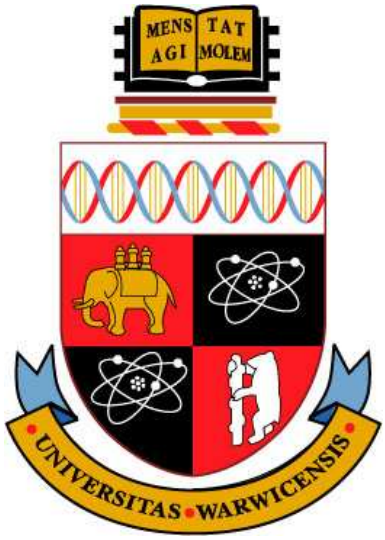


Modelling Via Normalisation for Parametric and Nonparametric Inference

by

Michalis Kolossiatis

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy in Statistics

Department of Statistics

September 2009

THE UNIVERSITY OF WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

I would first like to thank my two supervisors, Professor Mark F. J. Steel and Dr Jim E. Griffin. Their support and guidance during the last four years has been invaluable.

I would also like to thank my two examiners, Professor Jim Q. Smith and Professor Peter Müller for their suggestions in improving this work.

I am also very grateful to the members of the Department of Statistics of the University of Warwick for providing a friendly, supporting and stimulating environment.

I feel more than obliged to thank my family. Without their continuous support, love and encouragement I would not have been able to make it to the end.

I would also like to deeply thank all my friends for their support and for making this journey much more enjoyable.

Finally, I would like to thank the Centre for Research in Statistical Methodology (CRiSM), the Engineering and Physical Sciences Research Council (EPSRC) and the Ministry of Finance of the Republic of Cyprus for the financial support provided.

Declarations

I declare that the contents of this thesis are based on my own research in accordance with the regulations of the University of Warwick. The work in this thesis is original, unless where indicated by references. This thesis has not been submitted for examination at any other university.

Abstract

Bayesian nonparametric modelling has recently attracted a lot of attention, mainly due to the advancement of various simulation techniques, and especially Monte Carlo Markov Chain (MCMC) methods. In this thesis I propose some Bayesian nonparametric models for grouped data, which make use of dependent random probability measures. These probability measures are constructed by normalising infinitely divisible probability measures and exhibit nice theoretical properties. Implementation of these models is also easy, using mainly MCMC methods. An additional step in these algorithms is also proposed, in order to improve mixing. The proposed models are applied on both simulated and real-life data and the posterior inference for the parameters of interest are investigated, as well as the effect of the corresponding simulation algorithms. A new, n-dimensional distribution on the unit simplex, that contains many known distributions as special cases, is also proposed. The univariate version of this distribution is used as the underlying distribution for modelling binomial probabilities. Using simulated and real data, it is shown that this proposed model is particularly successful in modelling overdispersed count data.

Abbreviations

AIC	Akaike information criterion
BB	beta-binomial
BIC	Bayesian information criterion
cdf	cumulative distribution function
DP	Dirichlet process
EPPF	exchangeable product partition formula
iff	if and only if
MC	Monte Carlo
MCMC	Monte Carlo Markov chain
MDP	mixture of Dirichlet processes
MH	Metropolis-Hastings
MLE	maximum likelihood estimate
N-IGP	normalised inverse-Gaussian process
NRM	normalised random measure
pdf	probability distribution function
PDP	product of Dirichlet processes
RJMCMC	reversible jump MCMC
RPM	random probability measure
RWMH	random walk Metropolis-Hastings
SF	stochastic frontier
w.p.	with probability

Notation

The following notation is used throughout this thesis, unless otherwise stated. We usually use normal font type for scalars and **bold** font type for vectors, unless otherwise stated.

$\mathbb{N}$	The set of natural numbers
$\mathbb{R}$	The set of real numbers
$\emptyset$	The null set
δ_x	Dirac measure
B^c	The compliment set of a set B
$\|x\|$	The absolute value of a number x
$\stackrel{d}{=}$	Equality in distribution, i.e. identically distributed
A^{-1}	The inverse of a matrix A
$1_{()}$	The indicator function

Chapter 1

Introduction

Bayesian nonparametric modelling has recently attracted a lot of attention, partly because of the advancement of simulation methods, and especially Monte Carlo Markov chain (MCMC) methods. These models offer a flexible prior specification of the distribution of some data and can therefore be particularly useful in cases where it is preferred not to impose much prior structure on the distribution of those data. Bayesian nonparametric models can also be used in a variety of ways in modelling two or more correlated data sets (for example spatial data).

1.1 Bayesian Nonparametric Modelling

The term "Bayesian nonparametric model" refers to a probability model with infinitely many parameters (Bernardo and Smith, 1994), which results in inference which is directly comparable to classical nonparametric models. These methods have attracted a lot of attention recently, especially because of the recent advances in some simulation techniques, and especially Monte Carlo Markov chain methods, which facilitate the simulation of the posterior distributions of the parameters of interest. These models can be particularly useful in cases when there is uncertainty about the underlying distribution of some data, so modelling this distribution in a flexible way is desirable. As a result, they can be naturally applied to density estimation and regression models.

There are many classes of Bayesian nonparametric models. For the density estimation problem, i.e. the problem of estimating the underlying distribution(s) of some data, it is assumed that the data, say $Y_1, Y_2, \ldots, Y_n$, come from a distribution F or, more generally, each $Y_i \sim F_i$. In the Bayesian nonparametric setting, one considers these distributions also as random and assigns prior distributions to them. Some examples of these models are species sampling models, introduced

by Pitman (1996), Pólya trees (introduced by Ferguson (1974) and developed by Lavine (1992, 1994)), Bernstein polynomials and the very general class of Random Probability Measures (RPMs - see e.g. Crauel (2002)). We also note an interesting and very rich subclass of the RPM, the normalised random measures, which will be described in Section 1.2.3. For the regression problem $y_i = g(\boldsymbol{x}_i) + \epsilon_i,\ i = 1, 2, \ldots, n$ (where the bold symbols denote vectors), many approaches include some collection of distributions, say $B = \{f_1, f_2, \ldots\}$, and write $g = \sum_{k=1}^{\infty} b_k f_k$, for some basis coefficients $b_1, b_2, \ldots$. Popular choices for this collection B include spline, Fourier and wavelet models. For a more detailed review of the aforementioned (and more) Bayesian nonparametric methods, see Müller and Quintana (2004).

At this point something more about the RPMs needs to be said, since they are not only a very rich class of models, but also the one mostly used in practise. As the name indicates, a RPM is a probability measure that is itself taken to be random. Alternatively, as stated in Ferguson (1974), it can be thought of as a random variable whose values are probability measures. A more formal definition is the following:

Definition 1. *Let X be a Polish space and $\mathcal{B}$ denote its $\sigma-$algebra. Let also $(\Omega, \mathcal{F}, P)$ denote a probability space. A map*

$$\mu \ : \ \mathcal{B} \times \Omega \ \rightarrow [0, 1]$$

$$(B, \omega) \ \rightarrow \mu_\omega(B)$$

satisfying

1. *$\forall\ B \in \mathcal{B},\ \mu_\omega(B)$ (as a function of ω) is measurable,*
2. *for P-almost every $\omega \in \Omega$, $\mu_\omega(B)$ (as a function of B) is a Borel probability measure*

is said to be a random probability measure on X.

Within the Bayesian framework, a prior distribution is assigned to a RPM (i.e. a prior distribution of the random distribution). The most widely used prior specification for this random probability measure in the literature is the Dirichlet process (DP), introduced by Ferguson (1973). Other choices include the normalised inverse-Gaussian process (N-IGP) advocated by Lijoi et al. (2005), the invariant DP (Dalal, 1979) and the aforementioned Pólya trees.

Finally, note that these random measures fit naturally in a standard hierarchical model, for example:

$$Y_i \sim f(Y_i; \boldsymbol{\theta}_i),\ i = 1, 2, \ldots, n$$

$$\boldsymbol{\theta}_i \overset{iid}{\sim} G$$

$$G \sim \text{RPM}(\boldsymbol{\psi})$$

$$\boldsymbol{\psi} \sim h(\boldsymbol{\psi}).$$

This setup can be useful in cases where the realisations of the RPM used are discrete distributions, whereas the data under consideration are continuous. This will be demonstrated using the DP as the underlying RPM in the next subsection.

1.1.1 The Dirichlet process

First, the Dirichlet (Dir) distribution is defined:

Definition 2. *An n-dimensional random variable* $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ *defined on the unit simplex is said to follow a Dirichlet distribution with parameters* $a_1, a_2, \ldots, a_{n+1} > 0$*, denoted Dir*$(a_1, a_2, \ldots, a_{n+1})$*, if its density with respect to the Lebesque measure is:*

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{\Gamma(a_1 + a_2 + \cdots + a_{n+1})}{\Gamma(a_1)\Gamma(a_2)\cdots\Gamma(a_{n+1})} \prod_{i=1}^{n} x_i^{a_i - 1}(1 - \sum_{j=1}^{n} x_j)^{a_{n+1}-1},\ x_1, x_2, \ldots, x_n \geq 0, \sum_{k=1}^{n} x_k \leq 1.$$

In the above definition Γ denotes the gamma function, $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$. Notice that the univariate Dirichlet distribution is the known beta (Be) distribution:

Definition 3. *A random variable* X *defined on* $[0, 1]$ *is said to follow a beta distribution with parameters* $a_1 \geq 0$ *and* $a_2 \geq 0,\ a_1 + a_2 > 0$*, denoted Be*$(a_1, a_2)$*, if its density with respect to the Lebesque measure is:*

$$f_X(x) = \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} x^{a_1 - 1}(1 - x)^{a_2 - 1},\ 0 \leq x \leq 1.$$

If $a_1 = 0$*, then* $X = 0$ *almost surely and if* $a_2 = 0,\ X = 1$ *almost surely.*

A simple definition of the Dirichlet process is then the following:

Definition 4. *A random probability function* F *is said to follow a Dirichlet process with parameters* M *and* H_0 *if for any partition* $(A_1, A_2, \ldots, A_k)$ *of the probability space* Ω*, such that all* $A_i \in \mathcal{F}$*, the* $\sigma-$*algebra of* Ω*, the vector of random probabilities* $(F(A_1), F(A_2), \ldots, F(A_k))$ *follows a Dirichlet distribution with parameters* $MH_0(A_1), MH_0(A_2), \ldots, MH_0(A_k)$*.*
Symbolically:

$$F \sim DP(M, H_0) \overset{def}{\Leftrightarrow} \forall \text{ partition } (A_1, A_2, \ldots, A_k) \text{ of } \Omega,\ A_1, A_2, \ldots, A_k \in \mathcal{F}$$

$$(F(A_1), F(A_2), \ldots, F(A_k)) \sim Dir(MH_0(A_1), MH_0(A_2), \ldots, MH_0(A_k)).$$

As shown in Lemma 1 of Ferguson (1973), the existence of such a process is verified by showing that the Kolmogorov consistency conditions hold.

As seen from the definition, there are two parameters characterizing the DP: H_0 and M.
H_0 is a distribution function and is called the base or centering distribution of the DP. It can be seen as the centre of the process, since

$$\forall\ B \in \mathcal{F},\ \mathrm{E}\left(F(B)\right) = H_0(B). \tag{1.1.1}$$

The parameter $M > 0$ is a scalar called the concentration or precision parameter of the process, and it controls the variability of the process around H_0, since

$$\forall\ B \in \mathcal{F}, \mathrm{Var}\left(F(B)\right) = \frac{H_0(B)\left(1 - H_0(B)\right)}{1 + M}. \tag{1.1.2}$$

So, intuitively, M can also be seen as a measure of our belief in the base distribution H_0.
In fact, in his seminal paper, Ferguson (1973) uses a non-null finite measure α as the parameter of the DP. Then, by considering the parametrisation $M = \alpha(\Omega)$ and $H_0 = \frac{\alpha}{\alpha(\Omega)}$, we get the definition above and a better understanding of the role played by this measure α.
The reason for the popularity of the Dirichlet process is its mathematical properties, which lead to algebraic and computationally convenient expressions, therefore allowing for relatively easy posterior inference when combined with MCMC techniques. These properties include simple expressions for the expectation and variance of its realisations, as seen in (1.1.1) and (1.1.2).

The DP can also be represented using a stick-breaking representation (Sethuraman and Tiwari, 1982; Sethuraman, 1994): If $F \sim \mathrm{DP}(M, H_0)$, then

$$F(\cdot) = \sum_{i=1}^{\infty} w_i \delta_{\theta_i^*}(\cdot),\ \text{where } \theta_i^* \overset{iid}{\sim} H_0,\ w_i = V_i \prod_{j<i}(1 - V_j),\ \text{where } V_i \overset{iid}{\sim} \mathrm{Be}(1, M) \tag{1.1.3}$$

where δ_x denotes the Dirac measure giving mass 1 to the value x.
As can be seen from (1.1.3), any realisation of the DP is, with probability 1, a discrete distribution. This is an obvious drawback when one wants to model data from continuous distributions. On the other hand, this discreteness allows for clustering the values of a random distribution following a DP:
Let $F \sim \mathrm{DP}(M, H_0)$, where H_0 is a continuous distribution and assume a sample $\theta_1, \theta_2, \ldots, \theta_n$ from F. The number of discrete θ_i (denoted by θ_i^*), will be $K \leq n$. The distribution of K is given in Escobar and West (1995):

$$P(K = k|M, n) = c_n(k) n! M^k \frac{\Gamma(M)}{\Gamma(M + n)},\ k = 1, 2, \ldots, n \tag{1.1.4}$$

where $c_n(k) = P(K = k|M = 1, n)$, not involving M.
In the above we want H_0 to be a continuous distribution, in order to have all the θ_i^* being different.

If, on the other hand, H_0 was discrete, we would again have discrete values, but now there is the possibility that some of the clusters created by the discreteness of the DP (not of H_0) would be located at the same values.

Notice also that, as (1.1.4) suggests, for higher values of the concentration parameter, higher probabilities are given to larger number of clusters. The intuition for this is that higher values of M indicate less variation from the base distribution H_0, i.e. more belief in H_0. As a result, more observations from the DP will actually be taken from this base distribution. The above observations are consistent with (and complemented by) equation (1.1.5) below.

Next, the Pólya-urn representation of the DP are presented, i.e. an expression of the possible allocations of a new observation from the DP, given previous observations from the same DP. This representation was noted by Blackwell and MacQueen (1973) and has also a simple form: having observed $\theta_1, \theta_2, \ldots, \theta_n$ from $F \sim \mathrm{DP}(M, H_0)$, the (posterior) distribution of a new observation θ_{n+1} is as follows:

$$\forall\ A \in \mathcal{F},\ P\left(\theta_{n+1} \in A | \theta_1, \theta_2, \ldots, \theta_n\right) = \frac{M}{M+n} H_0(A) + \frac{1}{M+n} \sum_{i=1}^{n} \delta_{\theta_i}(A). \tag{1.1.5}$$

This means that any new value will be set equal to one of the previous values θ_i (with probability $\frac{1}{M+n}$ for each θ_i) or will be a new draw from the base distribution (with probability $\frac{M}{M+n}$). Again, notice that for higher values of M, more clusters are expected to be created for a specific data size.

A similar expression to the Pólya-urn representation is the so-called Chinese restaurant representation (Aldous, 1985; Pitman, 1996). Before explaining this algorithm, let us define the indicator functions $s_i,\ i = 1, 2, \ldots, n$, such that

$$s_i = j \Leftrightarrow \theta_i \equiv \theta_j^*,\ j = 1, 2, \ldots, K,$$

where $(\theta_1, \theta_2 \ldots, \theta_n)$ is a sample from a $\mathrm{DP}(M, H_0)$-distributed random distribution F and $(\theta_1^*, \theta_2^*, \ldots, \theta_K^*)$ is the vector of discrete values (clusters) of these data.

The Chinese restaurant representation now gives the probabilities of all possible values of a new indicator s_{n+1}, corresponding to a new observation from $\mathrm{F}|\theta_1, \theta_2, \ldots, \theta_n$. It is clear that s_{n+1} takes values in the set $\{1, 2, \ldots, K+1\}$, where the first K values correspond to the already existing clusters and the last value corresponds to a new cluster being created. These probabilities are as follows:

$$P(s_{n+1} = j) = \begin{cases} \frac{n_j}{n+M} & , \quad j = 1, 2, \ldots, K \\ \frac{M}{n+M} & , \quad j = K+1, \end{cases}$$

where n_j is the size of cluster j, i.e. how many of the θ_i are assigned to this cluster (i.e equal to θ_j^*).

Using the Pólya-urn scheme for a single DP with parameter M, it is easy to derive the Exchangeable Product Partition Formula (EPPF) for this model:

$$p(\boldsymbol{s}|M) = M^k \frac{\Gamma(M)}{\Gamma(M+n)} \prod_{i=1}^{K} \Gamma(n_i) \tag{1.1.6}$$

where $\boldsymbol{s} = (s_1, s_2, \ldots, s_n)$ is the vector of all allocation parameters.

Finally, the property that, in fact, characterizes the DP, is its conjugacy: given $\theta_1, \theta_2, \ldots, \theta_n$ from $F \sim \text{DP}(M, H_0)$, the posterior distribution of F is again a DP with parameters $M + n$ and $H_0 + \sum_{i=1}^{n} \delta_{\theta_i}$:

$$F|\theta_1, \theta_2, \ldots, \theta_n \sim \text{DP}\left(M + n, H_0 + \sum_{i=1}^{n} \delta_{\theta_i}\right).$$

Apart from its obvious advantages stated above, there is the quite unpleasant feature of the DP that its realisations are always discrete distributions. As a result, modelling continuous data using the DP as the distribution of their distribution would be inappropriate.

The usual solution to this problem is to add an additional level in the model, by assuming that the data come from a continuous distribution with parameters $\boldsymbol{\theta}$ and set the distribution of the parameters $\boldsymbol{\theta}$ to follow a DP (Ferguson, 1983; Lo, 1984):

$$\begin{gathered} Y_i \sim f(Y_i; \boldsymbol{\theta}_i, \boldsymbol{\zeta}),\ i = 1, 2, \ldots, n \\ \boldsymbol{\theta}_i \overset{iid}{\sim} G \\ G \sim \text{DP}\left(M, H_0(\boldsymbol{\psi})\right) \\ M \sim h_1(M),\ \boldsymbol{\zeta} \sim h_2(\boldsymbol{\zeta}),\ \boldsymbol{\psi} \sim h_3(\boldsymbol{\psi}) \end{gathered} \tag{1.1.7}$$

where $\boldsymbol{\zeta}$ are any other parameters in the likelihood f not modeled using the DP and $\boldsymbol{\psi}$ are the parameters of the base distributions (if any).

This setup is referred to as mixture of Dirichlet process (MDP) model, and was introduced by Antoniak (1974). Note also that the distribution of each Y_i (given $\boldsymbol{\zeta}$) is given by convolving f with $G \sim$ DP:

$$f(Y_i; \boldsymbol{\zeta}) = \int f(Y_i; \boldsymbol{\theta}, \boldsymbol{\zeta}) dG(\boldsymbol{\theta}_i), \text{ where } G \sim DP(M, H_0(\boldsymbol{\psi}))$$

and this, together with the discrete nature of the realisations of the DP, will lead to a mixture model for Y_i (given $\boldsymbol{\zeta}$) (Antoniak, 1974).

1.1.2 Computational issues

For models with many parameters, the joint posterior distribution of all parameters is usually extremely complicated to calculate, let alone to simulate from. Bayesian nonparametric models usually

fall in this category. Therefore, most inference using these models involves advanced computational methods, and mostly Monte Carlo Markov Chain (MCMC) methods are used. Other simulation methods are also applicable, like sequential importance sampling (see, for example, MacEachern et al., 1999; Fearnhead, 2004) and variational inference methods (Beal and Ghahramani, 2003).

MCMC methods consist of constructing a Markov Chain (i.e. a chain where each updating step depends only on the previous iteration of the chain) that has the desired posterior distribution of all parameters in the model as its stationary distribution.

The mostly used MCMC method is the Gibbs sampling. According to this method, we start with some initial values for our parameters. Then, in each step of the chain, each parameter in the model is sequentially simulated from its full conditional distribution, i.e. the distribution of the specific parameter given the data and all the other parameters. In each case the values of the parameters of interest are recorded and at the end some initial part of the chain is discarded as burn-in. The rest of the output consists of samples from the joint posterior distribution of all parameters. From this output the values of a specific parameter can also be taken, and those values will be samples from the posterior distribution of this parameter. In this method, if a full conditional distribution is of known form, for example if it is a beta distribution, simulating from it is straightforward. If, on the other hand, simulating from a full conditional distribution directly is not possible, Metropolis-Hastings (MH) steps or slice sampling can be used instead.

MH updating steps consist of proposing a value for the parameter that is to be updated and calculating the acceptance probability of the proposed value. This acceptance probability takes into account both the full conditional distribution of the parameter and the probability of the transition from the existing value of the parameter to the one proposed. More specifically, the acceptance probability α for moving a parameter ϑ from its current value ϑ_0 to a new value ϑ' is:

$$\alpha(\vartheta_0, \vartheta') = \min\left\{1, \frac{f(\vartheta')}{f(\vartheta_0)} \frac{q(\vartheta', \vartheta_0)}{q(\vartheta_0, \vartheta')}\right\}.$$

where f is the full conditional distribution of ϑ and $q(a, b)$ is the transition probability from a value a to a value b, and depends on the method of proposing these new values. Then, with probability $\alpha(\vartheta_0, \vartheta')$, the value of ϑ is changed to ϑ', otherwise it remains unchanged: $\vartheta = \vartheta_0$.
Popular choices for proposing new values in a MH step include independence MH steps (when the proposed value is taken independently of the current value) and the random walk Metropolis-Hastings steps (RWMH), when the proposed value is the sum of the existing value and a value from a zero-mean random variate. Roberts and Rosenthal (2001) discuss monitoring of RWMH updating steps, in order to optimise mixing. In the following, RWMH steps will mostly be used when the full conditional distributions are not of known form.

Slice sampling is a parameter augmentation technique, in which the parametric space is extended to include some auxiliary variables (see, for example, Damien et al., 1999; Neal, 2003). Suppose that we want to sample from a distribution $f(x)$. If this distribution is difficult to sample from, we can extend $f(x)$ to $g(x,u)$, where u is an auxiliary variable and g is such that $\int g(x,u)du = f(x)$ and $g(u|x) = \frac{g(u,x)}{\int g(x,u)du}$ is a uniform distribution on a set (and therefore easy to sample from, for example, using inverse transformation methods). The joint density g is also chosen such that $g(x|u)$ is also easy to sample from. We can then iteratively sample from $g(x|u)$ and $g(u|x)$, and the value of x obtained will be a draw from $f(x)$. As a simple example, consider $f(x) = xe^{-x^2}$. We can then set $g(x,u) \propto x1_{\left(0<u<e^{-x^2}\right)}$, where the indicator function $1_{(\ )}$ takes the value 1 if the expression in the subscript is true, and 0 otherwise. Then $g(u|x)$ is the uniform distribution in $\left(0, e^{-x^2}\right)$, whereas $g(x|u) = x1_{(|x|<-\log(u))}$, and therefore easy to sample from.

Since the MDP Model (1.1.7) is the one mostly used in Bayesian nonparametric inference, it would be useful to present the basic methods of implementation using MCMC methods. This will also provide a first introductory insight to the algorithms that will be presented in the next sections, since more or less the same principles apply. For simplicity, it is assumed that each θ_i is scalar, there are no extra parameters $\boldsymbol{\zeta}$ in the likelihood and the parameters of the base distribution $\boldsymbol{\psi}$ are fixed. Usually, in the more general case where additional parameters are introduced in the model, their full conditional distributions are explicitly known and easy to sample from.

As mentioned before, due to the infinite number of point masses and weights of any realisation of the Dirichlet process (as seen by the stick-breaking representation), it is impossible to simulate from the DP directly. However, in cases where one is not directly interested in the unknown random distribution itself, but rather in the posterior distributions of some parameters of the model (which is very common in practice), a sample from those posterior distributions can be obtained using MCMC methods. It is also possible to get samples from the predictive distribution of a parameter whose distribution is DP-distributed.

There are two main approaches in simulating Model (1.1.7). One approach is to use marginal methods, which consist of integrating the unknown distribution out of the posterior distributions and using the Pólya-urn representation of the Dirichlet process. The second approach is to use conditional methods, which consider the DP as part of the MCMC algorithm.

Marginal methods

This is the method mostly used in the literature. Usually this algorithm uses a Gibbs sampler, especially when the likelihood $f(Y;\theta)$ and the base distribution $H_0(\theta)$ form a conjugate pair for θ (i.e. their product as a function of θ has the functional form of a known distribution). Most of the

algorithms concerning the marginal method are based on the seminal paper of Escobar and West (1995), which is itself based on the work of Escobar (1988, 1994). Very good descriptions of some marginal algorithms can be found in Escobar and West (1998) and MacEachern (1998).

In order to implement the Gibbs algorithm, the full conditional distributions of all parameters in the model, i.e. the distributions of each parameter given all the other parameters and the data $Y_1, Y_2, \ldots, Y_n$, need to be calculated. The conditional independence relationships between the parameters that the hierarchical structure of this model expresses also enhance these calculations. Regarding the full conditional distribution of each θ_i, the Pólya-urn representation of the DP (1.1.5) can be used in order to integrate out the unknown distribution. The exchangeability of θ_i in this expression leads to the following posterior distribution for θ_i:

$$p\left(\theta_i|\theta_{-i}, \boldsymbol{Y}, M, \boldsymbol{\psi}\right) = q_0 \frac{f(Y_i|\theta_i)h_0(\theta_i|\boldsymbol{\psi})}{\int f(Y_i|\theta_i)dH_0(\theta_i|\boldsymbol{\psi})} + \sum_{j\neq i} q_j \delta_{\theta_j}(\theta_i) \tag{1.1.8}$$

where $h_0(\theta_i|\boldsymbol{\psi}) = \frac{dH_0(\theta_i|\boldsymbol{\psi})}{d\theta_i}$ is the pdf of H_0, $\boldsymbol{Y}$ is the data set, $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$, $q_0 \propto M \int f(Y_i|\theta_i)dH_0(\theta_i|\boldsymbol{\psi})$, $q_j \propto f(Y_i|\theta_j)$ and the factors of proportionalities are the same for each weight and such that the weights add to 1. This structure indicates how easy it is to simulate each θ_i given all the others. What is needed is to calculate the weights q_0 and $q_j,\ j \neq i$ and then assign θ_i to one of the other θ_j with probabilities $q_j,\ j \neq i$, or draw a new value from the base distribution H_0, with probability q_0. The tricky part is calculating the integral appearing in the expression for q_0. On the other hand, when the base distribution H_0 and the likelihood f form a conjugate pair, this integral will be trivial.

As far as the precision parameter M is concerned, it was demonstrated in Escobar and West (1995) that it is better to consider it as a random variable and impute it in the MCMC algorithm. In the same article, the authors give a $\text{Ga}(\alpha, \beta)$ prior distribution to M and propose a very simple way to update this quantity, using a fine algebraic trick.

Definition 5. *A random variable X is said to follow a gamma distribution with parameters $\alpha > 0$ and $\beta > 0$, denoted by $Ga(\alpha, \beta)$, if its density with respect to the Lebesque measure is:*

$$f_X(x) = \frac{\beta^{-\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta},\ x > 0.$$

The full conditional distribution of M in the specific model is proportional to its prior and to the distribution of the number of cluster, K, shown in (1.1.4). Therefore, $f(M|\cdots) \propto M^{\alpha+K-1} e^{-M/\beta} \frac{\Gamma(M)}{\Gamma(M+n)}$ and the last fraction can be substituted using the formula

$$\frac{\Gamma(M)}{\Gamma(M+n)} = \frac{(M+n)B(M+1, n)}{M\Gamma(n)},$$

where $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx$ is the usual beta function. This leads to:

$$\begin{aligned} f(M|\cdots) &\propto M^{\alpha+K-1}e^{-M/\beta}\frac{(M+n)B(M+1,n)}{M} \\ &= M^{\alpha+K-2}e^{-M/\beta}(M+n)\int_0^1 x^M(1-x)^{n-1}dx. \end{aligned}$$

The last expression can now be seen as the marginal distribution of the joint distribution of M and an auxiliary variable in $(0,1)$, say ξ, where $f(M,\xi|\cdots) \propto M^{\alpha+K-2}e^{-M/\beta}(M+n)\xi^M(1-\xi)^{n-1}$. The full conditionals of each of those quantities are as follows:

$$\begin{aligned} f(M|\xi,\cdots) &\propto M^{\alpha+K-2}e^{-M(1/\beta-\log(\xi))}(M+n) \\ &= M^{\alpha+K-1}e^{-M(1/\beta-\log(\xi))} + nM^{\alpha+K-2}e^{-M(1/\beta-\log(\xi))}(M+n). \end{aligned}$$

(i.e a mixture of gamma distributions), and

$$f(\xi|M,\cdots) \equiv \text{Be}(M+1,n).$$

What one needs to do in order to simulate from M in each step in the MCMC algorithm is to simulate from both $f(\xi|M,\cdots)$ and $f(M|\xi,\cdots)$ sequentially. Each value of ξ can be discarded after M is simulated, as it is just an auxiliary variable, whereas the values of M will be kept and used in the rest of the algorithm and in posterior inference.

Finally, it is also straightforward to include an additional step in the MCMC algorithm that approximates the predictive distribution $p(Y_{n+1}|Y_1,Y_2,\ldots,Y_n)$:

$$\begin{aligned} p(Y_{n+1}|Y_1,Y_2,\ldots,Y_n) &= \int\int\cdots\int p(Y_{n+1},\theta_1,\theta_2,\ldots,\theta_n|Y_1,Y_2,\ldots,Y_n)d\theta_1 d\theta_2\ldots d\theta_n \\ &= \int\int\cdots\int p(Y_{n+1}|\theta_1,\ldots,\theta_n,Y_1,\ldots,Y_n)f(\theta_1,\ldots,\theta_n|Y_1,\ldots,Y_n)d\theta_1\ldots d\theta_n \\ &= \int\int\cdots\int p(Y_{n+1}|\theta_1,\theta_2,\ldots,\theta_n)f(\theta_1,\theta_2,\ldots,\theta_n|Y_1,Y_2,\ldots,Y_n)d\theta_1 d\theta_2\ldots d\theta_n. \end{aligned}$$

In words, the predictive distribution can be expressed as an expectation of the random vector $(\theta_1,\ldots,\theta_n)$ from its posterior distribution. This expression is very complicated to calculate analytically. On the other hand, this expectation can be approximated using Monte Carlo (MC) methods: in each step of the MCMC algorithm after the burn-in period, a sample from these θ_i's is obtained. These samples are actually samples from the posterior distribution of $(\theta_1,\ldots,\theta_n)$. So, in each step $p(Y_{n+1}|\theta_1^t,\theta_2^t,\ldots,\theta_n^t)$ (which will be a simple mixture distribution) is calculated, where the values for the θ_i^t's are the values in the t-th cycle of the chain. By then calculating the mean of those

expressions over all the MCMC cycles after burn-in, we get an approximation of the predictive distribution.

An improvement to this algorithm was proposed by MacEachern (1994) and is almost always used in this type of models, as it is easy to implement and improves the mixing of the chain. It exploits the clustering of the values of a sample from a random distribution following a DP, as mentioned before. Instead of using the actual values $(\theta_1, \theta_2, \ldots, \theta_n)$, it uses the reparametrisation $(\theta_1^*, \theta_2^*, \ldots, \theta_K^*, s_1, s_2, \ldots, s_n)$, where the quantities are as defined before. The reason for this reparametrisation is that now the discrete values θ_j^* are also updated, resulting in better mixing of the Markov chain. Note that the prior distribution of the θ_j^*'s is the base distribution H_0, so the full conditional distribution for each θ_i^* will be proportional to the product of H_0 and of the product of the likelihood of the data that are associated with it, $H_0(\theta_i^*) \prod_{j:s_j=i} p(Y_j; \theta_i^*)$. As for the updating of the indicators, this will be the same as (1.1.8), with s replacing θ.

Apart from the above method of improving the calculations, other tricks have been proposed (see e.g. MacEachern (1998)). Some of them are:

1. Collapsing of the state space: The idea here is that, since we use simulation to avoid difficult integrals, one should try to evaluate as many integrals as possible before starting the simulation. So, if possible, we integrate out some parameters from our model before the simulations.

2. Blocking: The basic idea is to update parameters that are *a posteriori* highly correlated together, and therefore improve the mixing of the chain.

3. Rao-Blackwellization: Proposed by Gelfand and Smith (1990), this technique suggests replacing values generated as part of the simulation with appropriate conditional expectations. In this way, one can benefit from conditional distributions, if they are of known form.

The nonconjugate case

This is the case where the likelihood f and the base distribution H_0 do not form a conjugate pair. As a result, calculations of some integrals become nontrivial. For example, in the update of θ_i (or of the s_i), we have the integral $\int f(y; \theta_i) dH_0(\theta_i)$. If f and H_0 are not conjugate, calculating these integrals can be very difficult or even impossible. Although in nonparametric models a variety of base distributions can be used, since these models are quite flexible and will adapt to the specific choice (for example, a DP prior with M small can be used), there are cases where it seems more logical to use a specific base distribution, which is not conjugate to the likelihood. This case is

discussed, among others, in West et al. (1994), MacEachern and Müller (1998), Neal (2000) and Jain and Neal (2005) (all in the case of the MDP model).

The first paper provides the first algorithm designed for the nonconjugate case. The authors here propose an MCMC scheme for simulating from the posterior distributions of the parameters in the model. Their solution to the problematic integral $\int f(Y_i|\theta_i)dH_0(\theta_i)$ is to approximate it using Monte Carlo approximation or, in the special case where only one of these MC samples is used, to replace it by $f(Y_i|\theta')$, where θ' is a draw from H_0. However, as MacEachern and Müller (1998) note, this approximation fails theoretically, as the resulting Markov chain might converge to a stationary distribution that is not the same as the posterior distribution. Another issue about this method is that it is not easy to evaluate the accuracy of the approximation, as this approximation occurs within acceptance probabilities.

In the second reference the authors propose the so-called "no gaps algorithm". This method consists of augmenting the vector of discrete values $(\theta_1^*, \theta_2^*, \ldots, \theta_K^*)$ to $(\theta_1^*, \theta_2^*, \ldots, \theta_K^*, \theta_{K+1}^*, \ldots, \theta_n^*)$. The name of this method comes from the fact that the first K values of the full vector of the θ_i^*'s correspond to those clusters that are associated with at least one observation. As a result, there are no gaps in the values of the indicators $s_i,\ i = 1, 2, \ldots, n$. Using this augmentation, the problematic integral disappears and it is replaced by simple likelihood evaluations, since now all the new clusters that might be used are associated with some value (θ_{K+1}^* to θ_n^*).

In Neal (2000), two methods are proposed. The first one involves MH proposals for the update of the allocations $s_i,\ i = 1, 2, \ldots, n$, whereas the second method is very similar to the "no gaps" algorithm of MacEachern and Müller (1998), but slightly more general.

Finally, in the last approach, Jain and Neal (2005) propose an MCMC algorithm, which in each iteration proposes mix or split steps. More specifically, each mixing proposal consists of merging two clusters of discrete values of θ into one, and each splitting proposal suggests splitting one cluster of θ^* into two separate ones. This algorithm is computationally expensive, but with good mixing properties and can be a good choice in the nonconjugate case, if the other approaches fail to reach equilibrium in a sensible amount of time.

Conditional methods

Conditional methods can be particularly useful when it is not easy to integrate the random probability measure out of the joint posterior distribution of all parameters in a model. The basic idea in these methods is to impute the RPM in the parameter space and update it in the MCMC algorithm, as well as the other parameters. However, this involves an infinite number of parameters, making it practically impossible. Consider, for example, the discrete RPMs of the form $\sum_{i=1}^{\infty} w_i \delta_{\theta_i^*}(\cdot)$. In

order to simulate from such a representation, an infinite number of weights w_j and point masses θ_j^* needs to be simulated. In practice, we need a finite version of this, and a suggested solution is using some form of truncation, for example replacing the infinite sum in (1.1.3) with $\sum_{i=1}^{N} w_i \delta_{\theta_i^*}(\cdot)$, for N large enough (Ishwaran and Zarepour, 2000). Although the error produced by such approximations can be controlled (Ishwaran and James, 2001), it would be preferable to avoid such approximations completely. In the case of the DP as the distribution of the RPM, updates of the latter can be performed using its stick-breaking representation (1.1.3). Papaspiliopoulos and Roberts (2008) show that the approximation can be completely avoided, using a technique called retrospective sampling, which will be demonstrated in the case of a DP-distributed RPM, as is the case for Model (1.1.7): Suppose we want to create a sample $\theta_1, \theta_2, \ldots, \theta_n$ from $F \sim \mathrm{DP}(M, H_0)$. According to (1.1.3), if we could create an infinite number of pairs (w_j, θ_j^*), we would then assign each $\theta_i = \theta_j^*$ with probability w_j. This could be done using $U_i \sim U(0,1)$ and setting $\theta_i = \theta_j^*$ iff:

$$\sum_{k=0}^{j-1} w_k < U_i \leq \sum_{k=0}^{j} w_k, \tag{1.1.9}$$

where $w_0 = 0$. It is clear that, for a finite number of draws from F, only a finite number of pairs of weights and point masses is needed. Retrospective sampling method, now, simply exchanges the order of simulation between the pairs (w_j, θ_j^*) and the U_i's: we create a finite number of these pairs, and then check (1.1.9) for each U_i simulated. If for some of those U_i's, (1.1.9) is not satisfied, we go back and simulate more of these pairs, until (1.1.9) is satisfied for some j.

Retrospective sampling for the simple MDP model:

By replacing the DP by its equivalent expression (1.1.3), the MDP model can be written as follows:

$$\begin{gathered} Y_i \sim f(Y_i; \theta_{s_i}^*, \boldsymbol{\zeta}),\ i = 1, 2, \ldots, n \\ s_i \sim \sum_{j=1}^{\infty} w_j \delta_j,\ \text{where } w_j = V_j \prod_{k<j} (1 - V_k),\ \text{where } V_i \overset{iid}{\sim} \mathrm{Be}(1, M) \\ \theta_j^* \sim H_0(\boldsymbol{\psi}) \\ M \sim h_1(M),\ \boldsymbol{\zeta} \sim h_2(\boldsymbol{\zeta}),\ \boldsymbol{\psi} \sim h_3(\boldsymbol{\psi}). \end{gathered} \tag{1.1.10}$$

As in the marginal method, it is assumed that we do not have any additional parameters $\boldsymbol{\zeta}$ in the likelihood and the parameters $\boldsymbol{\psi}$ are fixed.

As mentioned above, since we just need a finite number of $\theta_i \equiv \theta_{s_i}^*$, only a finite number of weights and point masses is needed. So, we start with a large number of them, and if at some point in the simulation more are needed, we create them retrospectively. Another issue in this algorithm is that, in the full conditional distribution of each indicator $s_i,\ i = 1, 2, \ldots, n$, the intractable expression

$\sum_{j=1}^{\infty} w_j f(Y_i; \theta_j^*)$ appears as a normalising constant. As a result, we cannot simulate from these distributions directly. Papaspiliopoulos and Roberts (2008) propose a MH updating step in order to overcome this problem. To sum up, the proposed iterative steps in the MCMC algorithm (given an initial allocation $\boldsymbol{s} = (s_1, s_2, \ldots, s_n)$ and setting $N = \max\{\boldsymbol{s}\}$) are the following:

1. Simulate $\theta_j^*,\ j = 1, 2, \ldots, N$ from their full conditional distributions.
2. Simulate $V_j,\ j = 1, 2, \ldots, N$ from their full conditional distributions and calculate the weights $w_j = V_j \prod_{m<j}(1 - V_m),\ j = 1, 2, \ldots, N$.
3. For $i = 1, 2, \ldots, n$, simulate $U_i \sim U(0, 1)$.
   (a) Check if (1.1.9) is satisfied for some $j \leq N$. If yes, propose to update s_i using a MH update. If this proposed step is accepted, perform the change, otherwise keep the same value for s_i.
   (b) If, on the other hand, (1.1.9) is not satisfied for any $j \leq N$, simulate a pair $(V_{N+1}, \theta_{N+1}^*)$ from its prior distribution. Calculate $w_{N+1} = V_{N+1} \prod_{m \leq N}(1 - V_m)$, set $N = N + 1$ and go Step (3a).
4. Set $N = \max\{\boldsymbol{s}\}$.
5. Update M from its full conditional distribution, having first marginalised over the pairs (w_j, θ_j^*) not associated with any observation.

The above full conditional distributions are simple expressions and can be seen in Proposition 1 of Papaspiliopoulos and Roberts (2008). Notice also that, without the marginalisation mentioned in the update of M, one will not be able to perform this step, as the number of unused pairs is infinite.

Another method of overcoming the infinite number of parameters appearing in the stick-breaking representation of some RPMs (for example, those following a DP) without confronting to any approximation is the slice sampler, as demonstrated in Walker (2007). Consider, for example, Model (1.1.10), with the same simplifications mentioned above. Using (1.1.3), it is straightforward to see that the conditional likelihood of each data Y_i will be

$$f(Y_i | \boldsymbol{w}, \boldsymbol{\theta^*}) = \sum_{j=1}^{\infty} w_j f(Y_i; \theta_j^*). \tag{1.1.11}$$

By introducing a latent parameter u_i, (1.1.11) can be written as:

$$f(Y_i, u_i | \boldsymbol{w}, \boldsymbol{\theta^*}) = \sum_{j=1}^{\infty} 1_{(u_i < w_j)} f(Y_i; \theta_j^*). \tag{1.1.12}$$

By integrating out u_i we get (1.1.11), whereas this expression can be used to derive the full conditional distribution of each u_i (a uniform distribution on $(0, w_j)$, where j is the cluster associated

with observation Y_i), when those parameters are embedded at the parametric space of our model. On the other hand, given these u_i's, the model will now have only a finite number of parameters, therefore allowing for exact simulation from their full conditional distributions. As for the unused pairs of (w_j, θ_j^*), they need not be considered in the MCMC algorithm, as they can be integrated out. More details are given in Walker (2007).

