nu_\phi}\Big)^{-n}\right] = \frac{\Gamma(\frac{\nu_\phi}{2} + \frac{P_i}{2})\Gamma(\frac{\nu_\phi + 2n}{2})}{\Gamma(\frac{\nu_\phi}{2})\Gamma(\frac{\nu_\phi + 2n}{2} + \frac{P_i}{2})} \tag{C.2.2}$$

Proof

$$\begin{aligned}
&\mathbb{E}\left[\Big(1 + \frac{W_i^T \Sigma_i^{-1} W_i}{\nu_\phi}\Big)^{-n}\right] \\
&= \int \Big(1 + \frac{W_i^T \Sigma_i^{-1} W_i}{\nu_\phi}\Big)^{-n} (\nu_\phi \pi)^{-\frac{P_i}{2}} \frac{\Gamma(\frac{\nu_\phi}{2} + \frac{P_i}{2})}{\Gamma(\frac{\nu_\phi}{2})|\Sigma_i|^{\frac{1}{2}}} \Big(1 + \frac{1}{\nu_\phi} W_i^T \Sigma_i^{-1} W_i\Big)^{-\frac{\nu_\phi + P_i}{2}} \mathrm{d}\, W_i \\
&= (\nu_\phi \pi)^{-\frac{P_i}{2}} \frac{\Gamma(\frac{\nu_\phi}{2} + \frac{P_i}{2})}{\Gamma(\frac{\nu_\phi}{2})|\Sigma_i|^{\frac{1}{2}}} \int \Big(1 + \frac{W_i^T \Sigma_i^{-1} W_i}{\nu_\phi}\Big)^{-\frac{(\nu_\phi + 2n) + P_i}{2}} \\
&= (\nu_\phi \pi)^{-\frac{P_i}{2}} \frac{\Gamma(\frac{\nu_\phi}{2} + \frac{P_i}{2})}{\Gamma(\frac{\nu_\phi}{2})|\Sigma_i|^{\frac{1}{2}}} \int \Big(1 + \frac{1}{\nu_\phi + 2n} W_i^T \big(\frac{\nu_\phi}{\nu_\phi + 2n}\Sigma_i\big)^{-1} W_i\Big)^{-\frac{(\nu_\phi + 2n) + P_i}{2}} \mathrm{d}\, W_i
\end{aligned}$$

According to the density function displayed in Equation (C.2.1), the above integration could be solved easily by using the normalizing constant of a P_i dimensional Student's t distribution $t_{\nu_\phi + 2n}(0, \frac{\nu_\phi}{\nu_\phi + 2n}\Sigma_i)$,

$$\begin{aligned}
&\mathbb{E}\left[\Big(1 + \frac{W_i^T \Sigma_i^{-1} W_i}{\nu_\phi}\Big)^{-n}\right] \\
&= (\nu_\phi \pi)^{-\frac{P_i}{2}} \frac{\Gamma(\frac{\nu_\phi}{2} + \frac{P_i}{2})}{\Gamma(\frac{\nu_\phi}{2})|\Sigma_i|^{\frac{1}{2}}} \times \Big((\nu_\phi + 2n)\pi\Big)^{\frac{P_i}{2}} \frac{\Gamma(\frac{\nu_\phi + 2n}{2})}{\Gamma(\frac{\nu_\phi + 2n}{2} + \frac{P_i}{2})} \Big|\frac{\nu_\phi}{\nu_\phi + 2n}\Sigma_i\Big|^{\frac{1}{2}} \\
&= \frac{\Gamma(\frac{\nu_\phi}{2} + \frac{P_i}{2})\Gamma(\frac{\nu_\phi + 2n}{2})}{\Gamma(\frac{\nu_\phi}{2})\Gamma(\frac{\nu_\phi + 2n}{2} + \frac{P_i}{2})}
\end{aligned}$$

. □

When $n = 1$, we have

$$C_2 = \mathbb{E}\left[\Big(1 + \frac{W_i^T \Sigma_i^{-1} W_i}{\nu_\phi}\Big)^{-1}\right] = \frac{\Gamma(\frac{\nu_\phi}{2} + \frac{P_i}{2})\Gamma(\frac{\nu_\phi}{2} + 1)}{\Gamma(\frac{\nu_\phi}{2})\Gamma(\frac{\nu_\phi}{2} + \frac{P_i}{2} + 1)} = \frac{\nu_\phi}{\nu_\phi + P_i} \tag{C.2.3}$$

Proposition C.2.2

$$\int \eta^T\eta(1+\eta^T\eta)^{-k}\mathrm{d}\,\eta = \frac{n\Gamma(\frac{3}{2})[\Gamma(\frac{1}{2})]^{n-1}\Gamma\big(k-\frac{3}{2}-\frac{1}{2}(n-1)\big)}{\Gamma(k)} \tag{C.2.4}$$

where η is n-dimensional vector

$$\eta = \begin{pmatrix}\eta_1\\ \eta_2\\ \vdots\\ \eta_n\end{pmatrix}$$

Proof

$$\begin{aligned}&\int \eta^T\eta(1+\eta^T\eta)^{-k}\mathrm{d}\,\eta = \int\cdots\int(\eta_1^2+\cdots+\eta_n^2)(1+\eta_1^2+\cdots+\eta_n^2)^{-k}\mathrm{d}\,\eta_1\cdots\mathrm{d}\,\eta_n\\ &=\int\cdots\int\Big\{\int\eta_1^2(1+\eta_1^2+\cdots+\eta_n^2)^{-k}\mathrm{d}\,\eta_1\\ &\qquad\qquad+(\eta_2^2+\cdots+\eta_n^2)\int(1+\eta_1^2+\cdots+\eta_n^2)^{-k}\mathrm{d}\,\eta_1\Big\}\mathrm{d}\,\eta_2\cdots\mathrm{d}\,\eta_n\end{aligned} \tag{C.2.5}$$

Denote

$$I_1 = \int\eta_1^2(1+\eta_1^2+\cdots+\eta_n^2)^{-k}\mathrm{d}\,\eta_1, \qquad I_2 = \int(1+\eta_1^2+\cdots+\eta_n^2)^{-k}\mathrm{d}\,\eta_1$$

As for I_1,

$$I_1 = (1+\eta_2^2+\cdots+\eta_n^2)^{-k}\int\eta_1^2\Big(1+\frac{\eta_1^2}{1+\eta_2^2+\cdots+\eta_n^2}\Big)^{-k}\mathrm{d}\,\eta_1 \tag{C.2.6}$$

Let $V = \frac{\eta_1}{(1+\eta_2^2+\cdots+\eta_n^2)^{1/2}}$, then

$$\eta_1 = \Big(1+\eta_2^2+\cdots+\eta_n^2\Big)^{\frac{1}{2}}V, \qquad \mathrm{d}\,\eta_1 = \Big(1+\eta_2^2+\cdots+\eta_n^2\Big)^{\frac{1}{2}}\mathrm{d}\,V \tag{C.2.7}$$

By substituting the above line into Equation (C.2.6), we obtain

$$I_1 = (1+\eta_2^2+\cdots+\eta_n^2)^{-k+\frac{3}{2}}\int V^2(1+V^2)^{-k}\mathrm{d}\,V$$

Taking the transformation

$$r = \frac{1}{1+V^2}, \tag{C.2.8}$$

the above line changes to be

$$\begin{aligned}I_1 &= (1+\eta_2^2+\cdots+\eta_n^2)^{-k+\frac{3}{2}}\int_0^1 r^{k-\frac{5}{2}}(1-r)^{\frac{1}{2}}\mathrm{d}r\\ &= (1+\eta_2^2+\cdots+\eta_n^2)^{-k+\frac{3}{2}}\,\mathrm{Beta}\Big(k-\frac{3}{2},\frac{3}{2}\Big)\\ &= (1+\eta_2^2+\cdots+\eta_n^2)^{-k+\frac{3}{2}}\frac{\Gamma(k-\frac{3}{2})\Gamma(\frac{3}{2})}{\Gamma(k)}\end{aligned} \tag{C.2.9}$$

As for I_2, we have

$$I_2 = (1+\eta_2^2+\cdots+\eta_n^2)^{-k}\int\Big(1+\frac{\eta_1^2}{1+\eta_2^2+\cdots+\eta_n^2}\Big)^{-k}\mathrm{d}\eta_1$$

By using the same transformation shown in Equation (C.2.7), the above line changes to be

$$I_2 = (1+\eta_2^2+\cdots+\eta_n^2)^{-k+\frac{1}{2}}\int(1+V^2)^{-k}\mathrm{d}\,V$$

Taking the following transformation

$$V^2 = U \implies V = U^{\frac{1}{2}},\ dV = U^{-\frac{1}{2}}\mathrm{d}\,U \tag{C.2.10}$$

we have

$$\begin{aligned} I_2 &= (1+\eta_2^2+\cdots+\eta_n^2)^{-k+\frac{1}{2}}\int_0^\infty U^{-\frac{1}{2}}(1+U)^{-k}\mathrm{d}\,U \\ &= (1+\eta_2^2+\cdots+\eta_n^2)^{-k+\frac{1}{2}}\,\mathrm{Beta}\Big(\frac{1}{2},k-\frac{1}{2}\Big) \\ &= (1+\eta_2^2+\cdots+\eta_n^2)^{-k+\frac{1}{2}}\frac{\Gamma(\frac{1}{2})\Gamma(k-\frac{1}{2})}{\Gamma(k)} \end{aligned} \tag{C.2.11}$$

Therefore, we have

$$\begin{aligned} &\int(\eta_1^2+\cdots+\eta_n^2)(1+\eta_1^2+\cdots+\eta_n^2)^{-k}\mathrm{d}\,\eta_1 \\ &= \frac{\Gamma(k-\frac{3}{2})\Gamma(\frac{3}{2})}{\Gamma(k)}(1+\eta_2^2+\cdots+\eta_n^2)^{-k+\frac{3}{2}}+\frac{\Gamma(\frac{1}{2})\Gamma(k-\frac{1}{2})}{\Gamma(k)}(\eta_2^2+\cdots+\eta_n^2)(1+\eta_2^2+\cdots+\eta_n^2)^{-k+\frac{1}{2}} \end{aligned} \tag{C.2.12}$$

Integrating the above line further with respect to η_2, we obtain

$$\begin{aligned} &\int\int(\eta_1^2+\cdots+\eta_n^2)(1+\eta_1^2+\cdots+\eta_n^2)^{-k}\mathrm{d}\,\eta_1\mathrm{d}\,\eta_2 \\ &= \frac{\Gamma(k-\frac{3}{2})\Gamma(\frac{3}{2})}{\Gamma(k)}\int(1+\eta_2^2+\cdots+\eta_n^2)^{-k+\frac{3}{2}}\mathrm{d}\,\eta_2 \\ &\qquad + \frac{\Gamma(\frac{1}{2})\Gamma(k-\frac{1}{2})}{\Gamma(k)}\int(\eta_2^2+\cdots+\eta_n^2)(1+\eta_2^2+\cdots+\eta_n^2)^{-k+\frac{1}{2}}\mathrm{d}\,\eta_2 \end{aligned}$$