1.1.3 The normalised inverse-Gaussian process

The normalised inverse-Gaussian process (N-IGP) seems to be a very good alternative to the Dirichlet process. It was introduced by Lijoi et al. (2005), similarly to the way Ferguson (1973) introduces the DP. One possible definition makes use of the normalised inverse-Gaussian distribution in the same way the Dirichlet distribution is used in the DP:

Definition 6. *A random variable X is said to have the inverse-Gaussian distribution with shape parameter $\alpha \geq 0$ and scale parameter $\gamma > 0$, symbolically $X \sim IG(\alpha, \gamma)$, if its density with respect to the Lebesque measure is the following:*

$$f_X(x) = \frac{\alpha}{\sqrt{2\pi}} x^{-3/2} \exp\left[-\frac{1}{2}\left(\frac{\alpha^2}{x} + \gamma^2 x\right) + \gamma\alpha\right],\ x \geq 0,\ \textit{for } \alpha > 0$$

and $X = 0$ almost surely for $\alpha = 0$.

In the following assume, without loss of generality, that $\gamma = 1$ (since it is a scale parameter).

Definition 7. *Let $X_1, X_2, \ldots, X_n$ be independent random variables with $X_i \sim IG(\alpha_i, 1)$, $i = 1, 2, \ldots, n$, with all $\alpha_i > 0$. Then, the random vector $\boldsymbol{W} = (W_1, W_2, \ldots, W_n)$, where $W_i = \frac{X_i}{\sum_{j=1}^n X_j}$, $i = 1, 2, \ldots, n$ is said to follow a normalised inverse-Gaussian distribution with parameters $\alpha_1, \alpha_2, \ldots, \alpha_n$. Symbolically, $\boldsymbol{W} \sim$ N-IG$(\alpha_1, \alpha_2, \ldots, \alpha_n)$.*

Another way to define the N-IG distribution is using the derived probability density function. More specifically, if $\boldsymbol{W} = (W_1, W_2, \ldots, W_n) \sim$ N-IG $(\alpha_1, \alpha_2, \ldots, \alpha_n)$, its pdf will be:

$$f_{\boldsymbol{W}}(\boldsymbol{w}) = \frac{e^{\sum_{i=1}^n \alpha_i} \prod_{i=1}^n \alpha_i}{2^{n/2-1}\pi^{n/2}} K_{-n/2}\left(\sqrt{\mathrm{A}_n(w_1, \ldots, w_n)}\right) \left(w_1^{3/2} w_2^{3/2} \cdots w_n^{3/2} \left[\mathrm{A}_n(w_1, \ldots, w_n)\right]^{n/4}\right)^{-1}$$

where $\mathrm{A}_n(w_1, \ldots, w_n) = \sum_{i=1}^n \frac{\alpha_i^2}{w_i}$ and K denotes the modified Bessel function of the third type.

Definition 8. *A random probability measure F is said to follow a normalised inverse-Gaussian process with parameters M and H_0 if for any partition $(A_1, A_2, \ldots, A_k)$ of the probability space Ω, such that all $A_i \in \mathcal{F}$, the $\sigma-$algebra of Ω, the vector of random probabilities $(F(A_1), F(A_2), \ldots, F(A_k))$*

follows a normalised inverse-Gaussian distribution with parameters $MH_0(A_1), MH_0(A_2), \ldots, MH_0(A_k)$. *Symbolically:*

$$F \sim \textit{N-IGP}(M, H_0) \overset{def}{\Leftrightarrow} \forall \textit{ partition } (A_1, A_2, \ldots, A_k) \textit{ of } \Omega,\ A_1, A_2, \ldots, A_k \in \mathcal{F},$$

$$(F(A_1), F(A_2), \ldots, F(A_k)) \sim \textit{N-IG}(MH_0(A_1), MH_0(A_2), \ldots, MH_0(A_k)) .$$

In the original definition of Lijoi et al. (2005), instead of M and H_0 there is only a nun-null finite measure α. For $M = \alpha(\Omega) > 0$ and $H_0(\cdot) = \frac{\alpha(\cdot)}{\alpha(\Omega)}$, we can see that this is just a reparametrisation. In the same article, the authors use Proposition 3.9.2 of Regazzini (2001), in order to show that the N-IGP is well defined.

The N-IGP has many similarities with the DP. The first obvious one is the parametrisation. Again, there is a distribution, H_0 and a positive scalar, M, and by studying the expressions of the expectation and variance of any realisation of the N-IGP, one will see that those two parameters have the same intuitive interpretation as the corresponding parameters of the DP. More specifically, if $F \sim \text{N-IGP}(M, H_0)$, we have:

$$\forall\ B \in \mathcal{F},\ \mathrm{E}\left(F(B)\right) = H_0(B) \text{ and } \mathrm{Var}\left(F(B)\right) = H_0(B)\left(1 - H_0(B)\right) M^2 e^M \Gamma(-2, M)$$

where $\Gamma(a, x) = \int_x^\infty e^{-t} t^{a-1} dt$ is the incomplete gamma function. As before, H_0 can be seen as the centre of the process and M as a measure of our belief in this centre. Another common property of the N-IGP and the DP is the almost sure discreteness of their realisations. This can be a problem when modelling continuous data, but again this can be resolved using mixtures of N-IG processes, as in the case of MDPs.

It is also worth mentioning that the N-IGP and the DP are the only known processes whose finite dimensional distributions are known explicitly.

The Pólya-urn representation of the N-IGP is also known. The structure is similar to the expression for the DP, with more complicated expressions for the weights:

Given data $\theta_1, \theta_2, \ldots, \theta_n$ from $F \sim \text{N-IGP}(M, H_0)$,

$$\forall\ A \in \mathcal{F},\ P\left(\theta_{n+1} \in A | \theta_1, \theta_2, \ldots, \theta_n\right) = w_0 H_0(A) + w_1 \sum_{j=1}^{K} \left(n_j - \frac{1}{2}\right) \delta_{\theta_j^*}(A)$$

where $\theta_j^*,\ j = 1, 2, \ldots, K$ are the discrete values of $\theta_1, \theta_2, \ldots, \theta_n,\ K$ is the number of those discrete values, $n_j = \#\{\theta_i = \theta_j^*\},\ j = 1, 2, \ldots, K$ is the number of θ_i's that are equal to the discrete value θ_j^*,

$$w_0 = \frac{\sum_{r=0}^{n} \begin{pmatrix} n \\ r \end{pmatrix} \left(-M^2\right)^{1-r} \Gamma\left(K + 1 + 2r - 2n, M\right)}{2n \sum_{r=0}^{n-1} \begin{pmatrix} n-1 \\ r \end{pmatrix} \left(-M^2\right)^{-r} \Gamma\left(K + 2 + 2r - 2n, M\right)}$$

and

$$w_1 = \frac{\sum_{r=0}^{n} \left(\begin{array}{c} n \\ r \end{array} \right) (-M^2)^{1-r} \Gamma(K + 2r - 2n, M)}{n \sum_{r=0}^{n-1} \left(\begin{array}{c} n-1 \\ r \end{array} \right) (-M^2)^{-r} \Gamma(K + 2 + 2r - 2n, M)}.$$

As before, the Chinese restaurant representation has a simple form. More specifically, using the same notation as above, together with the indicators $s_i, s_2, \ldots, s_n$, where $s_i = j \Leftrightarrow \theta_i = \theta_j^*$, the posterior probabilities for the assignment of a new value θ_{n+1} are as follows:

$$P(s_{n+1} = j) = \begin{cases} w_1(n_j - 1/2) & , \quad j = 1, 2, \ldots, K \\ w_0 & , \quad j = K + 1. \end{cases}$$

As Lijoi et al. (2005) discuss, the specific structure of the weights seems more sensible than the one of the DP, since this one also takes into account the total number of ties in the sample. The mechanism also indicates more elaborate allocation of weights in the clusters θ_j^*'s and, according to the authors, is more aggressive in detecting or reducing clusters for the data. They also discuss that, in general, the N-IGP is less informative than the DP prior.
On the other hand, unlike the DP, the stick-breaking representation for the N-IGP is not yet known, nor is this process conjugate.

Finally, note that the computational implementation for the models using the N-IGP is straightforward, and very similar to the corresponding models which use the DP. The process can again be integrated out, using its Pólya-urn representation.

1.2 Combining Inference

1.2.1 Literature review

Assume now that we want to model dependent data, denoted by Y's here. This dependence can be introduced in a variety of ways.

A first way to model such data is using mixture models. In general, mixture models are used in cases where we assign each mixture component to represent a different subgroup in a heterogenous population or as parsimonious models for flexible density estimation. In the Bayesian context two such models are given in Richardson and Green (1998) and Fernández and Green (2002). In the former the authors propose the mixture model $p(Y_i|K, \boldsymbol{w}, \boldsymbol{\lambda}) = \sum_{j=1}^{K} w_j f(Y_i|\lambda_j),\ i = 1, 2, \ldots, n$, where $Y_i,\ i = 1, 2, \ldots, n$ are the data and $\boldsymbol{w}$ and $\boldsymbol{\lambda}$ denote the vectors of all weights and component-specific parameters λ_j, respectively. They also assume that the number of mixture components, K,

is also allowed to vary. In the latter, the proposed mixture model for spatial data is $p(Y_i|K, \boldsymbol{w}, \boldsymbol{\lambda}) = \sum_{j=1}^{K} w_{ji} f(Y_i|\lambda_j),\ i = 1, 2, \ldots, n$, where $Y_i,\ i = 1, 2, \ldots, n$ are the data and $\boldsymbol{w}$ and $\boldsymbol{\lambda}$ denote the vectors of all weights and component-specific parameters λ_j, respectively. Spatial correlation is captured through the prior of the weights, and the authors propose two alternative choices for this prior: the logistic normal and the group continuous model. Again, the number of mixture components is considered random, so reversible jump MCMC (RJMCMC) methods (Green, 1995) are used in both models to simulate from the posterior distribution of all parameters of interest.

Apart from mixture models, dependence between data can also be introduced in a variety of ways, even within only the Bayesian nonparametric context. These models can be especially useful when we want to model some data (and the type of dependence among them) in a flexible way. The various models will be demonstrated using the DP as the distribution of the random distributions.

A first way of introducing dependence in the DP-distributed underlying distributions of the data, given covariates $\boldsymbol{x}$ (say, $F_{\boldsymbol{x}}$'s) is through their stick-breaking representation (1.1.3). More specifically, MacEachern (1999) introduces the Dependent Dirichlet process, where it is assumed that the weights $w_{i,\boldsymbol{x}}$ and/or the atoms $\theta^*_{i,\boldsymbol{x}}$ depend on covariates $\boldsymbol{x}$ and are thought to follow a stochastic process across the correlated $F_{\boldsymbol{x}}$'s (i.e. across the values of $\boldsymbol{x}$, for each $i = 1, 2, \ldots$). On the other hand, the vector of weights is assumed independent from the vector of atoms in each $F_{\boldsymbol{x}}$ (i.e. for each $\boldsymbol{x}$). Griffin and Steel (2006), on the other hand, again use covariates, and assume that the random variables V_i creating the weights in (1.1.3) depend on these covariates. Dependence is now introduced through the ordering of the covariates. They call this construction the order-based dependent Dirichlet process. Dunson and Park (2008) construct the so-called kernel stick-breaking prior, where they assume that, at each covariate value $\boldsymbol{x}$, $F_{\boldsymbol{x}}$ has a stick-breaking prior with atoms being RPMs (for example, DP-distributed) and $V_{i,\boldsymbol{x}}$ is the product of a beta-distributed random variable and of a covariate-dependent kernel values at random locations. Finally, note that the methods introduced in the last three articles can be naturally extended to other stick-breaking priors.

All the models presented above introduce dependence through the dependence of some covariates. Another way would be to impute this dependence through the dependence of some hyperparameters in the random distributions of the data. In this PhD thesis, however, another form of dependent data will be considered. More specifically, I will deal with grouped data, i.e. data that are clustered in distinct (usually few) categories. A natural way to model grouped data is to assume that the data are clustered in a few categories and that in each of this categories to assume a random underlying distribution. Dependence can then be introduced in those RPMs. As a first example,

consider the following structure:

$$Y_{ji} \sim F_j^*(Y_{ji}),\ i = 1, 2, \ldots, n_j,\ j = 1, 2, \ldots, J$$

$$F_j^* \sim \mathrm{DP}(M, H_{0,j}),\ \text{where } H_{0,j} \equiv f(\lambda_j),\ j = 1, 2, \ldots, J$$

$$(\lambda_1, \lambda_2, \ldots, \lambda_J) \sim p(\lambda_1, \lambda_2, \ldots, \lambda_J)$$

and the hyperparameters $\lambda_1, \lambda_2, \ldots, \lambda_J$ are assumed *a priori* not independent. This model is called a Product of Dirichlet Processes (PDP) and was introduced by Cifarelli and Regazzini (1978). In this model, as well as the hierarchical Dirichlet process presented below, the discrete nature of the realisations of the DP results in the grouping of the data $Y_{ji},\ i = 1, 2, \ldots, n_j$, for each $j = 1, 2, \ldots, J$. Apart from the clustering, dependence among the data is also introduced through the dependence of the correlated distributions F_j^*, due to the dependent structure of their hyperparameters $\lambda_j,\ j = 1, 2, \ldots, J$.

Teh et al. (2006) proposed a model called Hierarchical Dirichlet process, where it is assumed that all the RPMs follow the same DP prior and the base distribution of the latter is again modelled through a DP:

$$Y_{ji} \sim F_j^*(Y_{ji}),\ i = 1, 2, \ldots, n_j,\ j = 1, 2, \ldots, J$$

$$F_j^* \sim \mathrm{DP}(M, H),\ i = 1, 2, \ldots, J$$

$$H \sim \mathrm{DP}(M_0, H_0).$$

Another method of modelling grouped data is to assume that within each data set the data are identically distributed and independent from a random distribution $F_j^*,\ j = 1, 2, \ldots, J$, and additionally assume that each correlated random distribution F_j^* consists of a common part shared by all of the F_j^*'s, F_0, and an idiosyncratic part, F_j. A general model of the last form was given by Müller et al. (2004).

1.2.2 The model of Müller et al. (2004)

Assume that there are J data sets $Y_{ji},\ i = 1, 2, \ldots, N_j, j = 1, 2, \ldots, J$, each from a distribution F_j^*. If we can assume that each of the distributions F_j^*'s consists of a common part, shared by all of them, and an idiosyncratic part, the model of Müller et al. (2004) can be used:

$$Y_{ji} \sim f(Y_{ji}; \theta_{ji}, \boldsymbol{\psi}),\ i = 1, 2, \ldots, N_j,\ j = 1, 2, \ldots, J$$

$$\theta_{ji} \sim F_j^*,\ \text{ where } F_j^* = \varepsilon F_0 + (1 - \varepsilon) F_j,\ j = 1, 2, \ldots, J$$

$$F_j \overset{ind}{\sim} \mathrm{DP}(M_j, H(\boldsymbol{\lambda})),\ j = 0, 1, 2, \ldots, J \tag{1.2.13}$$

$$\pi(\varepsilon) \;=\; \pi_0\delta_0(\varepsilon) + \pi_1\delta_1(\varepsilon) + (1-\pi_0-\pi_1)\mathrm{Be}(a_\varepsilon, b_\varepsilon)$$

$$M_0, M_1, M_2, \ldots, M_J \overset{iid}{\sim} \mathrm{Ga}(a_0, b_0),\ \boldsymbol{\psi} \sim \pi(\boldsymbol{\psi}),\ \boldsymbol{\lambda} \sim \pi(\boldsymbol{\lambda})$$

where $0 \leq \pi_0 < 1,\ 0 \leq \pi_1 < 1-\pi_0$ and the rest are as defined before. The vector $\boldsymbol{\lambda}$ is used to denote the vector of unknown parameters of the base distribution H, and $\boldsymbol{\psi}$ any additional parameters in the likelihood f.

The component distributions $F_0, F_1, F_2, \ldots, F_J$ are assigned independent Dirichlet process prior, resulting in a flexible model, even though the base distribution is common in all and what distinguishes their prior distributions are the concentration parameters $M_j,\ j = 0, 1, \ldots, J$. The concentration parameters are themselves given a gamma prior distribution, as is often the case in the literature.

Dependence among the random distributions F_j^*'s is introduced by the common part F_0 and the common weight assigned to this common part, ε. This weight can also be seen as the level of borrowing strength across the different distributions. The model assigns a quite general prior for this weight, giving positive probability to the extreme events $\varepsilon = 0$ and $\varepsilon = 1$. The former case corresponds to the event that the correlated distributions have no common part (and therefore, they are not actually correlated!) and the latter corresponds to the case when all the distributions are the same. Note also that the fact that there is a single weight in all distributions F_j^*'s is not as restrictive as it looks, because of the flexible prior distribution of $F_0, F_1, \ldots, F_J$. In cases where two different allocations for ε and the component distributions $F_0, F_j,\ j = 1, 2, \ldots, J$ fit the data equally well, then the Bayesian approach will tend to favor the most parsimonious model over the more complicated one, i.e. the model with less parameters. This is a direct implementation of Ockham's razor (Jefferys and Berger, 1992) in posterior inference and, as Müller et al. (2004) argue, it can be justified by the fact that, in the more complicated model, the (roughly the same) prior probability must be distributed over a larger number of parameters, and therefore the marginal probabilities will be smaller.

Computational implementation of Model (1.2.13) can be achieved again using MCMC methods. In fact, one can take the algorithm developed for the simple MDP model and just change a few things. The basic difference (apart from the obvious updating of parameters not present in the MDP model, for example, ε) will be the use of a second set of binary indicators, say $r_{ji}, i = 1, 2, \ldots, n_j, j = 1, 2, \ldots, J$, denoting if a specific θ_{ji} belongs to the common part F_0 (if $r_{ji} = 0$) or to the idiosyncratic part F_j (if $r_{ji} = 1$). Using these indicators, together with the indicators $s_{ji}, i = 1, 2, \ldots, n_j, j = 1, 2, \ldots, J$, which will denote the specific cluster (now within F_0 or F_j, according to the value of the related r_{ji}), $\theta_{ji}, i = 1, 2, \ldots, n_j, j = 1, 2, \ldots, J$ can be reparameterised to $r_{ji}, s_{ji}, i = 1, 2, \ldots, n_j, j = 1, 2, \ldots, J$ and $\theta^*_{ji}, i = 1, 2, \ldots, K_j, j = 0, 1, 2, \ldots, J$, where K_j is the number of the discrete values $\theta^*_{ji}, j = 0, 1, 2, \ldots, J$ within component distribution $F_0, F_1, \ldots, F_J$

respectively. As before, updating those discrete values can increase the efficacy of the algorithm. As for posterior inference, the basic quantities of interest will be the predictive distributions for each data set, $p(Y_{j,n_j+1}|Y_{j,1},\ldots,Y_{j,n_j}), j=1,2,\ldots,J$, the corresponding predictive distributions in the common and the idiosyncratic component distributions and the posterior distributions of the M_j's and ε. As an illustration of the above, consider the model of Müller et al. (2004) for the case of normal likelihood and normal base distribution for two correlated distributions:

$$\begin{aligned}
Y_{ji} &\sim N(\mu_{ji}, S),\ i=1,2,\ldots,N_j,\ j=1,2 \\
\mu_{ji} &\sim F_j^*,\ \text{where } F_j^* = \varepsilon F_0 + (1-\varepsilon)F_j,\ j=1,2 \\
F_0 &\sim \mathrm{DP}(M_0, H),\ F_j \overset{ind}{\sim} \mathrm{DP}(M_j, H),\ \text{for } H \equiv N(m, B) \\
\pi(\varepsilon) &= \pi_0\delta_0(\varepsilon) + \pi_1\delta_1(\varepsilon) + (1-\pi_0-\pi_1)\mathrm{Be}(a_\varepsilon, b_\varepsilon) \\
M_0, M_1, M_2 &\overset{iid}{\sim} \mathrm{Ga}(a_0, b_0),\ S \sim \mathrm{IGa}(q, 1/qR) \\
(m, B) &\sim N(m_0, A) \times \mathrm{IGa}(c, 1/cC)
\end{aligned} \tag{1.2.14}$$

where $\mathrm{IGa}(a,b)$ denotes the inverse gamma distribution with with shape parameter a and scale parameter b and $N(\mu,\sigma^2)$ denotes the normal distribution with mean μ and variance σ^2, as defined below.

Definition 9. *A random variable X is said to follow an inverse gamma (IGa) distribution with parameters $\alpha>0$ and $\beta>0$, denoted IGa(α,β), if its density with respect to the Lebesque measure is:*

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp\{-\beta/x\},\ x>0.$$

The mean of this distribution is $\frac{\beta}{\alpha-1}$, if $\alpha>1$ and the variance is $\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$, if $\alpha>2$. It also holds that, if $X\sim\mathrm{IGa}(\alpha,\beta)$, then $1/X\sim\mathrm{Ga}(\alpha,1/\beta)$.

Definition 10. *A random variable X is said to follow a normal distribution with parameters $\mu\in\mathbb{R}$ and $\sigma^2>0$,, denoted by $N(\mu,\sigma^2)$, if its density with respect to the Lebesque measure is:*

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\},\ x\in\mathbb{R}.$$

The joint posterior distribution of all parameters, using the parametrisation of the discrete values (say, ϕ_{ji} here) and the indicators, will be the following:

$$\begin{aligned}
f(\mathbf{s},\mathbf{r},\boldsymbol{\phi},\varepsilon,m,B,S,M_0,M_1,M_2|\mathbf{Y}) \ &\propto\ \prod_{j,i} f(Y_{ji}|r_{ji},s_{ji},\boldsymbol{\phi},S) f(m)f(B)f(M_0,M_1,M_2)f(S)f(\varepsilon) \\
&\times\ \prod_{j,i} f(\phi_{ji}|m,B) \prod_{j,i} f(r_{ji}|\varepsilon) f(\mathbf{s}|\mathbf{r},M_0,M_1,M_2)
\end{aligned}$$

where the bold letters denote the vector of all parameters indicated (for example, $\boldsymbol{s}$ denotes all the indicator variables $s_{ji},\ i = 1, 2, \ldots, N_j,\ j = 1, 2$ and $\boldsymbol{Y}$ denotes all the data).
The full conditional distribution of each parameter (i.e. the distribution given all the other parameters) is as follows:

- $m|\cdots \sim N(\frac{m_0 B + A\sum_{j,i}\phi_{ji}}{AK+B}, \frac{AB}{AK+B})$,
  where $K = K_0 + K_1 + K_2$ is the total number of discrete values in all component distributions.

- $B|\cdots \sim \text{IGa}(c + K/2, 1/cC + 1/2\sum_{j,i}(\phi_{ji} - m)^2)$.

- $S|\cdots \sim \text{IGa}(q + N/2, 1/qR + 1/2\sum_{j,i}(Y_{ji} - \mu_{ji})^2)$, where $N = N_1 + N_2$.

- $\varepsilon|\cdots \propto \begin{cases} 0 & \text{, w.p. } \pi_0 1_{(\sum r_{ji} = N)} \\ 1 & \text{, w.p. } \pi_1 1_{(\sum r_{ji} = 0)} \\ Be(a_\varepsilon + N - \sum r_{ji}, b_\varepsilon + \sum r_{ji}) & \text{, w.p. } (1 - \pi_0 - \pi_1) B(a_\varepsilon + N - \sum r_{ji}, b_\varepsilon + \sum r_{ji})/B(a_\varepsilon, b_\varepsilon). \end{cases}$

- $f(M_0|\cdots) \propto M_0^{a_0 + K_0 - 1} e^{-M_0 b_0} \frac{\Gamma(M_0)}{\Gamma(M_0 + n_0)}$,
  $f(M_1|\cdots) \propto M_1^{a_0 + K_1 - 1} e^{-M_1 b_0} \frac{\Gamma(M_1)}{\Gamma(M_1 + n_1)}$ and
  $f(M_2|\cdots) \propto M_2^{a_0 + K_2 - 1} e^{-M_2 b_0} \frac{\Gamma(M_2)}{\Gamma(M_2 + n_2)}$,
  where n_j is the number of data allocated to component distribution $F_j,\ j = 0, 1, 2$.

- $\phi_{0l}|\cdots \sim N(\frac{mS + B\sum_{j,i:r_{ji}=0, s_{ji}=l} Y_{ji}}{S + Bn_{0l}}, \frac{SB}{S + Bn_{0l}}),\ l = 1, 2, \ldots, K_0$ and
  $\phi_{jl}|\cdots \sim N(\frac{mS + B\sum_{i:r_{ji}=1, s_{ji}=l} Y_{ji}}{S + Bn_{jl}}, \frac{SB}{S + Bn_{jl}}),\ l = 1, 2, \cdots, K_j,\ j = 1, 2$,
  where n_{ji} is the number of data allocated to the i-th cluster of component distribution $F_j,\ i = 1, 2, \ldots, K_j,\ j = 0, 1, 2$.

- $f(\mathbf{s}, \mathbf{r}|\cdots) \propto \prod_{j,i} f(Y_{ji}|r_{ji}, s_{ji}, \boldsymbol{\phi}, S) f(\mathbf{r}|\varepsilon) f(\mathbf{s}|\mathbf{r}, M_0, M_1, M_2)$
  $\Rightarrow P(s_{ji} = h, r_{ji} = l|\cdots) = \begin{cases} \pi_{jh} & , h = 1, 2, \ldots, K_j,\ l = 1 \\ \pi_{0h} & , h = 1, 2, \ldots, K_0,\ l = 0 \\ \pi_j^* & , h = K_j + 1,\ l = 1 \\ \pi_0^* & , h = K_0 + 1,\ l = 0 \end{cases}, \ i = 1, \ldots, N_j, j = 1, 2$
  where $\pi_{jh} \propto (1 - \varepsilon)\varphi(Y_{ji}; \phi_{jh}, S) n_{jh}^- / (M_j + n_j^-),\ \pi_{0h} \propto \varepsilon\varphi(Y_{ji}; \phi_{0h}, S) n_{0h}^- / (M_0 + n_0^-)$,
  $\pi_j^* \propto (1 - \varepsilon)\varphi(Y_{ji}; m, S + B) M_j / (M_j + n_j^-)$, and $\pi_0^* \propto \varepsilon\varphi(Y_{ji}; m, S + B) M_0 / (M_0 + n_0^-)$, where the superscript $-$ means that the corresponding quantity is taken without counting the quantity associated with the (ji) point, φ is the pdf of the normal distribution and the above probabilities are all proportional to the same constant, which is such that the probabilities sum up to 1. Finally, note that in the last two cases for $(s_{ji}, r_{ji}|\cdots)$, a new value should be created. This is a draw from $N(\frac{BY_{ji} + mS}{B+S}, \frac{BS}{B+S})$.

We can directly simulate from all the above full conditional distributions, except from the ones of the precision parameters M_0, M_1 and M_2. On the other hand, for each of those three parameters the simple trick explained in Escobar and West (1995) can be applied.
Finally, the predictive distributions for the two data sets are as follows:

$$\begin{aligned} p(Y_{j,N_j+1}|\mathbf{Y}) &= \varepsilon\frac{M_0}{M_0+n_0}N(m,B+S)+\varepsilon\frac{1}{M_0+n_0}\sum_{d=1}^{K_0}n_{0d}N(\phi_{0d},S) \\ &+ (1-\varepsilon)\frac{M_j}{M_j+n_j}N(m,B+S)+(1-\varepsilon)\frac{1}{M_j+n_j}\sum_{d=1}^{K_j}n_{jd}N(\phi_{jd},S),\ j=1,2. \end{aligned}$$

To sum up, the model introduced by Müller et al. (2004) is a very general model for correlated distributions which have a common and an idiosyncratic part. As a nonparametric mixture model, it is a very flexible model, although the weight of the common part is the same *a priori* for all the correlated distributions. It is also easily implemented using MCMC methods and its clear structure allows for direct posterior inference of the parameters of interest. On the other hand, the fact that there is a common weight ε and the same base distribution H does not seem very sensible and it might be worth considering indexing either or both of them by j.

1.2.3 Normalising random measures

It is well known that, under mild conditions, one can construct random probability measures by normalising other random measures (see, for example, James et al., 2005). This class of measures is called normalised random measures (NRMs) and is a particularly rich one. Apart from the Dirichlet process (Ferguson, 1973), it also contains the N-IGP (a normalised inverse-Gaussian process, see for example Lijoi et al., 2005) and the Pitman-Yor process. As an example, consider the normalisation of the gamma Process:

Definition 11. *Let Ω denote a probability space and $\mathcal{F}$ the $\sigma-$algebra of Ω. It is said that a random measure G follows a Gamma process (ΓP) with parameters M and H_0, where $M>0$ and H_0 is a probability measure iff for any partition $A=\{A_1,A_2,\ldots,A_k\}$ of Ω, such that all $A_i\in\mathcal{F}$, the random probabilities $G(A_1),G(A_2),\ldots,G(A_k)$ are mutually independent and each $G(A_i)$ follows a gamma distribution with shape parameter $MH_0(A_i)$ and scale parameter 1.*

Ferguson (1973) defined the Dirichlet process by:

$$F\sim \mathrm{DP}(M,H)\Leftrightarrow \forall\ B\in\mathcal{F},\ F(B)=\frac{G(B)}{G(\Omega)}=\frac{G(B)}{G(B)+G(B^c)},\ \text{where } G\sim\Gamma\mathrm{P}(M,H).$$

B^c denotes the compliment set of B, and the denominator of the expression on the right highlights the dependence of $G(B)$ and $G(\Omega)$, since they come from the same process. We also note that the

parameters of the DP, i.e. the concentration parameter M and the base distribution H are the same as the equivalent parameters of the underlying ΓP.
The basic idea is that one can exploit the infinite divisibility of some random measure, in order to create (by normalising this measure) random probability measures that have the same distribution, but are not independent. This idea will be demonstrated in Section 2.1.1.

1.3 My Contribution

My contribution to Bayesian nonparametric modelling consists of proposing a new, general method of constructing models with dependent random distributions for grouped data. Two examples of these models for modelling data from two different groups are given. Generalisations for more than two correlated groups for these models, as well as for the model proposed in Müller et al. (2004), are also investigated. These models are also embedded in the stochastic frontier setting and used to construct a model for the efficiency of firms. In implementing the proposed models, I observed some problems in mixing, so an additional split-merge step in the MCMC algorithms is proposed. This algorithm is seen to improve mixing of the chains and can also be used in a variety of models.

Apart from my contribution to nonparametric models, I also propose models for parametric inference. More specifically, I introduce a general class of n-dimensional distributions, which includes the Dirichlet and the inverse-Gaussian distribution as special cases. The general formulae for the moments and cross-moments for this class of distributions are also derived (Mathematica codes for these expressions will be soon made available on the web). I apply this distribution to the underlying probabilities of success for binomial data and use the derived structure to model overdispersed count data.

1.4 Outline

This PhD thesis will proceed as follows: In Section 2 I describe a general class of models with dependent random distributions, as well as a new way of constructing such models. This is demonstrated by constructing two models in the two-dimensional case. The intuition and the theoretical properties of the derived models are discussed and I give some general ideas and concepts for generalising those models, as well as the model of Müller et al. (2004) in higher dimensions (higher number of dependent random distributions). The computational implementation of the models presented in this section are described in Section 3, with an illustration of the related algorithms using three simulated data sets. In Section 4 some models are applied to real-life data. At first my basic proposed

model and the model of Müller et al. (2004) are applied to financial data. Next, I embed those two models, together with a modification of my model (N-IGP priors for the RPMs in each component, instead of DP priors) to the stochastic frontier (SF) setting. Finally, the three models are applied to hospital cost frontier data. In Chapter 5 a general class of n-dimensional distribution in the unit simplex is proposed. Some theoretical properties of this class of distributions are discussed and the formula for its moments is derived. I then consider the univariate version of this distribution as the underlying distribution of the probability of success of binomial data. The derived model can be then used to model overdispersed count data. Finally, this model is applied to both simulated and real data (mice fetal mortality data) and the results for the mice data are compared to the results of other models in the literature. In Chapter 6 I provide a summary of what was done in this thesis, as well as possible future directions.

Chapter 2

A General Class of Models for Correlated Distributions

In this section I consider a general class of models for pairs of correlated distributions. I focus on two of those models, which I will be mainly using in the rest of the thesis. Some more general models are also considered, which incorporate a higher number of dependent distributions.

2.1 The Models For Two Correlated Distributions

As mentioned in the introduction, a common way of inducing dependence between data from different studies is to assume that their underlying distributions are correlated. In order to add extra flexibility to these models, it also assumed that those distributions are random, therefore creating a Bayesian nonparametric model.

In the simplest case of two correlated random distributions, say F_1^* and F_2^*, a general model of this type is the following:

$$F_1^* = \varepsilon_1 F_0 + (1 - \varepsilon_1) F_1$$

$$F_2^* = \varepsilon_2 F_0 + (1 - \varepsilon_2) F_2$$

where F_0, F_1 and F_2 are independent random probability measures and $\varepsilon_1, \varepsilon_2$ are random variables in the unit interval. In this structure F_0 can be seen as the common part shared by F_1^* and F_2^*, whereas F_1 and F_2 can be interpreted as idiosyncratic parts. The two correlated distributions can then be naturally embedded at an intermediate level of a larger hierarchical model. If the DP is assigned as the prior distribution of F_0, F_1 and F_2, the discreteness of its realisations can be overcome

by assuming that the data follow a continuous distribution f with some of its parameters following F_1^* and F_2^*:

$$\begin{aligned} Y_{ji} &\sim f(Y_{ji};\theta_{ji},\boldsymbol{\psi}),\ i=1,2,\ldots,N_j,\ j=1,2 \\ \theta_{ji} &\sim F_j^*,\ \text{ where } F_j^* \ = \ \varepsilon_j F_0 \ + \ (1-\varepsilon_j)F_j,\ j=1,2 \\ F_j &\sim \mathrm{DP}(M_j, H(\boldsymbol{\lambda})),\ j=0,1,2 \\ \varepsilon_j &\sim \pi(\varepsilon_j),\ j=1,2 \\ M_j &\overset{ind}{\sim} \pi(M_j),\ j=0,1,2,\ \boldsymbol{\psi}\sim\pi(\boldsymbol{\psi}),\ \boldsymbol{\lambda}\sim\pi(\boldsymbol{\lambda}). \end{aligned} \tag{2.1.1}$$

In the above, Y_{ji} are data from two different groups of sizes N_1 and N_2, θ_{ji} are the parameters to be flexibly modelled using nonparametric, correlated distributions and $\boldsymbol{\psi}$ are (potential) additional parameters in the distribution of the data. Different concentration parameters for the three DPs are assumed, but the same centering distribution, H, with some parameters $\boldsymbol{\lambda}$. In this way, the two distributions F_1^* and F_2^* share information, not only through F_0, but also through the common base distribution H, and their common parameter $\boldsymbol{\lambda}$.

It can be easily seen that the model of Müller et al. (2004) for $J=2$ is a special case of Model (2.1.1), where $\varepsilon_1=\varepsilon_2$, and a certain prior distribution is given to the common weight. On the other hand, the form of the above model offers other attractive options.

One such option is to have two concentration parameters, by setting $M_1=M_2$, and a common weight ε, but with a $\mathrm{Be}(M_0,M_1)$ prior. In this way, F_1^* and F_2^* are identically, *a priori* Dirichlet Process-distributed:

$$F_1^*, F_2^* \sim \mathrm{DP}(M_0+M_1, H(\boldsymbol{\lambda})).$$

This will be shown in more details in Section 2.1.3.

In this case, borrowing strength between the dependent distributions is also achieved through the common weight and the common concentration parameter for the idiosyncratic part, M_1.

A question that arises naturally is whether having the prior for the weights depending on the precision parameters of the DP followed by the random probability measures is sensible. In other words, does the prior $\varepsilon\sim\mathrm{Be}(M_0,M_1)$ make sense? At first sight, it might seem that the answer is no. Also, one might argue that having the same base distribution H for all F_0, F_1 and F_2 is not sensible, either. However, the combination of those two seemingly peculiar facts might be explained as follows: the proportion of information carried from the common part of F_1^* and F_2^* regarding the common base distribution H should be positively associated with the proportion of this common part in the models. This proportion can be expressed by $\frac{M_0}{M_0+M_1}$, since M_0 and M_1 are parameters controlling how close we are to the base distribution (for example, if most observations come from

the common part, then the base distribution should be "close" to this common part, so M_0 should be significantly larger than M_1 and the ratio should be large). On the other hand, this proportion of information of F_0 to H must be also positively associated to the weight ε, since this is the weight of F_0 in both F_1^* and F_2^*. However, the prior mean of the weight is exactly the ratio $\frac{M_0}{M_0+M_1}$. As a result, the prior distribution of ε, involving M_0 and M_1, together with having the same base distribution for all F_0, F_1 and F_2 can be justified.

A second interesting model of the form (2.1.1) could be one similar to the one above, with the difference that now there are two weights, ε_1 and ε_2, which are identically distributed as $\text{Be}(M_0, M_1)$ *a priori*, but not independent. This model is constructed using the normalisation ideas of Section 1.2.3, i.e. using a different and quite general method, through which the prior distributions and the correlation structures between the parameters are set. This method is described in the next subsection where dependent and identically distributed Dirichlet processes are constructed.

2.1.1 The model via direct normalisation

Let $G_i \overset{ind}{\sim} \Gamma\text{P}(M_i, H), i = 1, 2, \ldots, k$ and $M = \sum_{i=1}^{k} M_i$, and define $G^*(B) = \sum_{i=1}^{k} G_i(B)$, $\forall\ B \in \mathcal{F}$. Then,

$$G^*(\cdot) \sim \Gamma\text{P}(M, H). \tag{2.1.2}$$

This property is called infinite divisibility and is inherited to the gamma process from the underlying gamma distribution. It states that a gamma process (actually, any realisation of it) can be divided in as many (other) gamma processes as one wants:

Definition 12. *A distribution F is called infinite divisible if and only if $\forall\ n \in \mathbb{N}$, $\exists$ a distribution F_n such that F is equal to the convolution of n times F_n.*
In other words, F is called infinitely divisible if and only if $\forall\ n \in \mathbb{N}$, it can be represented as the distribution of the sum $S_n = X_{1,n} + X_{2,n} + \cdots + X_{n,n}$, where $X_{1,n}, X_{2,n}, \ldots, X_{n,n}$ are independent random variables, each following the same distribution, say F_n.

By normalising G^*, it is found that $F^*(\cdot)$ follows a Dirichlet process:
Let $B \in \mathcal{F}$ and $F^* \sim \text{DP}(M, H)$, $F_i \overset{ind}{\sim} \text{DP}(M_i, H), i = 1, 2, \ldots, k$.

Then,

$$\begin{aligned} F^*(B) &= \frac{G^*(B)}{G^*(\Omega)} \\ &\overset{(2.1.2)}{=} \frac{\sum_{i=1}^k G_i(B)}{\sum_{j=1}^k G_j(\Omega)} \\ &= \sum_{i=1}^k \frac{G_i(\Omega)}{\sum_{j=1}^k G_j(\Omega)} \frac{G_i(B)}{G_i(\Omega)} \\ &= \sum_{i=1}^k \varepsilon_i F_i(B), \text{ where } \varepsilon_i = \frac{G_i(\Omega)}{\sum_{j=1}^k G_j(\Omega)}. \end{aligned}$$

So, any DP with parameters M and H can be written as a weighted sum of k independent DPs with the same base distribution H and precision parameters M_i, such that $\sum_{i=1}^k M_i = M$. The corresponding weights are given by $\varepsilon_i = \frac{G_i(\Omega)}{\sum_{j=1}^k G_j(\Omega)}$, depending only on the M_i (since each $G_i(\Omega)$ is distributed as $\mathrm{Ga}(M_i, 1)$). In fact, those weights follow a Dirichlet distribution:

$$(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k) \sim \mathrm{Dir}(M_1, M_2, \ldots, M_k). \tag{2.1.3}$$

Let now $F_0 \sim \mathrm{DP}(M_0, H),\ F_1, F_2 \sim \mathrm{DP}(M_1, H)$. By normalising the underlying gamma process of the sum of F_0 and F_1, we get:

$$F_1^* = \varepsilon_1 F_0 + (1 - \varepsilon_1) F_1$$

where F_1^* now follows a $\mathrm{DP}(M_0 + M_1, H)$ and $\varepsilon_1 \sim \mathrm{Be}(M_0, M_1)$. Similarly, by normalising the gamma processes corresponding to F_0 and F_2, we get:

$$F_2^* = \varepsilon_2 F_0 + (1 - \varepsilon_2) F_2$$

where F_2^* follows also a $\mathrm{DP}(M_0 + M_1, H)$ and $\varepsilon_2 \sim \mathrm{Be}(M_0, M_1)$.
So, F_1^* and F_2^* are identically DP-distributed, but obviously not independent, due to the common part F_0. The same holds for the two weights, which are both beta-distributed, but are not independent. In fact, notice that $\varepsilon_1 = \frac{G_0(\Omega)}{G_0(\Omega)+G_1(\Omega)}$ and $\varepsilon_2 = \frac{G_0(\Omega)}{G_0(\Omega)+G_2(\Omega)}$ and their joint distribution is:

$$f_{\varepsilon_1,\varepsilon_2}(\varepsilon_1, \varepsilon_2) = \frac{\Gamma(M_0 + 2M_1)}{\Gamma(M_0)[\Gamma(M_1)]^2} \frac{\varepsilon_1^{M_0+M_1-1}(1-\varepsilon_1)^{M_1-1}\varepsilon_2^{M_0+M_1-1}(1-\varepsilon_2)^{M_1-1}}{(\varepsilon_1 + \varepsilon_2 - \varepsilon_1\varepsilon_2)^{M_0+2M_1}},\ 0 < \varepsilon_1, \varepsilon_2 < 1.$$

Proof:

Let $x_1 = G_0(\Omega) \sim \mathrm{Ga}(M_0, 1), x_2 = G_1(\Omega) \sim \mathrm{Ga}(M_1, 1)$ and $x_3 = G_2(\Omega) \sim \mathrm{Ga}(M_2, 1)$. Consider now the reparametrised vector $(\varepsilon_1, \varepsilon_2, y)$, where $\varepsilon_1 = \frac{x_1}{x_1+x_2}, \varepsilon_2 = \frac{x_1}{x_1+x_3}, y = x_1$.
By applying the formula for the distribution of a transformed random vector, and then integrate out y, we arrive at the above result. □

Finally, it is also worth mentioning that by repeating the same procedure as above, but now with normalising an inverse-Gaussian process (which is also infinite divisible), one can construct the same model, but now with normalised inverse-Gaussian processes (Lijoi et al., 2005) as the priors of F_0, F_1 and F_2, as well as of F_1^* and F_2^*. In this case, the weights will have normalised inverse-Gaussian distributions as priors, with parameters M_0 and M_1, where the latter are the concentration parameters of the corresponding N-IGP priors of F_0 and F_1 (or F_2).

2.1.2 The basic proposed model

The main model of consideration and comparison with the Müller et al. (2004) model is a simplified version of the above model for the DP case, where a common weight ε is assumed. This simplification allows for more direct sharing of information between the two distributions (since the weights now are the same, and not just correlated). This sharing of information can be particularly useful in cases of few observations from one or both dependent distributions. On the other hand, unless someone is particularly interested in inferring the weights in both distributions, not much is lost by having the same weight, because of the nonparametric, flexible modelling of F_0, F_1 and F_2. Most of the posterior mass for the weight will be assigned to the minimum of the weights creating the data and a (usually small) proportion will be assigned to values very close to zero. In order to illustrate this, consider the following example:

Example 1:

$$Y_{1i} \overset{iid}{\sim} \frac{7}{10}N(1,1) + \frac{3}{10}N(-10,1),\ i = 1,2,\ldots,N_1$$

$$Y_{2i} \overset{iid}{\sim} \frac{3}{10}N(1,1) + \frac{7}{10}N(8,1),\ i = 1,2,\ldots,N_2.$$

The dependent DPs, as described above, are used as the prior distribution of the distributions of the means of the above normal distributions. In order to simulate from the posterior distributions of all the parameters of interest, MCMC methods are used, which are discussed in the next section. For now, I will focus on the implications of applying the basic proposed model to this type of data. The data sizes used were $N_1 = N_2 = 100$. As can be seen from Figure 2.1, the posterior distribution of the weight puts most of its mass on values around 0.3 (which is the minimum of 0.7 and 0.3). Under this value for ε, F_0 will be concentrated around the (correct) value 1 (the form of this posterior for F_0 will also depend on the base distribution H, but for simplicity let's assume that this is a fairly smooth and unimodal distribution, for example a normal distribution), F_2 will be concentrated around 8 and F_1 will be bimodal, with about 57% of the mass around 1 and the rest around -10:

$$\text{For } \varepsilon \simeq 0.3,\ F_0 = N(1,1),\ F_1 = \frac{4}{7}N(1,1) + \frac{3}{7}N(-10,1) \text{ and } F_2 = N(8,1).$$

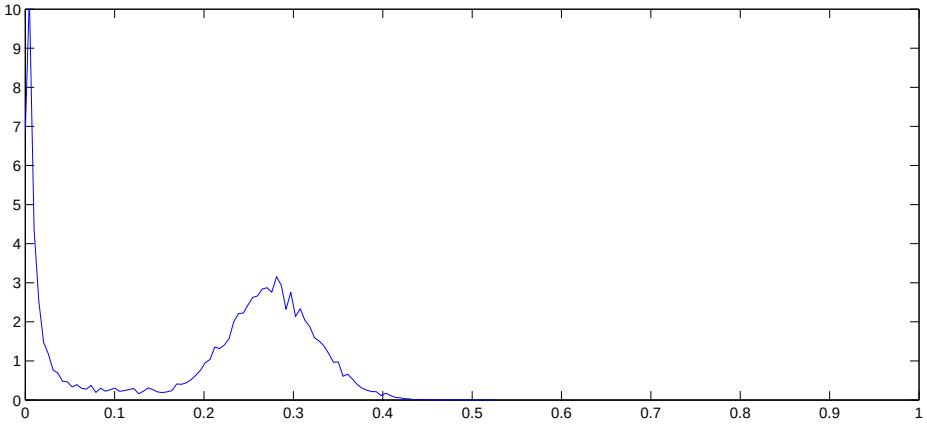


Figure 2.1: Kernel density estimate for the posterior of the weight ε for Model (2.1.4) for the first simulated data set.

The above interpretation for F_1 and F_2 are derived after we take out the common part F_0 from the two dependent distributions F_1^* and F_2^*. This assignment of the parameters creates four distinct clusters for the values of the means: one for each F_0 and F_2 and two for F_1. On the other hand, any value for the weight between 0 and 0.3 will produce a consistent assignment for F_0, F_1 and F_2 (in the sense that it locates correctly the mode of the common part, which is more crucial than locating the idiosyncratic parts), but with five clusters of values: one of F_0 and two for each F_j, $j = 1, 2$. The special case $\varepsilon = 0$, where it is assumed that F_1^* and F_2^* have no common part is also an economic option, as the number of clusters will only be four (two for each of F_1, F_2), and this is the reason why there is also posterior mass there:

$$\text{For } \varepsilon \simeq 0,\ F_0 = \emptyset,\ F_1 = \frac{7}{10}N(1,1) + \frac{3}{10}N(-10,1) \text{ and } F_2 = \frac{3}{10}N(1,1) + \frac{7}{10}N(8,1).$$

where $\emptyset$ denotes the empty set.

The reason that the posterior mass around 0 is usually smaller than at 0.3 is that the latter correctly discovers the common part and shares information more efficiently than in the former case, where there is no sharing of information regarding F_0. The other special case, $\varepsilon = 1$, where it is assumed that F_1^* and F_2^* are the same is not a valid one, since the form of the data suggests that this is not the case (if, on the other hand, the idiosyncratic distributions of the underlying distributions of the data were close, there would be some posterior mass for ε close to 1). The same holds for other values of the weight between 0.3 and 1. As a result of all the above, and as implied by the application of Ockham's razor, the posterior distribution for the weight will be mostly concentrated at 0.3, with another mode around 0. It is also evident that, as the number of data increases, the posterior mass

for ε at 0.3 will also increase. More details of the above and other concepts are presented in the analysis of simulated data in Chapter 3.

By embedding my proposed model for $J = 2$ in a hierarchical setting, similar to the one in Section 1.2.2 for the model of Müller et al. (2004), I get the following model:

$$Y_{ji} \sim f(Y_{ji}; \theta_{ji}, \boldsymbol{\psi}),\ i = 1, 2, \ldots, N_j,\ j = 1, 2$$

$$\theta_{ji} \sim F_j^*, \quad \text{where } F_j^* \ = \ \varepsilon F_0 \ + \ (1 - \varepsilon) F_j,\ j = 1, 2$$

$$F_0 \sim \text{DP}(M_0, H(\boldsymbol{\lambda})),\ F_1, F_2 \overset{iid}{\sim} \text{DP}(M_1, H(\boldsymbol{\lambda})) \tag{2.1.4}$$

$$\varepsilon \sim \text{Be}(M_0, M_1)$$

$$M_0, M_1 \overset{iid}{\sim} \text{Ga}(a_0, b_0),\ \boldsymbol{\psi} \sim \pi(\boldsymbol{\psi}),\ \boldsymbol{\lambda} \sim \pi(\boldsymbol{\lambda}).$$

Of course, other options for the priors of the concentration parameters are also available. Alternatively, one can also use the alternative pair of $x = M_0 + M_1$ and $y = \frac{M_0}{M_0 + M_1}$. In this setting y can be interpreted as the prior expectation of the weight ε and x as a precision parameter of it (since $\text{Var}(\varepsilon) = \frac{y(1-y)}{x+1}$). Based on the results of Theorem (2.1.1), y can also be interpreted as the prior correlation between $F_1^*(A)$ and $F_2^*(A)$ and x as a precision parameter of the prior distributions of $F_1^*(A)$ and $F_2^*(A)$. This reparametrisation is helpful when we have some prior beliefs about those two quantities, x and y, and so it is more reasonable to model them, instead of M_0 and M_1. For example, a $U(0, 1)$ prior for y and a $\text{Ga}(\lambda_1, \lambda_2)$ for x, for some hyperparameters $\lambda_1, \lambda_2 > 0$ can be used.

Properties of the proposed model

The proposed model has some very nice properties, both theoretical and computational, most of them a direct consequence of the way it was constructed. In this part the theoretical properties will be presented, whereas the computational implementation of the model is discussed in Chapter 3.
The marginal distributions of F_1^* and F_2^* are DP-distributed, because of the prior of the weight (which was inspired by the prior of the weights in the model via direct normalisation):

$$F_1^*(\cdot), F_2^*(\cdot) \overset{iid}{\sim} \text{DP}(M_0 + M_1, H). \tag{2.1.5}$$

Proof of (2.1.5):

LEt $A \in \mathcal{F}$. For $F_1^*(A)$, we have that:

$$\begin{aligned}
F_1^*(A) &= \varepsilon F_0(A) + (1-\varepsilon)F_1(A) \\
&= \frac{a}{a+b}\frac{G_0(A)}{G_0(\Omega)} + \frac{b}{a+b}\frac{G_1(A)}{G_1(\Omega)}, \text{ where } a \sim \text{Ga}(M_0,1), b \sim \text{Ga}(M_1,1) \\
&\stackrel{d}{=} \frac{G_0(A)+G_1(A)}{a+b}, \text{ since also } G_0(\Omega) \sim \text{Ga}(M_0,1), G_1(\Omega) \sim \text{Ga}(M_1,1) \\
&\stackrel{d}{=} \frac{G_0(A)+G_1(A)}{(G_0+G_1)(\Omega)}, \text{ since } (G_0+G_1)(\Omega) \stackrel{\text{d}}{=} a+b \\
&\sim \text{DP}(M_0+M_1, H(A))
\end{aligned}$$

In the above, $\stackrel{d}{=}$ denotes equality in distribution.

The same procedure can be used for $F_2^*(A)$. □

Next, using the distributions of $F_1^*, F_2^*, F_0, F_1, F_2$ and ε and the (conditional on M_0, M_1) independence of ε with the $F_j,\ j = 0,1,2$, it is straightforward to derive the following moment results:

Theorem 2.1.1. *Let Ω denote a probability space and $\mathcal{F}$ to be the $\sigma-$algebra of Ω. Let also $F_j^* = \varepsilon F_0 + (1-\varepsilon)F_j,\ j = 1,2,\ F_0 \sim DP(M_0, H),\ F_1, F_2 \stackrel{iid}{\sim} DP(M_1, H)$ and $\varepsilon \sim Be(M_0, M_1)$*
Then, $\forall\ A \in \mathcal{F}$,

$$E(F_1^*(A)) = E(F_2^*(A)) = H(A)$$

$$Var(F_1^*(A)) = Var(F_2^*(A)) = \frac{H(A)[1-H(A)]}{M_0+M_1+1}$$

$$Corr(F_1^*(A), F_2^*(A)) = \frac{M_0}{M_0+M_1}.$$

Proof:

The first two expressions are a direct result of the fact that both F_1^* and F_2^* are distributed as $\text{DP}(M_0+M_1, H)$.

For the last expression, I first obtain the covariance between the two:

$$\begin{aligned}
\text{Cov}(F_1^*(A), F_2^*(A)) &= \text{Cov}(\varepsilon F_0(A) + (1-\varepsilon)F_1(A), \varepsilon F_0(A) + (1-\varepsilon)F_2(A)) \\
&= \text{Var}(\varepsilon F_0(A)) + \text{Cov}(\varepsilon F_0(A), (1-\varepsilon)F_2(A)) \\
&\quad +\text{Cov}((1-\varepsilon)F_1(A), \varepsilon F_0(A)) + \text{Cov}((1-\varepsilon)F_1(A), (1-\varepsilon)F_2(A)) \\
&= \text{Var}(\varepsilon F_0(A)) + 2\text{Cov}(\varepsilon F_0(A), (1-\varepsilon)F_2(A)) + \text{Cov}((1-\varepsilon)F_1(A), (1-\varepsilon)F_2(A)) \\
&= \frac{M_0 H(A)(1-H(A))}{(M_0+M_1)(M_0+M_1+1)}.
\end{aligned}$$

In the above, I used the fact that F_1 and F_2 are identically distributed. For the calculations not shown, the basic first two moments of Dirichlet and beta distribution were used, as well as the independence of F_0, F_1, F_2 and ε, for example:
$\text{Var}(\varepsilon F_0(A)) = \text{E}(\varepsilon^2 F_0^2(A)) - (\text{E}(\varepsilon F_0(A)))^2 = \text{E}(\varepsilon^2)\text{E}(F_0^2(A)) - (\text{E}(\varepsilon))^2(\text{E}F_0(A))^2$.
Finally, by dividing the expression above with the product of the standard deviations of $F_1^*(A)$ and $F_2^*(A)$, we get the desired expression. □

The last expression is actually a pleasant result, as it indicates that the correlation between two realisations of F_1^* and F_2^* over the same set A do not depend on A itself. So, this expression can be thought of as "the" correlation between F_1^* and F_2^* (although there is not actually a strict definition of the correlation between two processes).

Next, the exchangeable product partition formula (EPPF), the Chinese restaurant and the Pólya-urn representations for Model (2.1.4) are derived. In order to do this, I first introduce two sets of indicators, r_{ji} and $s_{ji},\ i = 1, 2, \ldots, N_j,\ j = 1, 2$, where N_1 and N_2 are the data sizes from the two studies. The r_{ji} are binary indicators, taking values 0 and 1, depending on whether the underlying parameter θ_{ji}, associated with the $(j, i)-$th observation belongs to the common part or to the idiosyncratic part:

$$r_{ji} = \begin{cases} 0 & , \quad \text{if } \theta_{ji} \in F_0 \\ 1 & , \quad \text{if } \theta_{ji} \in F_j, \end{cases}$$

for $i = 1, 2, \ldots, N_j,\ j = 1, 2$.
Then, the indicators s_{ji} assign each observation Y_{ji} (or, equivalently, the underlying θ_{ji}) to one of the discrete clusters in each component distribution $F_j,\ j = 0, 1, 2$ (given the value of r_{ji}).

$$s_{ji} = k \ \Leftrightarrow \begin{cases} \theta_{ji} = \phi_{0k} & , \quad \text{if } r_{ji} = 0 \\ \theta_{ji} = \phi_{jk} & , \quad \text{if } r_{ji} = 1 \end{cases}$$

where $\phi_{ji},\ i = 1, 2, \ldots, K_j,\ j = 0, 1, 2$ are the discrete values in each F_j and K_j is the corresponding number of those clusters in use.
Notice that the indicators of the clusters make sense in this case, because of the discreteness of F_0, F_1 and F_2, and therefore there is a positive probability that any two observations within each of them being equal.

Proposition 2.1.1. *The EPPF for Model (2.1.4) is:*

$$p(\boldsymbol{s}, \boldsymbol{r}|\boldsymbol{M}) = \frac{\Gamma(M_0 + M_1)}{\Gamma(M_0 + M_1 + N)} M_0^{K_0} M_1^{K_1+K_2} \frac{\Gamma(M_1 + n_1 + n_2)\Gamma(M_1)}{\Gamma(M_1 + n_1)\Gamma(M_1 + n_2)} \prod_{j=0}^{2} \prod_{i=1}^{K_j} \Gamma(n_{j,i}) \tag{2.1.6}$$

where $\boldsymbol{s}$ denotes the vector of all s_{ji}, $\boldsymbol{r}$ is the vector of all r_{ji}, $\boldsymbol{M} = (M_0, M_1)$, $N = N_1 + N_2$ is the total data size, K_j is the number of clusters in component distribution j, $n_{j,i}$ is the number of data

allocated to the i-th cluster of component distribution F_j *and* $n_j = \sum_{i=1}^{K_j} n_{j,i}$ *is the number of data allocated to component* $j \in \{0, 1, 2\}$.

Proof:

The probability mass function of the indicator, given M_0 and M_1, and after having integrated out the weight is:

$$\begin{aligned}
p(\boldsymbol{s}, \boldsymbol{r}|M_0, M_1) &= \int_0^1 p(\boldsymbol{s}, \boldsymbol{r}, \varepsilon|\boldsymbol{M})d\varepsilon \\
&= \int_0^1 p(\boldsymbol{s}, \boldsymbol{r}|\varepsilon, \boldsymbol{M})f(\varepsilon|\boldsymbol{M})d\varepsilon \\
&= \int_0^1 p(\boldsymbol{s}|\varepsilon, \boldsymbol{r}, \boldsymbol{M})p(\boldsymbol{r}|\varepsilon)f(\varepsilon|\boldsymbol{M})d\varepsilon \\
&= p(\boldsymbol{s}|\boldsymbol{r}, \boldsymbol{M}) \int_0^1 p(\boldsymbol{r}|\varepsilon)f(\varepsilon|\boldsymbol{M})d\varepsilon \\
&= p(\boldsymbol{s}|\boldsymbol{r}, \boldsymbol{M}) \int_0^1 \varepsilon^{n_0}(1-\varepsilon)^{n_1+n_2} \frac{\Gamma(M_0+M_1)}{\Gamma(M_0)\Gamma(M_1)} \varepsilon^{M_0-1}(1-\varepsilon)^{M_1-1}d\varepsilon \\
&= p(\boldsymbol{s}|\boldsymbol{r}, \boldsymbol{M}) \frac{\Gamma(M_0+M_1)\Gamma(M_0+n_0)\Gamma(M_1+n_1+n_2)}{\Gamma(M_0)\Gamma(M_1)\Gamma(M_0+M_1+N)}.
\end{aligned}$$

Using the independence of s_{ji} in the three components (given the indicators r_{ji}) and applying expression (1.1.6) to each of them, the EPPF for Model (2.1.4) can be derived. □

This equation can now be used to derive the Chinese restaurant representations for this model:

Proposition 2.1.2. *Assume that we have data from both* F_1^* *and* F_2^* *in model (2.1.4), with corresponding indicators* $\mathbf{c}_{11}, \mathbf{c}_{12}, \ldots, \mathbf{c}_{1N_1} \sim F_1^*$, $\mathbf{c}_{21}, \mathbf{c}_{22}, \ldots, \mathbf{c}_{2N_2} \sim F_2^*$, *where* $\boldsymbol{c}_{ji} = (s_{ji}, r_{ji})$. *The Chinese restaurant representations will then be as follows:*

$P(s_{1,N_1+1} = K_0 + 1, r_{1,N_1+1} = 0|D, M_0, M_1) = \frac{M_0}{M_0+M_1+N}$,

$P(s_{1,N_1+1} = j, r_{1,N_1+1} = 0|D, M_0, M_1) = \frac{n_{0,j}}{M_0+M_1+N}$, $j = 1, 2, \ldots, K_0$,

$P(s_{1,N_1+1} = K_1 + 1, r_{1,N_1+1} = 1|D, M_0, M_1) = \frac{M_1(M_1+n_1+n_2)}{(M_1+n_1)(M_0+M_1+N)}$,

$P(s_{1,N_1+1} = j, r_{1,N_1+1} = 1|D, M_0, M_1) = \frac{n_{1,j}(M_1+n_1+n_2)}{(M_1+n_1)(M_0+M_1+N)}$, $j = 1, 2, \ldots, K_1$,

$P(s_{2,N_2+1} = K_0 + 1, r_{2,N_2+1} = 0|D, M_0, M_1) = \frac{M_0}{M_0+M_1+N}$,

$P(s_{2,N_2+1} = j, r_{2,N_2+1} = 0|D, M_0, M_1) = \frac{n_{0,j}}{M_0+M_1+N}$, $j = 1, 2, \ldots, K_0$,

$P(s_{2,N_2+1} = K_2 + 1, r_{2,N_2+1} = 1|D, M_0, M_1) = \frac{M_1(M_1+n_1+n_2)}{(M_1+n_2)(M_0+M_1+N)}$,

$P(s_{2,N_2+1} = j, r_{2,N_2+1} = 1|D, M_0, M_1) = \frac{n_{2,j}(M_1+n_1+n_2)}{(M_1+n_2)(M_0+M_1+N)}$, $j = 1, 2, \ldots, K_2$,

where D denotes the set of all data $(\mathbf{c}_{11}, \mathbf{c}_{12}, \ldots, \mathbf{c}_{1N_1}, \mathbf{c}_{21}, \mathbf{c}_{22}, \ldots, \mathbf{c}_{2N_2})$ and the rest are as defined in Proposition 2.1.1.

Proof:

Using the conditional probability formula, we get:

$p(\mathbf{c}\text{new}|\mathbf{c}_{11}, \mathbf{c}_{12}, \ldots, \mathbf{c}_{1N_1}, \mathbf{c}_{21}, \mathbf{c}_{22}, \ldots, \mathbf{c}_{2N_2}, M_0, M_1) = \frac{p(\mathbf{c}\text{new}, \mathbf{c}_{11}, \mathbf{c}_{12}, \ldots, \mathbf{c}_{1N_1}, \mathbf{c}_{21}, \mathbf{c}_{22}, \ldots, \mathbf{c}_{2N_2}, M_0, M_1)}{p(\mathbf{c}_{11}, \mathbf{c}_{12}, \ldots, \mathbf{c}_{1N_1}, \mathbf{c}_{21}, \mathbf{c}_{22}, \ldots, \mathbf{c}_{2N_2}, M_0, M_1)}$

$\overset{(2.1.6)}{\Rightarrow} p(\mathbf{c}\text{new}|\mathbf{c}_{11}, \mathbf{c}_{12}, \ldots, \mathbf{c}_{1N_1}, \mathbf{c}_{21}, \mathbf{c}_{22}, \ldots, \mathbf{c}_{2N_2}, M_0, M_1) = M_0^{K_0'-K_0} M_1^{K_1'+K_2'-K_1-K_2} \frac{\Gamma(M_0+M_1+N)}{\Gamma(M_0+M_1+N+1)} \times$

$$\frac{\Gamma(M_1+n_1'+n_2')\Gamma(M_1+n_1)\Gamma(M_1+n_2)}{\Gamma(M_1+n_1+n_2)\Gamma(M_1+n_1')\Gamma(M_1+n_2')} \frac{\prod_{j=0}^{2}\prod_{i=1}^{K_j'}\Gamma(n_{j,i}')}{\prod_{j=0}^{2}\prod_{i=1}^{K_j}\Gamma(n_{j,i})}.$$

where the superscript $'$ denotes the corresponding quantities when the new allocation is included. The rest follows immediately. □

Finally, the Pólya-urn representations for the same model can be now derived:

Proposition 2.1.3. *Assume that we have data $\theta_{1,1}, \theta_{1,2}, \ldots, \theta_{1,N_1}$ from F_1^* and $\theta_{2,1}, \theta_{2,2}, \ldots, \theta_{2,N_2}$ from F_2^*. The Pólya-urn representations for Model (2.1.4) will be as follows:*
$\forall\ A \in \mathcal{F}$,

$$\begin{aligned} P\left(\theta_{j,N_j+1} \in A|D\right) &= \frac{M_0(M_1+n_j)+M_1(M_1+n_1+n_2)}{(M_0+M_1+N)(M_1+n_j)} H_0(A) + \frac{1}{M_0+M_1+N}\sum_{i=1}^{K_0} n_{0i}\delta_{\theta_{0i}^*}(A) \\ &+ \frac{M_0+M_1+n_j}{(M_0+M_1+N)(M_1+n_j)}\sum_{i=1}^{K_j} n_{ji}\delta_{\theta_{ji}^*}(A),\ j=1.2. \end{aligned}$$

where D denotes the set of all data, $N = N_1 + N_2$ is the total data size, θ_{ji}^* are the discrete values (clusters) of the data in component distribution F_j, K_j is the number of these discrete values in each F_j, $n_{j,i}$ is the number of data allocated to the i-th cluster of F_j and $n_j = \sum_{i=1}^{K_j} n_{j,i}$ is the number of data allocated to F_j, $j \in \{0,1,2\}$.

Proof:

Straightforward, by adding the corresponding probabilities from Proposition 2.1.2. □

Another interesting and intriguing result arises when one uses x and y instead of M_0 and M_1 in equation (2.1.6):

$$p(\boldsymbol{s}, \boldsymbol{r}|x,y) = \frac{\Gamma(x)}{\Gamma(x+N)} x^{K_0+K_1+K_2} \prod_{j=0}^{2}\prod_{i=1}^{K_j}\Gamma(n_{j,i}) y^{K_0}(1-y)^{K_1+K_2} \frac{\Gamma(x(1-y)+n_1+n_2)\Gamma(x(1-y))}{\Gamma(x(1-y)+n_1)\Gamma(x(1-y)+n_2)}. \tag{2.1.7}$$

We can see now that the first part of (2.1.7) $\left(\frac{\Gamma(x)}{\Gamma(x+N)}x^{K_0+K_1+K_2}\prod_{j=0}^{2}\prod_{i=1}^{K_j}\Gamma(n_{j,i})\right)$, is the same as the $p(\boldsymbol{s}|x)$, if there was only one DP with precision parameter $x = M_0 + M_1$. The second part $\left(y^{K_0}(1-y)^{K_1+K_2}\right)$ is like "splitting" the discrete values from this joint DP to the common part and to the idiosyncratic parts, with corresponding probabilities $y = \frac{M_0}{M_0+M_1}$ and $1-y$. Finally, the last part is like "splitting" the data not allocated to the common part into the two idiosyncratic parts.

Comparison of my proposed model with the model of Müller et al. (2004)

The two models that will be considered and compared will be my basic proposed model (2.1.4) and the model proposed in Müller et al. (2004). Although constructed using different approaches, the two models look very similar. There are only two differences:

1. In the first model (2.1.4) there are only two concentration parameters, M_0 and M_1, whereas in the second there includes three concentration parameters, M_0, M_1 and M_2.
2. The prior of the common weight also depends on the concentration parameters in my proposed model. In the model of Müller et al. (2004) the prior of the weight only depends on some other hyperparameters.

The aforementioned differences between the two models seem to be minor. However, they result in some notable differences in their behaviour and their properties. The reason for that is exactly the way these two models were constructed. In general, one can argue that the model of Müller et al. (2004) is more flexible, since the construction of the prior distribution for ε is a more general one, and there is one extra parameter (M_2). On the other hand, the construction method used here is a more systematic one, and induces some nice properties for my model. In Model (2.1.4) these random distributions F_1^* and F_2^* are DP-distributed, whereas this is not true in the case of the other model (in general). As for the first two central moments and the correlation structure, the expressions are very simple and easy to use. The corresponding quantities for the model of Müller et al. (2004) are:

$$\mathrm{E}(F_1^*(A)) = \mathrm{E}(F_2^*(A)) = H(A)$$

$$\mathrm{Var}(F_1^*(A)) = \frac{H(A)[1-H(A)]}{(1+M_0)(1+M_1)}\left[(1+M_1)\left(\pi_1 + \frac{\pi_2 a_\varepsilon(a_\varepsilon+1)}{c_\varepsilon(c_\varepsilon+1)}\right) + (1+M_0)\left(\pi_0 + \frac{\pi_2 b_\varepsilon(b_\varepsilon+1)}{c_\varepsilon(c_\varepsilon+1)}\right)\right]$$

$$\mathrm{Var}(F_2^*(A)) = \frac{H(A)[1-H(A)]}{(1+M_0)(1+M_2)}\left[(1+M_2)\left(\pi_1 + \frac{\pi_2 a_\varepsilon(a_\varepsilon+1)}{c_\varepsilon(c_\varepsilon+1)}\right) + (1+M_0)\left(\pi_0 + \frac{\pi_2 b_\varepsilon(b_\varepsilon+1)}{c_\varepsilon(c_\varepsilon+1)}\right)\right]$$

$$\mathrm{Corr}(F_1^*(A), F_2^*(A)) = \frac{d_1\sqrt{(1+M_1)(1+M_2)}}{\sqrt{[(1+M_1)d_1 + (1+M_0)d_2][(1+M_2)d_1 + (1+M_0)d_2]}}$$

where $\pi_2 = 1-\pi_0-\pi_1,\ c_\varepsilon = a_\varepsilon + b_\varepsilon,\ d_1 = \pi_1 b_\varepsilon(b_\varepsilon + 2a_\varepsilon + 1) + (1-\pi_0)a_\varepsilon(a_\varepsilon+1)$ and $d_2 = \pi_0 a_\varepsilon(a_\varepsilon + 2b_\varepsilon + 1) + (1-\pi_1)b_\varepsilon(b_\varepsilon+1)$ Again, the correlation does not depend on A itself, but

the expressions here are more complicated.

The same holds for the Pólya-urn representations and the expression of the pdf for the indicators $\boldsymbol{s}, \boldsymbol{r}$, whose expressions are too complicated to state (although quite simple to derive).

Another nice feature of my model is the nice intuitive form of expression (2.1.7), i.e $p(\boldsymbol{s}, \boldsymbol{r}|x, y)$. On the other hand, since there are three M's in the model of Müller et al. (2004), the parameters x and y do not even have a natural interpretation.

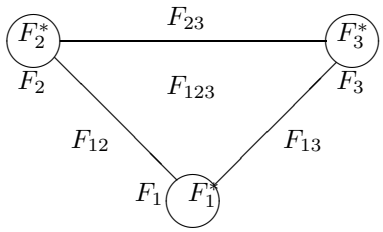
The issue of whether it makes sense to have the prior for the weight depending on the concentration parameters of the DP priors of F_0, F_1 and F_2 is discussed at the beginning of this chapter.

Finally, as will be explained in the next subsection, the basic proposed model, due to its method of construction, offers a more straightforward and systematic way of extending it to higher dimensions.

Apart from the two major models, one can consider some variations of them. The most straightforward is by assigning different weights ε_1 and ε_2 to F_1^* and F_2^*, respectively. This variation can be considered for both models. In the case of my proposed model, if this is done using the normalisation technique described above, we just get the model discussed in Section 2.1.1. On the other hand, the dependence of the weights and the dependence of the weights with F_0, F_1 and F_2 (which is a consequence of the construction method) causes some problems in the calculation of moments and other theoretical properties of the model. Another variation of the models could include introducing additional dependence through the hyperparameters of the prior of the weights, for example $\varepsilon_1, \varepsilon_2 \overset{iid}{\sim} \text{Be}(a_\varepsilon, b_\varepsilon),\ (a_\varepsilon, b_\varepsilon) \sim \pi(a_\varepsilon, b_\varepsilon)$. In such a setting, the weights are independent conditional on a_ε and b_ε, but marginally they are not independent.

2.2 Generalisations in Three Dimensions

2.2.1 General concepts



Graph 1: The basic structure of the 3-d models.

An interesting extension of the models analysed in the previous section is to cases of more than two correlated distributions. For example, consider the case of three distributions, F_1^*, F_2^* and F_3^*. One can model the correlation between them by considering a part which is shared in all three F_j^*'s, say F_{123} ("similar" to F_0 in the simpler case), and additionally, for each pair $F_i^*, F_j^*,\ i \neq j$,

assume a shared part, say F_{ji}. We can also assume an idiosyncratic part F_j for each corresponding $F_j^*,\ j = 1, 2, 3$ and, as in the 2-dimensional case, a nonparametric prior is assigned to each of the component distributions F_{123}, F_{ji}, F_j. So, each distribution will be a weighted sum of the above nonparametric distributions and dependence is introduced using the common ones, F_{123} and F_{ji}, and in some cases through also the weights. This basic structure can also be seen in Graph 1 (which is just to get the basic idea, not formally a graphical model).

Some general properties that such generalised models would preferably have are the following:

1. To provide a natural way of generalising some of the models presented in the previous section.
2. To assure first and second moments of the F_j^*'s that are not changing much when the number of components is increased (e.g. more than three component distributions).
3. **Dimensional Coherence**, meaning that, when we take a (proper) subset of our model, we would have the same model as if we had modelled it directly, using the same structure (but of course in a setting with fewer components).

A more formal definition of dimensional coherence, could be the following:
Define T_k as the class of all models for data of dimension k. By dimensional coherence we mean that, if a model for $F_1, F_2, \ldots, F_k$ belongs to T_k and C is a subset of $\{1, 2, \ldots, k\}$ of dimension $k' < k$, then the model for $\{F_i\}_{i \in C}$ belongs to $T_{k'}$.

However, in practice I noticed that for the third condition to hold, some form of interaction between the weights and the component distributions or between the weights and the precision parameters of the distributions of F's is needed *a priori*. Therefore, in some cases I decided to relax this requirement and just ask for dimensional coherence regarding the prior distribution of the weights.

In the rest of this chapter I will consider generalisations of the models to three groups and the use of Dirichlet process priors for the common and idiosyncratic parts. In other words, the models will be of the form:

$$\theta_{ji} \sim F_j^* = \sum_{k \in R_j} \varepsilon_k^{(j)} F_k,\ i = 1, 2, \ldots, N_j,\ j = 1, 2, 3$$

$$F_{123} \sim \mathrm{DP}(M_{123}, H(\boldsymbol{\lambda})), F_{12} \sim \mathrm{DP}(M_{12}, H(\boldsymbol{\lambda})), F_{13} \sim \mathrm{DP}(M_{13}, H(\boldsymbol{\lambda})), F_{23} \sim \mathrm{DP}(M_{12}, H(\boldsymbol{\lambda})),$$

$$F_1 \sim \mathrm{DP}(M_1, H(\boldsymbol{\lambda})), F_2 \sim \mathrm{DP}(M_2, H(\boldsymbol{\lambda})), F_3 \sim \mathrm{DP}(M_3, H(\boldsymbol{\lambda}))$$

$$\boldsymbol{\varepsilon}^{(j)} \sim \pi(\boldsymbol{\varepsilon}^{(j)}),\ j = 1, 2, 3 \tag{2.2.8}$$

$$\boldsymbol{M} \sim \pi(\boldsymbol{M}).$$

In the above $\boldsymbol{\varepsilon}^{(j)}$ is the vector of all weights involved in F_j^*, $\boldsymbol{M}$ is the vector of all M's and R_j denotes the subset of the powerset of $\{1, 2, 3\}$ of all sets that include the specific $j,\ j = 1, 2, 3$. As in

the previous models, the above structure can be embedded in a hierarchical setting, basically adding a likelihood function $f(Y;\theta,\boldsymbol{\psi})$ for our data and priors for $\boldsymbol{\psi}$ and $\boldsymbol{\lambda}$.

The ideas can be extended, of course, in higher dimensions, but this implies additional complexity arising from the increase in dimension, both in notation, as well as in theoretical calculations and computational burden.

As an illustration of the above, consider the indicator variables (s_{ji}, r_{ji}) in the two-dimensional models. The way they were defined was very simple and straightforward, especially for the r_{ji}, which were just binary quantities. In the case of the three-dimensional models, however, things become much more complicated, for example, the r_{ji} will now be case-specific. The easiest way of defining those indicators here is then:

$$r_{1i} = \begin{cases} 0 \text{ , if } \theta_{1i} \in F_{123} \\ 1 \text{ , if } \theta_{1i} \in F_{1} \\ 2 \text{ , if } \theta_{1i} \in F_{12} \\ 3 \text{ , if } \theta_{1i} \in F_{13} \end{cases}, \quad r_{2i} = \begin{cases} 0 \text{ , if } \theta_{2i} \in F_{123} \\ 1 \text{ , if } \theta_{2i} \in F_{2} \\ 2 \text{ , if } \theta_{2i} \in F_{12} \\ 3 \text{ , if } \theta_{2i} \in F_{23} \end{cases}, \quad r_{3i} = \begin{cases} 0 \text{ , if } \theta_{3i} \in F_{123} \\ 1 \text{ , if } \theta_{3i} \in F_{3} \\ 2 \text{ , if } \theta_{3i} \in F_{13} \\ 3 \text{ , if } \theta_{3i} \in F_{23} \end{cases} \text{ and}$$

$$s_{1i} = k \Leftrightarrow \theta_{1i} = \begin{cases} \phi_{123,k} \text{ , if } r_{1i} = 0 \\ \phi_{1,k} \text{ , if } r_{1i} = 1 \\ \phi_{12,k} \text{ , if } r_{1i} = 2 \\ \phi_{13,k} \text{ , if } r_{1i} = 3 \end{cases}, \quad s_{2i} = k \Leftrightarrow \theta_{2i} = \begin{cases} \phi_{123,k} \text{ , if } r_{2i} = 0 \\ \phi_{2,k} \text{ , if } r_{2i} = 1 \\ \phi_{12,k} \text{ , if } r_{2i} = 2 \\ \phi_{23,k} \text{ , if } r_{2i} = 3 \end{cases} \text{ and}$$

$$s_{3i} = k \Leftrightarrow \theta_{3i} = \begin{cases} \phi_{123,k} \text{ , if } r_{3i} = 0 \\ \phi_{3,k} \text{ , if } r_{3i} = 1 \\ \phi_{13,k} \text{ , if } r_{3i} = 2 \\ \phi_{23,k} \text{ , if } r_{3i} = 3 \end{cases}.$$

As before, the parameters $\phi_{ji,k}$ denote the discrete values in $F_{ji},\ i \in P_j$, where P_j is the powerset of $\{1,2,3\}\backslash\{j\},\ j = 1,2,3$.

2.2.2 The extension of my proposed model (2.1.4)

As mentioned above, the normalisation technique allows for straightforward construction of models similar to the ones presented above in three (or even more) dimensions. In the case of normalising gamma processes, Dirichlet processes are created, which can then be used as prior distributions of some correlated distributions:

Let $G_{123} \sim \Gamma\mathrm{P}(M_{123}, H),\ G_i \sim \Gamma\mathrm{P}(M_i, H),\ i = 1,2,3,\ G_{ji} \sim \Gamma\mathrm{P}(M_{ji}, H),\ i,j \in \{1,2,3\}, i \neq j$ and all of them being independent.

It is also assumed that $G_{ij} \equiv G_{ji}$ with $M_{ij} = M_{ji}$, which is a reasonable assumption in the

following setting. This assumption says that the interaction between any two distributions is the same both ways, which makes sense here, since we want to model the correlation between correlated distributions.

Also denote

$G_1^* = G_{123} + G_1 + G_{12} + G_{13}$,

$G_2^* = G_{123} + G_2 + G_{12} + G_{23}$ and

$G_3^* = G_{123} + G_3 + G_{13} + G_{23}$.

From the additivity of the ΓP, it holds that:

$G_1 \sim \Gamma\mathrm{P}(M_{123} + M_1 + M_{12} + M_{13}, H)$,

$G_2^* \sim \Gamma\mathrm{P}(M_{123} + M_2 + M_{12} + M_{23}, H)$ and

$G_3^* \sim \Gamma\mathrm{P}(M_{123} + M_3 + M_{13} + M_{23}, H)$.

By normalizing those G_j^*'s, we get:

$F_1^*(\cdot) = \frac{G_1^*(\cdot)}{G_1^*(\Omega)} \sim \mathrm{DP}(M_{123} + M_1 + M_{12} + M_{13}, H)$,
$F_2^*(\cdot) = \frac{G_2^*(\cdot)}{G_2^*(\Omega)} \sim \mathrm{DP}(M_{123} + M_2 + M_{12} + M_{23}, H)$ and
$F_3^*(\cdot) = \frac{G_3^*(\cdot)}{G_3^*(\Omega)} \sim \mathrm{DP}(M_{123} + M_3 + M_{13} + M_{23}, H)$.

On the other hand, by rewriting the F_j^*'s as weighted sums of the Dirichlet processes derived by normalising each of the ΓP-distributed G's, we get:

$$F_1^*(\cdot) = \sum_{j\in P_1} \frac{G_{1j(\cdot)}}{\sum_{k\in P_1} G_{1k(\Omega)}} = \sum_{j\in P_1} \underbrace{\frac{G_{1j(\cdot)}}{G_{1j}(\Omega)}}_{F_{1j}\sim \mathrm{DP}(M_{1j},H)} \underbrace{\frac{G_{1j}(\Omega)}{\sum_{k\in P_1} G_{1k}(\Omega)}}_{\varepsilon_{1j}^{(1)}}.$$

So,

$$F_1^* = \varepsilon_{123}^{(1)} F_{123} + \varepsilon_{12}^{(1)} F_{12} + \varepsilon_{13}^{(1)} F_{13} + (1 - \varepsilon_{123}^{(1)} - \varepsilon_{12}^{(1)} - \varepsilon_{13}^{(1)}) F_1 \tag{2.2.9}$$

where, for simplicity and visual consistency, I used the same subscripts that were used for the G's to the F's $\Big($for example, $F_{123}(\cdot) = \frac{G_{123}(\cdot)}{G_{123}(\Omega)}\Big)$.

It also holds that $(\varepsilon_{123}^{(1)}, \varepsilon_{12}^{(1)}, \varepsilon_{13}^{(1)}) \sim \mathrm{Dir}(M_{123}, M_{12}, M_{13}, M_1)$, and marginally each

$\varepsilon_{1j}^{(1)} \sim \mathrm{Be}(M_{1j}, \sum_{k\in P_1} M_{1k} - M_{1j})$ and each $F_{1j} \sim \mathrm{DP}(M_{1j}, H)$.

Note that the F's are independent here, whereas the weights are not.

Similarly, by normalising G_2^* and G_3^*, we get:

$$F_2^* = \varepsilon_{123}^{(2)} F_{123} + \varepsilon_{12}^{(2)} F_{12} + \varepsilon_{23}^{(2)} F_{23} + (1 - \varepsilon_{123}^{(2)} - \varepsilon_{12}^{(2)} - \varepsilon_{23}^{(2)}) F_2 \tag{2.2.10}$$

and

$$F_3^* = \varepsilon_{123}^{(3)} F_{123} + \varepsilon_{13}^{(3)} F_{13} + \varepsilon_{23}^{(3)} F_{23} + (1 - \varepsilon_{123}^{(3)} - \varepsilon_{13}^{(3)} - \varepsilon_{23}^{(3)}) F_3 \tag{2.2.11}$$

and similar prior distributions for the weights as for the weights in F_1^*.

The F's are the same in expressions (2.2.9)-(2.2.11) and they are mutually independent. On the

other hand, the weights in these expressions are case-specific, and that is why an extra superscript was used. They are not independent neither within nor across the three distributions F_1^*, F_2^*, F_3^*, but some of them are identically distributed (e.g. $\varepsilon_{123}^{(1)},\ \varepsilon_{123}^{(2)}$ and $\varepsilon_{123}^{(3)}$).

The distributions F_j^*'s are, of course, not independent. Dependence is expressed through the common parts F_{123} and F_{ji} and through the weights.

For $M_1 = M_2 = M_3,\ M_{12} = M_{13} = M_{23},\ F_j^* \sim \text{DP}(M_{123} + 2M_{12} + M_1, H),\ j = 1, 2, 3$, so they are identically distributed and dependent (as in the two-dimensional case). The weights are not independent in any case, but some of them are identically distributed (conditional on the M's). Also, $F_1 \overset{d}{=} F_2 \overset{d}{=} F_3$ and $F_{12} \overset{d}{=} F_{13} \overset{d}{=} F_{23}$.

More generally, the F_j^*'s are also identically distributed for any combination of M's > 0 that satisfies:

$$M_1 + M_{13} = M_2 + M_{23}$$

$$M_1 + M_{12} = M_3 + M_{23}$$

$$M_2 + M_{12} = M_3 + M_{13}.$$

The last equation above is derived from the first two, so it can be dropped.