According to Equation (C.2.11) for I_2 and Equation (C.2.12), the above line could be

changed to be

$$\begin{aligned}
&\int\int(\eta_1^2+\cdots+\eta_n^2)(1+\eta_1^2+\cdots+\eta_n^2)^{-k}\mathrm{d}\,\eta_1\mathrm{d}\,\eta_2\\
&=\frac{\Gamma(k-\frac{3}{2})\Gamma(\frac{3}{2})}{\Gamma(k)}\frac{\Gamma(\frac{1}{2})\Gamma(k-\frac{3}{2}-\frac{1}{2})}{\Gamma(k-\frac{3}{2})}(1+\eta_3^2+\cdots+\eta_n^2)^{-k+\frac{3}{2}+\frac{1}{2}}\\
&\qquad+\frac{\Gamma(\frac{1}{2})\Gamma(k-\frac{1}{2})}{\Gamma(k)}\left\{\frac{\Gamma(k-\frac{1}{2}-\frac{3}{2})\Gamma(\frac{3}{2})}{\Gamma(k-\frac{1}{2})}(1+\eta_3^2+\cdots+\eta_n^2)^{-k+\frac{1}{2}+\frac{3}{2}}\right.\\
&\qquad\qquad\left.+\frac{\Gamma(\frac{1}{2})\Gamma(k-\frac{1}{2}-\frac{1}{2})}{\Gamma(k-\frac{1}{2})}(\eta_3^2+\cdots+\eta_n^2)(1+\eta_3^2+\cdots+\eta_n^2)^{-k+\frac{1}{2}+\frac{1}{2}}\right.\\
&=2\,\frac{\Gamma(\frac{3}{2})\Gamma(\frac{1}{2})\Gamma(k-\frac{3}{2}-\frac{1}{2})}{\Gamma(k)}(1+\eta_3^2+\cdots+\eta_n^2)^{-k+\frac{3}{2}+\frac{1}{2}}\\
&\qquad+\frac{\left[\Gamma(\frac{1}{2})\right]^2\Gamma(k-2\times\frac{1}{2})}{\Gamma(k)}(\eta_3^2+\cdots+\eta_n^2)(1+\eta_3^2+\cdots+\eta_n^2)^{-k+2\times\frac{1}{2}}
\end{aligned}\tag{C.2.13}$$

Similarly, the integration with respect to η_3 of the above line is

$$\begin{aligned}
&\int\int\int(\eta_1^2+\cdots+\eta_n^2)(1+\eta_1^2+\cdots+\eta_n^2)^{-k}\mathrm{d}\,\eta_1\mathrm{d}\,\eta_2\mathrm{d}\,\eta_3\\
&=3\,\frac{\Gamma(\frac{3}{2})\Gamma(\frac{1}{2})\Gamma(\frac{1}{2})\Gamma(k-\frac{3}{2}-2\times\frac{1}{2})}{\Gamma(k)}(1+\eta_4^2+\cdots+\eta_n^2)^{-k+\frac{3}{2}+2\times\frac{1}{2}}\\
&\qquad+\frac{\left[\Gamma(\frac{1}{2})\right]^3\Gamma(k-3\times\frac{1}{2})}{\Gamma(k)}(\eta_4^2+\cdots+\eta_n^2)(1+\eta_4^2+\cdots+\eta_n^2)^{-k+3\times\frac{1}{2}}
\end{aligned}\tag{C.2.14}$$

After integrating out the first $n-1$ element of η, we have

$$\begin{aligned}
&\int\cdots\int(\eta_1^2+\cdots+\eta_n^2)(1+\eta_1^2+\cdots+\eta_n^2)^{-k}\mathrm{d}\,\eta_1\cdots\mathrm{d}\,\eta_{n-1}\\
&=(n-1)\,\frac{\Gamma(\frac{3}{2})\left[\Gamma(\frac{1}{2})\right]^{n-2}\Gamma(k-\frac{3}{2}-(n-2)\times\frac{1}{2})}{\Gamma(k)}(1+\eta_n^2)^{-k+\frac{3}{2}+(n-2)\times\frac{1}{2}}\\
&\qquad+\frac{\left[\Gamma(\frac{1}{2})\right]^{n-1}\Gamma(k-(n-1)\times\frac{1}{2})}{\Gamma(k)}\eta_n^2(1+\eta_n^2)^{-k+(n-1)\times\frac{1}{2}}
\end{aligned}\tag{C.2.15}$$

By integrating out the last term η_n from the above line, we obtain

$$\begin{aligned}
&\int\cdots\int(\eta_1^2+\cdots+\eta_n^2)(1+\eta_1^2+\cdots+\eta_n^2)^{-k}\mathrm{d}\,\eta_1\cdots\mathrm{d}\,\eta_n\\
&=(n-1)\,\frac{\Gamma(\frac{3}{2})\left[\Gamma(\frac{1}{2})\right]^{n-2}\Gamma(k-\frac{3}{2}-(n-2)\times\frac{1}{2})}{\Gamma(k)}\times\frac{\Gamma(\frac{1}{2})\Gamma\big(k-\frac{3}{2}-\frac{1}{2}(n-2)-\frac{1}{2}\big)}{\Gamma\big(k-\frac{3}{2}-\frac{1}{2}(n-2)\big)}\\
&\qquad+\frac{\left[\Gamma(\frac{1}{2})\right]^{n-1}\Gamma(k-(n-1)\times\frac{1}{2})}{\Gamma(k)}\times\frac{\Gamma(\frac{3}{2})\Gamma(k-(n-1)\times\frac{1}{2}-\frac{3}{2})}{\Gamma(k-(n-1)\times\frac{1}{2})}\\
&=n\,\frac{\Gamma(\frac{3}{2})\left[\Gamma(\frac{1}{2})\right]^{n-1}\Gamma(k-\frac{3}{2}-\frac{1}{2}(n-1))}{\Gamma(k)}
\end{aligned}\tag{C.2.16}$$

. □

Proposition C.2.3

$$\int (1+\eta^T\eta)^{-k}\mathrm{d}\,\eta = \frac{\Gamma(k-\frac{1}{2}n)[\Gamma(\frac{1}{2})]^n}{\Gamma(k)} \tag{C.2.17}$$

Proof

$$\int (1+\eta^T\eta)^{-k}\mathrm{d}\,\eta = \int\cdots\int\Big(\int (1+\eta_1^2+\cdots+\eta_n^2)^{-k}\mathrm{d}\,\eta_1\Big)\mathrm{d}\,\eta_2\cdots\mathrm{d}\,\eta_n$$

Recall the result for I_2 displayed in Equation (C.2.12), the above line changes to be

$$\int (1+\eta^T\eta)^{-k}\mathrm{d}\,\eta = \frac{\Gamma(k-\frac{1}{2})\Gamma(\frac{1}{2})}{\Gamma(k)}\int\cdots\int (1+\eta_2^2+\cdots+\eta_n^2)^{-k+\frac{1}{2}}\mathrm{d}\,\eta_2\cdots\eta_n$$

Apply the result for I_2 repeatedly on the above formula, we have

$$\begin{aligned}
&\int (1+\eta^T\eta)^{-k}\mathrm{d}\,\eta\\
&= \frac{\Gamma(k-\frac{1}{2})\Gamma(\frac{1}{2})}{\Gamma(k)}\times\frac{\Gamma(k-\frac{1}{2}\times 2)\Gamma(\frac{1}{2})}{\Gamma(k-\frac{1}{2})}\times\cdots\frac{\Gamma(k-\frac{1}{2}(n-1))\Gamma(\frac{1}{2})}{\Gamma(k-\frac{1}{2}(n-2))}\times\frac{\Gamma(k-\frac{1}{2}n)\Gamma(\frac{1}{2})}{\Gamma(k-\frac{1}{2}(n-1))}\\
&= \frac{\Gamma(k-\frac{1}{2}n)[\Gamma(\frac{1}{2})]^n}{\Gamma(k)}
\end{aligned} \tag{C.2.18}$$

. □

Proposition C.2.4

$$\int (\eta^T\eta)^2(1+\eta^T\eta)^{-k}\mathrm{d}\,\eta = \frac{n(n+2)\pi^{\frac{n}{2}}}{4}\frac{\Gamma(k-\frac{1}{2}n-2)}{\Gamma(k)} \tag{C.2.19}$$

Proof

$$\int (\eta^T\eta)^2(1+\eta^T\eta)^{-k}\mathrm{d}\,\eta = \int\cdots\int(\eta_1^2+\cdots\eta_n^2)^2(1+\eta_1^2+\cdots\eta_n^2)^{-k}\mathrm{d}\,\eta_1\cdots\mathrm{d}\,\eta_n$$

Firstly, we investigate the integration with respect to only η_1,

$$\begin{aligned}
&\int (\eta_1^2+\cdots\eta_n^2)^2(1+\eta_1^2+\cdots\eta_n^2)^{-k}\mathrm{d}\,\eta_1\\
&= (1+\eta_1^2+\cdots\eta_n^2)^{-k}\int\Big(\eta_1^4+(\eta_2^2+\cdots+\eta_n^2)^2+2\eta_1^2(\eta_2^2+\cdots\eta_n^2)\Big)\\
&\qquad\qquad\times\Big(1+\frac{\eta_1^2}{1+\eta_2^2+\cdots+\eta_n^2}\Big)^{-k}\mathrm{d}\,\eta_1\\
&= (1+\eta_2^2+\cdots\eta_n^2)^{-k}\int\eta_1^4\Big(1+\frac{\eta_1^2}{1+\eta_2^2+\cdots+\eta_n^2}\Big)^{-k}\mathrm{d}\,\eta_1\\
&\qquad+(\eta_2^2+\cdots+\eta_n^2)^2\int\Big(1+\eta_1^2+\cdots+\eta_n^2\Big)^{-k}\mathrm{d}\,\eta_1\\
&\qquad+2(\eta_2^2+\cdots+\eta_n^2)\int\eta_1^2\Big(1+\eta_1^2+\cdots+\eta_n^2\Big)^{-k}\mathrm{d}\,\eta_1
\end{aligned} \tag{C.2.20}$$

Let

$$I_3 = \int \eta_1^4 \Big(1 + \frac{\eta_1^2}{1+\eta_2^2+\cdots+\eta_n^2}\Big)^{-k} \mathrm{d}\,\eta_1 \tag{C.2.21}$$

For I_3 in Equation (C.2.21), by taking the transformation displayed in Equation (C.2.7) we have

$$I_3 = (1+\eta_2^2+\cdots+\eta_n^2)^{\frac{5}{2}} \int V^4 (1+V^2)^{-k} \mathrm{d}\,V$$

Make further transformation as shown in Equation (C.2.8), the above line changes to be

$$\begin{aligned} I_3 &= (1+\eta_2^2+\cdots+\eta_n^2)^{\frac{5}{2}} \int_0^1 r^{k-\frac{7}{2}} (1-r)^{\frac{3}{2}} \mathrm{d}\,r \\ &= \mathrm{Beta}\Big(k-\frac{5}{2}, \frac{5}{2}\Big) \\ &= (1+\eta_2^2+\cdots+\eta_n^2)^{\frac{5}{2}} \frac{\Gamma(k-\frac{5}{2})\Gamma(\frac{5}{2})}{\Gamma(k)} \end{aligned} \tag{C.2.22}$$

Recall the results of I_1 and I_2 shown in Equation (C.2.9) and (C.2.11), we have

$$\begin{aligned} &\int (\eta_1^2+\cdots\eta_n^2)^2 (1+\eta_1^2+\cdots\eta_n^2)^{-k} \mathrm{d}\,\eta_1 \\ &= \frac{\Gamma(k-\frac{5}{2})\Gamma(\frac{5}{2})}{\Gamma(k)} (1+\eta_2^2+\cdots+\eta_n^2)^{-k+\frac{5}{2}} \\ &\qquad + \frac{\Gamma(k-\frac{1}{2})\Gamma(\frac{1}{2})}{\Gamma(k)} (\eta_2^2+\cdots+\eta_n^2)^2 (1+\eta_2^2+\cdots+\eta_n^2)^{-k+\frac{1}{2}} \\ &\qquad\quad + 2\frac{\Gamma(k-\frac{3}{2})\Gamma(\frac{3}{2})}{\Gamma(k)} (\eta_2^2+\cdots+\eta_n^2)(1+\eta_2^2+\cdots+\eta_n^2)^{-k+\frac{3}{2}} \end{aligned}$$

We take further integrations of the above line with respect to the rest components,

$$\begin{aligned} &\int\cdots\int (\eta_1^2+\cdots\eta_n^2)^2 (1+\eta_1^2+\cdots\eta_n^2)^{-k} \mathrm{d}\,\eta_1\cdots\mathrm{d}\,\eta_n \\ &= \frac{\Gamma(k-\frac{5}{2})\Gamma(\frac{5}{2})}{\Gamma(k)} \int\cdots\int (1+\eta_2^2+\cdots+\eta_n^2)^{-k+\frac{5}{2}} \mathrm{d}\,\eta_2\cdots\mathrm{d}\,\eta_n \\ &+ \int\cdots\int \frac{\Gamma(k-\frac{1}{2})\Gamma(\frac{1}{2})}{\Gamma(k)} (\eta_2^2+\cdots+\eta_n^2)^2 (1+\eta_2^2+\cdots+\eta_n^2)^{-k+\frac{1}{2}} \mathrm{d}\,\eta_2\cdots\mathrm{d}\,\eta_n \\ &+ 2\frac{\Gamma(k-\frac{3}{2})\Gamma(\frac{3}{2})}{\Gamma(k)} \int\cdots\int (\eta_2^2+\cdots+\eta_n^2)(1+\eta_2^2+\cdots+\eta_n^2)^{-k+\frac{3}{2}} \mathrm{d}\,\eta_2\cdots\mathrm{d}\,\eta_n \end{aligned} \tag{C.2.23}$$

It is easy to recognize that the integration component of first term and third term in the

above line have the same form as those in Equation (C.2.17) and (C.2.4), thus

$$\begin{aligned}
&\int\cdots\int(\eta_1^2+\cdots\eta_n^2)^2(1+\eta_1^2+\cdots\eta_n^2)^{-k}\mathrm{d}\,\eta_1\cdots\mathrm{d}\,\eta_n\\
&=\frac{\Gamma\big(k-\frac{5}{2}-\frac{1}{2}(n-1)\big)\Gamma(\frac{5}{2})[\Gamma(\frac{1}{2})]^{n-1}}{\Gamma(k)}\\
&\qquad\qquad+2\frac{(n-1)[\Gamma(\frac{3}{2})]^2[\Gamma(\frac{1}{2})]^{n-2}\Gamma\big(k-\frac{3}{2}\times 2-\frac{1}{2}(n-2)\big)}{\Gamma(k)}\\
&\quad+\frac{\Gamma(k-\frac{1}{2})\Gamma(\frac{1}{2})}{\Gamma(k)}\int\cdots\int(\eta_2^2+\cdots+\eta_n^2)^2(1+\eta_2^2+\cdots+\eta_n^2)^{-k+\frac{1}{2}}\mathrm{d}\,\eta_2\cdots\mathrm{d}\,\eta_n
\end{aligned}\tag{C.2.24}$$

Applying the fact stated by the above equation on its own last term, we have

$$\begin{aligned}
&\int\cdots\int(\eta_1^2+\cdots\eta_n^2)^2(1+\eta_1^2+\cdots\eta_n^2)^{-k}\mathrm{d}\,\eta_1\cdots\mathrm{d}\,\eta_n\\
&=\frac{\Gamma\big(k-\frac{5}{2}-\frac{1}{2}(n-1)\big)\Gamma(\frac{5}{2})[\Gamma(\frac{1}{2})]^{n-1}}{\Gamma(k)}\\
&\qquad\qquad+2\frac{(n-1)[\Gamma(\frac{3}{2})]^2[\Gamma(\frac{1}{2})]^{n-2}\Gamma\big(k-\frac{3}{2}\times 2-\frac{1}{2}(n-2)\big)}{\Gamma(k)}\\
&\quad+\frac{\Gamma(k-\frac{1}{2})\Gamma(\frac{1}{2})}{\Gamma(k)}\Big\{\frac{\Gamma\big(k-\frac{1}{2}-\frac{5}{2}-\frac{1}{2}(n-2)\big)\Gamma(\frac{5}{2})[\Gamma(\frac{1}{2})]^{n-2}}{\Gamma(k-\frac{1}{2})}\\
&\qquad\qquad+2\frac{(n-2)[\Gamma(\frac{3}{2})]^2[\Gamma(\frac{1}{2})]^{n-3}\Gamma\big(k-\frac{1}{2}-\frac{3}{2}\times 2-\frac{1}{2}(n-3)\big)}{\Gamma(k-\frac{1}{2})}\\
&\quad+\frac{\Gamma(k-\frac{1}{2}-\frac{1}{2})\Gamma(\frac{1}{2})}{\Gamma(k-\frac{1}{2})}\int\cdots\int(\eta_3^2+\cdots+\eta_n^2)^2(1+\eta_2^2+\cdots+\eta_n^2)^{-k+\frac{1}{2}+\frac{1}{2}}\mathrm{d}\,\eta_3\cdots\mathrm{d}\,\eta_n\Big\}\\
&=2\frac{\Gamma\big(k-\frac{5}{2}-\frac{1}{2}(n-1)\big)\Gamma(\frac{5}{2})[\Gamma(\frac{1}{2})]^{n-1}}{\Gamma(k)}\\
&\qquad\qquad+2\frac{[\Gamma(\frac{3}{2})]^2[\Gamma(\frac{1}{2})]^{n-2}\Gamma\big(k-\frac{3}{2}\times 2-\frac{1}{2}(n-2)\big)}{\Gamma(k)}[(n-1)+(n-2)]\\
&\quad+\frac{[\Gamma(\frac{1}{2})]^2\Gamma(k-\frac{1}{2}\times 2)}{\Gamma(k)}\int\cdots\int(\eta_3^2+\cdots+\eta_n^2)^2(1+\eta_2^2+\cdots+\eta_n^2)^{-k+\frac{1}{2}\times 2}\mathrm{d}\,\eta_3\cdots\mathrm{d}\,\eta_n
\end{aligned}$$

By repeating this process until η_n,

$$\begin{aligned}
&\int\cdots\int(\eta_1^2+\cdots\eta_n^2)^2(1+\eta_1^2+\cdots\eta_n^2)^{-k}\mathrm{d}\,\eta_1\cdots\mathrm{d}\,\eta_n\\
&=(n-1)\frac{\Gamma\big(k-\frac{5}{2}-\frac{1}{2}(n-1)\big)\Gamma(\frac{5}{2})[\Gamma(\frac{1}{2})]^{n-1}}{\Gamma(k)}\\
&\qquad+n(n-1)\frac{[\Gamma(\frac{3}{2})]^2[\Gamma(\frac{1}{2})]^{n-2}\Gamma\big(k-\frac{3}{2}\times 2-\frac{1}{2}(n-2)\big)}{\Gamma(k)}\\
&\qquad+\frac{[\Gamma(\frac{1}{2})]^{n-1}\Gamma(k-\frac{1}{2}(n-1))}{\Gamma(k)}\int\eta_n^4(1+\eta_2^2)^{-k+\frac{1}{2}(n-1)}\mathrm{d}\,\eta_n
\end{aligned}\tag{C.2.25}$$

Taking the following transformation

$$r=\frac{1}{1+\eta_n^2}$$

we have

$$\int \eta_n^4(1+\eta_2^2)^{-k+\frac{1}{2}(n-1)}\mathrm{d}\,\eta_n = \int r^{k-\frac{1}{2}n-3}(1-r)^{\frac{3}{2}}\mathrm{d}\,r = \mathrm{Beta}(k-\frac{1}{2}n-2,\frac{5}{2}) \quad \text{(C.2.26)}$$

Substitute the above result into Equation (C.2.25),

$$\begin{aligned}
&\int\cdots\int(\eta_1^2+\cdots\eta_n^2)^2(1+\eta_1^2+\cdots\eta_n^2)^{-k}\mathrm{d}\,\eta_1\cdots\mathrm{d}\,\eta_n \\
&= \left(n\Gamma(\frac{5}{2})[\Gamma(\frac{1}{2})]^{n-1} + n(n-1)[\Gamma(\frac{3}{2})]^2[\Gamma(\frac{1}{2})]^{n-2}\right)\frac{\Gamma(k-\frac{1}{2}n-2)}{\Gamma(k)} \\
&= \frac{n(n+2)\pi^{\frac{n}{2}}}{4}\frac{\Gamma(k-\frac{1}{2}n-2)}{\Gamma(k)}
\end{aligned}$$

. □

Proposition C.2.5

$$\mathbb{E}\left[(W_i^T\Sigma_i^{-1}W_i)\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-2}\right] = \frac{\nu_\phi^2 P_i}{(\nu_\phi+P_i+2)(\nu_\phi+P_i)} \quad \text{(C.2.27)}$$

Proof

$$\begin{aligned}
&\mathbb{E}\left[(W_i^T\Sigma_i^{-1}W_i)\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-2}\right] = \int(W_i^T\Sigma_i^{-1}W_i)\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-2}p(W_i)\mathrm{d}\,W_i \\
&= C\int(W_i^T\Sigma_i^{-1}W_i)\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-\frac{\nu_\phi+4+P_i}{2}}\mathrm{d}\,W_i
\end{aligned}$$

By taking the transformation

$$X_i = \frac{1}{\sqrt{\nu_\phi}}\Sigma_i^{-\frac{1}{2}}W_i \implies W_i = \sqrt{\nu_\phi}\Sigma_i^{\frac{1}{2}}X_i,\ \mathrm{d}\,W_i = |\sqrt{\nu_\phi}\Sigma_i^{\frac{1}{2}}|\mathrm{d}\,X_i$$

we have

$$\begin{aligned}
&\mathbb{E}\left[(W_i^T\Sigma_i^{-1}W_i)\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-2}\right] \\
&\qquad = C\nu_\phi\big|\sqrt{\nu_\phi}\Sigma_i^{\frac{1}{2}}\big|\int X_i^TX_i(1+X_i^TX_i)^{-\frac{\nu_\phi+4+P_i}{2}}\mathrm{d}\,X_i
\end{aligned}$$

According to Equation (C.2.4), the above line changes to be

$$\begin{aligned}
&\mathbb{E}\left[(W_i^T\Sigma_i^{-1}W_i)\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-2}\right] \\
&= C\nu_\phi\big|\sqrt{\nu_\phi}\Sigma_i^{\frac{1}{2}}\big|P_i\,\frac{\Gamma(\frac{3}{2})\big[\Gamma(\frac{1}{2})\big]^{P_i-1}\Gamma\big(\frac{\nu_\phi+4+P_i}{2}-\frac{3}{2}-\frac{1}{2}(P_i-1)\big)}{\Gamma(\frac{\nu_\phi+4+P_i}{2})}
\end{aligned} \quad \text{(C.2.28)}$$

Note that C is the normalizing constant of multivariate Student's t distribution shown in Equation (C.2.1) and the following fact about Gamma function

$$\Gamma(z+1) = z\Gamma(z),\ \Gamma(\frac{1}{2}) = \sqrt{\pi}$$

Equation (C.2.28) changes to be

$$\mathbb{E}\left[(W_i^T\Sigma_i^{-1}W_i)\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-2}\right]=\frac{\nu_\phi^2 P_i}{(\nu_\phi+P_i+2)(\nu_\phi+P_i)}$$

. □

Proposition C.2.6

$$\mathbb{E}\left[W_{il}^T\Sigma_{il}^{-1}W_{il}\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-1}\right]=\frac{|L_{il}|\nu_\phi}{\nu_\phi+P_i} \tag{C.2.29}$$

Proof Since W_i follows a multivariate Student's t distribution $t_{\nu_\phi}(0,\Sigma_i)$, it is easy to obtain

$$\begin{aligned}
&\mathbb{E}\left[W_{il}^T\Sigma_{il}^{-1}W_{il}\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-1}\right]\\
&=\int W_{il}^T\Sigma_{il}^{-1}W_{il}\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-1}p(W_i)\mathrm{d}\,W_i\\
&=C\int W_{il}^T\Sigma_{il}^{-1}W_{il}\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-\frac{\nu_\phi+P_i}{2}-1}\mathrm{d}\,W_i\\
&=C\int\int W_{il}^T\Sigma_{il}^{-1}W_{il}\Big(1+\frac{W_{il}^T\Sigma_{il}^{-1}W_{il}}{\nu_\phi}+\frac{W_{-il}^T\Sigma_{-il}^{-1}W_{il}}{\nu_\phi}\Big)^{-\frac{\nu_\phi+P_i}{2}-1}\mathrm{d}\,W_{il}\mathrm{d}\,W_{-il}\\
&=C\int\Big(1+\frac{W_{-il}^T\Sigma_{-il}^{-1}W_{il}}{\nu_\phi}\Big)^{-\frac{\nu_\phi+P_i}{2}-1}\\
&\qquad\times\left\{\int W_{il}^T\Sigma_{il}^{-1}W_{il}\Big(1+\frac{W_{il}^T\Sigma_{il}^{-1}W_{il}}{\nu_\phi+W_{-il}^T\Sigma_{-il}^{-1}W_{-il}}\Big)^{-\frac{\nu_\phi+P_i}{2}-1}\mathrm{d}\,W_{il}\right\}\mathrm{d}\,W_{-il}
\end{aligned} \tag{C.2.30}$$

where W_{-il} denotes all the elements in W_i except those in W_{il} and Σ_{-il} denotes the covariance matrix for W_{-il}. Denote the dimension of W_{-il} by P_{-il}, it is obvious that $P_{-il}=P_i-|L_{il}|$.