Another special case would be to have $M_1 \overset{d}{=} M_2 \overset{d}{=} M_3$ and $M_{12} \overset{d}{=} M_{13} \overset{d}{=} M_{23}$. Marginally, $F_1 \overset{d}{=} F_2 \overset{d}{=} F_3,\ F_{12} \overset{d}{=} F_{13} \overset{d}{=} F_{23}$ and $F_1^* \overset{d}{=} F_2^* \overset{d}{=} F_3^*$. The weights are not independent, but some of them are marginally identically distributed (e.g. $\varepsilon_{12}^{(1)}$ and $\varepsilon_{13}^{(3)}$).The same results hold if we just assume that all the M's are independent and identically distributed (usually having a $\text{Ga}(a_0, b_0)$ distribution).

The notation used here is very convenient, since it keeps the same subscripts for the related F and G (and, of course, for the F_j^*'s and the normalised $G_j^*, j = 1, 2, 3$), but also associates the weights with the G_{ji} (for example, $\varepsilon_{123}^{(1)} = \frac{G_{123}(\Omega)}{G_{123}(\Omega)+G_{12}(\Omega)+G_{13}(\Omega)+G_1(\Omega)}$).

Alternative notation could extend the idea of Lavine (1992) for Pólya trees. For example, the vector of weights in F_1^* will be $\left(\varepsilon_{11}^{(1)}, \varepsilon_{10}^{(1)}, \varepsilon_{01}^{(1)}, \varepsilon_{00}^{(1)}\right)$, where the sum of those weights is 1. This notation would be convenient when splitting parts of our model to smaller ones, for example F_0 in the two-dimensional model into F_{01} and F_{02}. So, as we continue splitting the components, the new notation arises straightforwardly, especially when each component is split into exactly two parts, and in which case only zeros and ones are needed. This notation would also be convenient if one follows the opposite route and clusters some components that share some property (e.g. those created from splitting a specific component). On the other hand, this notation would not provide a direct association between the weights and the components (or the component distributions), as the notation used here does.

A far more important issue than the distributions (or the equality) of some of the concentration parameters in this structure is the distributions (or equality) of the weights. As in the 2-d case, directly using the F_j^*'s constructed using the normalisation method causes some algebraic complications, due to the dependence of the component distributions and the weights. I therefore tried some simpler models, starting with the much simpler case of assuming non case-specific weights (i.e. without superscripts) and independent of the F_i. Surprisingly, although simple, this structure appeared to be an inconvenient one. This is because Dirichlet priors for each triplet of weights is not appropriate, because of the common weights in the three triplets. A possible solution would have been to set $\varepsilon_{12} = \varepsilon_{13} = \varepsilon_{23}$ and set $(\varepsilon_{123}, 2\varepsilon_{12}) \sim \text{Dir}(\lambda_1, \lambda_2, \lambda_3)$ *a priori*. This solution, however, causes other problems, both intuitive (is the assumption that, *a priori*, an observation has equal probabilities of being assigned to two distinct component distributions sensible?) and algebraic (the factor 2 in the prior of the weights, actually, causes these problems).

As a result of the above, I adopted a slightly more complicated model: the weights were considered to be case-specific, independent of the F's, independent across the correlated distributions F_j^*'s and each triplet of them was assigned a Dirichlet prior. This prior of the weights will, of course, still depend on the concentration parameters. As will be shown, this kind of structure works very well. So, a model of this form is be the following:

$$\theta_{ji} \sim F_j^*, \quad \text{where } F_j^* \text{ are as in (2.2.9)-(2.2.11)}$$

$$F_{123} \sim \text{DP}(M_{123}, H(\boldsymbol{\lambda})), F_{12}, F_{13}, F_{23} \overset{iid}{\sim} \text{DP}(M_{12}, H(\boldsymbol{\lambda})), F_1, F_2, F_3 \overset{iid}{\sim} \text{DP}(M_1, H(\boldsymbol{\lambda})) \tag{2.2.12}$$

$$(\varepsilon_{123}^{(1)}, \varepsilon_{12}^{(1)}, \varepsilon_{13}^{(1)}), (\varepsilon_{123}^{(2)}, \varepsilon_{12}^{(2)}, \varepsilon_{23}^{(2)}), (\varepsilon_{123}^{(3)}, \varepsilon_{13}^{(3)}, \varepsilon_{23}^{(3)}) \overset{iid}{\sim} \text{Dir}(M_{123}, M_{12}, M_{12}, M_1)$$

$$M_{123}, M_{12}, M_1 \overset{iid}{\sim} \text{Ga}(a_0, b_0).$$

The above is a special case of Model (2.2.8) for $M_1 = M_2 = M_3$ and $M_{12} = M_{13} = M_{23}$, and all of them given the same $\text{Ga}(a_0, b_0)$ prior. The weights are independent across $F_j^*$and each triplet is given a Dirichlet prior distribution, with parameters the M's. Notice that the above conditions for the concentration parameters results in the same Dirichlet process distribution for the three correlated distributions:

$$F_1^*, F_2^*, F_3^* \overset{id}{\sim} \text{DP}(M_{123} + 2M_{12} + M_1, H(\boldsymbol{\lambda})).$$

Moments

Theorem 2.2.1. *Let Ω denote a probability space and $\mathcal{F}$ to be the $\sigma-$algebra of Ω. Let also F_j^*, $j = 1, 2, 3$ be distributed as in (2.2.12).*

Then, $\forall\ A \in \mathcal{F}$,

$$E(F_1^*(A)) = E(F_2^*(A)) = E(F_3^*(A)) = H(A)$$

$$Var(F_1^*(A)) = Var(F_2^*(A)) = Var(F_3^*(A)) = \frac{H(A)[1-H(A)]}{M_{123}+2M_{12}+M_1+1}$$

$$Corr(F_i^*(A), F_j^*(A)) = \frac{M_{123}+2M_{12}+M_1+1}{(M_{123}+2M_{12}+M_1)^2}\left(\frac{M_{123}^2}{1+M_{123}}+\frac{M_{12}^2}{1+M_{12}}\right),\ i \neq j.$$

Proof:

The first two results are straightforward, since all $F_j^* \sim \mathrm{DP}(M_{123}+2M_{12}+M_1, H)$

For the last result, it is enough to show that:

$$\mathrm{Cov}(F_i^*(A), F_j^*(A)) = \frac{H(A)(1-H(A))}{(M_{123}+2M_{12}+M_1)^2}\left(\frac{M_{123}^2}{1+M_{123}}+\frac{M_{12}^2}{1+M_{12}}\right),\ i \neq j.$$

For the last equation notice that, due to the independence of each triplet of weights and the independence of the weights and the F's, the only terms that will not be zero will be $\mathrm{Cov}(\varepsilon_{123}^{(i)}F_{123}, \varepsilon_{123}^{(j)}F_{123})$ and $\mathrm{Cov}(\varepsilon_{ij}^{(i)}F_{ij}, \varepsilon_{ij}^{(j)}F_{ij})$. Using again the independence between the parameters in this model, we get:

$$\begin{aligned}
\mathrm{Cov}(\varepsilon_{123}^{(i)}F_{123}, \varepsilon_{123}^{(j)}F_{123}) &= \mathrm{E}(\varepsilon_{123}^{(i)}F_{123}\varepsilon_{123}^{(j)}F_{123}) - \mathrm{E}(\varepsilon_{123}^{(i)}F_{123})\mathrm{E}(\varepsilon_{123}^{(j)}F_{123}) \\
&= \mathrm{E}(\varepsilon_{123}^{(i)})\mathrm{E}(\varepsilon_{123}^{(i)})\mathrm{E}(F_{123}^2) - \mathrm{E}(\varepsilon_{123}^{(i)})\mathrm{E}(\varepsilon_{123}^{(j)})\mathrm{E}^2(F_{123}) \\
&= \mathrm{E}(\varepsilon_{123}^{(i)})\mathrm{E}(\varepsilon_{123}^{(j)})\mathrm{Var}(F_{123}) \\
&= \frac{M_{123}^2}{(M_{123}+2M_{12}+M_1)^2}\frac{H(A)(1-H(A))}{1+M_{123}}
\end{aligned}$$

$$\begin{aligned}
\mathrm{Cov}(\varepsilon_{ij}^{(i)}F_{ij}, \varepsilon_{ij}^{(j)}F_{ij}) &= \mathrm{E}(\varepsilon_{ij}^{(i)}F_{ij}\varepsilon_{ij}^{(j)}F_{ij}) - \mathrm{E}(\varepsilon_{ij}^{(i)}F_{ij})\mathrm{E}(\varepsilon_{ij}^{(j)}F_{ij}) \\
&= \mathrm{E}(\varepsilon_{ij}^{(i)})\mathrm{E}(\varepsilon_{ij}^{(i)})\mathrm{E}(F_{ij}^2) - \mathrm{E}(\varepsilon_{ij}^{(i)})\mathrm{E}(\varepsilon_{ij}^{(j)})\mathrm{E}^2(F_{ij}) \\
&= \mathrm{E}(\varepsilon_{ij}^{(i)})\mathrm{E}(\varepsilon_{ij}^{(j)})\mathrm{Var}(F_{ij}) \\
&= \frac{M_{12}^2}{(M_{123}+2M_{12}+M_1)^2}\frac{H(A)(1-H(A))}{1+M_{12}}.
\end{aligned}$$

By adding those two, we get the desired result for the covariance, and by dividing with the square root of the product of the variances of $F_i^*(A)$ and $F_j^*(A)$, we get the above formula for the correlation. □

Note that, as in the two-dimensional model, the correlation between any pair of F_j^*'s is independent of the set A, which is indeed a very pleasant result.

Dimensional Coherence

Without loss of generality, consider F_1^* and F_2^* :

$$F_1^* = \varepsilon_{123}^{(1)}F_{123} + \varepsilon_{12}^{(1)}F_{12} + \varepsilon_{13}^{(1)}F_{13} + (1-\varepsilon_{123}^{(1)}-\varepsilon_{12}^{(1)}-\varepsilon_{13}^{(1)})F_1$$

$$F_2^* = \varepsilon_{123}^{(2)} F_{123} + \varepsilon_{12}^{(2)} F_{12} + \varepsilon_{23}^{(2)} F_{23} + (1 - \varepsilon_{123}^{(2)} - \varepsilon_{12}^{(2)} - \varepsilon_{23}^{(2)}) F_2.$$

The common parts (actually, the parts corresponding to common component distributions F_{123} and F_{12}) are
$\varepsilon_{123}^{(1)} F_{123} + \varepsilon_{12}^{(1)} F_{12}$ and $\varepsilon_{123}^{(2)} F_{123} + \varepsilon_{12}^{(2)} F_{12}$ and the two idiosyncratic ones are
$\varepsilon_{13}^{(1)} F_{13} + (1 - \varepsilon_{123}^{(1)} - \varepsilon_{12}^{(1)} - \varepsilon_{13}^{(1)}) F_1$ and $\varepsilon_{23}^{(2)} F_{23} + (1 - \varepsilon_{123}^{(2)} - \varepsilon_{12}^{(2)} - \varepsilon_{23}^{(2)}) F_2$, respectively.
Dimensional coherence here means that the common parts should be of the form $(\varepsilon_{123}^{(j)} + \varepsilon_{12}^{(j)}) F_0'$ and the other two of the form $(1 - \varepsilon_{123}^{(j)} - \varepsilon_{12}^{(j)}) F_j'$, $j = 1, 2$. We also want F_0', F_1' and F_2' to be Dirichlet process distributed. In other words, we want this "parametrisation" to be a special case of the corresponding two-dimensional model:

1. $\varepsilon_{123}^{(1)} F_{123} + \varepsilon_{12}^{(1)} F_{12} = (\varepsilon_{123}^{(1)} + \varepsilon_{12}^{(1)}) F_0'$
2. $\varepsilon_{123}^{(2)} F_{123} + \varepsilon_{12}^{(2)} F_{12} = (\varepsilon_{123}^{(2)} + \varepsilon_{12}^{(2)}) F_0'$
3. $\varepsilon_{13}^{(1)} F_{13} + (1 - \varepsilon_{123}^{(1)} - \varepsilon_{12}^{(1)} - \varepsilon_{13}^{(1)}) F_1 = (1 - \varepsilon_{123}^{(1)} - \varepsilon_{12}^{(1)}) F_1'$
4. $\varepsilon_{23}^{(2)} F_{23} + (1 - \varepsilon_{123}^{(2)} - \varepsilon_{12}^{(2)} - \varepsilon_{23}^{(2)}) F_2 = (1 - \varepsilon_{123}^{(2)} - \varepsilon_{12}^{(2)}) F_2'$.

Equivalently, we want:

1. $\frac{\varepsilon_{123}^{(1)}}{\varepsilon_{123}^{(1)} + \varepsilon_{12}^{(1)}} F_{123} + \frac{\varepsilon_{12}^{(1)}}{\varepsilon_{123}^{(1)} + \varepsilon_{12}^{(1)}} F_{12} = F_0' \sim \mathrm{DP}(M_0', H')$
2. $\frac{\varepsilon_{123}^{(2)}}{\varepsilon_{123}^{(2)} + \varepsilon_{12}^{(2)}} F_{123} + \frac{\varepsilon_{12}^{(2)}}{\varepsilon_{123}^{(2)} + \varepsilon_{12}^{(2)}} F_{12} = F_0' \sim \mathrm{DP}(M_0', H')$
3. $\frac{\varepsilon_{13}^{(1)}}{1 - \varepsilon_{123}^{(1)} - \varepsilon_{12}^{(1)}} F_{13} + \frac{1 - \varepsilon_{123}^{(1)} - \varepsilon_{12}^{(1)} - \varepsilon_{13}^{(1)}}{1 - \varepsilon_{123}^{(1)} - \varepsilon_{12}^{(1)}} F_1 = F_1' \sim \mathrm{DP}(M_1', H')$
4. $\frac{\varepsilon_{23}^{(2)}}{1 - \varepsilon_{123}^{(2)} - \varepsilon_{12}^{(2)}} F_{23} + \frac{1 - \varepsilon_{123}^{(2)} - \varepsilon_{12}^{(2)} - \varepsilon_{23}^{(2)}}{1 - \varepsilon_{123}^{(2)} - \varepsilon_{12}^{(2)}} F_2 = F_2' \sim \mathrm{DP}(M_2', H')$,

respectively.
Due to the priors of the weights, the above conditions are satisfied and $H' \equiv H, M_0' = M_{123} + M_{12}, M_1' = M_2' = M_{12} + M_1$. However, the common F_0' produced in the two first conditions is not exactly the same, because of the different weights. If the same triplet of weights was used in both F_1^* and F_2^*, then there would be no problem. However, since we want the dimensional coherence to hold for all three pairs among F_1^*, F_2^* and F_3^*, we will then have to assume the same weights for all three models, which, as mentioned above, causes other types of problems and was rejected already. This case highlights my previous comment that dimensional coherence is quite a strong assumption and trying to satisfy it can result in placing strict assumptions in the structure of the model. So, let us consider dimensional coherence of the prior of the weights:
Consider again the first two correlated distributions. Dimensional coherence of the weights means that the weights that would be the common ones and those that would be the idiosyncratic ones in

a lower-dimensional model would have a Dirichlet distribution. More exactly, a beta distribution, as now we will be in two dimensions. In other words, we want:

1. $\varepsilon_{123}^{(1)} + \varepsilon_{12}^{(1)} \sim \text{Be}(a_1, a_2)$ for $a_1, a_2 > 0$
2. $\varepsilon_{123}^{(2)} + \varepsilon_{12}^{(2)} \sim \text{Be}(a_3, a_4)$ for $a_3, a_4 > 0$.

It is easy to see that this holds in this case, because of the Dirichlet prior distribution for each triplet of weights, and $a_1 = a_3 = M_{123} + M_{12},\ a_2 = a_4 = M_{12} + M_1$.
It is also straightforward that, since there is only one weight for each component distribution, there is no need to check if the above holds for the weights of the idiosyncratic parts (as those will be just 1 - (the weights for the common part)).
Similar results are derived by considering the pairs F_1^* and F_3^* and F_2^* and F_3^*, so it can be said that dimensional coherence for the prior of the weights does hold in this model.

2.2.3 Extensions of the model of Müller *et al* (2004)

Similar to the model suggested in Müller et al. (2004), which is constructed in a general, but perhaps less systematic way, its extensions can also be constructed in many ways. In this subsection two of them will be considered, one having case-specific weights (which can be therefore compared with the model developed in the previous section) and a model with weights specific to the number of components included, i.e. ε_{123} is the same for all three F_j^*'s, and $\varepsilon_{12} = \varepsilon_{13} = \varepsilon_{23}$ (although this model will have the problems discussed in the previous subsections - I present it here for completeness). In both cases, zero probability is assigned to some weights being equal to 0 or 1, merely for algebraic and computational simplicity, as including those two probabilities increases the size of derived expressions considerably.

First extension

As in the previous model, it is assumed that the weights are case-specific, independent of the F's, and each triplet of weights is independent and follows a Dirichlet prior distribution:

$$F_1^* = \varepsilon_{123}^{(1)} F_{123} + \varepsilon_{12}^{(1)} F_{12} + \varepsilon_{13}^{(1)} F_{13} + (1 - \varepsilon_{123}^{(1)} - \varepsilon_{12}^{(1)} - \varepsilon_{13}^{(1)}) F_1 \tag{2.2.13}$$

$$F_2^* = \varepsilon_{123}^{(2)} F_{123} + \varepsilon_{12}^{(2)} F_{12} + \varepsilon_{23}^{(2)} F_{23} + (1 - \varepsilon_{123}^{(2)} - \varepsilon_{12}^{(2)} - \varepsilon_{23}^{(2)}) F_2 \tag{2.2.14}$$

$$F_3^* = \varepsilon_{123}^{(3)} F_{123} + \varepsilon_{13}^{(3)} F_{13} + \varepsilon_{23}^{(3)} F_{23} + (1 - \varepsilon_{123}^{(3)} - \varepsilon_{13}^{(3)} - \varepsilon_{23}^{(3)}) F_3 \tag{2.2.15}$$

$$\left(\varepsilon_{123}^{(1)}, \varepsilon_{12}^{(1)}, \varepsilon_{13}^{(1)}\right) \sim \text{Dir}(\alpha_{1,11}, \alpha_{1,10}, \alpha_{1,01}, \alpha_{1,00}) \tag{2.2.16}$$

$$\left(\varepsilon_{123}^{(2)}, \varepsilon_{12}^{(2)}, \varepsilon_{23}^{(2)}\right) \sim \text{Dir}(\alpha_{2,11}, \alpha_{2,10}, \alpha_{2,01}, \alpha_{2,00}) \tag{2.2.17}$$

$$\left(\varepsilon_{123}^{(3)}, \varepsilon_{13}^{(3)}, \varepsilon_{23}^{(3)}\right) \sim \text{Dir}(\alpha_{3,11}, \alpha_{3,10}, \alpha_{3,01}, \alpha_{3,00}). \tag{2.2.18}$$

A fully hierarchical model would then be the following:

$$\theta_{ji} \sim F_j^*, \text{ where } F_j^* \text{ is as in (2.2.13)-(2.2.15)}$$

$$F_{123} \sim \text{DP}(M_{123}, H(\boldsymbol{\lambda})), F_{12}, F_{13}, F_{23} \overset{iid}{\sim} \text{DP}(M_{12}, H(\boldsymbol{\lambda})), F_1, F_2, F_3 \overset{iid}{\sim} \text{DP}(M_1, H(\boldsymbol{\lambda})) \tag{2.2.19}$$

$$\left(\varepsilon_{123}^{(1)}, \varepsilon_{12}^{(1)}, \varepsilon_{13}^{(1)}\right), \left(\varepsilon_{123}^{(2)}, \varepsilon_{12}^{(2)}, \varepsilon_{23}^{(2)}\right), \left(\varepsilon_{123}^{(3)}, \varepsilon_{13}^{(3)}, \varepsilon_{23}^{(3)}\right) \text{ are as in (2.2.16)-(2.2.18) and independent}$$

$$M_{123}, M_{12}, M_1 \overset{iid}{\sim} \text{Ga}(a_0, b_0).$$

Model (2.2.19) is again a special case of Model (2.2.8) for $M_1 = M_2 = M_3$ and $M_{12} = M_{13} = M_{23}$, and all of them given the same $\text{Ga}(a_0, b_0)$ prior. The difference of this model from model (2.2.12) is the parameters in the prior distributions of the weights. In this model we have a more general allocation for these parameters $(\alpha_{j,11}, \alpha_{j,10}, \alpha_{j,01}, \alpha_{j,00},\ j = 1, 2, 3)$, whereas before those parameters were the concentration parameters of the priors of the F's. None the less, by having the same concentration parameters for all F_j (M_1) and the same for all F_{ji} (M_{12}), it is guaranteed that all the $F_j^*,\ j = 1, 2, 3$ are identically distributed (marginally, since each triplet of weights has the same prior). However, it cannot be guaranteed that they are also DP-distributed, a result that would have provided some nice properties.

Moments

Theorem 2.2.2. *Let Ω denote a probability space and $\mathcal{F}$ to be the $\sigma-$algebra of Ω. Let also $F_j^*,\ j = 1, 2, 3$ be distributed as in (2.2.19).*

Then, $\forall\ A \in \mathcal{F}$,

$$E(F_1^*(A)) = E(F_2^*(A)) = E(F_3^*(A)) = H(A)$$

$$Var(F_j^*(A)) = \frac{H(A)[1 - H(A)]}{\alpha_j^*(\alpha_j^* + 1)} \left[\frac{\alpha_{j,11}(\alpha_{j,11} + 1)}{1 + M_{123}} + \frac{\alpha_{j,10}(\alpha_{j,10} + 1) + \alpha_{j,01}(\alpha_{j,01} + 1)}{1 + M_{12}} + \frac{\alpha_{j,00}(\alpha_{j,00} + 1)}{1 + M_1}\right], \quad j = 1, 2, 3$$

$$Corr(F_1^*(A), F_2^*(A)) = \sqrt{\frac{\alpha_1^* + 1}{\alpha_1^*} \cdot \frac{\alpha_2^* + 1}{\alpha_2^*}} \frac{1}{\sqrt{d_1 d_2}} \left[\frac{\alpha_{1,11}\alpha_{2,11}}{1 + M_{123}} + \frac{\alpha_{1,10}\alpha_{2,10}}{1 + M_{12}}\right]$$

$$Corr(F_1^*(A), F_3^*(A)) = \sqrt{\frac{\alpha_1^* + 1}{\alpha_1^*} \cdot \frac{\alpha_3^* + 1}{\alpha_3^*}} \frac{1}{\sqrt{d_1 d_3}} \left[\frac{\alpha_{1,11}\alpha_{3,11}}{1 + M_{123}} + \frac{\alpha_{1,01}\alpha_{3,10}}{1 + M_{12}}\right]$$

$$Corr(F_2^*(A), F_3^*(A)) = \sqrt{\frac{\alpha_2^* + 1}{\alpha_2^*} \cdot \frac{\alpha_3^* + 1}{\alpha_3^*}} \frac{1}{\sqrt{d_2 d_3}} \left[\frac{\alpha_{2,11}\alpha_{3,11}}{1 + M_{123}} + \frac{\alpha_{2,01}\alpha_{3,01}}{1 + M_{12}}\right]$$

where $\alpha_j^* = \alpha_{j,11} + \alpha_{j,10} + \alpha_{j,01} + \alpha_{j,00}$ and

$d_j = \frac{\alpha_{j,11}(\alpha_{j,11}+1)}{1+M_{123}} + \frac{\alpha_{j,10}(\alpha_{j,10}+1)+\alpha_{j,01}(\alpha_{j,01}+1)}{1+M_{12}} + \frac{\alpha_{j,00}(\alpha_{j,00}+1)}{1+M_1},\ j = 1, 2, 3.$

Notice also that again, the correlation structure between $F_i^*, F_j^*,\ i \neq j$ is independent of the set A chosen.

Proof:

I first derive the above moments conditional on the weights and the precision parameters:

$$\mathrm{E}(F_1^*(A)|\boldsymbol{\varepsilon}) = \mathrm{E}(F_2^*(A)|\boldsymbol{\varepsilon}) = \mathrm{E}(F_3^*(A)|\boldsymbol{\varepsilon}) = H(A) \tag{2.2.20}$$

$$\mathrm{Var}(F_j^*(A)|\boldsymbol{\varepsilon}) = H(A)[1 - H(A)]c_j,\ j = 1, 2, 3 \tag{2.2.21}$$

$$\mathrm{Corr}(F_i^*(A), F_j^*(A)|\boldsymbol{\varepsilon}) = \frac{1}{\sqrt{c_i c_j}} \left[\frac{\varepsilon_{123}^{(i)} \varepsilon_{123}^{(j)}}{1 + M_{123}} + \frac{\varepsilon_{ij}^{(i)} \varepsilon_{ij}^{(j)}}{1 + M_{12}} \right],\ i, j = 1, 2, 3,\ i \neq j \tag{2.2.22}$$

where $\boldsymbol{\varepsilon}$ is the vector of all weights in the model and

$$c_1 = \frac{\varepsilon_{123}^{(1)2}}{1 + M_{123}} + \frac{\varepsilon_{12}^{(1)2} + \varepsilon_{13}^{(1)2}}{1 + M_{12}} + \frac{(1 - \varepsilon_{123}^{(1)} - \varepsilon_{12}^{(1)} - \varepsilon_{13}^{(1)})^2}{1 + M_1},$$

$$c_2 = \frac{\varepsilon_{123}^{(2)2}}{1 + M_{123}} + \frac{\varepsilon_{12}^{(2)2} + \varepsilon_{13}^{(2)2}}{1 + M_{12}} + \frac{(1 - \varepsilon_{123}^{(2)} - \varepsilon_{12}^{(2)} - \varepsilon_{13}^{(2)})^2}{1 + M_1} \text{ and}$$

$$c_3 = \frac{\varepsilon_{123}^{(3)2}}{1 + M_{123}} + \frac{\varepsilon_{13}^{(3)2} + \varepsilon_{23}^{(3)2}}{1 + M_{12}} + \frac{(1 - \varepsilon_{123}^{(3)} - \varepsilon_{13}^{(3)} - \varepsilon_{23}^{(3)})^2}{1 + M_1}.$$

Given the weights, the above calculations are straightforward, using the independence of the weights across the distributions F_j^*'s, the independence of the component distributions F's and the simple forms for the first two moments for the F's, since they are DP-distributed.

Next, expressions (2.2.20)- (2.2.22) (which are conditional on the weights) are used, together with some known theoretical results, in order to derive the formulae in Theorem 2.2.2:

It is well known that

$$\mathrm{E}(X) = \mathrm{E}(\mathrm{E}(X|Y)). \tag{2.2.23}$$

Applying it to (2.2.20), we have that

$$\begin{aligned} \mathrm{E}(F_1^*(A)) &= \mathrm{E}_{\boldsymbol{\varepsilon}_1}(\mathrm{E}(F_1^*(A)|\boldsymbol{\varepsilon}_1)),\ \text{where } \boldsymbol{\varepsilon}_1 = (\varepsilon_{1,11}, \varepsilon_{1,10}, \varepsilon_{1,01}) \\ &= \mathrm{E}_{\boldsymbol{\varepsilon}_1}(H(A)) \\ &= H(A),\ \text{since the expression } H(A) \text{ does not include the weights.} \end{aligned}$$

The same calculation can be done for the expectation of F_2^* and F_3^*.

For the variances, we can use the identity

$$\mathrm{Var}(X) = \mathrm{Var}(\mathrm{E}(X|Y)) + \mathrm{E}(\mathrm{Var}(X|Y)). \tag{2.2.24}$$

For example, for $j = 1$, we have:

$$\begin{aligned}
\mathrm{Var}(F_1^*(A)) &= \mathrm{Var}_{\boldsymbol{\varepsilon}_1}(\mathrm{E}(F_1^*(A)|\boldsymbol{\varepsilon}_1)) + \mathrm{E}_{\boldsymbol{\varepsilon}_1}(\mathrm{Var}(F_1^*(A)|\boldsymbol{\varepsilon}_1)) \\
&= \mathrm{E}_{\boldsymbol{\varepsilon}_1}(\mathrm{Var}(F_1^*(A)|\boldsymbol{\varepsilon}_1)), \quad \text{since } \mathrm{E}(F_1^*(A)|\boldsymbol{\varepsilon}_1) = H(A), \text{ not involving the weights} \\
&\overset{(2.2.21)}{=} \mathrm{E}_{\boldsymbol{\varepsilon}_1}\left(H(A)(1-H(A))(\frac{\varepsilon_{123}^{(1)2}}{1+M_{123}} + \frac{\varepsilon_{12}^{(1)2}+\varepsilon_{13}^{(1)2}}{1+M_{12}} + \frac{(1-\varepsilon_{123}^{(1)}-\varepsilon_{12}^{(1)}-\varepsilon_{13}^{(1)})^2}{1+M_1})\right) \\
&= H(A)(1-H(A))\left(\frac{\mathrm{E}(\varepsilon_{123}^{(1)2})}{1+M_{123}} + \frac{\mathrm{E}(\varepsilon_{12}^{(1)2})+\mathrm{E}(\varepsilon_{13}^{(1)2})}{1+M_{12}} + \frac{\mathrm{E}(1-\varepsilon_{123}^{(1)}-\varepsilon_{12}^{(1)}-\varepsilon_{13}^{(1)})^2}{1+M_1}\right) \\
&= \frac{H(A)[1-H(A)]}{\alpha_1^*(\alpha_1^*+1)}\left[\frac{\alpha_{1,11}(\alpha_{1,11}+1)}{1+M_{123}} + \frac{\alpha_{1,10}(\alpha_{1,10}+1)+\alpha_{1,01}(\alpha_{1,01}+1)}{1+M_{12}} + \frac{\alpha_{1,00}(\alpha_{1,00}+1)}{1+M_1}\right].
\end{aligned}$$

For the variances of $F_1^*(A)$ and $F_2^*(A)$ the same procedure can be followed.

Finally, for the covariances we can use the formula

$$\mathrm{Cov}(F_i^*(A), F_j^*(A)) = \mathrm{E}\left(\mathrm{Cov}(F_i^*(A), F_j^*(A)|\boldsymbol{\varepsilon})\right), \ i \neq j. \tag{2.2.25}$$

Proof of (2.2.25):

$$\begin{aligned}
\mathrm{Cov}(F_i^*(A), F_j^*(A)) &= \mathrm{E}(F_i^*(A) \cdot F_j^*(A)) - \mathrm{E}(F_i^*(A))\mathrm{E}(F_j^*(A)) \\
&= \mathrm{E}[\mathrm{E}(F_i^*(A) \cdot F_j^*(A)|\boldsymbol{\varepsilon})] - H^2(A) \\
&= \mathrm{E}(\mathrm{Cov}(F_i^*(A), F_j^*(A)|\boldsymbol{\varepsilon})) + \mathrm{E}\left(\mathrm{E}(F_i^*(A)|\boldsymbol{\varepsilon})\mathrm{E}(F_j^*(A)|\boldsymbol{\varepsilon})\right) - H^2(A) \\
&= \mathrm{E}(\mathrm{Cov}(F_i^*(A), F_j^*(A)|\boldsymbol{\varepsilon})) + H^2(A) - H^2(A) \\
&= \mathrm{E}(\mathrm{Cov}(F_i^*(A), F_j^*(A)|\boldsymbol{\varepsilon})). \qquad \square
\end{aligned}$$

So, using (2.2.20)-(2.2.22) and the priors of the weights, we get the stated result for the correlations. $\square$

Dimensional Coherence

The notion and method of study of dimensional coherence will be the same as before. Unfortunately, the issue that the produced common part from two different cases will not be exactly the same is also present here. This is due to the fact that, again, the weights are case-specific. So, again, dimensional coherence regarding the prior distribution of the weights is studied. For this model, we want to check whether sums like $\varepsilon_{123}^{(1)} + \varepsilon_{12}^{(1)}$ (or $\varepsilon_{123}^{(1)} + \varepsilon_{13}^{(1)}$) have also a Dirichlet distribution (which, in two dimensions, is the beta distribution). This is trivial, though, due to the divisibility of the Dirichlet distribution. In fact, the method to show this is the same as in the previous model, with the parameters $\alpha_{j,11}$ replacing M_{123}, $\alpha_{j,10}$ and $\alpha_{j,01}$ replacing M_{12} and $\alpha_{j,00}$ replacing M_1.

So, we can say that in this model, as in the previous model, dimensional coherence regarding the priors of the weights holds.

Second extension

In this extented model it is assumed that weights that are involved with the same number of components are the same in all three distributions. For example, three components are involved in each F_{123}, since F_{123} is the part that expresses the common part of all three correlated distributions, so it is assumed that the weight corresponding to F_{123} is the same in all F_j^*. As a result, there wiill only be three distinct weights in this model, one corresponding to F_{123}, the second corresponding to $F_{ij},\ i,j=1,2,3,\ i\neq j$ and the last corresponding to $F_j,\ j=1,2,3$. Again, it is assumed that the weights are independent of the component distributions (the F's), and a Dirichlet prior is assigned to them:

$$F_1^* = \varepsilon_{123}F_{123} + \varepsilon_{12}F_{12} + \varepsilon_{12}F_{13} + (1-\varepsilon_{123}-2\varepsilon_{12})F_1 \tag{2.2.26}$$

$$F_2^* = \varepsilon_{123}F_{123} + \varepsilon_{12}F_{12} + \varepsilon_{12}F_{23} + (1-\varepsilon_{123}-2\varepsilon_{12})F_2 \tag{2.2.27}$$

$$F_3^* = \varepsilon_{123}F_{123} + \varepsilon_{12}F_{13} + \varepsilon_{12}F_{23} + (1-\varepsilon_{123}-2\varepsilon_{12})F_3 \tag{2.2.28}$$

$$(\varepsilon_{123}, 2\varepsilon_{12}) \sim \text{Dir}(\alpha_{123}, \alpha_{12}, \alpha_1).$$

The weight corresponding to the idiosyncratic parts, F_j, can be found by $\varepsilon_1 = 1-\varepsilon_{123}-2\varepsilon_{12}$. Note also that, the factor 2 of the second weight in the prior distribution of the weights is added in order for the weights to sum to one.

Only three concentration parameters are used, M_{123}, M_{12} and M_1. In this case, this assures that the F_j^*'s are identically distributed. Unfortunately, in general they are not DP-distributed.

The basic problem of this model, however, is the counter-intuitive convention that the probability of an observation belonging to any of the two common component distributions (excluding the overall common F_{123}) is the same. This also causes the factor 2 in the prior of the weights, which causes many complications in the algebraic calculations. However, I decided to present this model for completeness purposes.

The hierarchical model constructed in the previous cases now becomes:

$$\theta_{ji} \sim F_j^*, \text{ where } F_j^* \text{ is as in (2.2.26)-(2.2.28)}$$

$$F_{123} \sim \text{DP}(M_{123}, H(\boldsymbol{\lambda})), F_{12}, F_{13}, F_{23} \overset{iid}{\sim} \text{DP}(M_{12}, H(\boldsymbol{\lambda})), F_1, F_2, F_3 \overset{iid}{\sim} \text{DP}(M_1, H(\boldsymbol{\lambda})) \tag{2.2.29}$$

$$(\varepsilon_{123}, 2\varepsilon_{12}) \sim \text{Dir}(\alpha_{123}, \alpha_{12}, \alpha_1)$$

$$M_{123}, M_{12}, M_1 \overset{iid}{\sim} \text{Ga}(a_0, b_0).$$

This model is a special case of Model (2.2.8) for (again) $M_1 = M_2 = M_3$ and $M_{12} = M_{13} = M_{23}$, and all of them given the same $\text{Ga}(a_0, b_0)$ prior distribution. The weights used are only two and they are the same across F_j^*, with the specific Dirichlet prior for them.

Moments

Theorem 2.2.3. *Let Ω denote a probability space and $\mathcal{F}$ to be the $\sigma-$algebra of Ω. Let also $F_j^*,\ j = 1,2,3$ be distributed as in (2.2.29).*

Then, $\forall\ A \in \mathcal{F}$,

$$E(F_1^*(A)) = E(F_2^*(A)) = E(F_3^*(A)) = H(A)$$

$$Var(F_j^*(A)) = \frac{H(A)[1-H(A)]}{\alpha^*(\alpha^*+1)}\left[\frac{\alpha_{123}(\alpha_{123}+1)}{1+M_{123}} + \frac{\alpha_{12}(\alpha_{12}+1)}{2(1+M_{12})} + \frac{\alpha_1(\alpha_1+1)}{1+M_1}\right],\ j = 1,2,3$$

$$Corr(F_i^*(A), F_j^*(A)) = \frac{\left[\frac{\alpha_{123}(\alpha_{123}+1)}{1+M_{123}} + \frac{\alpha_{12}(\alpha_{12}+1)}{4(1+M_{12})}\right]}{\left[\frac{\alpha_{123}(\alpha_{123}+1)}{1+M_{123}} + \frac{\alpha_{12}(\alpha_{12}+1)}{2(1+M_{12})} + \frac{\alpha_1(\alpha_1+1)}{1+M_1}\right]},\ i \neq\ j$$

where $\alpha^* = \alpha_{123} + \alpha_{12} + \alpha_1$.

Once again, the correlation between $F_i^*(A)$ and $F_j^*(A),\ i \neq j$ is independent of the set A chosen. Notice also that all F_j^* have the same expectation and variance and each pair of them has the same correlation. This is something trivial, since they are identically distributed, and with the same covariance structure among them!

Proof:

The proof is similar to the first extension. We first derive the moments conditional on the weights:

$$\mathrm{E}(F_1^*(A)) = \mathrm{E}(F_2^*(A)) = \mathrm{E}(F_3^*(A)) = H(A)$$

$$\mathrm{Var}(F_j^*(A)) = H(A)[1-H(A)]B,\ j = 1,2,3$$

$$\mathrm{Corr}(F_i^*(A), F_j^*(A)) = \frac{1}{B}\left[\frac{\varepsilon_{123}^2}{1+M_{123}} + \frac{\varepsilon_{12}^2}{1+M_{12}}\right],\ i \neq\ j$$

where $B = \frac{\varepsilon_{123}^2}{1+M_{123}} + \frac{2\varepsilon_{12}^2}{1+M_{12}} + \frac{(1-\varepsilon_{123}-2\varepsilon_{12})^2}{1+M_1}$.

The above formulae are very easy to show, using the independence of the F's and since the weights are considered fixed.

The relationships (2.2.23)- (2.2.25) can then be used (the first two formulae hold in general, whereas (2.2.25) can be proven to hold in this case, too) and the stated results follow immediately. □

Dimensional Coherence

Since the weight for all the idiosyncratic parts is common, and since this is true for the parts between any two component distributions (F_{12}, F_{13}, F_{23}), in this case there is no problem of identifiability of the common part between any two of $F_j^*,\ j = 1,2,3$. For example, the common part between F_1^* and F_2^* will be $(\varepsilon_{123})F_{123} + \varepsilon_{12}F_{12} = (\varepsilon_{123} + \varepsilon_{12})F_0$, for some F_0. For dimensional coherence, it

should hold that $F_0 = \frac{\varepsilon_{123}}{\varepsilon_{123}+\varepsilon_{12}} F_{123} + \frac{\varepsilon_{12}}{\varepsilon_{123}+\varepsilon_{12}} F_{12}$ is DP-distributed. Given the DP prior of the $F's$, this holds if and only if

$$\frac{\varepsilon_{123}}{\varepsilon_{123}+\varepsilon_{12}} \sim \text{Be}(M_{123}, M_{12}).$$

Unfortunately, the distribution of $\frac{\varepsilon_{123}}{\varepsilon_{123}+\varepsilon_{12}}$, cannot be found in closed form, and it is (of course) not a beta distribution. The reason for this is the factor 2 in the prior of the weights. As a result, dimensional coherence cannot be established. What's more, the same reason prevents us from deriving the distribution of, say, $\varepsilon_{123} + \varepsilon_{12}$, so dimensional coherence of the prior distribution of the weights does not hold, either.

Finally, let us just mention that the Pólya-urn representations for all three extended models presented here are easily derived. This is done using the corresponding Chinese restaurant representations, which are also easy to derive.

2.3 Summary

In this chapter a general class of correlated distributions that can be naturally applied to modelling grouped data was introduced. A model of this form can be constructed by normalising dependent random measures. This was demonstrated using the gamma process as distribution of the underlying random measure, and the derived dependent distributions are DP-distributed. Apart from this model, a similar but simpler model is considered, and both these models are compared with the model introduced in Müller et al. (2004). The two new models and the model of Müller et al. (2004), although constructed using different methods, were also very similar. The proposed models are intuitively appealing, since they are constructed in a systematic way, and also have good theoretical properties (for example, the expressions for the first two moments are very simple ones). Notice also that, due to the way they are constructed, the proposed models (especially the one constructed using the normalisation method) can be naturally generalised in higher dimensions (meaning higher number of dependent random distributions). Such an extension, together with two generalisations of the model of Müller et al. (2004) are considered in the case of three dependent distributions. Some theoretical properties of the extended models are considered, with particular attention to dimensional coherence. The last condition is not easy to satisfy, and none of the three models did satisfy it. However, I was able to establish dimensional coherence of the prior distribution of the weights in the extension of my model and in one of the extensions of the Müller et al. (2004) model.

Chapter 3

Computational Implementation

In this section I deal with the computational implementation of the models presented in the previous section, and especially Model (2.1.4) and the model arising from direct normalisation. I first discuss the implementation of my basic model (2.1.4) and suggest an additional step that can be used in most of the algorithms, in which we propose splitting or merging some clusters in the components. Some techniques presented here can also be used in the simulation of the posterior distributions of the parameters in the model proposed in Müller et al. (2004). Implementation for the three models in three dimensions presented are also discussed, as well as for a model similar to my basic proposed model, but with N-IGP priors, instead of DPs. Finally, in the case of the model via direct normalisation, a data augmentation scheme will allow us to perform slice sampling for simulating from the full conditional distributions of some parameters in a natural way.

3.1 General Concepts

In order to demonstrate the computational implementation of these models, I will assume that the likelihood $f(Y_{ji};\theta_{ji},S)$ is a normal distribution with mean μ_{ji} and variance S and that $\mu_{ji} \sim F_j^*$, where F_j^* are as defined in each model. The base distribution H is also a normal distribution, say $N(m,B)$. The mean m is assigned a normal prior with parameters m_0 and A, the variance B is assigned an inverse gamma distribution with shape parameter c and scale parameter $(cC)^{-1}$, $\text{IGa}(c,(cC)^{-1})$. Finally, the variance S also has an inverse gamma distribution with parameters q and $(qR)^{-1}$. Notice that the priors used for $S, m,$ and B are the same as the ones used in Müller et al. (2004).

As in the model of Müller et al. (2004), the main parameters of interest will be the weight

(or weights) of the common and idiosyncratic parts, the concentration parameters of the Dirichlet processes and the predictive distributions for the correlated distributions $F_j^*,\ j = 1, 2$, as well as those of the component distributions $F_j,\ j = 0, 1, 2$.

As mentioned in the introduction, as in almost all the cases of Bayesian nonparametric models, in order to simulate from the posterior distribution of the parameters of interest, we use MCMC methods, with mostly Gibbs and Metropolis-Hastings sampling steps. We also use the marginal approach for simulating from Dirichlet processes, described in Section 1.1.2. The conjugacy of the likelihood and the base distribution means that the standard samplers for conjugate DPs can be used. The alternative approach, would be the conditional method, also described in Section 1.1.2. Additionally, in order to increase the efficiency of the MCMC algorithm, in each cycle of the algorithm the clusters of discrete values of the μ_{ji} are also updated, as suggested in MacEachern (1998).

3.2 The Proposed Algorithm for Model (2.1.4)

My basic model (2.1.4), together with the distributions mentioned above, becomes:

$$Y_{ji} \sim N(\mu_{ji}, S),\ i = 1, 2, \ldots, N_j,\ j = 1, 2$$

$$\mu_{ji} \sim F_j^*,\ \text{where } F_j^* = \varepsilon F_0 + (1 - \varepsilon) F_j$$

$$F_0 \sim DP(M_0, H),\ F_j \overset{iid}{\sim} DP(M_1, H),\ \text{where } H \equiv N(m, B)$$

$$\varepsilon \sim Be(M_0, M_1)$$

$$M_0, M_1 \overset{iid}{\sim} Ga(a_0, b_0),\ (m, B) \sim N(m_0, A) \times \text{IGa}(c, 1/cC),\ S \sim \text{IGa}(q, 1/qR).$$

Let $\phi_{ji},\ i = 1, 2, \ldots, K_j,\ j = 0, 1, 2$ denote the discrete values of the μ_{ji} in each component distribution $F_j,\ j = 0, 1, 2$. As in Müller et al. (2004), the auxiliary indicator variables $s_{ji},$ and r_{ji} are used:

$$r_{ji} = \begin{cases} 0 \text{ , if } \mu_{ji} \in F_0 \\ 1 \text{ , if } \mu_{ji} \in F_j, \end{cases} \qquad s_{ji} = k \ \Leftrightarrow \begin{cases} \mu_{ji} = \phi_{0k} \text{ , if } r_{ji} = 0 \\ \mu_{ji} = \phi_{jk} \text{ , if } r_{ji} = 1. \end{cases} , \ i = 1, 2, \ldots, N_j,\ j = 1, 2.$$

Instead of using the parameters μ_{ji} and ϕ_{ji}, we can therefore use the equivalent parametrisation $s_{ji}, r_{ji}, \phi_{ji}$.

The full set of parameters in this model is $(\mathbf{s}, \mathbf{r}, \boldsymbol{\phi}, \varepsilon, M_0, M_1, m, B, S)$, where the bold symbols denote the vector of all indicated parameters (e.g. $\mathbf{s}$ denotes all $s_{ji},\ i = 1, 2, \ldots, N_j,\ j = 1, 2$). Let also $\mathbf{Y}$ denote the vector of observations $(Y_{11}, Y_{12}, \ldots, Y_{1,N_1}, Y_{21}, Y_{22}, \ldots, Y_{2,N_2})$, n_0, n_1 and n_2 are the number of observations assigned in each component distribution F_0, F_1 and F_2, respectively

and $n_{ji},\ i = 1, 2, \ldots, K_j,\ j = 0, 1, 2$ are the number of data allocated to cluster i in component distribution $F_j,\ j = 0, 1, 2$. Then,

$$\begin{aligned} f(\mathbf{s}, \mathbf{r}, \boldsymbol{\phi}, \varepsilon, m, B, S, M_0, M_1 | \mathbf{Y}) &\propto \prod_{j,i} f(Y_{ji} | r_{ji}, s_{ji}, \boldsymbol{\phi}, S) f(m) f(B) f(M_0, M_1) f(S) f(\varepsilon | M_0, M_1) \\ &\times \prod_{j,i} f(\phi_{ji} | m, B) \prod_{j,i} f(r_{ji} | \varepsilon) f(\mathbf{s} | \mathbf{r}, M_0, M_1). \end{aligned}$$

The full conditional distribution of each parameter (i.e. the distribution given all the other parameters) is as follows:

- $m | \cdots \sim N(\frac{m_0 B + A \sum_{j,i} \phi_{ji}}{AK+B}, \frac{AB}{AK+B})$,
  where $K = K_0 + K_1 + K_2$ is the total number of discrete values in all components.

- $B | \cdots \sim \text{IGa}(c + K/2, 1/cC + 1/2 \sum_{j,i} (\phi_{ji} - m)^2)$.

- $S | \cdots \sim \text{IGa}(q + N/2, 1/qR + 1/2 \sum_{j,i} (Y_{ji} - \mu_{ji})^2)$, where $N = N_1 + N_2$.

- $\varepsilon | \cdots \sim Be(M_0 + N - \sum_{j,i} r_{ji}, M_1 + \sum_{j,i} r_{ji})$.

- $f(M_0, M_1 | \cdots) \propto M_0^{a_0 + K_0 - 1} e^{-M_0 [b_0 - \log(\varepsilon)]} M_1^{a_0 + K_1 + K_2 - 1} e^{-M_1 [b_0 - \log(1-\varepsilon)]} \frac{\Gamma(M_1)\Gamma(M_1 + M_0)}{\Gamma(M_1 + n_0)\Gamma(M_1 + n_1)\Gamma(M_1 + n_2)}$.
  So,
  $f(M_0 | \cdots) \propto M_0^{a_0 + K_0 - 1} e^{-M_0 [b_0 - \log(\varepsilon)]} \frac{\Gamma(M_0 + M_1)}{\Gamma(M_0 + n_0)}$ and
  $f(M_1 | \cdots) \propto M_1^{a_0 + K_1 + K_2 - 1} e^{-M_1 [b_0 - \log(1-\varepsilon)]} \frac{\Gamma(M_1)\Gamma(M_0 + M_1)}{\Gamma(M_1 + n_1)\Gamma(M_1 + n_2)}$.

- $\phi_{0l} | \cdots \sim N(\frac{mS + B \sum_{j,i: r_{ji} = 0, s_{ji} = l} Y_{ji}}{S + B n_{0l}}, \frac{SB}{S + B n_{0l}}),\ l = 1, 2, \ldots, K_0$ and
  $\phi_{jl} | \cdots \sim N(\frac{mS + B \sum_{i: r_{ji} = 1, s_{ji} = l} Y_{ji}}{S + B n_{jl}}, \frac{SB}{S + B n_{jl}}),\ l = 1, 2, \cdots, K_j,\ j = 1, 2.$

- $P(s_{ji} = h, r_{ji} = l | \cdots) = \begin{cases} \pi_{jh}\ , h = 1, 2, \ldots, K_j,\ l = 1 \\ \pi_{0h}\ , h = 1, 2, \ldots, K_0,\ l = 0 \\ \pi_j^*\ , h = K_j + 1,\ l = 1 \\ \pi_0^*\ , h = K_0 + 1,\ l = 0 \end{cases}, \ i = 1, 2, \ldots, N_j, j = 1, 2.$

  where $\pi_{jh} \propto (1 - \varepsilon) \varphi(Y_{ji}; \phi_{jh}, S) n_{jh}^- / (M_1 + n_j^-),\ \pi_{0h} \propto \varepsilon \varphi(Y_{ji}; \phi_{0h}, S) n_{0h}^- / (M_0 + n_0^-)$,
  $\pi_j^* \propto (1 - \varepsilon) \varphi(Y_{ji}; m, S + B) M_1 / (M_1 + n_j^-)$, and $\pi_0^* \propto \varepsilon \varphi(Y_{ji}; m, S + B) M_0 / (M_0 + n_0^-)$, where the superscript $-$ means that the corresponding quantity are taken without counting the quantity associated with the (ji) point, φ is the pdf of the normal distribution and the above probabilities are all proportional to the same constant, which is such that the probabilities sum up to 1. Finally, note that in the last two cases for $(s_{ji}, r_{ji} | \cdots)$, a new value should be created. This is a draw from $N(\frac{BY_{ji} + mS}{B+S}, \frac{BS}{B+S})$.

We can directly simulate from all the above full conditional distributions, except from the ones of the precision parameters M_0 and M_1, for which RWMH steps can be used, either to each one or together. More specifically, since those quantities are defined on the positive real line, the proposals of the RWMH are applied to their logarithms.
Finally, the predictive distributions for the two data sets are as follows:

$$\begin{aligned} p(Y_{j,N_j+1}|\mathbf{Y}) &= \varepsilon\frac{M_0}{M_0+n_0}N(m,B+S)+\varepsilon\frac{1}{M_0+n_0}\sum_{d=1}^{K_0}n_{0d}N(\phi_{0d},S) \\ &+ (1-\varepsilon)\frac{M_1}{M_1+n_j}N(m,B+S)+(1-\varepsilon)\frac{1}{M_1+n_j}\sum_{d=1}^{K_j}n_{jd}N(\phi_{jd},S),\ j=1,2. \end{aligned}$$

One full cycle of the MCMC algorithm consists of updating each of those parameters from their full conditional distribution. Note that most of the steps will be the same as in the algorithm in Müller et al. (2004). The differences will be in the full conditional distributions of ε, M_0 and M_1 and the fact that M_2 needs to be replaced by M_1 in all the other full conditionals, since $M_2 = M_1$ here.

All the full conditionals are of known form, apart from the ones for M_0 and M_1. Whereas Gibbs sampling steps can be used for the rest, MH steps can be used for the last two:
I first tried to use random walk MH steps (more precisely, applied to the logarithms of them):
For M_0, we propose $\log(M_0') = \log(M_0) + \zeta_0 \ \Leftrightarrow M_0' = M_0 e^{\zeta_0}$, where $\zeta_0 \sim N(0,\sigma^2)$ and accept M_0' with probability

$$\alpha(M_0, M_0') = \min\left\{1, \left(\frac{M_0'}{M_0}\right)^{a+K_0} e^{(M_0-M_0')(b-\log(\varepsilon))}\frac{\Gamma(M_1+M_0')\Gamma(M_0+n_0)}{\Gamma(M_1+M_0)\Gamma(M_0'+n_0)}\right\}.$$

Otherwise, keep M_0.
Similarly, propose $\log(M_1') = \log(M_1) + \zeta_1 \ \Leftrightarrow M_1' = M_1 e^{\zeta_1}$, $\zeta_1 \sim N(0,\sigma^2)$ and accept it with probability

$$\alpha(M_1, M_1') = \min\left\{1, \left(\frac{M_1'}{M_1}\right)^{a+K_1+K_2} e^{(M_1-M_1')(b-\log(1-\varepsilon))}\frac{\Gamma(M_1')\Gamma(M_0+M_1')\Gamma(M_1+n_1)\Gamma(M_1+n_2)}{\Gamma(M_1)\Gamma(M_0+M_1)\Gamma(M_1'+n_1)\Gamma(M_1'+n_2)}\right\}.$$

Otherwise, keep M_1.
However, in practice I encountered a situation where the value of either M_0 or M_1 got stuck at zero, since then all the proposed values will again be zero. This is a problem of the accuracy of the program used, since in theory those two quantities are always positive. A possible solution to this is to truncate the values for M_0 and M_1 produced at a very low level. In particular, when a value for any of those two parameters produced was less than 10^{-8}, I was setting it to be 10^{-8}. In practise, this solution worked well.
If we alternatively work with $x = M_0 + M_1$ and $y = \frac{M_0}{M_0+M_1}$, the corresponding full conditional

distributions will be:

$$f(y|\cdots) \propto y^{K_0}(1-y)^{K_1+K_2}\left(\frac{\varepsilon}{1-\varepsilon}\right)^{xy}\frac{\Gamma(x(1-y))}{\Gamma(xy+n_0)\Gamma(x(1-y)+n_1)\Gamma(x(1-y)+n_2)}$$

$$f(x|\cdots) \propto x^{K_0+K_1+K_2+d_1-1}e^{-d_2x}[\varepsilon^y(1-\varepsilon)^{1-y}]^x\frac{\Gamma(x)\Gamma(x(1-y))}{\Gamma(xy+n_0)\Gamma(x(1-y)+n_1)\Gamma(x(1-y)+n_2)}.$$

The two full conditionals are not of known form, so random walk Metropolis-Hastings steps are used:

Propose $\text{logit}(y') = \text{logit}(y) + \zeta_1 \Leftrightarrow y' = \frac{ye^{\zeta_1}}{1-y+ye^{\zeta_1}}$, $\zeta_1 \sim \text{N}(0,\sigma_1^2)$ (i.e. random walk on the logit of y, where $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$, $0<p<1$) and accept it with probability

$$\alpha(y,y') = \min\left\{1, \left(\frac{y'}{y}\right)^{K_0+1}\left(\frac{1-y'}{1-y}\right)^{K_1+K_2+1}\left(\frac{\varepsilon}{1-\varepsilon}\right)^{x(y'-y)}\frac{\Gamma(x(1-y'))\Gamma(xy+n_0)\Gamma(x(1-y)+n_1)\Gamma(x(1-y)+n_2)}{\Gamma(x(1-y))\Gamma(xy'+n_0)\Gamma(x(1-y')+n_1)\Gamma(x(1-y')+n_2)}\right\}.$$

Propose $\log(x') = \log(x) + \zeta_2 \Leftrightarrow x' = xe^{\zeta_2}$, $\zeta_2 \sim \text{N}(0,\sigma_2^2)$, (i.e. random walk on the logarithm of x) and accept it with probability

$$\alpha(x,x') = \min\left\{1, \left(\frac{x'}{x}\right)^{K_0+K_1+K_2+d_1}e^{d_2(x-x')}[\varepsilon^y(1-\varepsilon)^{1-y}]^{x'-x}\frac{\Gamma(x')\Gamma(x'(1-y))\Gamma(xy+n_0)\Gamma(x(1-y)+n_1)\Gamma(x(1-y)+n_2)}{\Gamma(x)\Gamma(x(1-y))\Gamma(x'y+n_0)\Gamma(x'(1-y)+n_1)\Gamma(x'(1-y)+n_2)}\right\}.$$

Example 1 (continued):

In order to better understand the differences between my proposed model (2.1.4) and the model of Müller et al. (2004), and also get a better insight of the proposed MCMC algorithms, I applied them to the data mentioned in Section 2.1.2, and assuming normal likelihood and normal base distribution (i.e. as in Model (1.2.14)). The kernel density estimates for the posterior distribution of ε were first plotted, together with the trace plots of the MCMC output for this parameters in Figures 3.1 (Müller et al. (2004) model) and 3.2 (Model (2.1.4)).

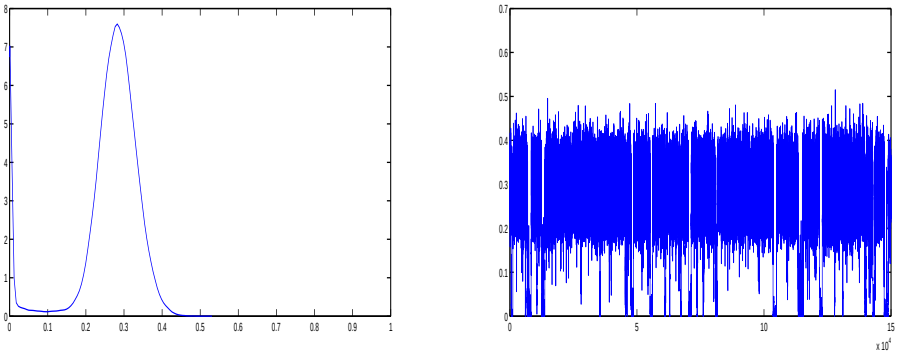


Figure 3.1: Kernel density estimate (left) and trace plot (right) for the posterior of ε for the model of Müller et al. (2004) for the first simulated data set.

From the two kernel density plots, it is clear that the posterior distribution of ε is bimodal at 0 and a value very close to 0.3, as expected from the discussion in Section 2.1.3. The mass at 0 is about 9.9% for Model (1.2.13) and about 74.1% for Model (2.1.4). Notice that the kernel density estimate for my model should have had a significantly larger mode close to 0, but this mode was

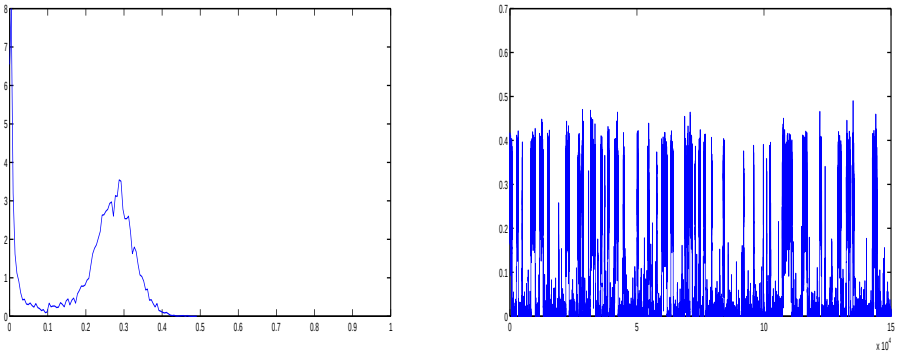


Figure 3.2: Kernel density estimate (left) and trace plot (right) for the posterior of ε for Model (2.1.4) for the first simulated data set.

reduced here, for better inspection of the mode at 0.3. From the corresponding trace plots, it can be seen that the chain jumps between those two modes of the weight, and perhaps not so often as it would have been ideal to (seen especially in the case of my proposed model).

The trace plots indicate a possible mixing problem in the algorithms used. It would be therefore desirable to try and improve the mixing of the chain by increasing the frequency of the jumps between the two modes of ε. In order to do so, an additional step in the algorithm for Model (2.1.4) is proposed (which could again be used in most models of this type, for example for the model of Müller et al. (2004)), a novel mix-split step. In this step we propose to either split a cluster from the common part F_0 to two clusters, one in each idiosyncratic part F_1 and F_2, or to merge two clusters, one from each of F_1 and F_2 to a common cluster in F_0.

3.3 The Mix-Split Step

The basic form of this extra step consists of first choosing whether we will propose a mix or a split move (with probability 1/2 each) and then perform it. If a split step is chosen, we uniformly choose a cluster from F_0 and propose to split it into two clusters, one in F_1 and one in F_2 (or move it to either F_1 or F_2, if this cluster contains only data from the first or second data set, respectively). If a merge step is chosen, we uniformly choose a cluster from F_1, or an empty cluster, and a cluster from F_2, or an empty cluster, and we propose to merge those two clusters (or move a cluster, if in one of the two cases an empty cluster is chosen) to a common cluster in F_0.

This split-merge step is a Metropolis-Hastings update, so the acceptance probability in each case needs to be calculated. These probabilities will depend on the method of mix-split selected (the basic one discussed thoroughly below or the alternative one, introduced after the description of the

basic algorithm), on whether a split or a merge step is selected and on the existing and proposed allocation of the indicator parameters $s_{ji}, r_{ji},\ i = 1, 2, \ldots, N_j,\ j = 1, 2$.

In the following, let K_0, K_1 and K_2 denote the number of clusters in components F_0, F_1 and F_2, respectively, m_{01} and m_{02} denote the number of data from each data set associated with a chosen cluster in F_0 in a split step and let m_1, m_2 be the number of data from each data set associated with the chosen clusters in F_1, F_2 respectively in a merge step. Let also n_1 and n_2 denote the current (i.e. before the proposed mix or split step) number of data assigned in each idiosyncratic component distribution, F_1 and F_2, respectively.

The algorithm for the (basic) mix-split step and the corresponding acceptance probabilities $\alpha(\mathbf{c}, \mathbf{c}')$, where $\mathbf{c} = (\mathbf{r}, \mathbf{s})$ is the complete vector of indicators are as follows:

Basic Mix-Split Method:

1. Choose split or merge, each w.p. 1/2.
2. If a split step is selected:
   (a) If $K_0 = 0$, we do nothing (we exit the split/merge step), since there is no cluster to split (or move to either F_0 or F_1).
   (b) Else, we choose a cluster from the common part (F_0) uniformly. We then propose to:
   - move this cluster to one of the two idiosyncratic parts ($F_1,\ F_2$), if the data associated with the chosen cluster come only from the first or the second data set, respectively.
   - split this cluster to two clusters, one in each of the idiosyncratic parts, if the related data come from both data sets. In such a case, the data from the first group will be moved to the new cluster in F_1 and the data from the second group will be moved to the new cluster in F_2.

   Next, we calculate the acceptance probability (according to the case above that applies) and accept the split with this probability. If the step is accepted, we transfer the data from the first data set associated to the selected cluster (if any) to a (new) cluster to the first idiosyncratic part and the data from the second data set allocated to the cluster to be split (again, if any) to a (new) cluster to the second idiosyncratic part. Analytically,we have:
   i. If we propose to move a cluster from F_0 to F_1, say the cluster corresponding to the d-th discrete value in $F_0,\ \phi_{0d}$, the acceptance probability will be:
   $\alpha(\mathbf{c}, \mathbf{c}') = \min\left\{1, \frac{M_1}{M_0}\frac{\Gamma(M_1+n_1+n_2+m_{01})\Gamma(M_1+n_2)}{\Gamma(M_1+n_1+n_2)\Gamma(M_1+n_2+m_{01})}\frac{K_0}{(K_1+2)(K_2+1)-1}\right\}$.
   If the step is approved, we change the indicators of the data associated with the

removed cluster as follows:

If $r_{1i} = 0,\ s_{1i} = d,$ we set $r_{1i} = 1$ and $s_{1i} = K_1 + 1$ (i.e. we create a new cluster in F_1).

Additionally, since we destroy the selected cluster from F_0, we must adjust the s_{ji} of the rest of the observations from both data sets allocated in F_0. What we actually do is to reduce each s_{ji} by 1, if $s_{ji} > d$ and $r_{ji} = 0$.

We also transfer the value of the selected cluster to the new cluster created, K_0 is reduced by 1 and K_1 is increased by 1.

ii. If we propose to move a cluster from F_0 to F_2, i.e. when $m_{01} = 0$, similar things as above apply. The acceptance probability will now be:
Case (2bii) : $\alpha(\mathbf{c}, \mathbf{c}') = \min\left\{1, \frac{M_1}{M_0}\frac{\Gamma(M_1+n_1+n_2+m_{02})\Gamma(M_1+n_2)}{\Gamma(M_1+n_1+n_2)\Gamma(M_1+n_2+m_{02})}\frac{K_0}{(K_1+1)(K_2+2)-1}\right\}$.

iii. If we propose to split a cluster to both F_1 and F_2, say the cluster corresponding to the d-th discrete value in $F_0,\ \phi_{0d}$, the acceptance probability will be:
$\alpha(\mathbf{c}, \mathbf{c}') = \min\left\{1, \frac{M_0}{M_1^2}\frac{\Gamma(M_1+n_1+n_2-m_{01}-m_{02})\Gamma(M_1+n_1)\Gamma(M_1+n_2)}{\Gamma(M_1+n_1+n_2)\Gamma(M_1+n_1-m_{01})\Gamma(M_1+n_2-m_{02})}\frac{\Gamma(m_{01}+m_{02})}{\Gamma(n_1)\Gamma(m_{02})}\times\right.$
$\sqrt{\frac{(m_{01}B+S)(m_{02}B+S)}{S[(m_{01}+m_{02})B+S]}}\frac{K_0}{(K_1+2)(K_2+2)-1}\exp\{-\frac{1}{2}[\frac{(m_{01}+m_{02})m^2-2m(\sum Y_1'+\sum Y_2')-\frac{B}{S}(\sum Y_1'+\sum Y_2')^2}{(m_{01}+m_{02})B+S}$
$\left.-\frac{m_{01}m^2-2m\sum Y_1'-\frac{B}{S}(\sum Y_1')^2}{m_{01}B+S} - \frac{m_{02}m^2-2m\sum Y_2'-\frac{B}{S}(\sum Y_2')^2}{m_{02}B+S}]\}\right\}$.

If the split is accepted, we change the indicators as follows:

If $r_{1i} = 0,\ s_{1i} = d,$ we set $r_{1i} = 1$ and $s_{1i} = K_1 + 1$ (i.e. we create a new cluster in F_1).

If $r_{2i} = 0,\ s_{2i} = d,$ we set $r_{2i} = 1$ and $s_{2i} = K_2 + 1$ (i.e. we create a new cluster in F_2).

We also reduce the s_{ji} that are larger than d (and with corresponding $r_{ji} = 0$) by 1.

We set the new clusters equal to the value of the split cluster, K_0 is reduced by 1, and both K_1 and K_2 are increased by 1.

Otherwise, we do nothing.

3. If a merge step is selected:

   (a) If $K_1 = K_2 = 0$, we exit, since there are no clusters to merge (or move to F_0).

   (b) Otherwise, if only $K_1 = 0$, we propose to move a cluster from the second idiosyncratic part to the common one. In other words, we propose merging a cluster from F_2 with an empty cluster from F_1.

   In this case, we uniformly choose a cluster from the second idiosyncratic part (corresponding to the discrete, say, ϕ_{2d}) and move it to the common part. We accept this step

with probability:
$\alpha(\mathbf{c},\mathbf{c}') = \min\left\{1, \frac{M_0}{M_1}\frac{(K_1+1)(K_2+1)-1}{K_0+1}\right\}$.
If we accept the step, we do the following:
If $r_{2i}=1,\ s_{2i}=d$, we set $r_{2i}=0,\ s_{2i}=K_0+1$ (i.e. we create a new cluster in F_0).
Additionally, if $r_{2i}=1,\ s_{2i}>d$, we reduce s_{2i} by 1.
We set the new cluster in F_0 equal to ϕ_{2d}, reduce K_2 by 1 and increase K_0 by 1.
If the step is rejected, we do nothing.

(c) Similarly, if only $K_2=0$ (with also the same acceptance probability).

(d) If both K_1 and K_2 are positive, we uniformly choose a cluster from F_1 or an empty cluster (in which case we just move a cluster from F_2 to F_0), i.e each cluster (and the empty cluster) is chosen with probability $\frac{1}{K_1+1}$. We similarly choose a cluster from F_2 or an empty cluster. The possibility of merging an empty cluster from F_1 or F_2 is needed in order to have reversibility of the MCMC. We must also note that, if two empty clusters are chosen, we repeat the above draw, since this merging is prohibited (again in order to have a reversible MCMC algorithm).

i. If we (only) choose an empty cluster from F_1, we propose to transfer the selected cluster from F_2 to F_0 and accept it with probability:
$\alpha(\mathbf{c},\mathbf{c}') = \min\left\{1, \frac{M_0}{M_1}\frac{\Gamma(M_1+n_1+n_2-m_2)\Gamma(M_1+n_2)}{\Gamma(M_1+n_1+n_2)\Gamma(M_1+n_2-m_2)}\frac{(K_1+1)(K_2+1)-1}{K_0+1}\right\}$.
If the step is accepted, we perform the transfer as in (b), otherwise we exit.

ii. Similarly as in (c), if an empty cluster from F_2 is chosen. In this case the acceptance probability will be:

$\alpha(\mathbf{c},\mathbf{c}') = \min\left\{1, \frac{M_0}{M_1}\frac{\Gamma(M_1+n_1+n_2-m_1)\Gamma(M_1+n_1)}{\Gamma(M_1+n_1+n_2)\Gamma(M_1+n_1-m_1)}\frac{(K_1+1)(K_2+1)-1}{K_0+1}\right\}$.

iii. If two existing clusters are chosen, corresponding to, say, (ϕ_{1d},ϕ_{2b}) we propose merging the two clusters in a common cluster in F_0. The acceptance probability is:
$\alpha(\mathbf{c},\mathbf{c}') = \min\left\{1, \frac{M_0}{M_1^2}\frac{\Gamma(M_1+n_1+n_2-m_1-m_2)\Gamma(M_1+n_1)\Gamma(M_1+n_2)}{\Gamma(M_1+n_1+n_2)\Gamma(M_1+n_1-m_1)\Gamma(M_1+n_2-m_2)}\frac{\Gamma(m_1+m_2)}{\Gamma(n_1)\Gamma(m_2)}\times\right.$
$\sqrt{\frac{(m_1B+S)(m_2B+S)}{S[(m_1+m_2)B+S]}}\frac{(K_1+1)(K_2+1)-1}{K_0+1}\exp\{-\frac{1}{2}[\frac{(m_1+m_2)m^2-2m(\sum Y_1'+\sum Y_2')-\frac{B}{S}(\sum Y_1'+\sum Y_2')^2}{(m_1+m_2)B+S}$
$\left.-\frac{m_1m^2-2m\sum Y_1'-\frac{B}{S}(\sum Y_1')^2}{m_1B+S}-\frac{m_2m^2-2m\sum Y_2'-\frac{B}{S}(\sum Y_2')^2}{m_2B+S}]\}\right\}$,
and if the step is accepted, we transfer the data associated with them to one common (new) cluster in the common part, as follows:
If $r_{1i}=1,\ s_{2i}=d$, we set $r_{2i}=0,\ s_{2i}=K_0+1$.
If $r_{2i}=1,\ s_{2i}=b$, we set $r_{2i}=0,\ s_{2i}=K_0+1$.
If $r_{1i}=1,\ s_{1i}>d$, we decrease s_{1i} by 1.

If $r_{2i} = 1,\ s_{2i} > b$, we decrease s_{2i} by 1.

We set the value of the created cluster equal to either of the two selected clusters, decrease K_1 and K_2 by 1 and increase K_0 by 1.

Otherwise, we do nothing.

The reason for including empty clusters when randomly picking clusters in the merge step is to guarantee the reversibility of the MCM Chain. This is true because the act of merging an existing cluster from, say F_1, with an empty cluster (i.e. moving a cluster from F_1 to F_0) is the reverse of moving a cluster from F_0 to F_1, which will happen if we propose to split a cluster in F_0 that is associated only with data from F_1^*.

Notice also that in this algorithm we might do nothing at some steps. For example if we choose to propose a split step, but we do not have any clusters in the common component. A possible variation of this method could be to first examine the configuration of the clusters in F_0, F_1 and F_2 and then propose a mix or a split step, based on this configuration. For example, if there are no clusters in F_0, then with probability 1 we propose a mix step. If all components are non-empty, we choose randomly split/merge, each w.p. 1/2. The algorithm in this case will follow the same basic structure and updating schemes when a split or a merge step is accepted, and it is therefore very similar to the algorithm above.

The details of the derivation of the acceptance probabilities for the basic algorithm are presented in Section A1 of the Appendix. This subsection in the Appendix also discusses the differences in the acceptance probabilities if the alternative mix-split method is applied (i.e. when the number of clusters in the components are examined before proposing a split or a merge step). Finally, notice also that these probabilities are calculated after integrating out the weight ε and the discrete values of the clusters ϕ_{ji} (in order to improve the efficiency of this step), so those quantities must be updated just after this split/merge step.

3.3.1 Mix-split step for the model of Müller et al. (2004):

As mentioned before, the above step can be also used in the case of simulating from the posterior distributions of the parameters of the model of Müller et al. (2004). The method will be the same, as will also be the transition probabilities (since they only depend on the method of proposing mix and split steps). The only difference will be the ratio of probabilities of the indicators with and without the proposed step, $\frac{f(\boldsymbol{c}')}{f(\boldsymbol{c})}$:

Following the same procedure as before, it can be seen that

$$f(\boldsymbol{c}|\cdots) \propto \int f(\boldsymbol{c}, \varepsilon|\boldsymbol{M})d\varepsilon \int f(\boldsymbol{Y}|\boldsymbol{\phi}, \boldsymbol{c}, S)f(\boldsymbol{\phi}|m, B)d\boldsymbol{\phi}.$$

The second integral $\int f(\boldsymbol{Y}|\boldsymbol{\phi},\boldsymbol{c},S)f(\boldsymbol{\phi}|m,B)d\boldsymbol{\phi}$ will be the same as before, whereas the first integral can be easily seen to be equal to $f(\boldsymbol{s}|\boldsymbol{r},\boldsymbol{M})\int_0^1 f(\boldsymbol{r}|\varepsilon)f(\varepsilon)d\varepsilon$.

It is the part that is different from the previous model, due to the different prior of the weight and the fact that there are three precision parameters here. In this case, this integral becomes:

$$\begin{aligned}
\int f(\boldsymbol{c},\varepsilon|\boldsymbol{M})d\varepsilon &= f(\boldsymbol{s}|\boldsymbol{r},\boldsymbol{M})\int_0^1 f(\boldsymbol{r}|\varepsilon)f(\varepsilon)d\varepsilon \\
&= \prod_{i:r_{1i}=1} f(s_{1i}|M_1)\prod_{i:r_{2i}=1} f(s_{2i}|M_2)\prod_{j,i:r_{ji}=0} f(s_{ji}|M_0)\int_0^1 f(\boldsymbol{r}|\varepsilon)f(\varepsilon)d\varepsilon \\
&= \cdots \\
&= M_0^{K_0}M_1^{K_1}M_2^{K_2}\frac{\Gamma(M_0)}{\Gamma(M_0+n_0)}\frac{\Gamma(M_1)}{\Gamma(M_1+n_1)}\frac{\Gamma(M_2)}{\Gamma(M_2+n_2)}\prod_{i=1}^{K_0}\Gamma(n_{0,i})\prod_{i=1}^{K_1}\Gamma(n_{1,i})\prod_{i=1}^{K_2}\Gamma(n_{2,i}) \\
&\quad\times\left[\pi_0\delta_{\mathbf{1}}(\boldsymbol{r})+\pi_1\delta_{\mathbf{0}}(\boldsymbol{r})+(1-\pi_0-\pi_1)\frac{B(a_\varepsilon+N-\sum r_{ji},b_\varepsilon+\sum r_{ji})}{B(a_\varepsilon,b_\varepsilon)}\right]
\end{aligned}$$

where the terminology used is the same as in Section 1.2.2. The ratios $\frac{f(\boldsymbol{c}')}{f(\boldsymbol{c})}$ are now easily calculated, using the above equation and the results of the previous model about the integral with respect to the ϕ's:

Split proposal:

1. If $m_{01}=0$, i.e. we move a cluster from F_0 to F_2:
   $\frac{f(\boldsymbol{c}_{split})}{f(\boldsymbol{c})}=\frac{M_2}{M_0}\frac{\Gamma(M_2+n_2)\Gamma(M_0+n_0)}{\Gamma(M_2+n_2+m_{02})\Gamma(M_0+n_0-m_{02})}\frac{g(\boldsymbol{r}_{split})}{g(\boldsymbol{r})}$.
2. If $m_{02}=0$, i.e. we move a cluster from F_0 to F_1:
   $\frac{f(\boldsymbol{c}_{split})}{f(\boldsymbol{c})}=\frac{M_1}{M_0}\frac{\Gamma(M_1+n_1)\Gamma(M_0+n_0)}{\Gamma(M_1+n_1+m_{01})\Gamma(M_0+n_0-m_{01})}\frac{g(\boldsymbol{r}_{split})}{g(\boldsymbol{r})}$.
3. If $m_{01}>0,$ and $m_{02}>0$, i.e. we split a cluster from F_0 to both F_1 and F_2:
   $\frac{f(\boldsymbol{c}_{split})}{f(\boldsymbol{c})}=\frac{M_1M_2}{M_0}\frac{\Gamma(M_0+n_0)\Gamma(M_1+n_1)\Gamma(M_2+n_2)}{\Gamma(M_0+n_0-m_{01}-m_{02})\Gamma(M_1+n_1+m_{01})\Gamma(M_2+n_2+m_{02})}\frac{\Gamma(m_{01})\Gamma(m_{02})}{\Gamma(m_{01}+m_{02})}\frac{g(\boldsymbol{r}_{split})}{g(\boldsymbol{r})}\sqrt{\frac{S[(m_{01}+m_{02})B+S]}{(m_{01}B+S)(m_{02}B+S)}}\times$

   $\exp\{\frac{1}{2}[\frac{(m_{01}+m_{02})m^2-2m(\sum Y_1'+\sum Y_2')-\frac{B}{S}(\sum Y_1'+\sum Y_2')^2}{(m_{01}+m_{02})B+S}-\frac{m_{01}m^2-2m\sum Y_1'-\frac{B}{S}(\sum Y_1')^2}{m_{01}B+S}-\frac{m_{02}m^2-2m\sum Y_2'-\frac{B}{S}(\sum Y_2')^2}{m_{02}B+S}]\}$.

Merge proposal:

1. If we move a cluster from F_2 to F_0 (in this case, $m_1=0$):
   $\frac{f(\boldsymbol{c}_{merge})}{f(\boldsymbol{c})}=\frac{M_0}{M_2}\frac{\Gamma(M_0+n_0)\Gamma(M_2+n_2)}{\Gamma(M_0+n_0+m_2)\Gamma(M_2+n_2-m_2)}\frac{g(\boldsymbol{r}_{merge})}{g(\boldsymbol{r})}$.
2. If we move a cluster from F_1 to F_0 ($m_2=0$):
   $\frac{f(\boldsymbol{c}_{merge})}{f(\boldsymbol{c})}=\frac{M_0}{M_1}\frac{\Gamma(M_0+n_0)\Gamma(M_1+n_1)}{\Gamma(M_0+n_0+m_1)\Gamma(M_1+n_1-m_1)}\frac{g(\boldsymbol{r}_{merge})}{g(\boldsymbol{r})}$.
3. If we merge a cluster from F_2 and a cluster from F_1 to F_0 :
   $\frac{f(\boldsymbol{c}_{merge})}{f(\boldsymbol{c})}=\frac{M_0}{M_1M_2}\frac{\Gamma(M_0+n_0)\Gamma(M_1+n_1)\Gamma(M_2+n_2)}{\Gamma(M_0+n_0+m_1+m_2)\Gamma(M_1+n_1-m_1)\Gamma(M_2+n_2-m_2)}\frac{\Gamma(m_1+m_2)}{\Gamma(m_1)\Gamma(m_2)}\frac{g(\boldsymbol{r}_{merge})}{g(\boldsymbol{r})}\sqrt{\frac{(m_1B+S)(m_2B+S)}{S[(m_1+m_2)B+S]}}\times$

   $\exp\{-\frac{1}{2}[\frac{(m_1+m_2)m^2-2m(\sum Y_1'+\sum Y_2')-\frac{B}{S}(\sum Y_1'+\sum Y_2')^2}{(m_1+m_2)B+S}-\frac{m_1m^2-2m\sum Y_1'-\frac{B}{S}(\sum Y_1')^2}{m_1B+S}-\frac{m_2m^2-2m\sum Y_2'-\frac{B}{S}(\sum Y_2')^2}{m_2B+S}]\}$.

In all the above m_{01}, m_{02}, m_1 and m_2 are the same as in the case of my model. I also introduced the notation $g(\boldsymbol{r}) = \left[\pi_0\delta_{\mathbf{1}}(\boldsymbol{r}) + \pi_1\delta_{\mathbf{0}}(\boldsymbol{r}) + (1-\pi_0-\pi_1)\frac{B(a_\varepsilon+n_0, b_\varepsilon+n_1+n_2)}{B(a_\varepsilon, b_\varepsilon)}\right]$ in order to avoid too long expressions here.

Finally, the acceptance probabilities are, in each case, the minimum of 1 and of the number resulting from multiplying the ratio calculated above with the ratio of the appropriate transition probabilities q.

3.4 Simulated Data

In this section I present some of the algorithms discussed above for fitting Model (2.1.4) and the model of Müller et al. (2004) applied to three simulated data sets.

3.4.1 The data

Example 1 (continued):

The first data set is the one used before in order to illustrate some properties of the models and the corresponding algorithms:

$$p(Y_{1i}) = \frac{7}{10}N(1,1) + \frac{3}{10}N(-10,1),\ i = 1,2,\ldots,100$$

$$p(Y_{2i}) = \frac{3}{10}N(1,1) + \frac{7}{10}N(8,1),\ i = 1,2,\ldots,100.$$

Example 2:

The second data set was taken from the following distributions:

$$Y_{1i} \overset{iid}{\sim} \frac{5}{10}N(1,1) + \frac{5}{10}N(-10,1),\ i = 1,2,\ldots,100$$

$$Y_{2i} \overset{iid}{\sim} \frac{7}{10}N(1,1) + \frac{3}{10}N(8,1),\ i = 1,2,\ldots,100.$$

Example 3:

The last data set is taken from the same underlying distributions as in the previous example, with the difference that now there are 200 data from each F_j^*:

$$Y_{1i} \overset{iid}{\sim} \frac{5}{10}N(1,1) + \frac{5}{10}N(-10,1),\ i = 1,2,\ldots,200$$

$$Y_{2i} \overset{iid}{\sim} \frac{7}{10}N(1,1) + \frac{3}{10}N(8,1),\ i = 1,2,\ldots,200.$$

The reason for considering this data set is to examine the intuition that the more data there are, the closer our results will be to those data and the distributions that created them. This will be mostly exhibited in the posterior distributions of $\varepsilon, M_0, M_1, M_2$ (for the model of Müller et al. (2004)) and x and y (for Model (2.1.4)).

3.4.2 Computations

I simulated from the posterior distributions of the parameters of interest using the MCMC algorithms presented above. The main issue about the implementation of these algorithms is the additional split/merge step, and its effect on the mixing of the chains.

Example 1:

First, consider the model of Müller et al. (2004). Figure 3.3 shows the trace plots for the weight ε with (right) and without (left) the mix-split step. From this graph, it seems that mixing is improved when this extra step is applied, although not too much. We arrive at the same conclusion by looking at the rate of jumps between the two modes in the posterior of ε (0 and 0.3) of our chain, since the mix-split step increases this percentage from 1.12% to 1.22%. The acceptance rate of the split steps was 3.94% and the corresponding rate for the merge steps was 3.92%.

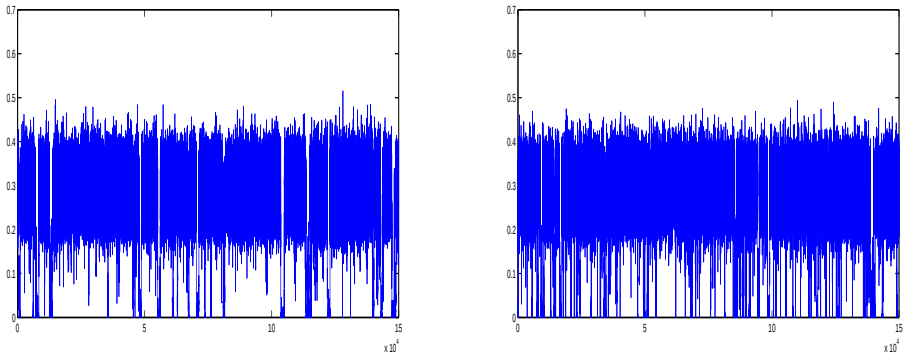


Figure 3.3: Trace plot for the posterior of ε without (left) and with the extra mix-split step (right) for the model of Müller et al. (2004) for the first simulated data set.