Taking the following transformation

$$X_{il}=\frac{1}{(\nu_\phi+W_{-il}^T\Sigma_{-il}^{-1}W_{-il})^{\frac{1}{2}}}\Sigma_{il}^{-\frac{1}{2}}W_{il} \tag{C.2.31}$$

$$\implies W_{il}=(\nu_\phi+W_{-il}^T\Sigma_{-il}^{-1}W_{-il})^{\frac{1}{2}}\Sigma_{il}^{\frac{1}{2}}X_{il},\ \mathrm{d}\,W_{il}=(\nu_\phi+W_{-il}^T\Sigma_{-il}^{-1}W_{-il})^{\frac{|L_{il}|}{2}}|\Sigma_{il}|^{\frac{1}{2}}\mathrm{d}\,X_{il}$$

the formula in Equation (C.2.30) could be rewritten as

$$\begin{aligned}
&\mathbb{E}\left[W_{il}^T\Sigma_{il}^{-1}W_{il}\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-1}\right]\\
&=C\nu_\phi^{1+\frac{1}{2}|L_{il}|}\big|\Sigma_{il}\big|^{\frac{1}{2}}\int\left(1+\frac{W_{-il}^T\Sigma_{-il}^{-1}W_{-il}}{\nu_\phi}\right)^{-\frac{\nu_\phi}{2}-\frac{P_{-il}}{2}}\\
&\qquad\times\int X_{il}^TX_{il}\Big(1+X_{il}^TX_{il}\Big)^{-\frac{\nu_\phi+P_i}{2}-1}\mathrm{d}\,X_{il}\mathrm{d}\,W_{-il}
\end{aligned}$$

According to Equation (C.2.4), the above line changes to be

$$\begin{aligned}
&\mathbb{E}\left[W_{il}^T\Sigma_{il}^{-1}W_{il}\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-1}\right] \\
&= C\nu_\phi^{1+\frac{1}{2}|L_{il}|}\big|\Sigma_{il}\big|^{\frac{1}{2}}|L_{il}|\frac{\Gamma(\frac{3}{2})\big[\Gamma(\frac{1}{2})\big]^{|L_{il}|-1}\Gamma(\frac{\nu_\phi}{2}+\frac{P_{-il}}{2})}{\Gamma(\frac{\nu_\phi}{2}+\frac{P_i}{2}+1)} \\
&\qquad\qquad \times\int\left(1+\frac{W_{-il}^T\Sigma_{-il}^{-1}W_{-il}}{\nu_\phi}\right)^{-\frac{\nu_\phi}{2}-\frac{P_{-il}}{2}}\mathrm{d}\,W_{-il}
\end{aligned} \tag{C.2.32}$$

By using the normalizing constant of $\frac{P_{-il}}{2}$-dimensional multivariate Student's t distribution $t_{\nu_\phi}\big(0,\Sigma_{-il}\big)$, the integration has the following result

$$\int\left(1+\frac{W_{-il}^T\Sigma_{-il}^{-1}W_{-il}}{\nu_\phi}\right)^{-\frac{\nu_\phi}{2}-\frac{P_{-il}}{2}}\mathrm{d}\,W_{-il} = \big(\nu_\phi\pi\big)^{\frac{P_{-il}}{2}}\frac{\Gamma(\frac{\nu_\phi}{2})|\Sigma_{-il}|^{\frac{1}{2}}}{\Gamma(\frac{\nu_\phi}{2}+\frac{P_{-il}}{2})} \tag{C.2.33}$$

Substituting Equation (C.2.33) into Equation (C.2.32), we have

$$\mathbb{E}\left[W_{il}^T\Sigma_{il}^{-1}W_{il}\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-1}\right] = \frac{|L_{il}|\nu_\phi}{\nu_\phi+P_i}$$

. □

Proposition C.2.7

$$\mathbb{E}\left[(W_{il}^T\Sigma_{il}^{-1}W_{il})^2\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-2}\right] = \frac{\nu_\phi^2|L_{il}|(|L_{il}+2|)}{(\nu_\phi+P_i)(\nu_\phi+P_i+2)} \tag{C.2.34}$$

Proof Since W_i follows a multivariate Student's t distribution $t_{\nu_\phi}(0,\Sigma_i)$, we have

$$\begin{aligned}
&\mathbb{E}\left[(W_{il}^T\Sigma_{il}^{-1}W_{il})^2\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-2}\right] = \\
&= C\int(W_{il}^T\Sigma_{il}^{-1}W_{il})^2\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-\frac{\nu_\phi+P_i}{2}-2}\mathrm{d}\,W_i \\
&= C\int\Big(1+\frac{W_{-il}^T\Sigma_{-il}^{-1}W_{il}}{\nu_\phi}\Big)^{-\frac{\nu_\phi+P_i}{2}-2} \\
&\qquad\times\left\{\int(W_{il}^T\Sigma_{il}^{-1}W_{il})^2\Big(1+\frac{W_{il}^T\Sigma_{il}^{-1}W_{il}}{\nu_\phi+W_{-il}^T\Sigma_{-il}^{-1}W_{-il}}\Big)^{-\frac{\nu_\phi+P_i}{2}-2}\mathrm{d}\,W_{il}\right\}\mathrm{d}\,W_{-il}
\end{aligned} \tag{C.2.35}$$

Taking the transformation displayed in Equation (C.2.31),

$$\begin{aligned}
&\mathbb{E}\left[(W_{il}^T\Sigma_{il}^{-1}W_{il})^2\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-2}\right] = \\
&= C\nu_\phi^{2+\frac{1}{2}|L_{il}|}\big|\Sigma_{il}\big|^{\frac{1}{2}}\int\left(1+\frac{W_{-il}^T\Sigma_{-il}^{-1}W_{-il}}{\nu_\phi}\right)^{-\frac{\nu_\phi}{2}-\frac{P_{-il}}{2}} \\
&\qquad\qquad\times\int(X_{il}^TX_{il})^2\Big(1+X_{il}^TX_{il}\Big)^{-\frac{\nu_\phi+P_i}{2}-2}\mathrm{d}\,X_{il}\mathrm{d}\,W_{-il}
\end{aligned}$$

According to Equation (C.2.19), the above line changes to be

$$\begin{aligned}
&\mathbb{E}\left[(W_{il}^T\Sigma_{il}^{-1}W_{il})^2\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-2}\right] = \\
&= C\nu_\phi^{2+\frac{1}{2}|L_{il}|}\big|\Sigma_{il}\big|^{\frac{1}{2}}\frac{|L_{il}|(|L_{il}+2|)}{4}(\pi)^{\frac{|L_{il}|}{2}}\frac{\Gamma(\frac{\nu_\phi}{2}+\frac{P_{-il}}{2})}{\Gamma(\frac{\nu_\phi+P_i}{2}+2)} \\
&\qquad\qquad \times\int\left(1+\frac{W_{-il}^T\Sigma_{-il}^{-1}W_{-il}}{\nu_\phi}\right)^{-\frac{\nu_\phi}{2}-\frac{P_{-il}}{2}}\mathrm{d}\,W_{-il} \qquad \text{(C.2.36)}
\end{aligned}$$

Due to the fact stated in Equation (C.2.33), we have

$$\mathbb{E}\left[(W_{il}^T\Sigma_{il}^{-1}W_{il})^2\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-2}\right] = \frac{\nu_\phi^2|L_{il}|(|L_{il}+2|)}{(\nu_\phi+P_i)(\nu_\phi+P_i+2)}$$

. □

Proposition C.2.8

$$\mathbb{E}\left[(W_{il}^T\Sigma_{il}^{-1}W_{il})(W_{ij}^T\Sigma_{ij}^{-1}W_{ij})\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-2}\right] = \frac{|L_{il}||L_{ij}|\nu_\phi^2}{(\nu_\phi+P_i+2)(\nu_\phi+P_i)} \qquad \text{(C.2.37)}$$

Proof Since W_i follows a multivariate Student's t distribution $t_{\nu_\phi}(0,\Sigma_i)$, we have

$$\begin{aligned}
&\mathbb{E}\left[(W_{il}^T\Sigma_{il}^{-1}W_{il})(W_{ij}^T\Sigma_{ij}^{-1}W_{ij})\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-2}\right] = \\
&= C\int(W_{il}^T\Sigma_{il}^{-1}W_{il})(W_{ij}^T\Sigma_{ij}^{-1}W_{ij})\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-\frac{\nu_\phi+P_i}{2}-2}\mathrm{d}\,W_i \\
&= C\int(W_{ij}^T\Sigma_{ij}^{-1}W_{ij})\Big(1+\frac{W_{-il}^T\Sigma_{-il}^{-1}W_{-il}}{\nu_\phi}\Big)^{-\frac{\nu_\phi+P_i}{2}-2} \\
&\qquad\qquad \times\int(W_{il}^T\Sigma_{il}^{-1}W_{il})\left(1+\frac{W_{il}^T\Sigma_{il}^{-1}W_{il}}{\nu_\phi+W_{-il}^T\Sigma_{-il}^{-1}W_{-il}}\right)^{-\frac{\nu_\phi+P_i}{2}-2}\mathrm{d}\,W_{il}\mathrm{d}\,W_{-il}
\end{aligned} \qquad \text{(C.2.38)}$$

Take the transformation in Equation (C.2.31) for W_{il}, we have

$$\begin{aligned}
&\mathbb{E}\left[(W_{il}^T\Sigma_{il}^{-1}W_{il})(W_{ij}^T\Sigma_{ij}^{-1}W_{ij})\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-2}\right] = \\
&= C\nu_\phi^{\frac{|L_{il}|}{2}+1}|\Sigma_{il}|^{\frac{1}{2}}\int(W_{ij}^T\Sigma_{ij}^{-1}W_{ij})\Big(1+\frac{W_{-il}^T\Sigma_{-il}^{-1}W_{-il}}{\nu_\phi}\Big)^{-\frac{\nu_\phi}{2}-\frac{P_{-il}}{2}-1} \\
&\qquad\qquad \times\int X_{il}^TX_{il}(1+X_{il}^TX_{il})^{-\frac{\nu_\phi+P_i}{2}-2}\mathrm{d}\,X_{il}\mathrm{d}\,W_{-il}
\end{aligned} \qquad \text{(C.2.39)}$$