Next, we look at the same data, applied to Model (2.1.4). As can be seen from Figure 3.4, the extra mix-split step (7.0% acceptance of split steps and 6.9% acceptance of merge steps) enhances the mixing of the chain even more than for the previous model. This can also be seen by the difference in the percentage of jumps between the two modes of the posterior of ε (0.23% without and 0.74% with the extra mix-split step).

Example 2:

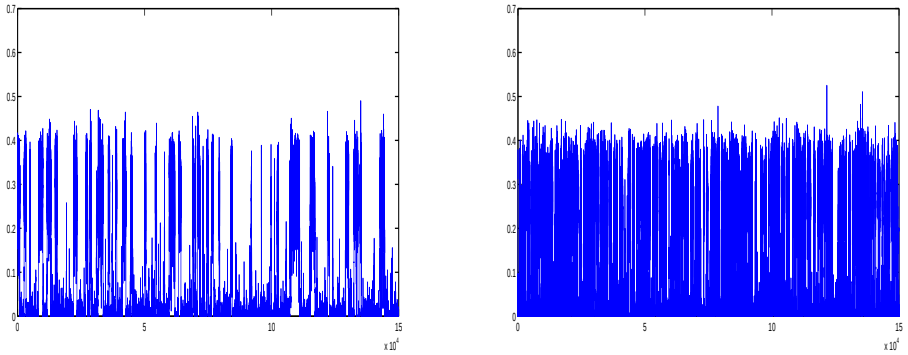


Figure 3.4: Trace plot for the posterior of ε without (left) and with the extra mix-split step (right) for Model (2.1.4) for the first simulated data set.

For the Müller et al. (2004) model, 3.4% of split and the same percentage for merge steps were accepted. The improvement of mixing is illustrated in Figure 3.5, and it is particularly interesting to note that without the mix-split step, the mode at 0 is not visited at all! Therefore, we can imagine that a case where this extra split-merge step could be particularly important is when one of the two modes is really small and the basic algorithm (i.e. without this extra step) might miss it completely. Finally, note that when the extra split/merge step is used, we have jumps between the two modes in 0.011% of the steps in the MCM Chain, whereas without this step this percentage is, as mentioned above, zero.

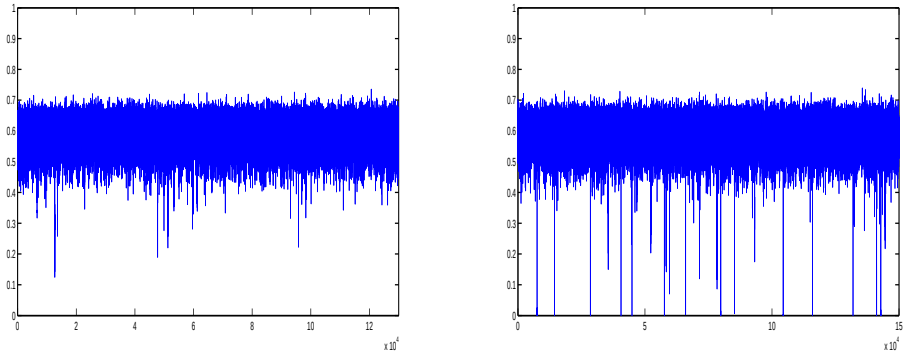


Figure 3.5: Trace plot for the posterior of ε without (left) and with the extra mix-split step (right) for the model of Müller et al. (2004) for the second simulated data set.

The difference in mixing with and without the mix-split step for Model (2.1.4) is seen in Figure 3.6. Again, there is evidently a substantial improvement in mixing caused by the 3.4% acceptance rate of split steps and 3.4% acceptance of merge steps. This is another case where the extra mix-split

step will be helpful, since the transitions of the chain between the two modes for ε without it are not so frequent (only 0.0013% of all the steps, whereas when the extra step is applied, this percentage becomes 0.22%).

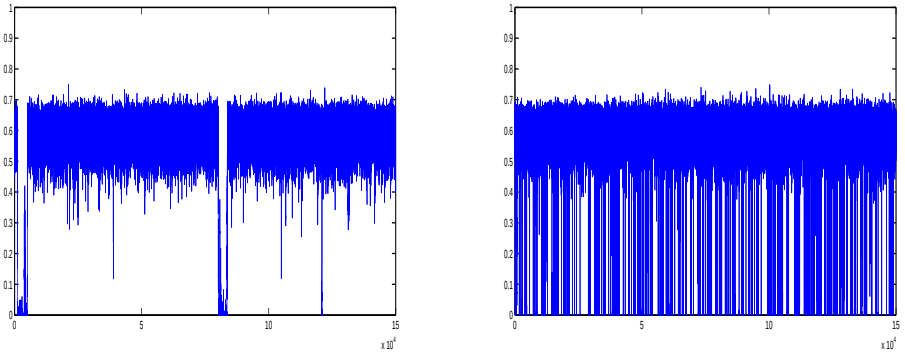


Figure 3.6: Trace plot for the posterior of ε without (left) and with the extra mix-split step (right) for Model (2.1.4) for the second simulated data set.

Another point to be made here is that trace plots like the one without the mix-split step can be deceiving and lead to wrong inference, especially when a small number of posterior samples is taken. This would be the case, for example, if one only takes the first 30000 or from the 60000th to the 100000th step of our chain, since then the mode at 0 will be overestimated. This is another reason for applying the mix-split step to the algorithms and improve the mixing of the corresponding chains.

3.4.3 Posterior inference

In this subsection I present the results for the posterior distributions of the parameters of interest in both my basic proposed model and the model of Müller et al. (2004). Since it is clear from above that including the additional mix-split step improves mixing, I will only present the results with this step included in the algorithms.

Example 1:

We first apply the model of Müller et al. (2004).

The kernel density estimate of the posterior distribution of ε is shown in Figure 3.7. As mentioned above, this distribution is bimodal: a larger mode at the minimum of the values of the weights that created the data (i.e. 0.3 and 0.7) and a smaller one around 0. In this case, only 8.3% of the posterior mass is close to zero (less that 0.01). As explained in Section 2, those two cases (i.e. $\varepsilon \simeq 0.3$ and 0) are the two most parsimonious models that sufficiently describe the data. In the case where ε is very close to 0, we have: $F_1 \equiv \frac{7}{10}N(1,1) + \frac{3}{10}N(-10,1)$, $F_2 \equiv \frac{3}{10}N(1,1) + \frac{7}{10}N(8,1)$ and F_0 is the

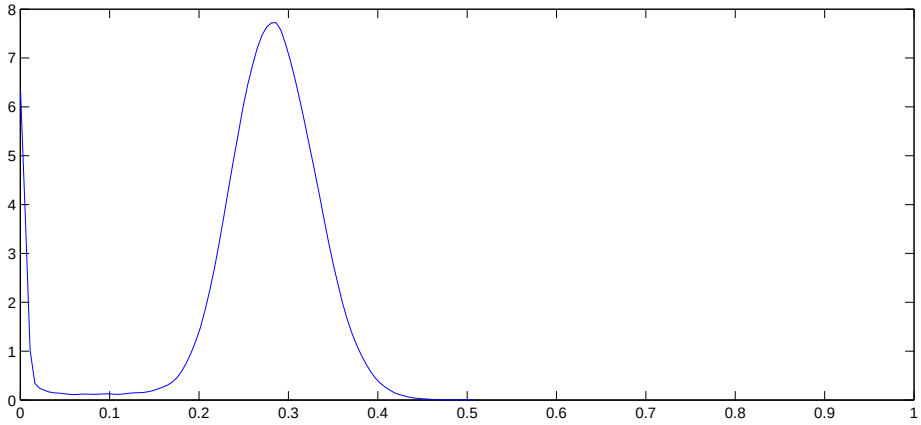


Figure 3.7: Kernel density estimate for the posterior of ε for the model of Müller et al. (2004) for the first simulated data set.

empty set.

In the case of $\varepsilon \simeq 0.3$, we have: $F_1 \equiv \frac{4}{7}N(1,1) + \frac{3}{7}N(-10,1), F_2 \equiv N(8,1)$ and $F_0 \equiv N(1,1)$. This can be seen in the predictive densities of F_1, F_2 and F_0 in Figure 3.8. Notice that, due to

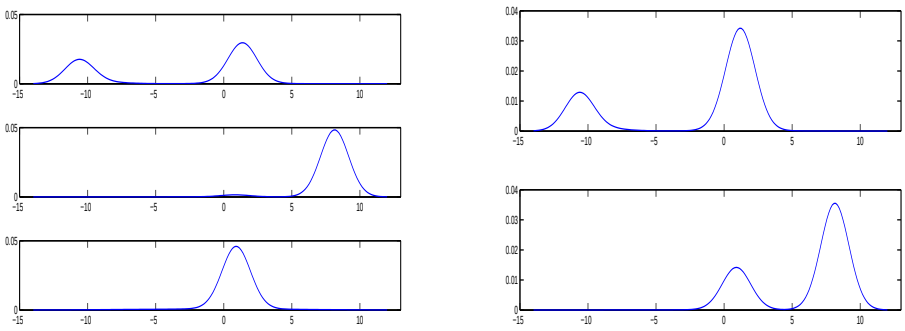


Figure 3.8: Predictive densities for the component distributions F_1 (top), F_2 (middle) and F_0 (bottom) (left) and of F_1^* (top) and F_2^* (bottom) (right) for the model of Müller et al. (2004) for the first simulated data set.

the much higher probability of the second case ($\varepsilon \simeq 0.3$), these predictive densities will reflect the representation of F_0, F_1 and F_2 induced in this case. The effect of the other possibility ($\varepsilon = 0$) can be seen in the tiny mode in the predictive density of F_2 around 1.

In the same figure the predictive densities of F_1^* and F_2^* were also plotted. As one would expect, those predictive distributions are very close (but of course, not exactly the same) to the distributions

that created the data, since the data size is large enough.

In Figure 3.9 I have plotted the posterior distributions of the three concentration parameters M_0, M_1, M_2. It is also interesting to see the interaction between those parameters with the number of clusters K_j in each component distribution F_j, $j = 0, 1, 2$, so those quantities were plotted in the same graph. As one would expect, for higher values of the concentration parameter, higher probabilities are given to larger number of clusters, as suggested in equation (1.1.4).

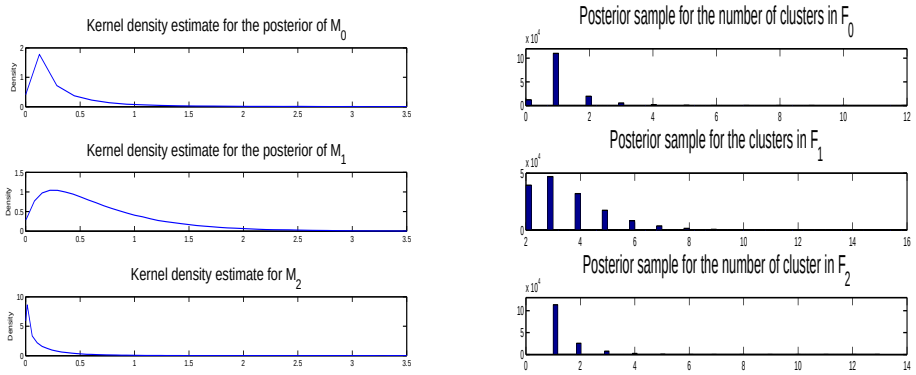


Figure 3.9: Posterior distributions of M_0, M_1 and M_2 (left) and posterior samples for K_0, K_1 and K_2 (right) for the model of Müller et al. (2004) for the first simulated data set.

Finally, the posterior distributions for S, m and B are shown in Figure 3.10, whereas the mean, median and 95% credible intervals (C.I.) for all the above parameters (except ε, since its posterior distribution is bimodal and, therefore, it does not make much sense talking about its mean and quantiles) are shown in Table 3.1.

	M_0	M_1	M_2	K_0	K_1	K_2	S	m	B
Mean	0.250	0.695	0.175	1.183	3.535	1.347	1.067	-0.478	32.124
Median	0.0843	0.545	0.0773	1	3	1	1.060	-0.473	27.024
2.5th percentile	0.00019	0.0516	0.00016	-	-	-	0.828	-4.376	11.276
97.5th percentile	1.510	2.207	0.894	-	-	-	1.342	3.440	82.897

Table 3.1: Mean, median and 95% C.I. for the parameters in the model of Müller et al. (2004) for the first data set (Note: I omit the 2.5-th and 97.5-th quantiles for the K_j's as they are discrete quantities).

Next, Model (2.1.4) was applied to the same data. The posterior of the weight is given in Figure 3.11, and it is seen that the better mixing of this chain results in a more visually appealing kernel density estimate, reducing the previously huge mode at 0. The modes of this distribution are

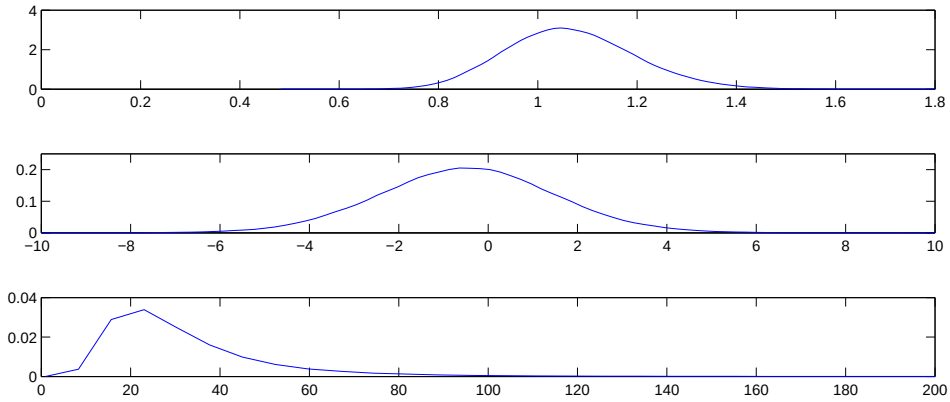


Figure 3.10: Posterior distributions for S (top), m (middle) and B (bottom) for the model of Müller et al. (2004) for the first simulated data set.

the same as with the model of Müller et al. (2004), however in this case we see a higher mode at 0 (57.9% of the posterior sample for ε was less than 0.01) and a less clear (although still obvious) discrimination of the two modes. This different behaviour, compared to the previous model, is definitely the interaction between the weight and the M's, both *a priori* and *a posteriori*. The

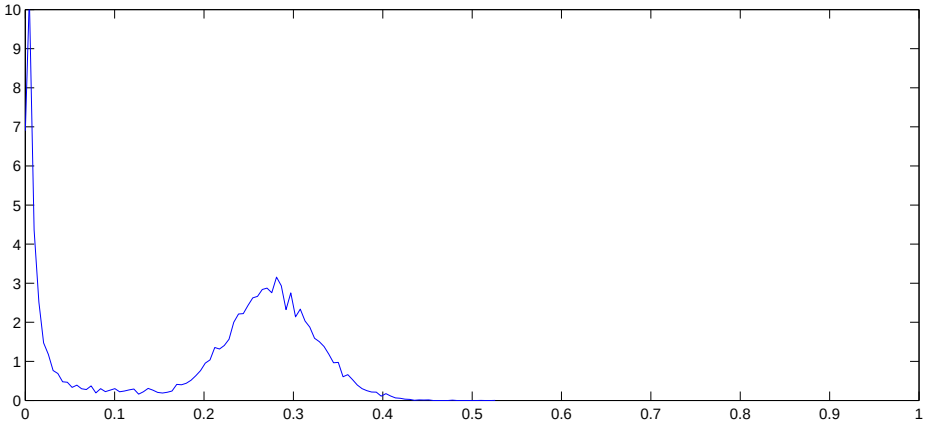


Figure 3.11: Kernel density estimate for the posterior of ε for Model (2.1.4) for the first simulated data set.

interpretation of the component distributions F_j for the two modes of ε is the same as before. As a result of the different weights of those two modes, the predictive densities of the F's (Figure 3.12, left) are different, but not excessively. On the contrary, the predictive densities of F_1^* and F_2^*

(Figure 3.12, right) are the same as before, which is what one would expect due to the large enough number of data.

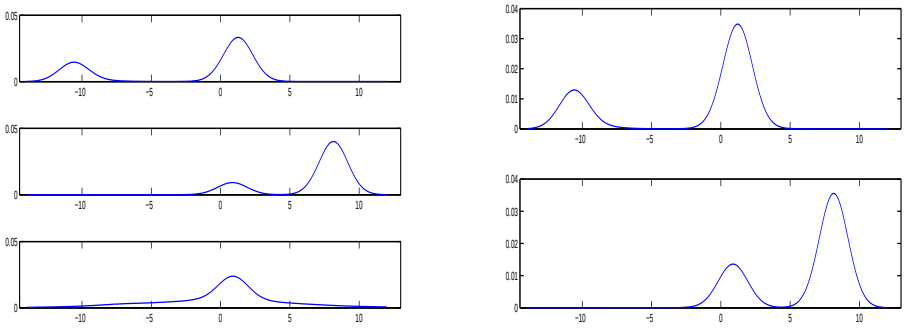


Figure 3.12: Predictive densities for the component distributions F_1 (top), F_2 (middle) and F_0 (bottom) (left) and of F_1^* (top) and F_2^* (bottom) (right) for the basic proposed model for the first simulated data set.

Next, I plotted the posterior distributions of M_0 and M_1, as well as those of the reparametrisation $y = \frac{M_0}{M_0+M_1}$, $x = M_0 + M_1$. The results are shown in Figure 3.13, and the posteriors of the cluster sizes K_j, $j = 0, 1, 2$ are shown in Figure 3.14. The posterior distributions of the other three parameters, S, m and B were similar as before, so I omit plotting them. Again, we see a

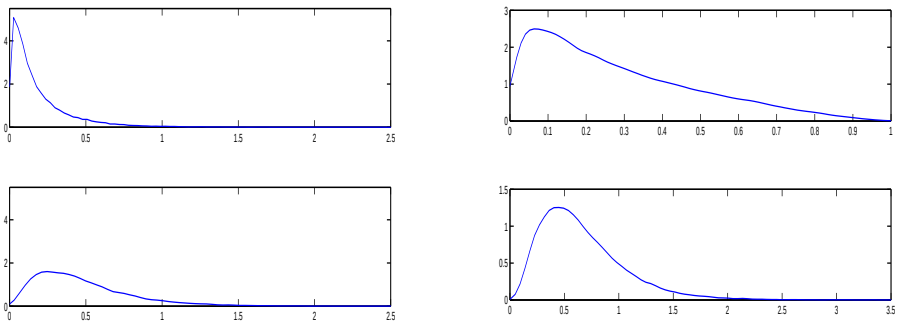


Figure 3.13: Posterior distributions of M_0 (top) and M_1 (bottom) (left) and of y (top) and x (bottom) (right) for Model (2.1.4) for the first simulated data set.

positive correlation between the value of M and the number of clusters in each component. The mean, median and 95% C.I. for the parameters in this model are shown in Table 3.2.

Apart from the obvious difference that we only have two M's here, instead of 3 in the model of

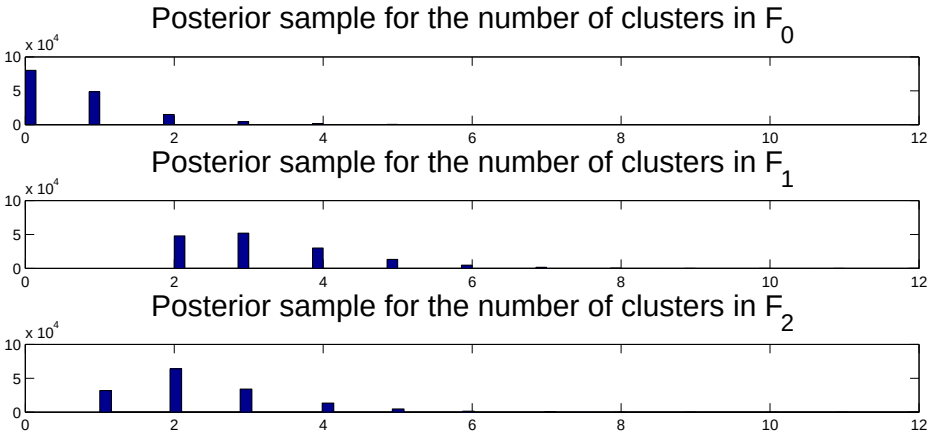


Figure 3.14: Posterior distributions of K_0, K_1 and K_2 for Model (2.1.4) for the first simulated data set.

	M_0	M_1	y	x	K_0	K_1	K_2	S	m	B
Mean	0.188	0.695	0.281	0.663	0.667	3.218	2.350	1.081	-0.241	31.416
Median	0.117	0.405	0.229	0.585	1	3	2	1.076	-0.228	26.473
2.5th percentile	0.008	0.074	0.0163	0.150	-	-	-	0.833	-4.139	11.013
97.5th percentile	0.771	1.274	0.779	1.626	-	-	-	1.358	3.618	81.055

Table 3.2: Mean, median and 95% C.I. for the parameters in Model (2.1.4) for the first data set (I omit the 2.5-th and 97.5-th quantiles for the K_j's as they are discrete quantities).

Müller et al. (2004), we see that the main differences here, compared to Table 3.1, are regarding these M's. These differences are also due to the fact that there is interaction between those quantities and the weight ε. In this spirit, it is interesting to note that the posterior mean for y is very close to 0.3, which makes sense since y is the mean of ε *a priori*, and also the posterior distribution of y is left-skewed, in order to accommodate the mode of ε at 0.

Example 2:

First I apply Model (1.2.14) to these data.

As before, the posterior distribution of ε (Figure 3.15) is bimodal: a larger mode at 0.54 and a smaller one around 0, containing only 0.19% of the posterior mass of ε. Normally, instead of around 0.6, the larger mode should have been at the minimum of the weights that created the data, i.e. 0.5. However, due to the not so large data size, in fact the number of Y_{1i} that can be assigned to the $N(1,1)$ cluster was 54. As a result, in practise it is like having data $Y_{1i} \overset{iid}{\sim} \frac{54}{100}N(1,1)+\frac{46}{10}N(-10,1)$, $i = 1, 2, \ldots, 100$ and therefore the smaller of the two weights is 0.54.

As before, the two values, $\varepsilon = 0.6$ and 0 are the two most parsimonious models that sufficiently de-

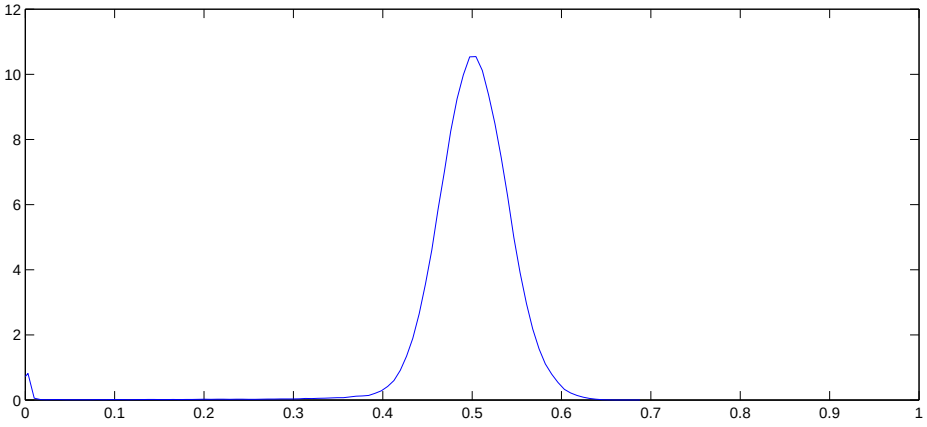


Figure 3.15: Posterior distribution of ε for the model of Müller et al. (2004) for the second simulated data set.

scribe the data. In the case where ε is very close to 0, we have: $F_1 \equiv \frac{54}{100}N(1,1)+\frac{46}{100}N(-10,1)$, $F_2 \equiv \frac{7}{10}N(1,1) + \frac{3}{10}N(8,1)$ and F_0 is the empty set. In the case of $\varepsilon \simeq 0.6$, we will more or less have: $F_1 \equiv N(-10,1), F_2 \equiv \frac{16}{46}N(1,1) + \frac{30}{46}N(8,1)$ and $F_0 \equiv N(1,1)$.

This can be seen in the predictive densities of F_1, F_2 and F_0 in Figure 3.16. Again, due to the high probability of the case $\varepsilon = 0.6$, these predictive densities are almost identical to the ones corresponding in this case. The effects of the case $\varepsilon = 0$, can again be seen in the small mode in the predictive density of F_2 around 1.

In the right part of the same figure I plotted the predictive densities of F_1^* and F_2^*, which are very close to the empirical distributions of the data (and not exactly the distributions that created the data, as explained above). In Figure 3.17 the kernel density estimates of the posterior distributions of the three concentration parameters, M_0, M_1, M_2, are shown.

Finally, Table 3.3 states the mean, median and 95% credible intervals (C.I.) for all parameters, except ε.

	M_0	M_1	M_2	K_0	K_1	K_2	S	m	B
Mean	0.126	0.137	0.222	1.166	1.107	1.404	0.828	0.564	31.641
Median	0.0539	0.0060	0.101	1	1	1	0.823	0.585	25.266
2.5th perc	0.00012	0.00013	0.00022	-	-	-	0.671	-3.686	10.036
97.5th perc	0.654	0.711	1.125	-	-	-	1.015	4.705	91.403

Table 3.3: Mean, median and 95% C.I. for the parameters in the model of Müller et al. (2004) for the second data set.

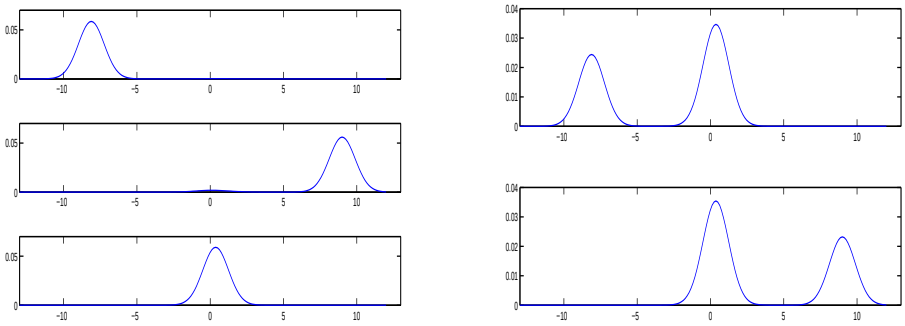


Figure 3.16: Predictive densities for the component distributions F_1 (top), F_2 (middle) and F_0 (bottom) (left) and of F_1^* (top) and F_2^* (bottom) (right) for the model of Müller et al. (2004) for the second simulated data set.

As can be seen from this table, the differences between the M's are very small, so we cannot actually see the association between higher values of the concentration parameters and higher number of clusters.

Consider now Model (2.1.4) for the same data. The mean, median and 95% C.I. for the parameters in this model are shown in Table 3.4.

	M_0	M_1	y	x	K_0	K_1	K_2	S	m	B
Mean	0.218	0.183	0.529	0.402	1.249	1.212	1.402	0.825	0.560	31.031
Median	0.149	0.128	0.542	0.326	1	1	1	0.821	0.580	24.901
2.5th perc	0.0096	0.088	0.0596	0.046	-	-	-	0.666	-3.644	9.845
97.5th perc	0.833	0.672	0.951	1.191	-	-	-	1.012	4.655	89.424

Table 3.4: Mean, median and 95% C.I. for the parameters in Model (2.1.4) for the second data set.

The posterior distribution of ε is given in Figure 3.18. Two modes at 0 and 0.6 are present and, as for the first data set, the mode at 0 is larger than in the case of the model of Müller et al. (2004) (only 0.19% were less than 0.01), but not as large as in the case of the previous data set (4.4% of the posterior sample for ε was less than 0.01 here).

The predictive densities are shown in Figure 3.19 for the component distributions F_j (left) and the correlated distributions F_j^* (right). The latter are as one would expect, in particular resembling the histogram of the data. The former are in accordance with the two modes (and the corresponding weights) of the weight and the interpretation of the component distributions in each of them, in a similar fashion as when the model of Müller et al. (2004) was applied to the same data. More specifically, for F_1 there is only one mode at -8, for F_2 there is a large mode at 10 and a much

smaller one at 1 (caused when ε is close to 0) and for F_0 a mode at 1. Both of those graphs are almost the same as the corresponding ones in the case of the model of Müller et al. (2004).

Next, in Figure 3.20 the posterior distributions of M_0 and M_1 are plotted, as well as those of the reparametrisation $y = \frac{M_0}{M_0+M_1},\ x = M_0 + M_1$.
We also see that the posterior mean for y is close to the larger posterior mode of ε at 0.6, which is a result of the fact that y is the prior mean of ε. Also, since the mode at 0 is a very small one, we cannot see much skewness in the posterior of y, which would otherwise account on accommodating the mode of ε at 0.

Example 3:
For the model of Müller et al. (2004) we see that the largest mode in the posterior of ε has moved to the more "correct" value of 0.5, whereas the mode at 0 is not affected (Figure 3.21).
For $\varepsilon \simeq 0$, we have: $F_1 \equiv \frac{5}{10}N(1,1)+\frac{5}{10}N(-10,1),\ F_2 \equiv \frac{7}{10}N(1,1)+\frac{3}{10}N(8,1)$ and F_0 is the empty set. In the case of $\varepsilon \simeq 0.5$, we have: $F_1 \equiv N(-10,1), F_2 \equiv \frac{2}{5}N(1,1)+\frac{3}{5}N(8,1)$ and $F_0 \equiv N(1,1)$.
This can be seen in the predictive densities of F_1, F_2 and F_0 in Figure 3.22.
Notice also that the predictive densities of F_1^* and F_2^* (not shown) will now be very close not only to the empirical distributions of the data, but also to the distributions that created the data, since here the data sizes are large enough.
Finally, in Table 3.5 the mean, median and 95% credible intervals (C.I.) for all the main parameters (except ε) are shown.

	M_0	M_1	M_2	K_0	K_1	K_2
Mean	0.126	0.127	0.490	1.182	1.165	2.860
Median	0.0517	0.0557	0.369	1	1	3
2.5th perc	0.00011	0.00012	0.0306	-	-	-
97.5th perc	0.650	0.711	1.632	-	-	-

Table 3.5: Mean, median and 95% C.I. for the parameters in the model of Müller et al. (2004) for the third data set.

In the case of Model (2.1.4), we have the same modes for ε (0 and 0.5), with the difference being at the relative mass of those modes (Figure 3.23). Specifically, the mode at 0 is substantially larger than in the case of the previous model, with 20.9% of the posterior values of ε being less than 0.01 (compared to the corresponding 1.0% in the case of the model of Müller et al. (2004) for the same data).
The predictive densities for the component distributions F_j are shown in Figure 3.24, and reflect what one would expect those component distributions to be under the two different cases, $\varepsilon = 0$ and $\varepsilon = 0.5$ and the weights of these cases. Notice that, due to the difference in the mass of ε at

0 from the previous model, the predictive distributions for F_1 and F_2 are different than before. As for the predictive densities of F_1^* and F_2^*, again those are the same as the distributions creating the data.

Finally, the mean, median and 95% C.I. for the parameters are shown in Table 3.6.

	M_0	M_1	y	x	K_0	K_1	K_2
Mean	0.200	0.290	0.392	0.490	1.109	1.600	2.570
Median	0.139	0.241	0.373	0.426	1	1	2
2.5th percentile	0.0098	0.0386	0.329	0.097	-	-	-
97.5th percentile	0.749	0.817	0.853	1.239	-	-	-

Table 3.6: Mean, median and 95% C.I. for the parameters in Model (2.1.4) for the third data set.

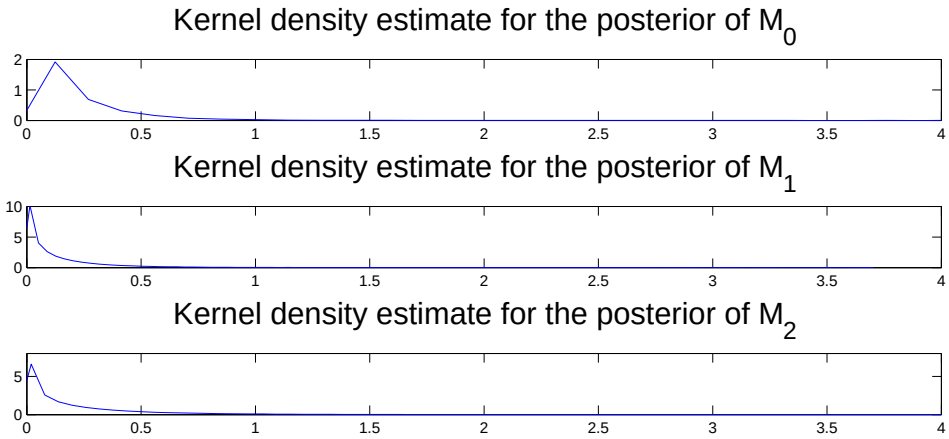


Figure 3.17: Posterior distributions of M_0, M_1 and M_2 for the model of Müller et al. (2004) for the second simulated data set.

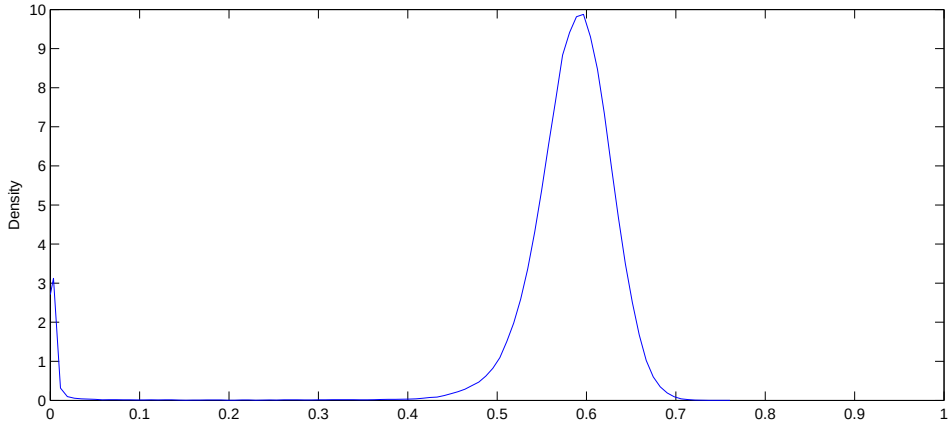


Figure 3.18: Kernel density estimate for the posterior of ε for Model (2.1.4) for the second simulated data set.

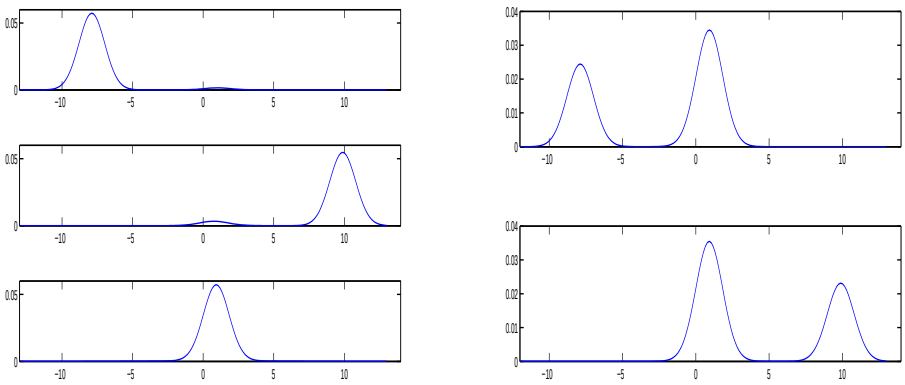


Figure 3.19: Predictive densities for the component distributions F_1 (top), F_2 (middle) and F_0 (bottom) (left) and of F_1^* (top) and F_2^* (bottom) (right) for Model (2.1.4) for the second simulated data set.

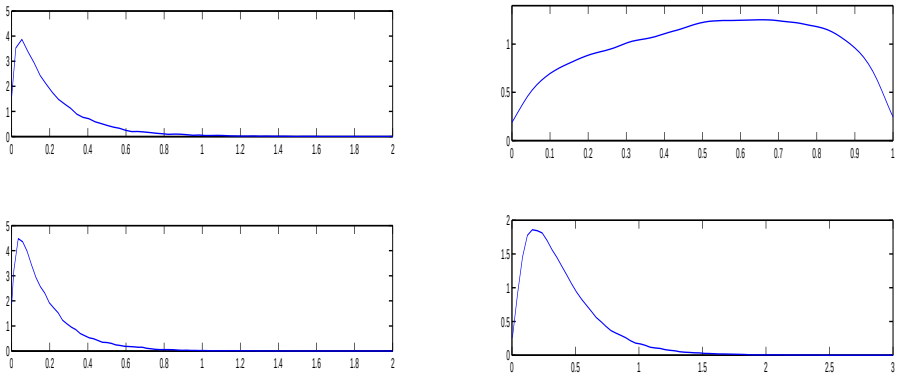


Figure 3.20: Posterior distributions of M_0 (top) and M_1 (bottom) (left) and of y (top) and x (bottom) (right) for Model (2.1.4) for the second simulated data set.

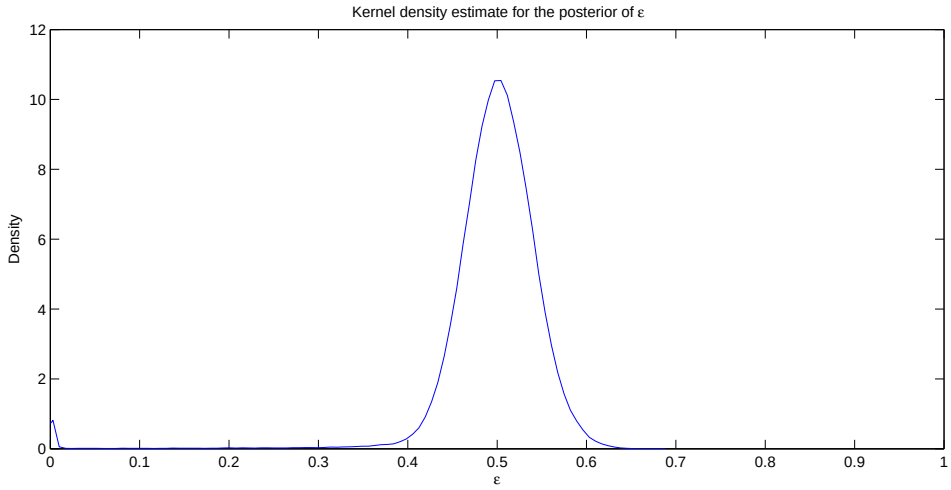


Figure 3.21: Posterior distribution of ε for the model of Müller et al. (2004) for the third simulated data set.

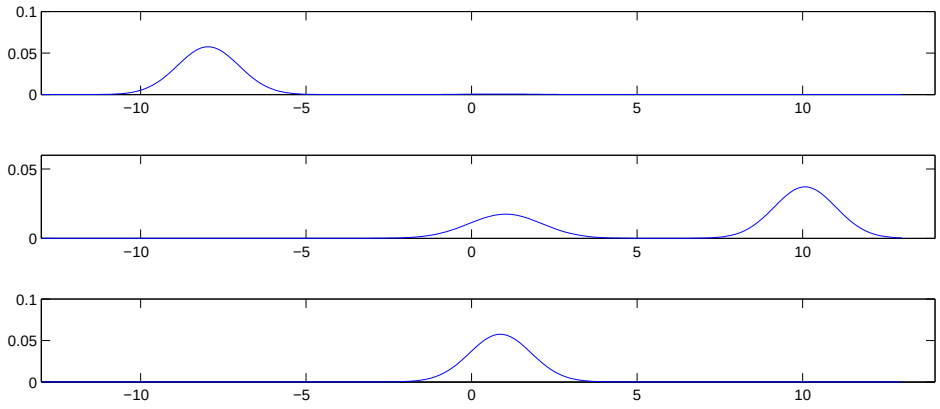


Figure 3.22: Predictive densities for the component distributions F_1 (top), F_2 (middle) and F_0 (bottom) for the model of Müller et al. (2004) for the third simulated data set.

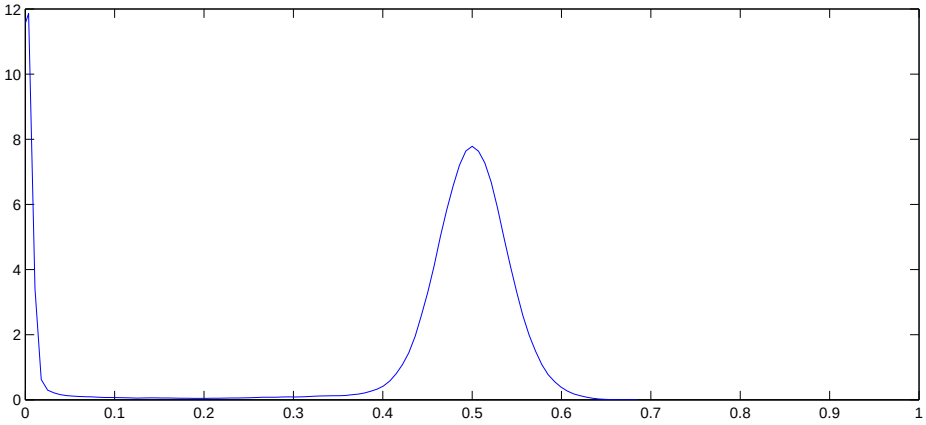


Figure 3.23: Kernel density estimate for the posterior of ε for Model (2.1.4) for the third simulated data set.

3.5 Algorithms For the Extended Models

The MCMC algorithm described in Section 3.2 can be extended accordingly and applied to each of the three-dimensional models presented in section 2.2. The hyperparameters can be chosen to express any prior beliefs concerning the parameters of the model, and the additional techniques used to improve the performance of the algorithm in the two-dimensional case, for example updating the

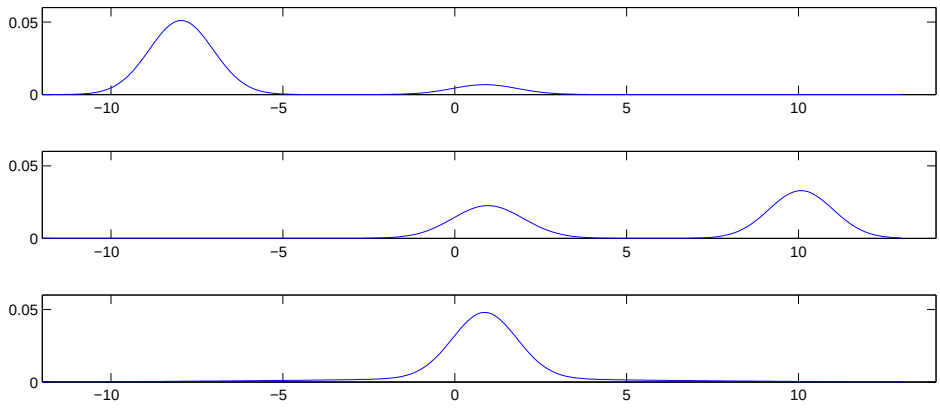


Figure 3.24: Predictive densities for the component distributions F_1 (top), F_2 (middle) and F_0 (bottom) for the basic proposed model for the third simulated data set.

discrete values ϕ_{ji}, can be directly applied here, too.

On the other hand, for the extended models, apart from the additional computational burden, there is also an additional complexity in terms of notation. For example, in the previous case we used binary indicators r_{ji}, taking the value 0 if the observation Y_{ji} was in the common part of the correlated distributions, and 1 otherwise. Here, apart from the fact that each r_{ji} can take four possible values (since there are four components in each distribution F_j^*), we also have the issue that the same value for $r_{ji},\ r_{j'i'},\ j \neq j'$ might correspond to different components, since we do not have the same components in all three distributions. However, these issues result in a small increase in the overall complexity of the algorithm and should not be overemphasised.

Finally, another comment to be made here is that, in order to have dimensional coherence for the prior distributions of some parameters, we might consider matching their hyperparameters with the hyperparameters of the simpler models. For example, consider the concentration parameters in the higher dimensional version of my proposed model (2.1.4), i.e. Model (2.2.12). If we want to have some form of dimensional coherence, and since the equivalent to one concentration parameter in Model (2.1.4) is the sum of two concentration parameters in the extended model, we might set $a_0 = a_1/2$ and $b_0 = b_1$, where a_0, b_0 are the hyperparameters in (2.2.12) and a_1, b_1 are the hyperparameters for M_0 and M_1 in (2.1.4). In this way, for example $M_{123} + M_{12} \sim Ga(2a_0, b_0) \equiv Ga(a_1, b_1)$, due to the additive property of the gamma distribution.

3.6 Simulating the Model Via Direct Normalisation and the Slice Sampler

In this subsection a modification of the algorithms presented above is discussed, using slice samplers in some steps in the MCMC algorithm. The model produced by directly using the normalisation technique to gamma processes is ideal in presenting this method.
The aforementioned model, within the context mentioned in Section 3.1 is the following:

$$Y_{ji} \sim N(\mu_{ji}, S),\ i = 1, 2, \ldots, N_j,\ j = 1, 2$$

$$\mu_{ji} \sim F_j^*,\ \text{where } F_j^* = \varepsilon_j F_0 + (1 - \varepsilon_j) F_j$$

$$F_0 \sim DP(M_0, H),\ F_1, F_2 \overset{iid}{\sim} DP(M_1, H),\ \text{where } H \equiv N(m, B)$$

$$\varepsilon_1, \varepsilon_2 \sim Be(M_0, M_1),\ \text{but not independent}$$

$$M_0, M_1 \overset{iid}{\sim} Ga(a_0, b_0),\ (m, B) \sim N(m_0, A) \times \text{IGa}(c, 1/cC),\ S \sim \text{IGa}(q, 1/qR).$$

In the above model, the dependence of the two weights is due to the term $G_0(\Omega)$ which appears in the expressions of both, and their joint distribution is given at the end of page 30.
On this occasion, however, a different parametrisation is used:
Set $\gamma_0 = G_0(\Omega), \gamma_1 = G_1(\Omega)$ and $\gamma_2 = G_2(\Omega)$. Then, $\gamma_0 \sim Ga(M_0, 1), \gamma_1 \sim Ga(M_1, 1)$ and $\gamma_2 \sim Ga(M_1, 1)$ and are mutually independent. Then, ε_1 can be replaced by $\frac{\gamma_0}{\gamma_0+\gamma_1}$ and ε_2 by $\frac{\gamma_0}{\gamma_0+\gamma_2}$ and the above model becomes:

$$Y_{ji} \sim N(\mu_{ji}, S),\ i = 1, 2, \ldots, N_j,\ j = 1, 2$$

$$\mu_{ji} \sim F_j^*,\ \text{where } F_j^* = \frac{\gamma_0}{\gamma_0 + \gamma_j} F_0 + \frac{\gamma_j}{\gamma_0 + \gamma_j} F_j,\ j = 1, 2$$

$$F_0 \sim DP(M_0, H),\ F_j \overset{iid}{\sim} DP(M_1, H),\ \text{where } H \equiv N(m, B) \tag{3.6.1}$$

$$\gamma_0 \sim Ga(M_0, 1), \gamma_1 \sim Ga(M_1, 1) \text{ and } \gamma_2 \sim Ga(M_1, 1) \text{ and mutually independent}$$

$$M_0, M_1 \overset{iid}{\sim} Ga(a_0, b_0),\ (m, B) \sim N(m_0, A) \times \text{IGa}(c, 1/cC),\ S \sim \text{IGa}(q, 1/qR).$$

We can simulate this model using an MCMC algorithm similar to the one used for Model (2.1.4). The differences will be:

1. ε in the code for Model (2.1.4) should be replaced by $\varepsilon_1 = \frac{\gamma_0}{\gamma_0+\gamma_1}$ or $\varepsilon_2 = \frac{\gamma_0}{\gamma_0+\gamma_2}$, depending on whether we are updating parameters related to F_1^* or F_2^*.

2. The full conditionals of M_0 and M_1, due to the different prior of the weights (actually the priors of the γ's), which also involve the M's. In this case, we have:
   $f(M_0, M_1|\cdots) \propto f(M_0)f(M_1)f(\boldsymbol{s}|\boldsymbol{r}, M_0, M_1)f(\gamma_0|M_0)f(\gamma_1|M_1)f(\gamma_2|M_1)$
   $\Rightarrow f(M_0, M_1|\cdots) \propto f(M_0)f(M_1)f(K_0|M_0)f(K_1|M_1)f(K_2|M_1)f(\gamma_0|M_0)f(\gamma_1|M_1)f(\gamma_2|M_1)$.
   So, for M_0 we get: $f(M_0|\cdots) \propto M_0^{a_0+K_0-1}e^{-b_0M_0}\gamma_0^{M_0}\frac{1}{\Gamma(M_0+n_0)}$
   and for M_1 : $f(M_1|\cdots) \propto M_1^{a_0+K_1+K_2-1}e^{-b_0M_1}\gamma_1^{M_1}\gamma_2^{M_1}\frac{1}{\Gamma(M_1+n_1)\Gamma(M_1+n_2)}$
   where K_j and n_j are as before.
   Since the above distributions are not of any standard form, Metropolis-Hastings updating steps can be used to simulate from them.

3. The full conditional distributions of the $\gamma' s$, which we will simulate using slice sampling methods.

The joint full conditional distribution for the γ's will be:

$$\begin{aligned}
f(\gamma_0,\gamma_1,\gamma_2|\cdots) &\propto f(\gamma_0|M_0)f(\gamma_1|M_1)f(\gamma_2|M_1)f(\boldsymbol{r}|\gamma_0,\gamma_1,\gamma_2)\\
&\propto \gamma_0^{M_0-1}e^{-\gamma_0}\gamma_1^{M_1-1}e^{-\gamma_1}\gamma_2^{M_1-1}e^{-\gamma_2}\prod_{i:r_{1i}=1}\left(\frac{\gamma_1}{\gamma_0+\gamma_1}\right)\times\\
&\quad\prod_{i:r_{1i}=0}\left(\frac{\gamma_0}{\gamma_0+\gamma_1}\right)\prod_{i:r_{2i}=1}\left(\frac{\gamma_2}{\gamma_0+\gamma_2}\right)\prod_{i:r_{2i}=0}\left(\frac{\gamma_0}{\gamma_0+\gamma_2}\right)\\
&\propto \gamma_0^{M_0-1}e^{-\gamma_0}\gamma_1^{M_1-1}e^{-\gamma_1}\gamma_2^{M_1-1}e^{-\gamma_2}\times\\
&\quad\left(\frac{\gamma_1}{\gamma_0+\gamma_1}\right)^{\sum r_{1i}}\left(\frac{\gamma_0}{\gamma_0+\gamma_1}\right)^{N_1-\sum r_{1i}}\left(\frac{\gamma_2}{\gamma_0+\gamma_2}\right)^{\sum r_{2i}}\left(\frac{\gamma_0}{\gamma_0+\gamma_2}\right)^{N_2-\sum r_{2i}}.
\end{aligned}$$

By introducing the auxiliary variables $U_i \sim U(0,1),\ i = 1, 2, \ldots, 7$, the above expression can be written as follows:
$f(\gamma_0,\gamma_1,\gamma_2|\cdots) \propto \gamma_0^{M_0-1}\gamma_1^{M_1-1}\gamma_2^{M_1-1}\int\int\cdots\int I\left(U_1 < \left(\frac{\gamma_0}{\gamma_0+\gamma_1}\right)^{N_1-\lambda_1}\right) I\left(U_2 < \left(\frac{\gamma_1}{\gamma_0+\gamma_1}\right)^{\lambda_1}\right)\times$
$I\left(U_3 < \left(\frac{\gamma_0}{\gamma_0+\gamma_2}\right)^{N_2-\lambda_2}\right) I\left(U_4 < \left(\frac{\gamma_2}{\gamma_0+\gamma_2}\right)^{\lambda_2}\right) I\left(U_5 < e^{-\gamma_0}\right) I\left(U_6 < e^{-\gamma_1}\right) I\left(U_7 < e^{-\gamma_2}\right) dU_1 dU_2\cdots dU_7$
where I denotes the indicator function, $\lambda_1 = \sum r_{1i}$ is the number of data allocated to the first idiosyncratic part and $\lambda_2 = \sum_i r_{2i}$ is the number of data allocated to the second idiosyncratic part.
By simple calculations, we find:

- $U_1|\cdots \sim U\left(0, \left(\frac{\gamma_0}{\gamma_0+\gamma_1}\right)^{N_1-\lambda_1}\right)$
- $U_2|\cdots \sim U\left(0, \left(\frac{\gamma_1}{\gamma_0+\gamma_1}\right)^{\lambda_1}\right)$
- $U_3|\cdots \sim U\left(0, \left(\frac{\gamma_0}{\gamma_0+\gamma_2}\right)^{N_2-\lambda_2}\right)$
- $U_4|\cdots \sim U\left(0, \left(\frac{\gamma_2}{\gamma_0+\gamma_2}\right)^{\lambda_2}\right)$

- $U_5|\cdots \sim U\left(0, e^{-\gamma_0}\right)$
- $U_6|\cdots \sim U\left(0, e^{-\gamma_1}\right)$
- $U_7|\cdots \sim U\left(0, e^{-\gamma_2}\right)$
- $f(\gamma_0|\cdots) \propto \gamma_0^{M_0-1} I\left(max\left\{\frac{\gamma_1 U_1^{1/(N_1-\lambda_1)}}{1-U_1^{1/(N_1-\lambda_1)}}, \frac{\gamma_2 U_3^{1/(N_2-\lambda_2)}}{1-U_3^{1/(N_2-\lambda_2)}}\right\} < \gamma_0 < min\left\{\frac{\gamma_1\left(1-U_2^{1/\lambda_1}\right)}{U_2^{1/\lambda_1}}, \frac{\gamma_2\left(1-U_4^{1/\lambda_2}\right)}{U_4^{1/\lambda_2}}, -\log(U_5)\right\}\right)$
- $f(\gamma_1|\cdots) \propto \gamma_1^{M_1-1} I\left(\frac{\gamma_0 U_2^{1/\lambda_1}}{1-U_2^{1/\lambda_1}} < \gamma_1 < min\left\{\frac{\gamma_0\left(1-U_1^{1/(N_1-\lambda_1)}\right)}{U_1^{1/(N_1-\lambda_1)}}, -\log(U_6)\right\}\right)$
- $f(\gamma_2|\cdots) \propto \gamma_2^{M_1-1} I\left(\frac{\gamma_0 U_4^{1/\lambda_2}}{1-U_4^{1/\lambda_2}} < \gamma_2 < min\left\{\frac{\gamma_0\left(1-U_3^{1/(N_2-\lambda_2)}\right)}{U_3^{1/(N_2-\lambda_2)}}, -\log(U_7)\right\}\right)$.

Sampling from the above distributions is easy using inversion sampling, i.e. drawing a value $V \sim U(0,1)$ and then setting $t = F^{-1}(V)$, where $t \in \{U_1, U_2, \ldots, U_7, \gamma_0, \gamma_1, \gamma_2\}$ and F is the corresponding cdf. Also note that, in the case of the γ's, the boundaries of its pdf are first calculated and then, for example, $\gamma_0 = F^{-1}(V) = \left(\alpha_0^{M_0} + V\left(\beta_0^{M_0} - \alpha_0^{M_0}\right)\right)^{1/M_0}$, where α_0, β_0 are those boundaries for γ_0. For the other two γ's, the formula will be the same, with M_1 instead of M_0 and with the corresponding boundaries.

Finally, notice that some care is needed in the case when one or more of the quantities λ_1, λ_2, $N_1 - \lambda_1$ and $N_2 - \lambda_2$ is zero. In this case, the relationships indicated in the corresponding indicator functions (the ones involving the quantities being zero) are redundant. So, that relationship can just be omitted and proceed with the rest. If such a relationship is involved in the left margin of the pdf of γ_1 or γ_2, this margin is set to 0. For example, if $\lambda_1 = 0$, the upper limit of the full conditional distribution of γ_0 will be the minimum of $\frac{\gamma_2\left(1-U_4^{1/\lambda_2}\right)}{U_4^{1/\lambda_2}}$ and $-\log(U_5)$ and the lower limit of the full conditional distribution of γ_1 will be 0. Similarly, if $N_2 = \lambda_2$, γ_0 will be defined on values greater than $\frac{\gamma_1 U_1^{1/(N_1-\lambda_1)}}{1-U_1^{1/(N_1-\lambda_1)}}$ and γ_2 will be defined for values less than $-\log(U_7)$.

Of course, one can alternatively use Metropolis-Hastings steps for updating the γ's, or equivalently, the weights $\varepsilon_1, \varepsilon_2$.

3.6.1 The mix-split step

As before, an additional mix-split step can be added in the MCMC algorithm, as it will improve mixing.

The procedure will be the same as before, and in order to calculate the acceptance probabilities, we

must calculate $f(\boldsymbol{c}|\cdots)$:

$$\begin{aligned} f(\boldsymbol{c}|\cdots) &= \int\int f(\boldsymbol{c},\gamma_0,\gamma_1,\gamma_2,\boldsymbol{\phi})d\varepsilon d\boldsymbol{\phi} \\ &\propto \int\int f(\boldsymbol{Y}|\boldsymbol{\phi},\boldsymbol{c},S)f(\boldsymbol{c},\gamma_0,\gamma_1,\gamma_2|\boldsymbol{\phi},\cdots)f(\boldsymbol{\phi}|m,B)d\gamma_0,d\gamma_1,d\gamma_2 d\boldsymbol{\phi} \\ &= \int\int f(\boldsymbol{Y}|\boldsymbol{\phi},\boldsymbol{c},S)f(\boldsymbol{c},\gamma_0,\gamma_1,\gamma_2|\boldsymbol{M})f(\boldsymbol{\phi}|m,B)d\boldsymbol{\phi}d\gamma_0,d\gamma_1,d\gamma_2 \\ &= \int f(\boldsymbol{c},\gamma_0,\gamma_1,\gamma_2|\boldsymbol{M})d\gamma_0,d\gamma_1,d\gamma_2 \int f(\boldsymbol{Y}|\boldsymbol{\phi},\boldsymbol{c},S)f(\boldsymbol{\phi}|m,B)d\boldsymbol{\phi}. \end{aligned}$$

In the above, I used the usual notation.

The problem is the first integral, which, apart from the factor $f(\boldsymbol{s}|\boldsymbol{r},\boldsymbol{M})$ which can be taken outside the integral, can be evaluated as:

$\int_0^\infty\int_0^\infty\int_0^\infty \left(\frac{\gamma_0}{\gamma_0+\gamma_1}\right)^{N_1-\lambda_1}\left(\frac{\gamma_1}{\gamma_0+\gamma_1}\right)^{\lambda_1}\left(\frac{\gamma_0}{\gamma_0+\gamma_2}\right)^{N_2-\lambda_2}\left(\frac{\gamma_2}{\gamma_0+\gamma_2}\right)^{\lambda_2} e^{-\gamma_0-\gamma_1-\gamma_2}\gamma_0^{M_0-1}\gamma_1^{M_1-1}\gamma_2^{M_1-1}d\gamma_0 d\gamma_1 d\gamma_2.$

The reparametrisation $x_1=\frac{\gamma_0}{\gamma_0+\gamma_1}, x_2=\frac{\gamma_0}{\gamma_0+\gamma_2}$ and $x_3=\gamma_0$ yields the equivalent integral:

$\int_0^1\int_0^1\int_0^\infty x_1^{N_1-\lambda_1-M_1-1}(1-x_1)^{\lambda_1+M_1-1}x_2^{N_2-\lambda_2-M_1-1}(1-x_2)^{\lambda_2+M_1-1}x_3^{M_0+2M_1-1}e^{x_3-\frac{x_3}{x_1}-\frac{x_3}{x_2}}dx_1dx_2dx_3.$

One of the problems of the integral is that one cannot guarantee for example that $N_1-\lambda_1-M_1$ is positive.

However, by integrating out x_3, we get:

$$\Gamma(M_0+M_1)\int_0^1\int_0^1 x_1^{N_1-\lambda_1+M_0+M_1-1}(1-x_1)^{\lambda_1+M_1-1}x_2^{N_2-\lambda_2+M_0+M_1-1}(1-x_2)^{\lambda_2+M_1-1} \times (x_1+x_2-x_1x_2)^{-M_0-2M_1}\,dx_1dx_2.$$

This is proportional (the constant of proportionality is known) to the expectation of $(x_1+x_2-x_1x_2)^{-M_0-2M_1}$, where

$x_1 \overset{iid}{\sim} Be(N_1-\lambda_1+M_0+M_1,\lambda_1+M_1), x_2 \overset{iid}{\sim} Be(N_2-\lambda_2+M_0+M_1,\lambda_2+M_1)$ and independent.

So, one way to deal with the above integral is by approximating it using Monte Carlo approximations, i.e. using $E\left((x_1+x_2-x_1x_2)^{-M_0-2M_1}\right)=\frac{1}{n}\sum_{i=1}^n\left(x_1^{(i)}+x_2^{(i)}-x_1^{(i)}x_2^{(i)}\right)^{-M_0-2M_1}$,

where $(x_1^{(i)},x_2^{(i)}) \overset{iid}{\sim} Be(N_1-\lambda_1+M_0+M_1,\lambda_1+M_1)\times Be(N_2-\lambda_2+M_0+M_1,\lambda_2+M_1)$ and $x_1^{(i)},x_2^{(i)}$ are independent for all $i=1,2,\ldots,n$. This method is relatively easy to formulate. On the other hand, it is an approximation, and there is a trade-off between the accuracy of this approximation and the speed of the constructed algorithm. If more samples $x_1^{(i)},x_2^{(i)}$ are used, we will get a better approximation, but the running time will increase considerably.

The mix-split step can therefore be performed as before (using any of the two mix-split methods described), with the integral shown here being approximated using Monte Carlo simulations. The other quantities in the mix-split acceptance probabilities are the same as before.

An alternative way to proceed would be not to integrate out γ_0,γ_1 and γ_2 in the mix-split step.In such a case, we consider updating $\gamma_0,\gamma_1,\gamma_2$, together with the update of the indicators r_{ji} and s_{ji}

induced by the mix-split step, and only the cluster locations ϕ_{ji} are integrated out. So, in the MCMC algorithm, we need to update the ϕ_{ji}'s just after the mix-split step, whereas the $\gamma's$ can still be updated using the slice sampler, in addition to the proposed update of them in the mix-split step.

The difficult part in this approach is not calculating an integral, but finding a sensible way to propose values for the γ's, which will be consistent with the proposed mixing or splitting of clusters. One way would be to set $\varepsilon_1' = \frac{\gamma_0'}{\gamma_0'+\gamma_1'} = \frac{n_1'}{N_1}$ and $\varepsilon_2' = \frac{\gamma_0'}{\gamma_0'+\gamma_2'} = \frac{n_2'}{N_2}$, where the superscript $'$ denotes the situation after (and if) the mix/split step is accepted. We then need to think of an updating proposal for either γ_0, γ_1 or γ_2, for example $\gamma_0' \sim Ga(M_0, 1)$ (i.e. like its prior). It is clear that, in this proposal, the only stochastic proposal concerning the γ's is the last one, the update of γ_0. The updating of $\varepsilon_1, \varepsilon_2$ is deterministic, given the indicators. This simplifies the calculation of the acceptance probabilities, however it can lead to poor mixing of the algorithm. A slight variation of this would be to use some distribution for updating also ε_1 and ε_2, centered at $\frac{n_1'}{N_1}$ and $\frac{n_2'}{N_2}$ respectively.

Alternatively, one could propose MH updates for γ_0, γ_1 and γ_2, independently of the proposed split/merge of the clusters (but still within the mix-split step). Again, the calculation of the acceptance probabilities is easy, but the mixing also depends on the proposed values for the γ's.

3.6.2 An alternative slice sampler

An alternative slice sampler for updating γ_0, γ_1 and γ_2 can be constructed using the identity

$$\int_0^\infty e^{-at} dt = 1/a.$$

So,

$$f(\gamma_0, \gamma_1, \gamma_2 | \cdots) \propto \gamma_0^{M_0-1} e^{-\gamma_0} \gamma_1^{M_1-1} e^{-\gamma_1} \gamma_2^{M_1-1} e^{-\gamma_2} \int\int \ldots \int \gamma_0 e^{-(\gamma_0+\gamma_1)U_1} \ldots \gamma_0 e^{-(\gamma_0+\gamma_1)U_{N_1-\lambda_1}} \times$$
$$\gamma_1 e^{-(\gamma_0+\gamma_1)U_{N_1-\lambda_1+1}} \ldots \gamma_1 e^{-(\gamma_0+\gamma_1)U_{N_1}} \gamma_0 e^{-(\gamma_0+\gamma_2)U_{N_1+1}} \ldots \gamma_0 e^{-(\gamma_0+\gamma_2)U_{N_1+N_2-\lambda_2}} \times$$
$$\gamma_2 e^{-(\gamma_0+\gamma_2)U_{N_1+N_2-\lambda_2+1}} \ldots \gamma_1 e^{-(\gamma_0+\gamma_2)U_{N_1+N_2}} dU_1 dU_2 \ldots dU_{N_1+N_2}.$$

Therefore, consider the augmented vector of parameters $(\gamma_0, \gamma_1, \gamma_2, U_1, U_2, \ldots, U_{N_1+N_2})$ with full conditional distribution:

$$f(\gamma_0, \gamma_1, \gamma_2, U_1, \ldots, U_{N_1+N_2} | \cdots) \propto \gamma_0^{M_0+N_1-\lambda_1+N_2-\lambda_2-1} e^{-\gamma_0} \gamma_1^{M_1+\lambda_1-1} e^{-\gamma_1} \gamma_2^{M_1+\lambda_2-1} e^{-\gamma_2} \times$$
$$e^{-(\gamma_0+\gamma_1)U_1} \ldots e^{-(\gamma_0+\gamma_1)U_{N_1-\lambda_1}} e^{-(\gamma_0+\gamma_1)U_{N_1-\lambda_1+1}} \ldots e^{-(\gamma_0+\gamma_1)U_{N_1}} e^{-(\gamma_0+\gamma_2)U_{N_1+1}} \ldots e^{-(\gamma_0+\gamma_2)U_{N_1+N_2-\lambda_2}} \times$$
$$e^{-(\gamma_0+\gamma_2)U_{N_1+N_2-\lambda_2+1}} \ldots e^{-(\gamma_0+\gamma_2)U_{N_1+N_2}}.$$

In this case, the full conditional distributions are of known form:

$U_i|\cdots \sim \text{Exp}(\gamma_0+\gamma_1),\ i=1,2,\ldots,N_1$

$U_i|\cdots \sim \text{Exp}(\gamma_0+\gamma_2),\ i=N_1+1,N_1+2,\ldots,N_1+N_2$

$\gamma_0|\cdots \sim Ga(M_0+N_1-\lambda_1+N_2-\lambda_2, 1+\sum U_i)$, where the sum is taken over all i

$\gamma_1|\cdots \sim Ga(M_1+\lambda_1, 1+\sum U_i)$, where the sum is taken over $i=1,2,\ldots,N_1$

$\gamma_2|\cdots \sim Ga(M_1+\lambda_2, 1+\sum U_i)$, where the sum is taken over $i=N_1+1,N_1+2,\ldots,N_1+N_2$.

In the above $\text{Exp}(\theta)$ denotes the exponential distribution with mean $1/\theta$:

Definition 13. *A random variable X is said to follow an exponential distribution with parameter $\theta>0$, denoted by Exp(θ), if its density with respect to the Lebesque measure is:*

$$f_X(x)=\theta e^{-\theta x},\ x>0.$$

Regardless of the fact that now there are N_1+N_2 auxiliary variables, instead of just seven before, the simulation time is not increased substantially, since the full conditional distributions of these auxiliary variables are of known form, and therefore easy to sample from. On the other hand, this slice method has the advantage of a better mix-split step. The idea is that, by incorporating these auxiliary variables in the mix-split step, the problematic integral $I=\int f(\boldsymbol{c},\gamma_0,\gamma_1,\gamma_2|\boldsymbol{M})d\gamma_0,d\gamma_1,d\gamma_2$ will be replaced by

$I'=\int f(\boldsymbol{c},\gamma_0,\gamma_1,\gamma_2,\boldsymbol{U}|\boldsymbol{M})d\gamma_0,d\gamma_1,d\gamma_2$, where $\boldsymbol{U}$ is the vector of all $U_i,\ i=1,\ldots,N_1+N_2$. This integral can now be solved analytically:

$I'\propto \int f(\boldsymbol{r}|\gamma_0,\gamma_1,\gamma_2)f(\gamma_0,\gamma_1,\gamma_2|\boldsymbol{M})f(\boldsymbol{U}|\gamma_0,\gamma_1,\gamma_2)d\gamma_0,d\gamma_1,d\gamma_2$

$$=\int_0^\infty\int_0^\infty\int_0^\infty \gamma_0^{M_0+N_1+N_2-\lambda_1-\lambda_2-1}\gamma_1^{M_1+\lambda_1-1}\gamma_2^{M_1+\lambda_2-1}e^{-\gamma_0(1+\sum U_i)}e^{-\gamma_1(1+\sum U_j)}e^{-\gamma_2(1+\sum U_l)}d\gamma_0d\gamma_1d\gamma_2$$

$$=\int_0^\infty \gamma_0^{M_0+N_1+N_2-\lambda_1-\lambda_2-1}e^{-\gamma_0(1+\sum U_i)}d\gamma_0\int_0^\infty \gamma_1^{M_1+\lambda_1-1}e^{-\gamma_1(1+\sum U_j)}d\gamma_1\int_0^\infty \gamma_2^{M_1+\lambda_2-1}e^{-\gamma_2(1+\sum U_l)}d\gamma_2$$

$$\Rightarrow I'\propto \frac{\Gamma(M_0+N_1+N_2-\lambda_1-\lambda_2)}{(1+\sum U_i)^{M_0+N_1+N_2-\lambda_1-\lambda_2}}\frac{\Gamma(M_1+\lambda_1)}{(1+\sum U_j)^{M_1+\lambda_1}}\frac{\Gamma(M_1+\lambda_2)}{(1+\sum U_l)^{M_1+\lambda_2}}.$$

Here, the first sum is taken over $i=1,2,\ldots,N_1,N_1+1,\ldots,N_1+N_2$, the second one over $j=1,2,\ldots,N_1$ and the last one over $l=N_1+1,N_1+2,\ldots,N_1+N_2$. The proportionality constant is, as before, $f(\boldsymbol{s}|\boldsymbol{r},\boldsymbol{M})$. If we include the expression for it, we will have:

$$I'=\frac{M_0^{k_0}M_1^{k_1+k_2}\Gamma(M_0)\Gamma^2(M_0)}{(1+\sum U_i)^{M_0+N_1+N_2-\lambda_1-\lambda_2}(1+\sum U_j)^{M_1+\lambda_1}(1+\sum U_l)^{M_1+\lambda_2}}\prod_{ji}\Gamma(n_{ji}),$$

where n_{ji} are the cluster sizes.

It is also clear that this trick is based exactly on the definition of these auxiliary variables U_i. Notice also that a similar approach using the auxiliary variables in the first slice sampler is not possible.

To sum up, this proposed algorithm is superior to the ones proposed before, since now the acceptance probabilities of each mix or split proposal can be exactly calculated. In this way, no MC estimation is needed, which results in both error due to approximation and slowing of the algorithm, nor we need to also propose updating of additional parameters, as when values for the γ_i were proposed. The fact that we now have an extended parameter space, with the addition of the U_i's does not change much in terms of coding burden and computational cost, since they are only used in the update of the γ_i and in the mix-split step, and even then only through the corresponding sums. Since their full conditional distributions are exponential distributions, it is computationally not expensive. Also notice that, since the size of these auxiliary variables is equal to the data size, there will not be any additional problems of variant dimensionality of the parametric space.

Finally, I tested the above algorithms with the following simulated data (say, fourth simulated data set):

$$Y_{1i} \sim \frac{3}{10}N(-10,1) + \frac{7}{10}N(1,1),\ i = 1,2,\ldots,120$$

$$Y_{2i} \sim \frac{7}{10}N(8,1) + \frac{3}{10}N(1,1),\ i = 1,2,\ldots,120.$$

As one would expect, the second slice sampler with the mix-split step where we can calculate exactly the acceptance probabilities (say, method A) performed best, giving the expected results, with good mixing of the chain and in a reasonable amount of time. On the other hand, the method using Monte Carlo estimates (method C) takes a lot more time to run, because of these approximations, and the results were not as good as in the last method shown. This could be improved by taking more than 20000 MC samples for each integral (in each cycle of the MCMC), but this would cause even higher running time. As for the second method proposed, i.e. not integrating out γ_0, γ_1 and γ_2 (method B), the results were better (in terms of mixing) and the algorithm was running faster than method C, when gamma proposals for those parameters were used, but still not as good as method A. Another issue is that the method A does not require any kind of tuning, as is the case for the number of MC samples in method C and of the method of proposing values for the γ's in method B.

As a result, I present the results when method A was used. The posterior distributions and trace plots for the two weights are shown in Figure 3.25. It can be seen that we have a large mode at the value indicated by the data (i.e. 0.3 for the first and 0.7 for the second data set), a smaller mode at 0 and a very small mode at 1. The mode at one might seem unjustified at first, but it makes sense if one observes from the trace plots that it only occurs when the other weights is zero. This zero weight in one of the two component distributions translates into no sharing of information between the two correlated distributions, therefore allowing more flexibility to the other weight to take values

close to the correct weight (here: 0.7), 0 or 1. The mass at 1 would normally correspond to 0, if there was not this non-identifiability caused by the fact that the other weight is zero.

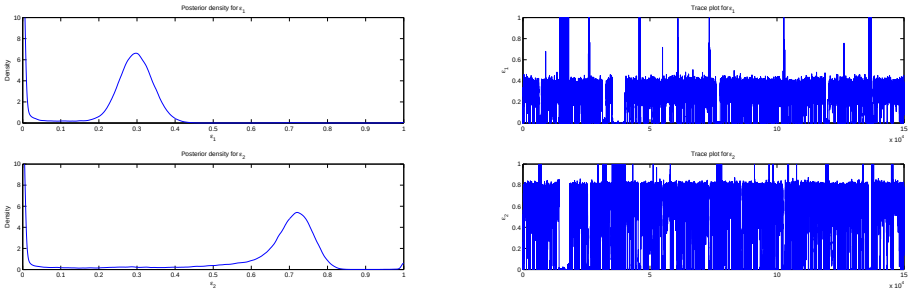


Figure 3.25: Posterior distributions (left) and trace plots (right) for ε_1 (top), ε_2 (bottom) for the fourth simulated data set, based on results using method A.

3.7 The Model With Normalised Inverse-Gaussian Process Priors

In this section I discuss the simulation of a model similar to my proposed model (2.1.4), but with N-IGP priors for the component distributions F_0, F_1 and F_2, instead of DP priors. The full hierarchical model, corresponding to the setting of Section 3.1 above, will be:

$$
\begin{gathered}
Y_{ji} \sim N(\mu_{ji}, S),\ i = 1, 2, \ldots, N_j,\ j = 1, 2 \\
\mu_{ji} \sim F_j^*,\ \text{where } F_j^* = \varepsilon F_0 + (1-\varepsilon)F_j,\ j = 1, 2 \\
F_0 \sim \text{N-IGP}(M_0, H),\ F_1, F_2 \overset{iid}{\sim} \text{N-IGP}(M_1, H),\ \text{where } H \equiv N(m, B) \qquad (3.7.2) \\
\varepsilon \sim \text{N-IG}(M_0, M_1) \\
M_0, M_1 \overset{iid}{\sim} Ga(a_0, b_0),\ (m, B) \sim N(m_0, A) \times \text{IGa}(c, 1/cC),\ S \sim \text{IGa}(q, 1/qR)
\end{gathered}
$$

where N-IGP denotes the normalised inverse-Gaussian process and N-IG denoted the normalised inverse-Gaussian distribution. Notice also that for the specific distribution of the weight, it is guaranteed that F_1^* and F_2^* are also marginally N-IGP-distributed, with parameters M_0+M_1 and H. The differences of the above model from Model (2.1.4) are the priors of the $F_j,\ j = 0, 1, 2$ and the prior of the weight. Therefore, the differences in the MCMC algorithm will be:

1. The update of ε.
2. The updates of M_0, M_1, since the prior of the weight involves those two quantities.

3. The updates of the indicator $s_{ji}, r_{ji},\ i = 1, 2, \ldots, N_j,\ j = 1, 2$.

4. The acceptance probabilities of the mix-split step (if we decide to include this extra step).

The first differences do not cause much additional trouble. In particular:

$$\begin{aligned} f(\varepsilon|\cdots) &\propto f(\varepsilon|M_0, M_1) f(\boldsymbol{r}|\varepsilon) \\ &\propto \frac{K_{-1}\left(\sqrt{A_2(\varepsilon, M_0, M_1)}\right)}{\varepsilon^{3/2}(1-\varepsilon)^{3/2}\sqrt{A_2(\varepsilon, M_0, M_1)}} \varepsilon^{N_1+N_2-\sum r_{ji}} (1-\varepsilon)^{\sum r_{ji}} \\ &= \frac{K_{-1}\left(\sqrt{A_2(\varepsilon, M_0, M_1)}\right)}{\sqrt{M_0^2(1-\varepsilon) + M_1^2\varepsilon}} \varepsilon^{N_1+N_2-\sum r_{ji}-1} (1-\varepsilon)^{\sum r_{ji}-1} \end{aligned}$$

where $A_2(\varepsilon, M_0, M_1) = \frac{M_0^2}{\varepsilon} + \frac{M_1^2}{1-\varepsilon}$, K_{-1} is the modified Bessel function of the third type and the sums of r_{ji} are taken over all j, i.

We can see here that, unlike the beta prior for the weight in the case of the DP priors, here we do not have conjugacy of the full conditional of ε. So, MH updates are used.

The full conditionals of the concentration parameters will be:

$$\begin{aligned} f(M_0, M_1|\cdots) &\propto f(M_0, M_1) f(\varepsilon|M_0, M_1) f(\boldsymbol{s}|\boldsymbol{r}, M_0, M_1) \\ &\propto M_0^{a_0-1} e^{-\frac{M_0}{b_0}} M_1^{a_0-1} e^{-\frac{M_1}{b_0}} \frac{K_{-1}\left(\sqrt{A_2(\varepsilon, M_0, M_1)}\right) M_0 M_1 e^{M_0+M_1}}{\sqrt{A_2(\varepsilon, M_0, M_1)}} \times \\ &\quad \frac{\Gamma(M_0) M_0^{K_0}}{\Gamma(M_0+n_0)} \frac{\Gamma(M_1) M_1^{K_1}}{\Gamma(M_1+n_1)} \frac{\Gamma(M_1) M_1^{K_2}}{\Gamma(M_1+n_2)} \\ &= M_0^{a_0+K_0} e^{-(1/b_0-1)M_0} M_1^{a_0+K_1+K_2} e^{-(1/b_0-1)M_1} \frac{K_{-1}\left(\sqrt{A_2(\varepsilon, M_0, M_1)}\right)}{\sqrt{A_2(\varepsilon, M_0, M_1)}} \times \\ &\quad \frac{\Gamma(M_0)}{\Gamma(M_0+n_0)} \frac{\Gamma(M_1)}{\Gamma(M_1+n_1)} \frac{\Gamma(M_1)}{\Gamma(M_1+n_2)}. \end{aligned}$$

So,

$$f(M_0|\cdots) \propto M_0^{a_0+K_0} e^{-(b_0-1)M_0} \frac{K_{-1}\left(\sqrt{A_2(\varepsilon, M_0, M_1)}\right)}{\sqrt{A_2(\varepsilon, M_0, M_1)}} \frac{\Gamma(M_0)}{\Gamma(M_0+n_0)}$$

and

$$f(M_1|\cdots) \propto M_1^{a_0+K_1+K_2} e^{-(b_0-1)M_1} \frac{K_{-1}\left(\sqrt{A_2(\varepsilon, M_0, M_1)}\right)}{\sqrt{A_2(\varepsilon, M_0, M_1)}} \frac{\Gamma(M_1)}{\Gamma(M_1+n_1)} \frac{\Gamma(M_1)}{\Gamma(M_1+n_2)}.$$

These expressions are not of a standard distribution form, so we use MH steps to update M_0 and M_1. All three full conditionals above include the modified Bessel function of the third type. This is a standard special function, with in-built functions in Matlab, which seem to work fine.

Unfortunately, this is not the case for the incomplete gamma function $\Gamma(a, x)$, when $a < 0$. Such functions appear in the quantities w_0 and w_1, involved in the Pólya-urn representation of the

N-IGP, as shown in Section 1.1.3. As a result, the probabilities that appear in the updating steps for each pair of indicators (s, r) cannot be calculated efficiently. Therefore, I was not able to code the above model in Matlab, at least not using the marginal method that I was using up to this point. A solution to this is to use slice sampling ideas, and more specifically the ideas appearing in Walker (2007), Kalli et al. (2008) and Griffin and Walker (2009):
Consider RPMs that can be written in an infinite sum expression of the form

$$G = \sum_{j=1}^{\infty} w_j \delta_{\theta_j}$$

where the weights w_j are positive quantities summing up to 1. An example of such a RPM is the DP (Sethuraman and Tiwari (1982) and Sethuraman (1994)) and, more generally, the class of normalised random measures (see, for example, James et al. (2005)), which also includes the N-IGP. Then a mixture model (Lo, 1984) with such a RPM as the mixing distribution will be of the form:

$$f_G(y) = \int h(y|\theta) dG(\theta) = \sum_{j=1}^{\infty} w_j h(y|\theta_j), \tag{3.7.3}$$

where h is a density function (usually continuous).
The basic idea is that one can replace the weights w_j appearing in the last expression with the indicator function of an auxiliary, uniformly distributed random variable being less than this weight, $I(U < w_j)$. This is the same as extending $f_G(y)$ above to $f_G(y, U) = \sum_{j=1}^{\infty} 1_{(U<w_j)} h(y|\theta_j)$. In the case of NRMs (in which the realisations of the N-IGP belong), the weights can be written as $w_j = \frac{J_j}{\sum_{i=1}^{\infty} J_i}$, $j = 1, 2, \ldots$, where J_i, $i = 1, 2, \ldots$ are jump sizes from a Lévy process, with finite sum. As a result, we can now write $f_G(y, u) = \frac{1}{J} \sum_{j=1}^{\infty} 1_{(U<J_j)} h(y|\theta_j)$, where $J = \sum_{i=1}^{\infty} J_i$. The authors then propose two more auxuliary variables: s, which is an indicator of the cluster in which the observation is assigned (i.e. the same as the indicators s_{ji} used throughout so far) and v, which is an exponentially-distributed auxiliary variable, used to make the algorithm more efficient (Nieto-Barajas et al. (2004)). The joint posterior of those parameters will be:

$$f_G(y, U, v, s) = e^{-vJ} 1_{(U<J_s)} h(y|\theta_s).$$

The marginal distribution of y from this joint distribution will be, of course, (3.7.3).

The above method is a slice sampler and, since in each cycle of the MCMC only a finite number of those weights will be needed, simulating from the posterior distributions of all parameters (except, of course, G itself) is possible. Notice also that here the RPM G is not integrated out, so this method is a conditional algorithm.

As an example of this method, consider simulating from the posterior distributions of the parameters in Model (3.7.2). For this model three v-like auxiliary variables are needed, say v_0, v_1 and v_2,

each corresponding to component distribution $F_j,\ j=0,1,2$, respectively. We also need three vectors of jumps, say $\boldsymbol{J_0}, \boldsymbol{J_1}$ and $\boldsymbol{J_2}$ (where $\boldsymbol{J_j} = (J_{j1}, J_{j2}, \ldots),\ j=0,1,2$), again each corresponding to component distribution $F_j,\ j=0,1,2$. Let also $\boldsymbol{J}$ denote the vector of all those jumps in all component distributions and $J_j = \sum_{i=1}^{\infty} J_{ji},\ j=0,1,2$. Notice also that in the case of the N-IGP, the *a priori* Lévy density of these jumps is

$$w(x) = \frac{M}{\sqrt{\pi}} x^{-3/2} e^{-x}. \tag{3.7.4}$$

An upper truncation point for all the jumps J_{ji}, say L, is also introduced. Using this L there will only be a finite number of jumps involved in this algorithm, say K_0^*, K_1^* and K_2^* for each component distribution. As for the rest of the jumps, we can integrate them out.