Recall the fact in Equation (C.2.4), we have

$$\begin{aligned}
&\mathbb{E}\left[(W_{il}^T\Sigma_{il}^{-1}W_{il})(W_{ij}^T\Sigma_{ij}^{-1}W_{ij})\Big(1+\frac{W_i^T\Sigma_i^{-1}W_i}{\nu_\phi}\Big)^{-2}\right] = \\
&= C\nu_\phi^{\frac{|L_{il}|}{2}+1}|\Sigma_{il}|^{\frac{1}{2}}\frac{|L_{il}|}{2}\pi^{\frac{|L_{il}|}{2}}\frac{\Gamma(\frac{\nu_\phi}{2}+\frac{P_{-il}}{2}+1)}{\Gamma(\frac{\nu_\phi}{2}+\frac{P_i}{2}+2)} \\
&\qquad \times\int(W_{ij}^T\Sigma_{ij}^{-1}W_{ij})\Big(1+\frac{W_{-il}^T\Sigma_{-il}^{-1}W_{-il}}{\nu_\phi}\Big)^{-\frac{\nu_\phi}{2}-\frac{P_{-il}}{2}-1}\mathrm{d}\,W_{-il}
\end{aligned} \qquad \text{(C.2.40)}$$

Let $I_4 = \int (W_{ij}^T \Sigma_{ij}^{-1} W_{ij})\Big(1 + \frac{W_{-il}^T \Sigma_{-il}^{-1} W_{-il}}{\nu_\phi}\Big)^{-\frac{\nu_\phi}{2} - \frac{P_{-il}}{2} - 1} \mathrm{d}\, W_{-il}$, we have

$$I_4 = \int \Big(1 + \frac{W_{-ilj}^T \Sigma_{-ilj}^{-1} W_{-ilj}}{\nu_\phi}\Big)^{-\frac{\nu_\phi}{2} - \frac{P_{-il}}{2} - 1} \times \int (W_{ij}^T \Sigma_{ij}^{-1} W_{ij})\Big(1 + \frac{W_{ij}^T \Sigma_{ij}^{-1} W_{ij}}{\nu_\phi + W_{-ilj}^T \Sigma_{-ilj}^{-1} W_{-ilj}}\Big)^{-\frac{\nu_\phi}{2} - \frac{P_{-il}}{2} - 1} \mathrm{d}\, W_j \mathrm{d}\, W_{-ilj} \quad \text{(C.2.41)}$$

where W_{-ilj} denotes the the rest elements in W_i with W_{il} and W_{ij} being removed and thus to be a $P_{-ilj} = P_i - |L_{il}| - |L_{ij}|$ dimensional variable; Σ_{-ilj} is the variance matrix for W_{-ilj}. Similar transformation trick as shown in Equation (C.2.31) is applied on W_{ij}, i.e.

$$X_{ij} = \frac{1}{(\nu_\phi + W_{-ilj}^T \Sigma_{-ilj}^{-1} W_{-ilj})^{\frac{1}{2}}} \Sigma_{ij}^{-\frac{1}{2}} W_{ij}$$

Thus,

$$I_4 = \nu_\phi^{\frac{|L_{ij}|}{2}+1} |\Sigma_{ij}|^{\frac{1}{2}} \int \Big(1 + \frac{W_{-ilj}^T \Sigma_{-ilj}^{-1} W_{-ilj}}{\nu_\phi}\Big)^{-\frac{\nu_\phi}{2} - \frac{P_{-ilj}}{2}} \times \int X_{ij}^T X_{ij} (1 + X_{ij}^T X_{ij})^{-\frac{\nu_\phi}{2} - \frac{P_{-il}}{2} - 1} \mathrm{d}\, X_{ij} \mathrm{d}\, W_{-ilj} \quad \text{(C.2.42)}$$

Applying the fact in Equation (C.2.4) again, we have

$$I_4 = \nu_\phi^{\frac{|L_{ij}|}{2}+1} |\Sigma_{ij}|^{\frac{1}{2}} \frac{|L_{ij}|}{2} \pi^{\frac{|L_{ij}|}{2}} \frac{\Gamma(\frac{\nu_\phi}{2} + \frac{P_{-ilj}}{2})}{\Gamma(\frac{\nu_\phi}{2} + \frac{P_{-il}}{2} + 1)} \int \Big(1 + \frac{W_{-ilj}^T \Sigma_{-ilj}^{-1} W_{-ilj}}{\nu_\phi}\Big)^{-\frac{\nu_\phi}{2} - \frac{P_{-ilj}}{2}} \mathrm{d}\, W_{-ilj} \quad \text{(C.2.43)}$$

Applying the similar transformation trick

$$X_{-ilj} = \frac{1}{\sqrt{\nu_\phi}} \Sigma_{-ilj}^{-\frac{1}{2}} W_{-ilj}$$

I_4 turns to be

$$I_4 = \nu_\phi^{\frac{|L_{ij}|}{2}+1} |\Sigma_{ij}|^{\frac{1}{2}} \frac{|L_{ij}|}{2} \pi^{\frac{|L_{ij}|}{2}} \frac{\Gamma(\frac{\nu_\phi}{2} + \frac{P_{-ilj}}{2})}{\Gamma(\frac{\nu_\phi}{2} + \frac{P_{-il}}{2} + 1)} \nu_\phi^{\frac{P_{-ilj}}{2}} |\Sigma_{-ilj}|^{\frac{1}{2}} \times \int (1 + X_{-ilj}^T X_{-ilj})^{-\frac{\nu_\phi}{2} - \frac{P_{-ilj}}{2}} \mathrm{d}\, X_{-ilj} \quad \text{(C.2.44)}$$

Recall the fact in Equation (C.2.17), we have

$$I_4 = \frac{|L_{ij}|}{2} (\nu_\phi \pi)^{\frac{P_{-il}}{2}} \nu_\phi |\Sigma_{-il}|^{\frac{1}{2}} \frac{\Gamma(\frac{\nu_\phi}{2})}{\Gamma(\frac{\nu_\phi}{2} + \frac{P_{-il}}{2} + 1)} \quad \text{(C.2.45)}$$

Substitute the above result into Equation (C.2.40), we have

$$\mathbb{E}\Big[(W_{il}^T \Sigma_{il}^{-1} W_{il})(W_{ij}^T \Sigma_{ij}^{-1} W_{ij})\Big(1 + \frac{W_i^T \Sigma_i^{-1} W_i}{\nu_\phi}\Big)^{-2}\Big] = \frac{|L_{il}||L_{ij}|\nu_\phi^2}{(\nu_\phi + P_i + 2)(\nu_\phi + P_i)}$$

. □

C.3 Expectations for Block $\left\{\nu_\kappa, \sigma_\varepsilon, \{\kappa_{ijk}\}_{k=1,\ldots,K_{ij};(i,j)\in\mathcal{J}_\mathcal{I}}\right\}$

Assume random variable $T_{ijk} \sim t_{\nu_\kappa}$, then its probability density function is

$$p(T_{ijk}) = D\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-\frac{\nu_\kappa+1}{2}} \tag{C.3.1}$$

where

$$D = \frac{\Gamma(\frac{\nu_\kappa}{2}+\frac{1}{2})}{\sqrt{\nu_\kappa\pi}\Gamma(\frac{\nu_\kappa}{2})} \tag{C.3.2}$$

Proposition C.3.1

$$\mathbb{E}\Big[\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-k}\Big] = \frac{\sqrt{\nu_\kappa+2k}}{\sqrt{\nu_\kappa}}\frac{\Gamma(\frac{\nu_\kappa}{2}+\frac{1}{2})\Gamma(\frac{\nu_\kappa}{2}+k)}{\Gamma(\frac{\nu_\kappa}{2})\Gamma(\frac{\nu_\kappa}{2}+\frac{1}{2}+k)} \tag{C.3.3}$$

Proof According to the pdf illustrated in Equation (C.3.1),

$$\mathbb{E}\Big[\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-k}\Big] = D\int\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-\frac{\nu_\kappa+2k+1}{2}}\mathrm{d}\,T_{ijk}$$

By using the normalizing constant of a standard Student's t distribution $t_{\nu_\kappa+2k}$, the above integration turns out to be

$$\begin{aligned}\mathbb{E}\Big[\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-k}\Big] &= D\frac{\sqrt{\nu_\kappa\pi}\Gamma(\frac{\nu_\kappa+2k}{2})}{\Gamma(\frac{\nu_\kappa+2k}{2}+\frac{1}{2})}\\ &= \frac{\Gamma(\frac{\nu_\kappa}{2}+\frac{1}{2})\Gamma(\frac{\nu_\kappa}{2}+k)}{\Gamma(\frac{\nu_\kappa}{2})\Gamma(\frac{\nu_\kappa}{2}+\frac{1}{2}+k)}\end{aligned}$$

. □

When $k=1$, we have

$$\mathbb{E}\Big[\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-1}\Big] = \frac{\nu_\kappa}{\nu_\kappa+1} \tag{C.3.4}$$

When $k=2$, we have

$$\mathbb{E}\Big[\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-2}\Big] = \frac{\nu_\kappa(\nu_\kappa+2)}{(\nu_\kappa+1)(\nu_\kappa+3)} \tag{C.3.5}$$

Proposition C.3.2

$$\mathbb{E}\Big[T_{ijk}^m\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-k}\Big] = \nu_\kappa^{\frac{m}{2}}\frac{\Gamma(\frac{\nu_\kappa}{2}+\frac{1}{2})}{\Gamma(\frac{\nu_\kappa}{2})\Gamma(\frac{1}{2})}\frac{\Gamma(\frac{\nu_\kappa}{2}+k-\frac{m}{2})\Gamma(\frac{m}{2}+\frac{1}{2})}{\Gamma(\frac{\nu_\kappa}{2}+k+\frac{1}{2})} \tag{C.3.6}$$

Proof According to the pdf shown in Equation (C.3.1),

$$\mathbb{E}\Big[T_{ijk}^m\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-k}\Big] = D\int T_{ijk}^m\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-\frac{\nu_\kappa+2k+1}{2}}\mathrm{d}\,T_{ijk}$$

By taking the transformation $Z_{ijk} = \frac{T_{ijk}}{\sqrt{\nu_\kappa}}$, the above integration changes to be

$$\mathbb{E}\Big[T_{ijk}^m\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-k}\Big] = D\int \nu_\kappa^{\frac{m}{2}+\frac{1}{2}} Z_{ijk}^m\Big(1+Z_{ijk}^2\Big)^{-\frac{\nu_\kappa+2k+1}{2}}\,\mathrm{d}\,Z_{ijk}$$

Taking transformation again by $X_{ijk} = \frac{1}{1+Z_{ijk}^2}$, the integration problem turns out to be

$$\mathbb{E}\Big[T_{ijk}^m\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-k}\Big] = D\int_0^1 \nu_\kappa^{\frac{m}{2}+\frac{1}{2}} X_{ijk}^{\frac{\nu_\kappa}{2}+k-\frac{m}{2}-1}(1-X_{ijk})^{\frac{m}{2}+\frac{1}{2}-1}\mathrm{d}\,X_{ijk}$$

According to definition of Beta function and the normalizing constant shown in Equation (C.3.2),

$$\mathbb{E}\Big[T_{ijk}^m\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-k}\Big] = \nu_\kappa^{\frac{m}{2}}\frac{\Gamma(\frac{\nu_\kappa}{2}+\frac{1}{2})}{\Gamma(\frac{\nu_\kappa}{2})\Gamma(\frac{1}{2})}\frac{\Gamma(\frac{\nu_\kappa}{2}+k-\frac{m}{2})\Gamma(\frac{m}{2}+\frac{1}{2})}{\Gamma(\frac{\nu_\kappa}{2}+k+\frac{1}{2})}$$

. □

When $m=2, k=1$, we have

$$\mathbb{E}\Big[T_{ijk}^2\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-1}\Big] = \frac{\nu_\kappa}{\nu_\kappa+1} \tag{C.3.7}$$

When $m=2, k=2$, we have

$$\mathbb{E}\Big[T_{ijk}^2\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-2}\Big] = \frac{\nu_\kappa^2}{(\nu_\kappa+3)(\nu_\kappa+1)} \tag{C.3.8}$$

When $m=4, k=2$, we have

$$\mathbb{E}\Big[T_{ijk}^4\Big(1+\frac{T_{ijk}^2}{\nu_\kappa}\Big)^{-2}\Big] = \frac{3\nu_\kappa^2}{(\nu_\kappa+3)(\nu_\kappa+1)} \tag{C.3.9}$$

C.4 Simulation Results

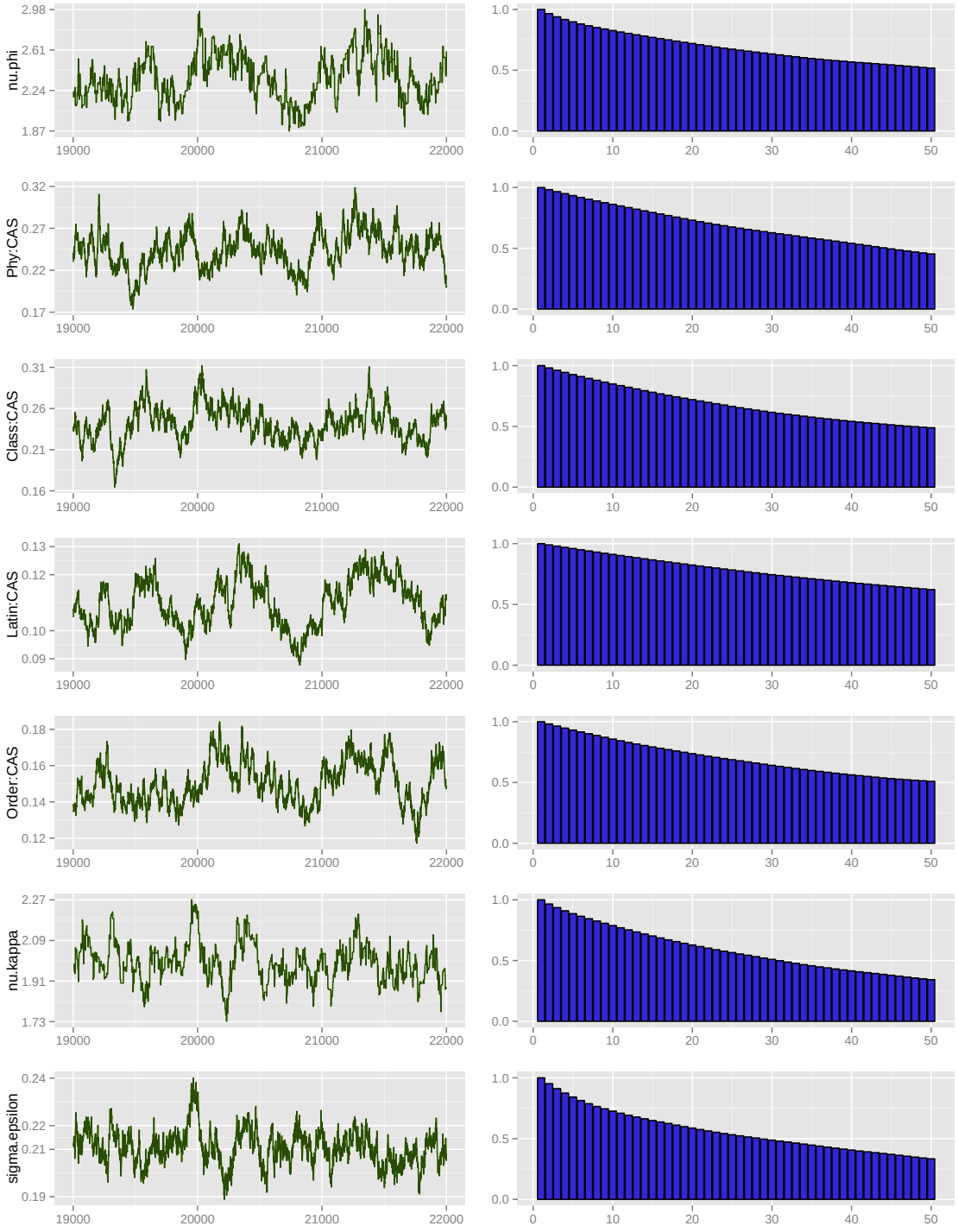


Figure C.1: Left column: Trace plots for the last 3000 posterior samples in the Markov chain given by using the modified MCMCglmm method to simulate $\{\nu_\kappa, \sigma_\varepsilon\}$ and $\left\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\right\}$; Right column: corresponding auto-correlation plots.

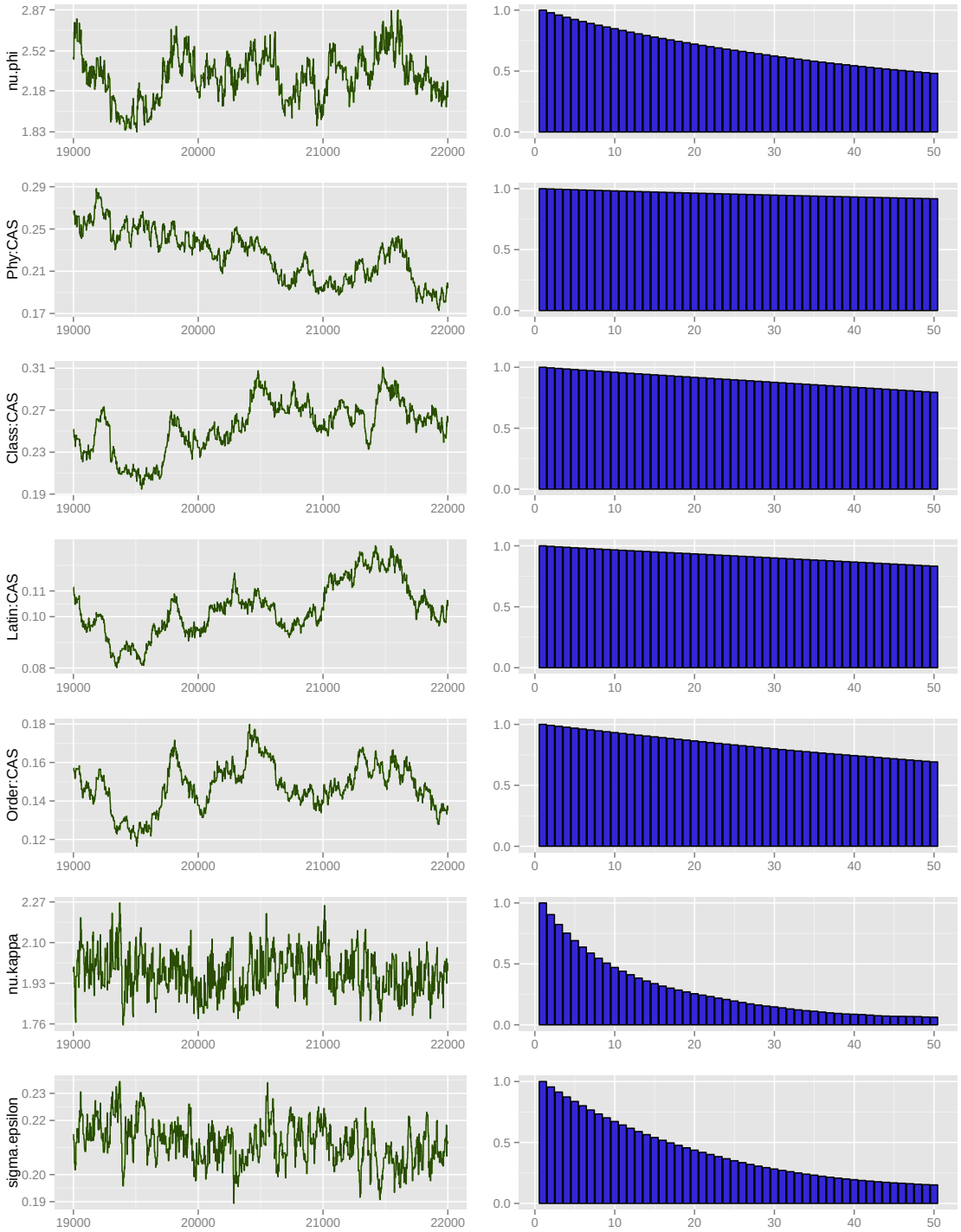


Figure C.2: Left column: Trace plots for the last 3000 posterior samples in the Markov chain given by using the RWMH to simulate the marginalized conditional distributions of $\{\nu_\kappa, \sigma_\varepsilon\}$ and $\left\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\right\}$; Right column: corresponding auto-correlation plots

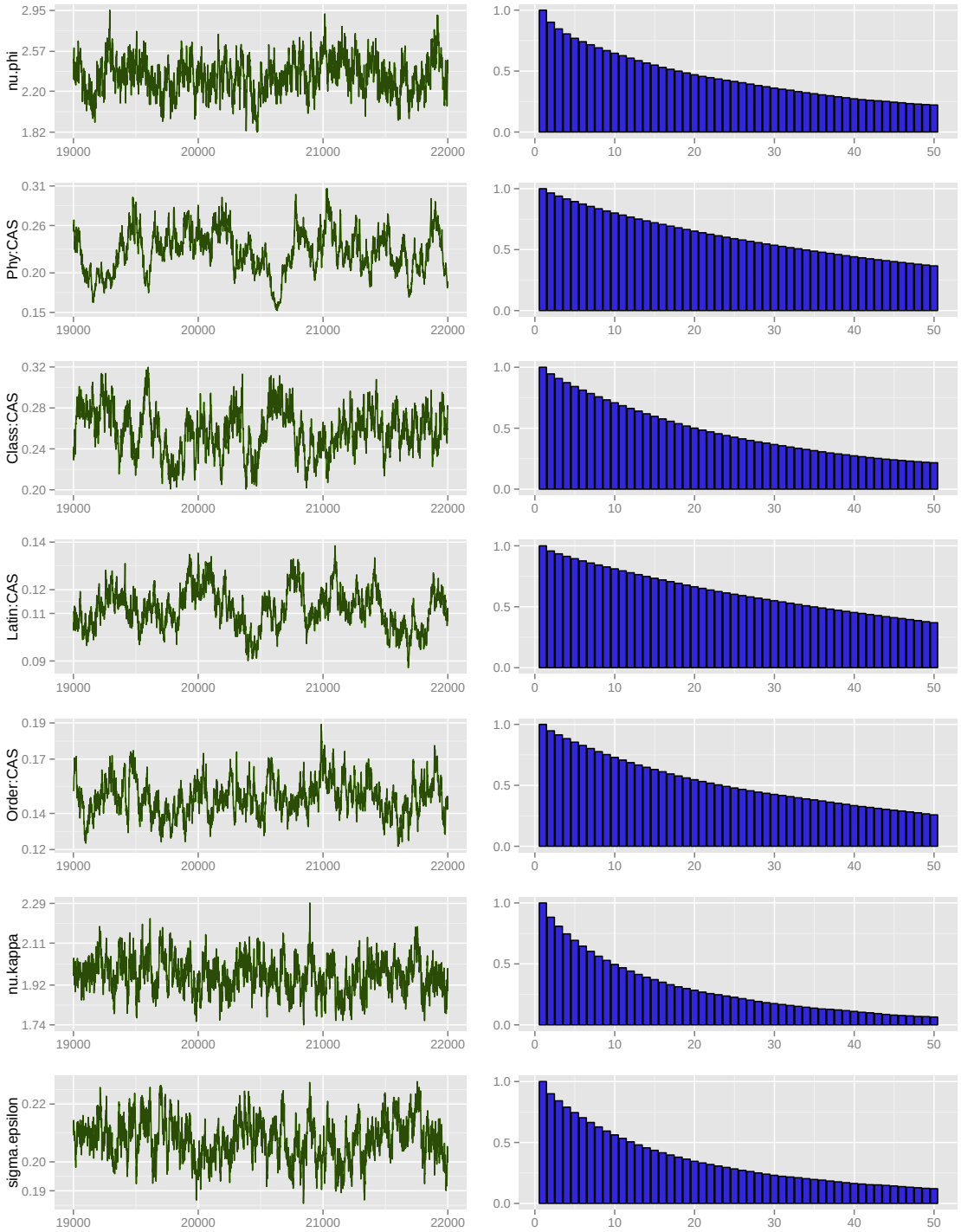


Figure C.3: Left column: Trace plots for the last 3000 posterior samples in the Markov chain given by using the HMC sampler to simulate the marginalized conditional distributions of $\{\nu_\kappa, \sigma_\varepsilon\}$ and $\left\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\right\}$; Right column: corresponding auto-correlation plots

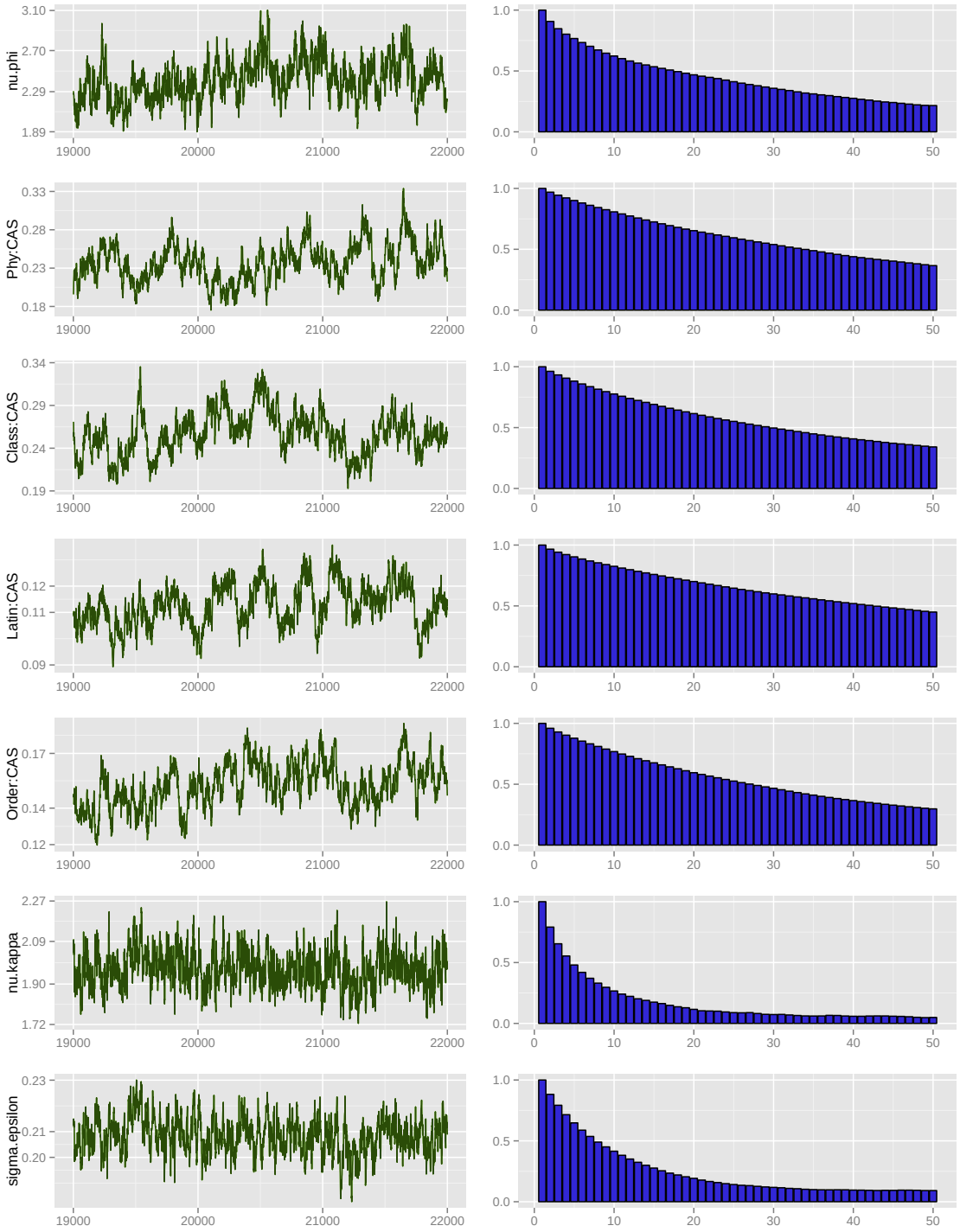


Figure C.4: Left column: Trace plots for the last 3000 posterior samples in the Markov chain given by using the NUTS sampler to simulate the marginalized conditional distributions of $\{\nu_\kappa, \sigma_\varepsilon\}$ and $\left\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\right\}$; Right column: corresponding auto-correlation plots

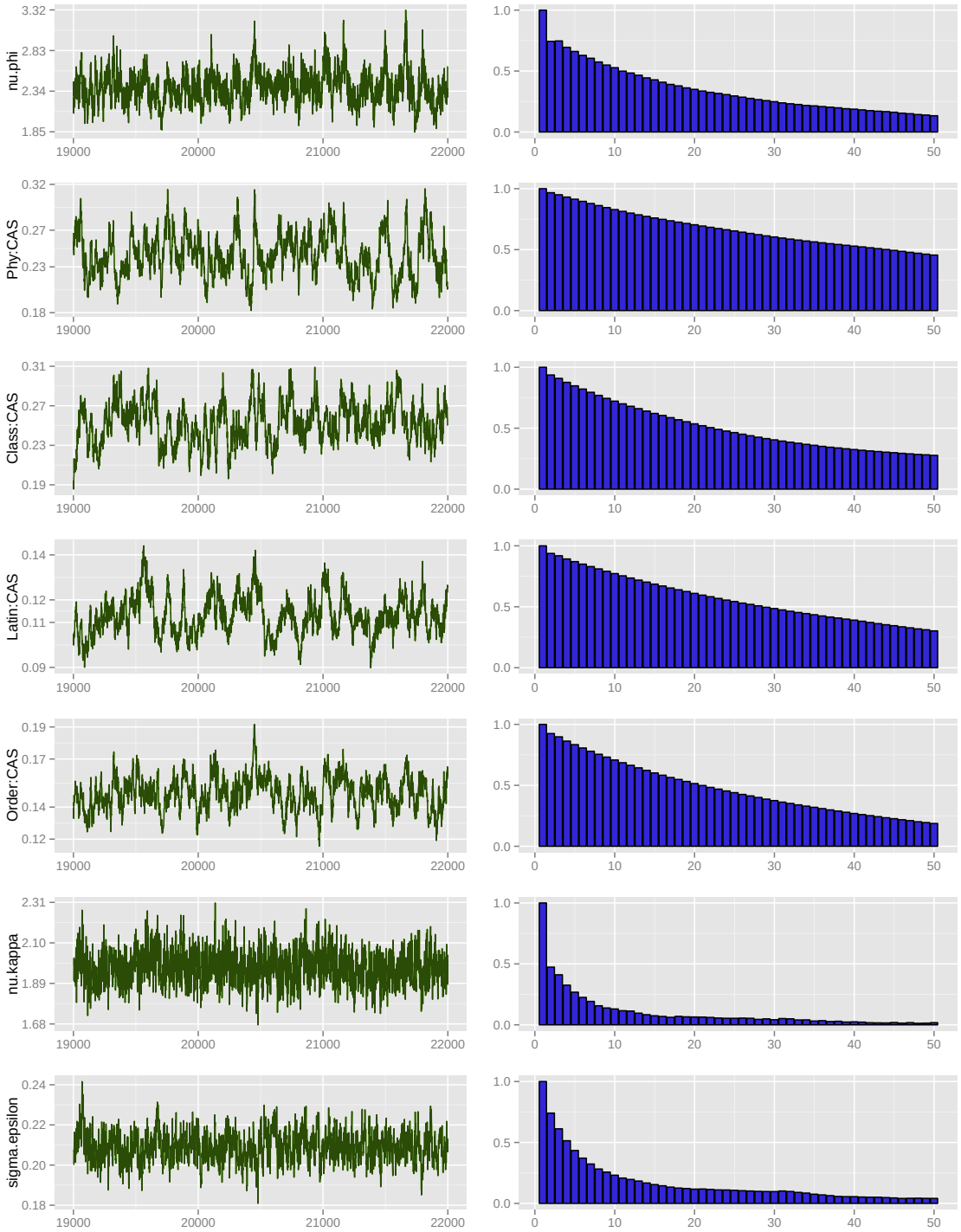


Figure C.5: Left column: Trace plots for the last 3000 posterior samples in the Markov chain given by using the RMHMC sampler to simulate the marginalized conditional distributions of $\{\nu_\kappa, \sigma_\varepsilon\}$ and $\left\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\right\}$; Right column: corresponding auto-correlation plots

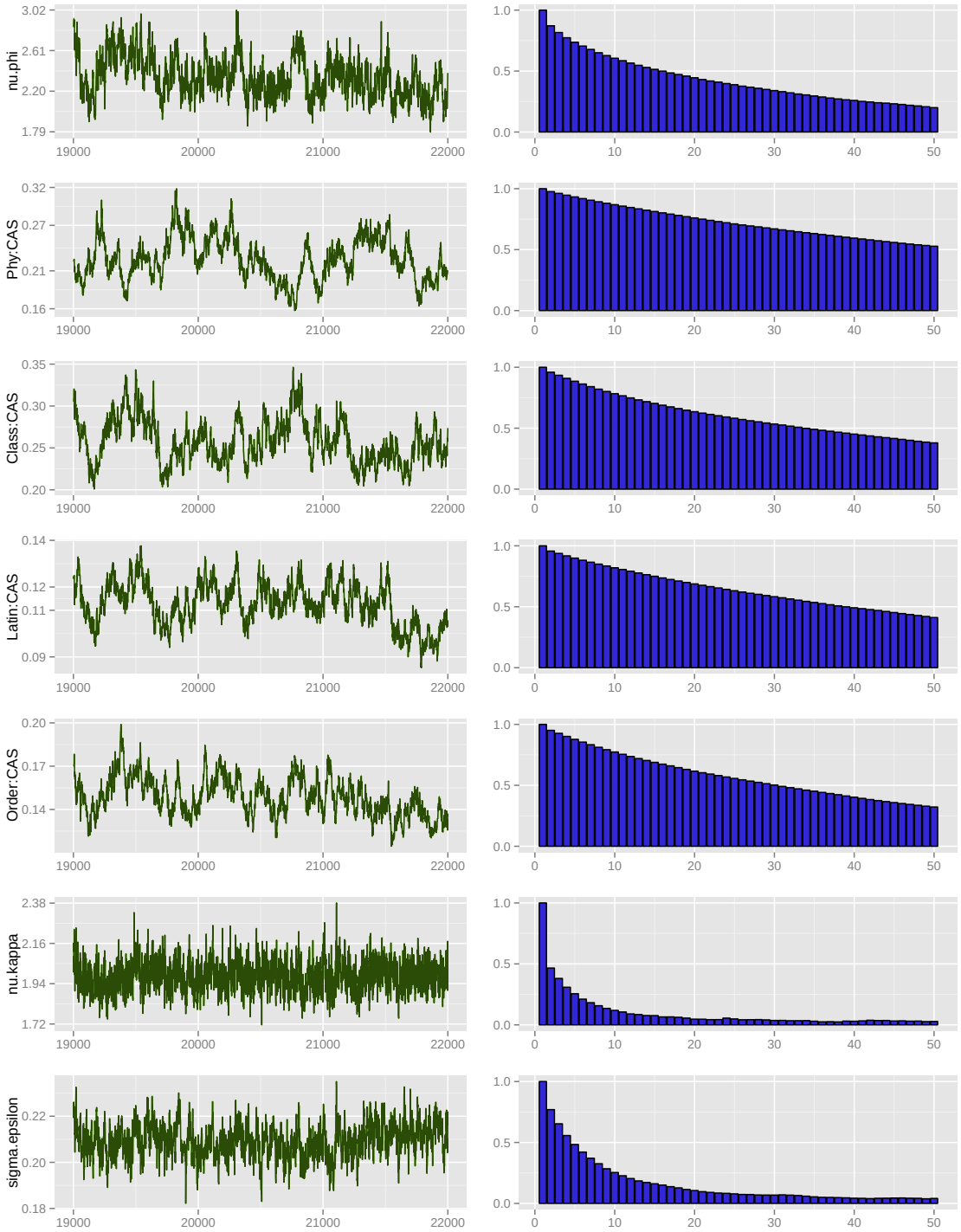


Figure C.6: Left column: Trace plots for the last 3000 posterior samples in the Markov chain given by using the HMC with stochastic step-size sampler to simulate the marginalized conditional distributions of $\{\nu_\kappa, \sigma_\varepsilon\}$ and $\left\{\nu_\phi, \{\sigma_{\xi l}\}_{l=1:L}\right\}$; Right column: corresponding auto-correlation plots

C.5 ESS

	ESS					
parameters	MCMCglmm	RWMH	HMC	NUTS	RMHMC	HMC-S
ν_ϕ	154	174	396	370	546	399
$\sigma_{\xi 1}$	176	21	222	220	170	139
$\sigma_{\xi 2}$	182	48	366	264	339	224
$\sigma_{\xi 3}$	149	77	328	276	361	266
$\sigma_{\xi 4}$	103	38	203	183	254	195
ν_κ	242	815	672	1217	2632	2720
σ_ε	248	433	627	948	1330	1588
σ_α	1923	1284	1931	2016	2188	1734
$\sigma_{\beta 1}$	1934	2337	2495	1213	1515	1614
$\sigma_{\beta 2}$	316	266	286	178	205	291
$\sigma_{\beta 3}$	1074	956	858	916	517	906
$\sigma_{\beta 4}$	1040	960	981	1027	1057	1006
μ	11323	12729	15683	9816	11522	11978

Table C.1: ESS of 20000 Posterior samples from 6 sampling methods. The first column is the original modified MCMCglmm without the marinalized distirbutions. The rest of the columns represent sampling methods with the marginalized distributions.

Appendix D

Stan's Model Code

```
fish_code <- '
data{
  int<lower = 0> N;
  int<lower = 0> M_CAS;
  int<lower = 0> M_Phylum_division;
  int<lower = 0> M_Class;
  int<lower = 0> M_Order;
  int<lower = 0> M_Latin;
  int<lower = 0> M_t1i;
  int<lower = 0> M_t2i;
  int<lower = 0> M_t3i;
  int<lower = 0> M_t4i;
  real y[N];
  // index of random effects
  int CAS[N];
  int Phylum_division[N];
  int Class[N];
  int Order[N];
  int Latin[N];
  int t1i[N];
  int t2i[N];
  int t3i[N];
  int t4i[N];
}
parameters{
  real mu;
  real alpha[M_CAS];
  real beta_Pd[M_Phylum_division];
  real beta_C[M_Class];
  real beta_O[M_Order];
  real beta_L[M_Latin];
  real lamda[M_CAS];
```

```
  real xi_1[M_t1i];
  real xi_2[M_t2i];
  real xi_3[M_t3i];
  real xi_4[M_t4i];
  real kappa[N];
  real<lower = 0> sigma_alpha;
  real<lower = 0> sigma_epsilon;
  real<lower = 0> sigma_beta_Pd;
  real<lower = 0> sigma_beta_C;
  real<lower = 0> sigma_beta_O;
  real<lower = 0> sigma_beta_L;
  real<lower = 0> sigma_xi_1;
  real<lower = 0> sigma_xi_2;
  real<lower = 0> sigma_xi_3;
  real<lower = 0> sigma_xi_4;
  real<lower = 1> nu_phi;
  real<lower = 1> nu_kappa;
}
transformed parameters {
  real theta[N];
  real sy[N];
  for(n in 1:N){
     theta[n] <- mu + alpha[CAS[n]] + beta_Pd[Phylum_division[n]]
     + beta_C[Class[n]] + beta_O[Order[n]] + beta_L[Latin[n]] + (xi_1[t1i[n]]
     + xi_2[t2i[n]] + xi_3[t3i[n]] + xi_4[t4i[n]])/sqrt(lamda[CAS[n]]);
     sy[n] <-  sigma_epsilon/sqrt(kappa[n]);
   }
}
model {
  mu ~ normal(0, 10);
  alpha ~ normal(0, sigma_alpha); //vectorized
  beta_Pd ~ normal(0, sigma_beta_Pd); //vectorized
  beta_C ~ normal(0, sigma_beta_C); //vectorized
  beta_O ~ normal(0, sigma_beta_O); //vectorized
  beta_L ~ normal(0, sigma_beta_L); //vectorized
  lamda ~ gamma(nu_phi/2, nu_phi/2); //vectorized
  xi_1 ~ normal(0, sigma_xi_1); //vectorized
  xi_2 ~ normal(0, sigma_xi_2); //vectorized
  xi_3 ~ normal(0, sigma_xi_3); //vectorized
  xi_4 ~ normal(0, sigma_xi_4); //vectorized
  kappa ~ gamma(nu_kappa/2, nu_kappa/2); //vectorized
  y ~ normal(theta, sy);
  increment_log_prob( -2*log(nu_phi) - 2*log(nu_kappa) - log(sigma_epsilon));
}
'
```

Bibliography

Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010.

Douglas Bates and Dirk Eddelbuettel. Fast and elegant numerical linear algebra using the rcppeigen package. *Journal of Statistical Software*, 52(5):1–24, 2013.

James O Berger and José M Bernardo. On the development of reference priors. *Bayesian statistics*, 4(4):35–60, 1992a.

James O Berger and José M Bernardo. *Reference priors in a variance components problem*. Springer, 1992b.

James O Berger, Jos M Bernardo, and Manuel Mendoza. *On priors that maximize expected information*. Purdue University. Department of Statistics, 1988.

James O Berger, José M Bernardo, and Dongchu Sun. The formal definition of reference priors. *The Annals of Statistics*, pages 905–938, 2009.

Jose M Bernardo. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 113–147, 1979.

MJ Betancourt. Generalizing the no-u-turn sampler to riemannian manifolds. *arXiv preprint arXiv:1304.1920*, 2013.

George EP Box and George C Tiao. *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons, 2011.

Sung-Hyuk Cha. Comprehensive survey on distance/similarity measures between probability density functions. *City*, 1(2):1, 2007.

Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4):327–335, 1995.

Peter S Craig. Exploring novel ways of using species sensitivity distributions to establish pnecs for industrial chemicals: Final report to project steering group 3 april 2013. 2013.

Michael J Daniels. A prior for the variance in hierarchical models. *Canadian Journal of Statistics*, 27(3):567–578, 1999.

Timothy A Davis. User guide for cholmod: a sparse cholesky factorization and modification package. *Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL, USA*, 2008.

AP Dawid. Invariant prior distributions. *Encyclopedia of Statistical Sciences*, 1983.

Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics letters B*, 195(2):216–222, 1987.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*, volume 2. Taylor & Francis, 2014.

Andrew Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3): 515–534, 2006.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741, 1984.

Edward I George and Robert McCulloch. On obtaining invariant prior distributions. *Journal of statistical planning and inference*, 37(2):169–179, 1993.

Jayanta K Ghosh, Mohan Delampady, and Tapas Samanta. *An introduction to Bayesian analysis: theory and methods*. Springer Science & Business Media, 2007.

Malay Ghosh et al. Objective priors: An introduction for frequentists. *Statistical Science*, 26(2):187–202, 2011.

Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

Mark Girolami and Ben Calderhead. Riemann manifold langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

Jarrod D Hadfield et al. Mcmc methods for multi-response generalized linear mixed models: the mcmcglmm R package. *Journal of Statistical Software*, 33(2):1–22, 2010.

Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations*, volume 31. Springer, 2006.

JOHN Hartigan. Invariant prior distributions. *The Annals of Mathematical Statistics*, pages 836–845, 1964.

W Keith Hastings. Monte carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *arXiv preprint arXiv:1111.4246*, 2011.

Piet Hut, Jun Makino, and Steve McMillan. Building a better leapfrog. *The Astrophysical Journal*, 443:L93–L96, 1995.

Edwin T Jaynes. Prior probabilities. *Systems Science and Cybernetics, IEEE Transactions on*, 4(3):227–241, 1968.

Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

Harold Jeffreys. *The theory of probability*. Oxford University Press, 1998.

Ian H Jermyn. Invariant Bayesian estimation on manifolds. *Annals of statistics*, pages 583–605, 2005.

Jorge V José and Eugene J Saletan. *Classical dynamics: a contemporary approach*. Cambridge University Press, 1998.

Robert E Kass. The geometry of asymptotic inference. *Statistical Science*, pages 188–219, 1989.

Robert E Kass and Larry Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370, 1996.

Jeroen SW Lamb and John AG Roberts. Time-reversal symmetry in dynamical systems: a survey. *Physica D: Nonlinear Phenomena*, 112(1):1–39, 1998.

Benedict Leimkuhler and Sebastian Reich. *Simulating Hamiltonian dynamics*, volume 14. Cambridge University Press, 2004.

Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *Information Theory, IEEE Transactions on*, 52(10):4394–4412, 2006.

Ruitao Liu, Arijit Chakrabarti, Tapas Samanta, Jayanta K Ghosh, Malay Ghosh, et al. On divergence measures leading to jeffreys and other reference priors. *Bayesian Analysis*, 9(2):331–370, 2014.

Pierre Simon marquis de Laplace. *Théorie analytique des probabilités English Version*. V. Courcier, 1820.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

Harvey Motulsky. *GraphPad Prism: User's Guide*. GraphPAD, 1995.

R Neal. Mcmc using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, pages 113–162, 2011.

Radford M Neal. Markov chain Monte Carlo methods based on slicing the density function. *Preprint*, 1997.

Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

Gareth O Roberts, Jeffrey S Rosenthal, et al. Optimal scaling for various metropolis-hastings algorithms. *Statistical science*, 16(4):351–367, 2001.

Glenn Shafer et al. *A mathematical theory of evidence*, volume 1. Princeton university press Princeton, 1976.

David Spiegelhalter, Andrew Thomas, Nicky Best, and Wally Gilks. Bugs 0.5: Bayesian inference using gibbs sampling manual (version ii). *MRC Biostatistics Unit, Cambridge*, 1996.

Stan Development Team. Rstan: the R interface to stan, version 2.5.0, 2014a. URL `http://mc-stan.org/rstan.html`.

Stan Development Team. Stan: A c++ library for probability and sampling, version 2.5.0, 2014b. URL `http://mc-stan.org/`.

R Statistical Package. R: A language and environment for statistical computing. *Vienna, Austria: R Foundation for Statistical Computing*, 2009.

Luke Tierney. A note on metropolis-hastings kernels for general state spaces. *Annals of Applied Probability*, pages 1–9, 1998.

Ziyu Wang, Shakir Mohamed, and De Nando. Adaptive Hamiltonian and Riemann manifold Monte Carlo. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1462–1470, 2013.

Keying Ye. Bayesian reference prior analysis on the ratio of variances for the balanced one-way random effect model. *Journal of statistical planning and inference*, 41(3):267–280, 1994.