The joint full conditional distribution of all parameters in the model is as follows:

$$\begin{aligned} f(\boldsymbol{s},\boldsymbol{r},\boldsymbol{\phi},\boldsymbol{U},\boldsymbol{J},v_0,v_1,v_2,S,m,B,\varepsilon,M_0,M_1|\boldsymbol{Y}) \quad &\propto \quad f(\boldsymbol{Y}|\boldsymbol{s},\boldsymbol{r},\boldsymbol{\phi},S) f(\varepsilon|M_0,M_1) f(M_0,M_1) f(\boldsymbol{\phi}|m,B) \\ &\times \quad f(\boldsymbol{J_0}|M_0) f(\boldsymbol{J_1}|M_1) f(\boldsymbol{J_2}|M_1) f(\boldsymbol{U}) f(\boldsymbol{r}|\varepsilon) \\ &\times \quad f(\boldsymbol{s}|\boldsymbol{r},\boldsymbol{U},\boldsymbol{J}) \prod_{j=0}^{2} f(v_j|\boldsymbol{s},\boldsymbol{r},\boldsymbol{J_j}) f(S) f(m,B) \end{aligned}$$

where the last product is proportional to $\prod_{j=0}^{2} \frac{v_j^{n_j-1}}{\Gamma(n_j)} e^{-v_j J_j}$, $\boldsymbol{U}$ is the vector of all $U_{ji}, i=1,2,\ldots,N_j,\ j=1,2$ and the rest is as before (see e.g. Section 3.2).

The full conditional distributions of all parameters will be as follows:

- $f(\varepsilon|\cdots) \propto \varepsilon^{n_0-1}(1-\varepsilon)^{n_1+n_2-1} \frac{K_{-1}\left(\sqrt{\frac{M_0^2}{\varepsilon}+\frac{M_1^2}{1-\varepsilon}}\right)}{\sqrt{M_0^2(1-\varepsilon)+M_1^2\varepsilon}}$.
- $f(M_0|\cdots) \propto M_0^{a_0+K_0^*} \frac{K_{-1}\left(\sqrt{\frac{M_0^2}{\varepsilon}+\frac{M_1^2}{1-\varepsilon}}\right)}{\sqrt{M_0^2(1-\varepsilon)+M_1^2\varepsilon}} e^{-M_0\left(\int_L^{\infty} q(x)dx + \int_0^L (1-e^{-v_0 x}) q(x)dx - 1 + 1/b_0\right)}$.
- $$\begin{aligned} f(M_1|\cdots) \quad &\propto \quad M_1^{a_0+K_1^*+K_2^*} \frac{K_{-1}\left(\sqrt{\frac{M_0^2}{\varepsilon}+\frac{M_1^2}{1-\varepsilon}}\right)}{\sqrt{M_0^2(1-\varepsilon)+M_1^2\varepsilon}} \times \\ &\quad e^{-M_1\left(\int_L^{\infty} q(x)dx + \int_0^L (1-e^{-v_1 x}) q(x)dx + \int_L^{\infty} q(x)dx + \int_0^L (1-e^{-v_2 x}) q(x)dx - 1 + 1/b_0\right)}. \end{aligned}$$
- For each pair $(s_{ji}, r_{ji}),\ i=1,2,\ldots,N_j,\ j=1,2$, we have:
  $$P(s_{ji}=k, r_{ji}=l|\cdots) \propto \begin{cases} \varepsilon\varphi(Y_{ji};\phi_{0l},S) 1_{(U_{ji}<J_{0l})} \frac{v_0}{n_0^-} & , k=1,2,\ldots,K_0^*,\ l=0 \\ (1-\varepsilon)\varphi(Y_{ji};\phi_{jl},S) 1_{(U_{ji}<J_{jl})} \frac{v_j}{n_j^-}, & k=1,2,\ldots,K_j^*,\ l=1. \end{cases}$$
- For each $v_j,\ j=0,1,2$, we have:
  $f(v_j|\cdots) \propto v_j^{n_j-1} e^{-v_j \sum_{k=1}^{K_j^*} J_{jk}} e^{-\int_0^L (1-e^{-v_j x}) w(x)dx)}$

- For all $i = 1, 2, \ldots, N_i,\ j = 1, 2,\ U_{ji}|\cdots \sim \begin{cases} U(0, J_{0,s_{ji}}) & , \text{ if } r_{ji} = 0 \\ U(0, J_{j,s_{ji}}) & , \text{ if } r_{ji} = 1. \end{cases}$

- For the jumps J_{jk} with at least one observation allocated to it (i.e. $n_{kl} > 0$), we have: $J_{jk}|\cdots \sim \text{Ga}(n_{jk} - 0.5, 1 + v_j),\ j = 0, 1, 2.$
  For the jumps J_{kl} with no observations allocated to them, see Kalli et al. (2008).

- The full conditional distributions of m, B, S and $\boldsymbol{\phi}$ and the corresponding updating schemes are as in the previous models.

In the above, K_{-1} is the modified Bessel function of the third type, φ is the pdf of a normal distribution and $q(x) = w(x)/M$, where w(x) is as in (3.7.4).
We can update $U_{ji}, (s_{ji}, r_{ji})$ and the jumps with observations allocated to them using Gibbs sampling. For ε, M_0, M_1 and v_j we can use MH updating steps.

Finally, it is worth mentioning a result that arose when I was trying to calculate the acceptance probabilities of the mix-split step. This result served as my motivation for some interesting results that appear in the next chapter.
Specifically, when trying to integrate out the weight ε, the following integral appears:

$$\begin{aligned} I_1 &= \int_0^1 f(\boldsymbol{r}|\varepsilon) f(\varepsilon|M_0, M_1) d\varepsilon \\ &\propto \int_0^1 \frac{K_{-1}\left(\sqrt{A_2(\varepsilon, M_0, M_1)}\right)}{\sqrt{M_0^2(1-\varepsilon) + M_1^2\varepsilon}} \varepsilon^{N_1+N_2-\sum r_{ji}-1}(1-\varepsilon)^{\sum r_{ji}-1} d\varepsilon. \end{aligned}$$

The last integral cannot be solved analytically, except for very special cases. On the other hand, one might notice that I_1 can be written in moment form:

$$\begin{aligned} I_1 &= E_\varepsilon\left(\varepsilon^{N_1+N_2-\lambda}(1-\varepsilon)^{\lambda}\right) \\ &= E_\varepsilon\left(\varepsilon^{N_1+N_2-\lambda}\sum_{k=0}^{\lambda}\binom{\lambda}{k}(-1)^k\varepsilon^k\right) \\ &= \sum_{k=0}^{\lambda}\binom{\lambda}{k}(-1)^k E_\varepsilon\left(\varepsilon^{N_1+N_2+k-\lambda}\right) \end{aligned}$$

where $\lambda = \sum_{j,i} r_{ji}$ and ε follows a N-IG distribution.
So, what we just need is an expression for the moments of a N-IG-distributed random variable. Fortunately, I was able to derive these expressions, as well as the expressions for the moments for a more general class of distributions. The results are in Chapter 5, together with some notes on implementing these expressions in Mathematica.

3.8 Summary

In this chapter I described the MCMC algorithms used in order to simulate from the posterior distribution of the parameters of the models in Section 2. For my basic proposed model the algorithm was similar to the one described in Müller et al. (2004), whereas for the model via direct normalisation slice sampling methods were used. By considering simulated data sets, I then observed some problems in the mixing of the chains, so an additional step in each of these algorithms was proposed. This extra step consists of splitting a cluster to two others or merging two clusters together. It was shown that this extra step indeed improves mixing, especially in specific cases that were highlighted. The idea of this extra step is quite general, therefore it can potentially be applied to other MCMC algorithms, as well. I also discussed implementation of the models for three correlated distributions and of the model similar to my basic proposed model, only this time with N-IGP priors, instead of DP priors. For the latter model, the slice sampler of Griffin and Walker (2009) for simulating RPMs that have an infinite sum representation was used.

Chapter 4

Applications

In this chapter I apply some of the models presented in the previous chapters to real-life data. First, some financial data are considered, which are modelled as coming from two correlated distributions, using my basic proposed model and the model of Müller et al. (2004). Next, I embed those two models, as well a model similar to the basic proposed model, but with N-IGP (instead of DP) priors for the correlated distributions, in the stochastic frontiers setting. The three derived models are then used in analysing some hospital cost frontier data.

4.1 Financial Data

4.1.1 Description of data

First, I apply some of the models discussed in Chapter 2, and especially my basic proposed model (2.1.4) and the equivalent Müller et al. (2004) model, i.e. Model (1.2.13), to financial data. More specifically, the data consist of the daily returns of two stocks of Dow Jones 30, Alcoa Inc. and Exxon Mobil Corp., for the period from the 11th of November, 1999, up to and including the 4th of November, 2003, as found in StatLib (http://lib.stat.cmu.edu/). This means that $J = 2$ and the data sizes are $N_1 = N_2 = 1000$.

4.1.2 Description of the models and the MCMC algorithms

In this approach, it is assumed *a priori* that the marginal distributions of the daily returns of the two stocks are identically distributed and dependent, as implied by the models used. Another approach in modelling the correlation between these data could be to use skewed multivariate distributions, for example see Ferreira and Steel (2007) and references therein.

For both models normal likelihood, normal base distribution and the same priors for the hyperparameters S, m and B as in Model (1.2.14) are assumed. For the model of Müller et al. (2004) the following values are used: $\pi_0 = \pi_1 = 0.1$ and $a_\varepsilon = b_\varepsilon = 1$ (so that $\text{Be}(a_\varepsilon, b_\varepsilon) \equiv U(0,1)$). For both models we set $a_0 = 0.5,\ b_0 = 2$, i.e. the concentration parameters follow a gamma distribution with mean 1 and variance 2. This prior favors small values for the M's, which is often the case in models of this type. The hyperparameters for m are set to $m_0 = 0$ and $A = 10$, which seem sensible after looking at the data (alternatively, m_0 could had been assigned the overall mean of the data), and the large variance ensures that the prior for m is not very informative. Next, a quite vague prior is assigned to the variance S of the likelihood, with $q = 0.01$ and $R = 10000$, i.e. an inverse gamma prior with parameters 0.01 and 0.01. Finally, for the variance B of the base distribution we set $c = 2.1$ and $C = 0.5$, resulting in an inverse gamma distribution with mean 2 and variance 7.5.

In the simulations performed the auxiliary indicator variables r_{ji} and s_{ji} were used, we worked with the concentration parameters M_0 and M_1 (and M_2, in the case of the model of Müller et al. (2004)), instead of their sum $x = M_0 + M_1$ and the ratio $y = \frac{M_0}{M_0+M_1}$ and the discrete values ϕ_{ji} were also updated. We will simulate from the two models both with and without the (basic) mix-split step of Section 3.3, in order to examine the effect of this additional step in the performance of the algorithm. A burn-in period of 50000 iterations was used (actually, even 40000 iterations seemed a sufficient burn-in period), and in each simulation an additional 150000 iterations were performed.

4.1.3 Results

The main parameters of interest here are the common weight ε and the concentration parameters of F_0, F_1 and F_2. It is also interesting to look at the predictive distribution in each of F_1^* and F_2^*, as well as the predictives of the component distributions F_0, F_1 and F_2. Secondary parameters of interest could be m, B and S, as well as the clusters sizes K_0, K_1 and K_2. The graphs of the posterior distributions of the parameters of interest were created using kernel density estimators with normal kernels and a choice of bandwidth based on trial and error.

Let us first consider the model of Müller et al. (2004), with and without the mix-split step. The first thing to notice is the effect of the extra mix-split step. As can be seen from the trace plots in Figure 4.1, mixing for the parameter ε is better when this extra step is applied, as one would expect. Although the improvement in mixing is not huge here, in the following I will report the results with this extra step. Looking at the kernel density estimate of the posterior distribution of the weight ε (Figure 4.2), we see a large mode around 0.7 and two smaller ones around 0 and 0.1 (the sizes of the smaller weights is actually slightly different without the mix-split step). The predictive densities of F_1^* and F_2^* are the same in both cases, and resemble the histogram of the actual data displayed

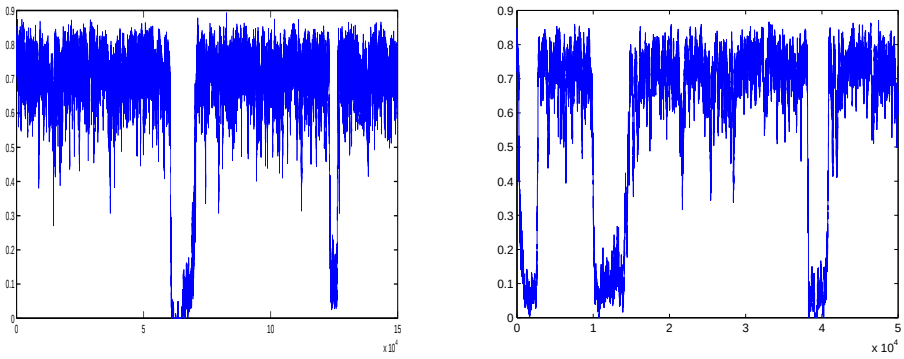


Figure 4.1: Trace plots for ε in Model (1.2.13), with (right) and without (left) the mix-split step.

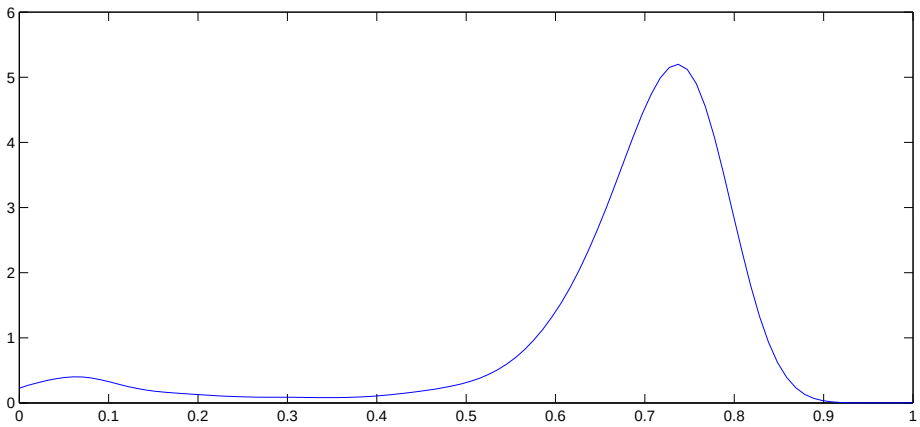


Figure 4.2: Posterior distribution of ε, Model (1.2.13).

next to them in Figure 4.3. Notice that, although one data point from the second data set (Exxon Mobil Corp) takes the value 7.88, in the predictive density of F_2^* we do not get a significant mode around that point (not shown in the graph). This is because only one observation is allocated to that (potential) cluster, and because the specific value is very far from the posterior base distribution. The lack of mode at that point allows us to focus on the values of the predictives in the interval (-5,7), for a better visual inspection of the results. The same reasoning applies when displaying the predictive densities of F_0, F_1, F_2, as well as in the other models.

The predictive densities of the component distributions are shown in Figure 4.4. For F_0, this predictive density is slightly positively skewed with one large mode around 0 and heavier than normal tails. F_2, on the other hand, seems fairly symmetric, with the same mode around 0. The

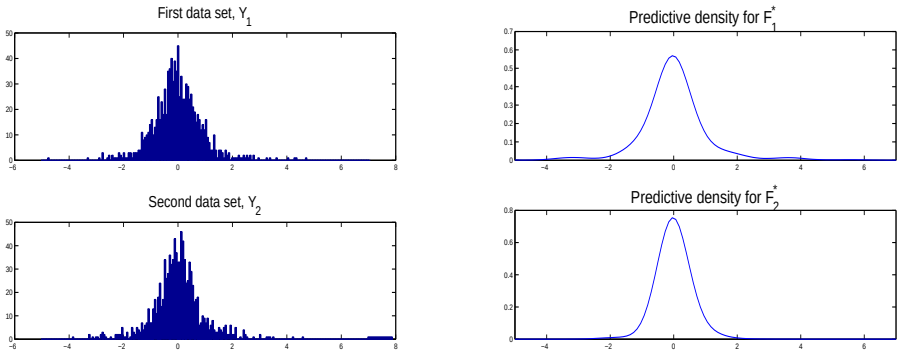


Figure 4.3: Histograms of the data (left) and predictive densities for F_1^* and F_2^* (right).

predictive density for F_1 exhibits some different features. The mode at 0 exists, but now we have two smaller modes, one around -3 and one around 3.5. The predictive density for F_1 also exhibits positive skewness. The differences in the predictive densities of F_1 and F_2 highlight the differences in the marginal distributions between the two stocks, which seem to be more extreme in the intervals $(-4, -2)$ and $(2, 5)$.

Next, consider the results for the posterior distributions of the precision parameters M_0, M_1 and M_2 (Figure 4.5). It is seen that M_1 takes significantly larger values than M_0 and M_2. This difference between M_2 and M_1 simply means that (*a posteriori*) F_1 is closer to the (common) base distribution H than F_2, and can also be seen as another aspect in which the two idiosyncratic parts F_1 and F_2 differ. Also notice that the posterior distributions of all three concentration parameters look very similar to a gamma distribution, indicating a prior-posterior accordance for these parameters. For comparison purposes, the prior distributions for the M's were also plotted in the same axes.
By looking at Figure 4.6, it is clear that there are many more clusters in F_1 than in F_0 or F_2. This result is in accordance with the fact that M_1 is larger than the other two (see, e.g. equation 10 in Escobar and West (1995)).
Finally, the posterior mean, median and and 95% credible interval for the posterior distributions of m, S and B are shown in Table 4.1.

Next, we investigate the use of my proposed model, Model (2.1.4), when applied to the same data set, again with and without the extra mix-split step in the MCMC algorithm.
In this case, the improvement of mixing caused by the extra mix-split step is more obvious, although the percentage of accepted split or merge steps was almost the same as before (10.5%). As seen in Figure 4.7, there is a substantial improvement in the mixing of ε when the split/merge step is used

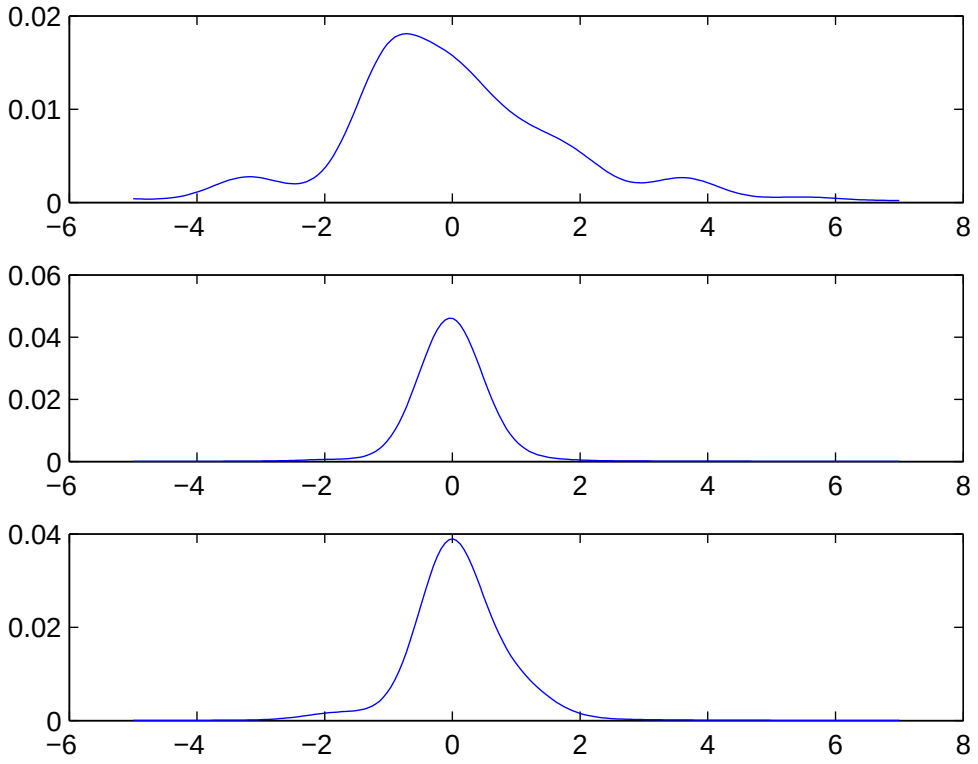


Figure 4.4: Predictive densities for F_1 (top), F_2 (middle) and F_0 (bottom) for Model (1.2.13).

in the algorithm. This improvement is more evident than in the case of the previous model, and I will therefore report the results when this step was included in the code. The posterior of ε is seen in the next graph (Figure 4.8) and it has a big mode around 0.7, and a smaller one around 0.1. The predictive densities for F_1^* and F_2^* were very similar to the ones in the case of Model (1.2.14). This is a sensible result, since we have 1000 data from each data set, therefore dominating the predictive distributions.

The predictives for the component distributions F_0, F_1 and F_2 (Figure 4.9) were also very similar to the previous ones, especially the predictive densities for F_0 and F_1. The predictive density for F_0 is also unimodal at 0 and slightly positively skewed. For F_1 there are three modes, a large one around -1 and two smaller around -3 and 3.5. The predictive density for F_2 is again unimodal around 0, but now with a little negative skewness, unlike the corresponding distribution for the model of Müller et al. (2004).

Figure 4.10 (left) shows the posterior distributions of M_0 and M_1 for this model. Notice that,

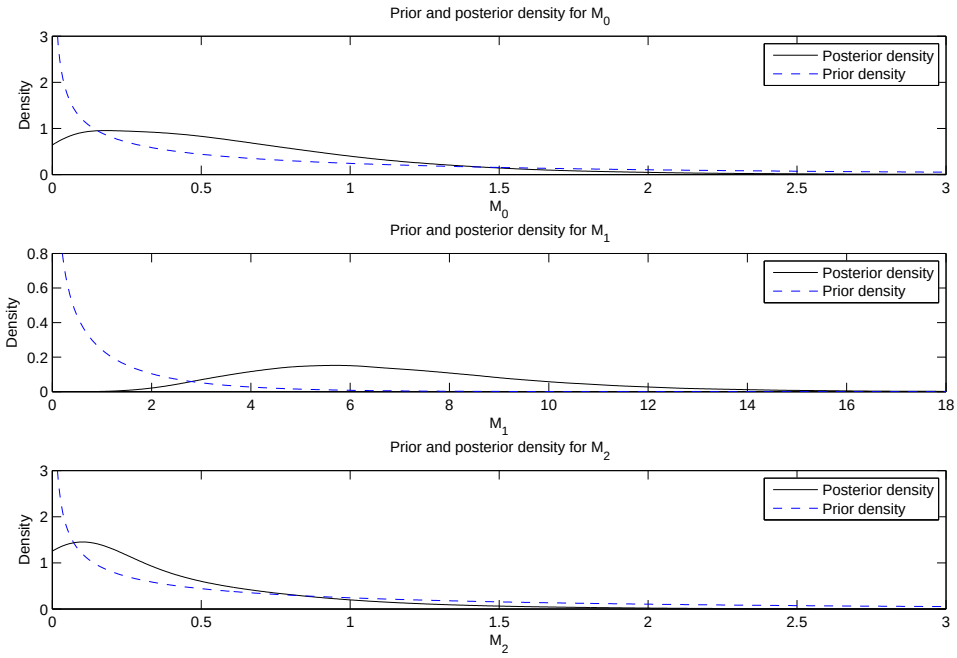


Figure 4.5: Prior (dashed line) and posterior (solid lines) distributions of M_0, M_1 and M_2 in Model (1.2.13).

without the extra step the results were similar. However, mixing was not as good as in the improved algorithm (i.e. like in the case of ε). These posteriors look different from the ones for the precision parameters M_0, M_1, M_2 used in the model of Müller et al. (2004). An apparent reason for that is that there are only two such parameters, rather than three. Another difference is the prior distribution of the weight ε, which involves the precision parameters only in model (2.1.4). On the other hand, the posterior distributions are gamma-like, indicating a prior-posterior accordance. It is also worth mentioning that M_0 takes smaller values than M_1.

In this model, it is also interesting to look at the posterior distributions of the reparametrisation $x = M_0 + M_1,\ y = \frac{M_0}{M_0+M_1}$. y can be seen as the prior mean of ε, as well as the prior correlation between $F_1^*(A)$ and $F_2^*(A),\ \forall\ A \in \mathcal{F}$. On the other hand, x can be thought of as a precision parameters of the prior distributions of $F_1^*(A)$ and $F_2^*(A)$. As seen in Figure 4.10 (right), the posterior of x is like a gamma distribution, with mean around 2.50, and y is a negatively skewed distribution, with mean around 0.20. In the same axes the prior distribution of all these quantities were also plotted, for prior-posterior comparison purposes.

Next, Figure 4.11 shows samples from the posterior distributions of the number of clusters in

	S	m	B
Mean	0.237	0.749	14.26
Median	0.236	0.740	13.38
2.5th percentile	0.189	-0.828	7.78
97.5th percentile	0.284	2.38	25.79

Table 4.1: Posterior mean, medians and 95% credible intervals for S, m and B in Model (1.2.13).

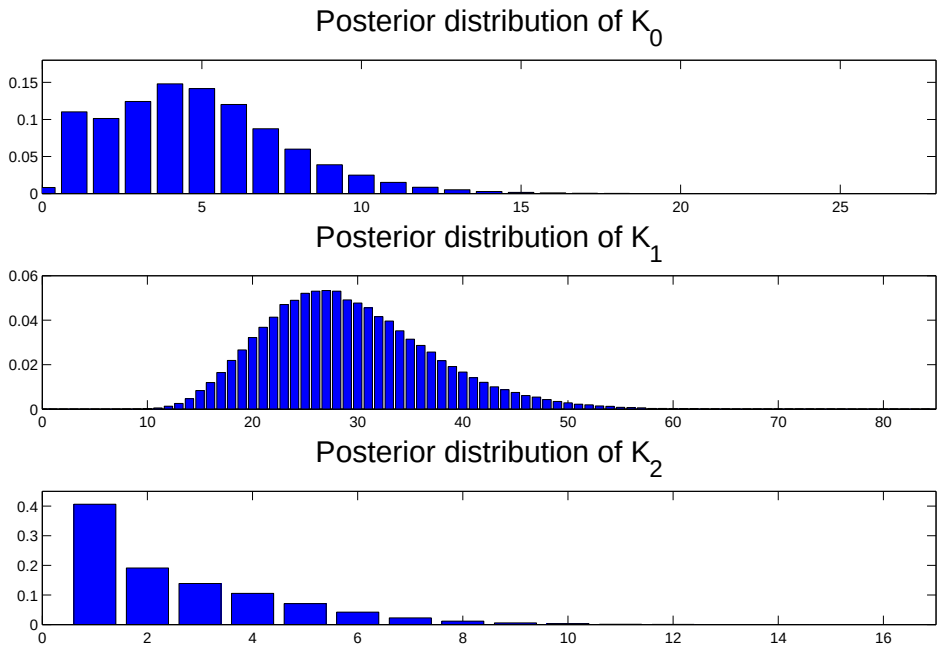


Figure 4.6: Posterior distributions of K_0, K_1 and K_2 for the Müller et al. (2004) model.

each component distribution, K_0, K_1 and K_2.

Finally, the posterior distributions of S, m and B for this model were very similar to the case of Model (1.2.13).

I conclude by summarizing the findings and mentioning the differences when the two models (the model of Müller et al. (2004) and model (2.1.4)) were fitted to the same data. The comments will be about the results with the mix-split step applied to the MCMC algorithms, since it has been shown that this extra step results in better mixing of the chain.

The predictive densities for the correlated distributions F_1^* and F_2^* were the same, as they are dominated by the large data size. This is more or less true for the predictive densities of the component distributions F_0 and F_2. On the contrary, the predictive density for F_1 was not the

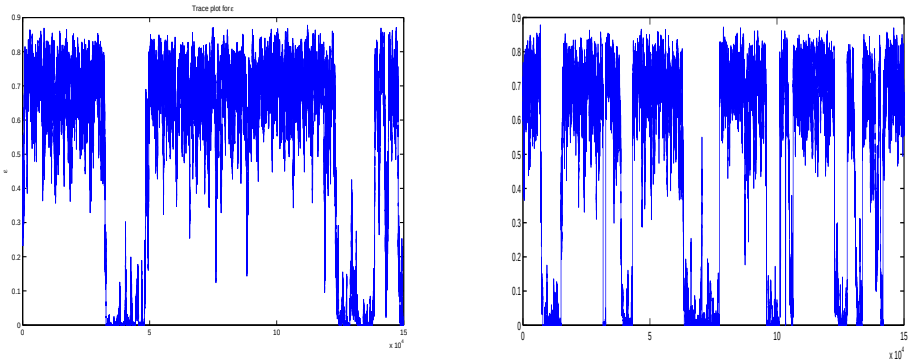


Figure 4.7: Trace plots for ε in Model (2.1.4) with (right) and without (left) the mix-split step.

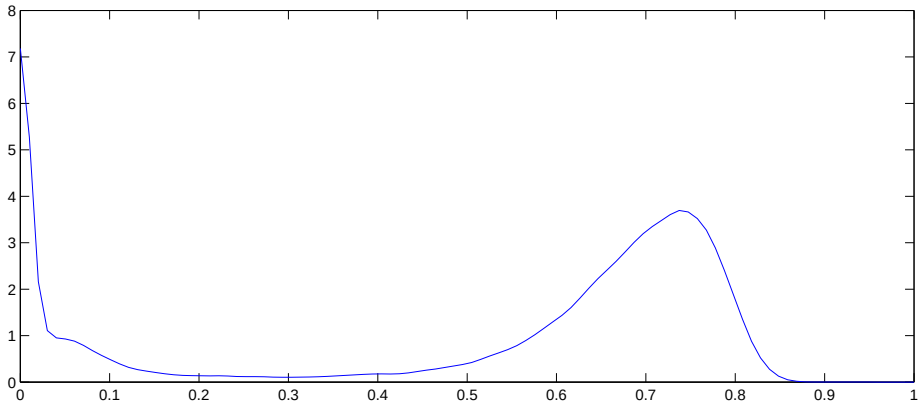


Figure 4.8: Posterior distribution of ε in Model (2.1.4).

same. However, the differences were mostly a matter of the sizes of the clusters for each component distribution.

Regarding the posterior distribution of the weight ε, the posterior modes are more or less the same. There are slight differences in the relative sizes of those modes. When the extra mix-split step was used, mixing of ε was good for both models.

There were some differences in the posterior distributions of the concentration parameters, as well. As mentioned before, direct comparison of those quantities is not straightforward, since in one model there are three of those parameters, whereas in the other one, only two. Apart from that, one can see that M_0 has more or less the same posterior in both models, whereas M_1 is much larger in the model of Müller et al. (2004) than in Model (2.1.4). A possible reason for this is that in the former case there are two concentration parameters for the idiosyncratic parts, M_1 and M_2, whereas in the

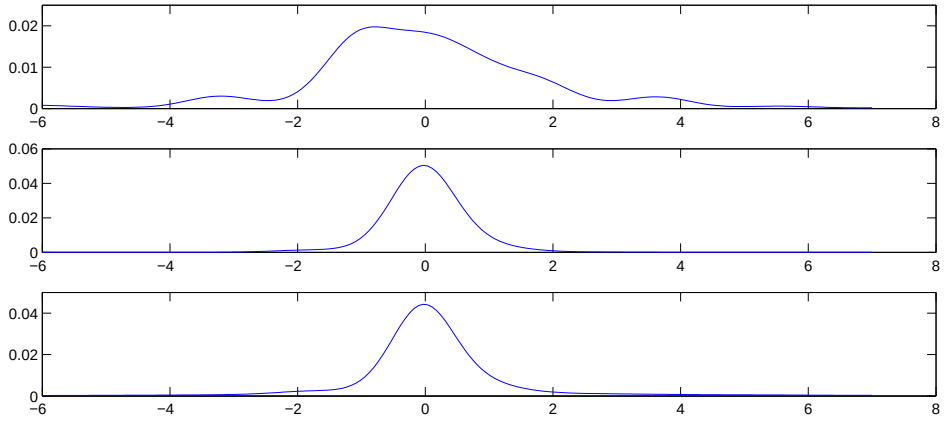


Figure 4.9: Predictive densities for F_1 (top), F_2 (middle) and F_0 (bottom) for Model (2.1.4).

latter case, there is only M_1. So, in one sense, M_1 in Model (2.1.4) accounts for both M_1 and M_2 of Model (1.2.13), and since M_2 is much smaller than M_1 there, this causes the "common" M_1 in Model (2.1.4) to take smaller values than M_1 in the other model.

These differences are consistent with the differences in the number of clusters in the component distributions, K_0, K_1 and K_2. We see that K_0 is slightly larger in Model (2.1.4), K_1 is much larger for Model (1.2.13) and K_2 is significantly larger for (2.1.4). In both cases K_1 is much larger than K_0 and K_2, K_0 is larger than K_2 for Model (1.2.13), whereas the opposite holds for model (2.1.4). The differences in the posterior medians of the precision parameters and of the number of clusters, as well as the correspondence of the magnitude of those two sets of parameters, can also be seen in Table 4.2.

Model	M_0	M_1	M_2	K_0	K_1	K_2
(1.2.13)	0.49	6.46	0.19	5	28	2
(2.1.4)	0.41	1.86	-	3	17	6

Table 4.2: Posterior median values for some parameters of interest for Models (1.2.13) and (2.1.4) applied to the financial data.

4.1.4 Comparison of the two models

In this subsection the two models are compared, both in terms of their predictive power and using Bayes factors. The results indicate that the two models perform similarly well for the specific data.

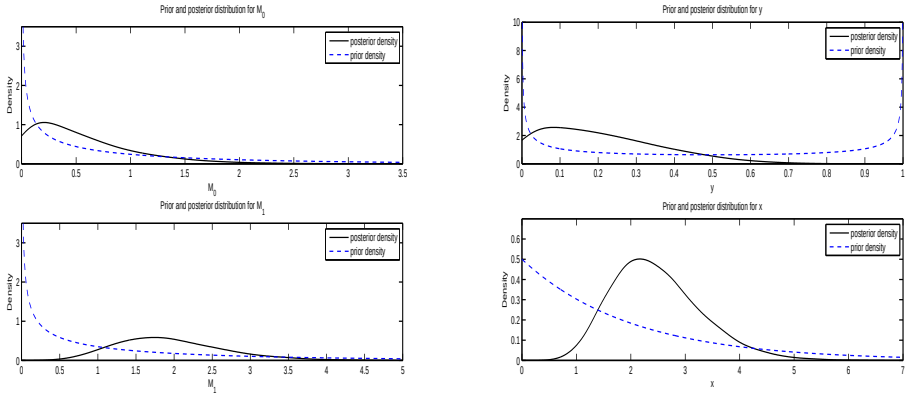


Figure 4.10: Prior (dashed line) and posterior (solid line) distributions of M_0 (top), M_1 (bottom)(left) and y (top), x (bottom)(right) in Model (2.1.4).

Predictive power

In order to quantify the predictive power of the models, an additional 500 data from each data set (Alcoa Inc. and Exxon Mobil Corp. daily returns) were used. The test statistic will be

$$T = -\frac{1}{1000}\sum_{j=1}^{2}\sum_{i=1001}^{1500}\log(p(Y_{ji}|\boldsymbol{Y})),$$

where $\boldsymbol{Y}$ is the vector of data used in the models and $Y_{ji},\ i = 1001, 1002, \ldots, 1500, j = 1, 2$ are the additional data used for assessing the predictive power of each model. This statistic is using the logarithmic score function, $\log S(p, \omega) = \log(p(\omega))$ (Good (1952), Gneiting and Raftery (2007)), applied to the posterior distribution $f(\cdot|\boldsymbol{Y})$. It is clear that, the smaller the value of T, the better is a model in terms of predicting future observations. Another point to be made here is that the MCMC output can be used in calculating T in each case. More specifically, since $p(Y_{ji}|\boldsymbol{Y}) = \int_\Theta p(Y_{ji}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{Y})d\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is the full vector of parameters in the model, $p(Y_{ji}|\boldsymbol{Y})$ can be estimated with $\frac{1}{M}\sum_{t=1}^{M} p(Y_{ji}|\boldsymbol{\theta}^{(t)})$, where $\boldsymbol{\theta}^{(t)},\ t = 1, 2, \ldots, M$ are the MCMC posterior samples and M is the length of this chain.

Applying T to the model of Müller et al. (2004), I have found a value of 1.5885, whereas the equivalent value for model (2.1.4) is 1.5854. This means that the latter model performs slightly better than the former one in terms of predictive power.

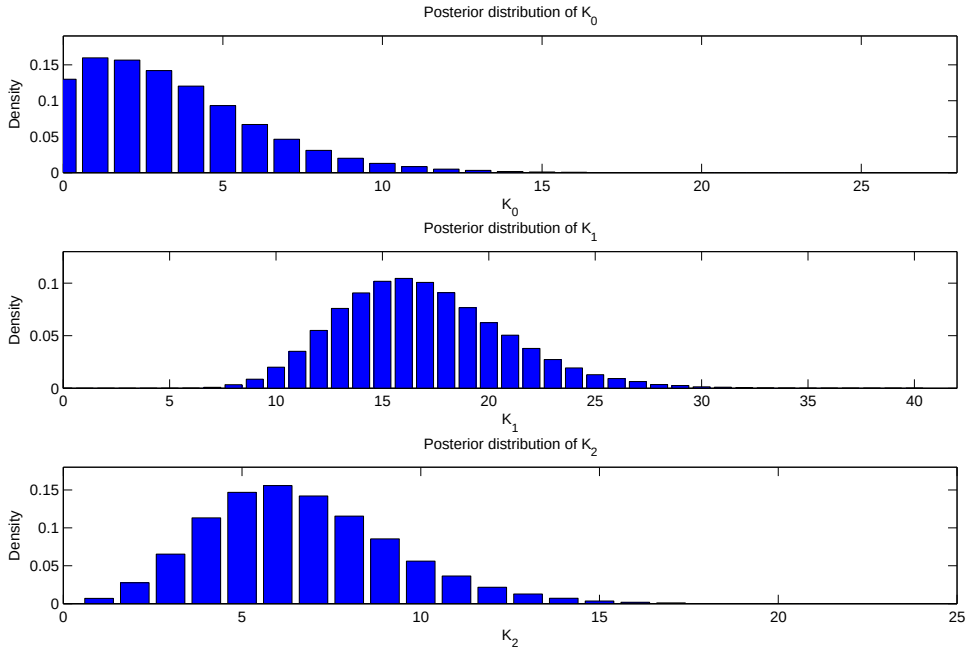


Figure 4.11: Posterior distributions of K_0 (top), K_1 (middle) and K_2 (bottom) in Model (2.1.4).

Bayes factors

The second method of assessing the relative performance of the two models is the Bayes factor: The Bayes factor of a model H_1 against a model H_2 is the ratio of posterior to prior odds:

$$B = \frac{p(\boldsymbol{Y}|H_1)}{p(\boldsymbol{Y}|H_2)},$$

where $\boldsymbol{Y}$ is the vector of data (Jeffreys, 1939). It is obvious that values of B much larger than 1 suggest that the model H_1 fits the data better than H_2, whereas the opposite is true if B is much less than 1. Good (1952) also noted that the logarithm of the Bayes factor can be seen as the difference of the logarithmic score function of the two models.

In order to estimate the two probabilities, use the estimator $\hat{p}_4$ of Newton and Raftery (1994) is used, with $\delta = 0.01$ and $m = 150000$ MCMC iterations (after the burn-in). The quantity $p(\boldsymbol{x}|\theta^{(i)})$ appearing there is just $\frac{1}{(2\pi S)^{(N_1+N_2)/2}} e^{-\frac{1}{2}\sum_{ji}(Y_{ji}-\mu_{ji})^2}$, following the notation of Section 3.2.
In order to calculate $\hat{p}_4$, I run into very small values (due to the large data size), so I used the logarithm of it. Each of these logarithms was derived using a simple iterative scheme on the appropriate modification of equation (16) of Newton and Raftery (1994) (in order to have a similar equation with

$\log(\hat{p}_4)$ instead of $\hat{p}_4$), and then use the relation $\log(B) = \log(p(\boldsymbol{Y}|H_1)) - \log(p(\boldsymbol{Y}|H_2))$ mentioned before.

The results indicate that $B \simeq 0.14$, where H_1 denotes Model (1.2.13) and H_2 denotes and Model (2.1.4). This means that my proposed model, apart from predicting future observations slightly better than the model of Müller et al. (2004), it is also slightly better in explaining the current data.

4.2 Stochastic Frontier Data

In this section I apply some of the models presented before to cost frontiers for some US hospital data.

4.2.1 Stochastic frontier models

Stochastic frontier (SF) models were introduced by Aigner et al. (1977) and Meeusen and van den Broeck (1977), in order to model the efficiency of firms. Such a frontier can be either a production or a cost frontier. The former corresponds to the maximum amount of output that can be produced from a specific set of inputs, whereas the latter represents the minimum cost of producing a certain level of output, given specific input prices. These frontiers represent the theoretical scenario where the productivity and efficiency of a specific firm are optimal, for example when there is absolutely no loss of profit from producing a specific product due to bad managerial decisions (for example producing too much of this product) or reduced effort from the employees at the factory. On the other hand, those frontiers do not account for losses due to factors that are beyond the firm's control, such as destroyed crops caused by bad weather. The last set of factors implies that observed output or cost will be distributed around this maximum output or minimum cost, therefore creating stochastic (instead of fixed) frontiers.

In real life, however, obtaining this optimal efficiency is very rare, and the inefficiency of a firm leads to lower output (when dealing with an output frontier) or higher cost (when dealing with a cost frontier). A natural way to measure this discrepancy is by estimating the difference of the optimal and the actual output (or cost) observed. SF models offer a natural way to do this, by assuming (say, for a cost frontier) that

$$Y_{it} = \alpha + \boldsymbol{X}_{it}'\boldsymbol{\beta} + u_i + v_{it},\ i = 1, 2, \ldots, n,\ t = 1, 2, \ldots, T \tag{4.2.1}$$

where Y_{it} is the logarithm of cost and $\boldsymbol{X}_{it}$ is a vector of outputs for firm i in time period t, α is an intercept, $\boldsymbol{\beta}$ is the vector of covariates and T is the time horizon. It is evident from (4.2.1) that we

have panel data here, i.e. observations over time (and for the same time periods) for each firm.

The key parametrisation in this type of models is that of the errors v_{it} and u_i. The first set of those errors, v_{it}, accounts for the uncertainty due to factors outside the firms's power (for example weather conditions, measurement errors or machine performance), and can be either positive or negative. On the other hand, u_i are the firm-specific disturbances and represent the losses in efficiency due to factors within the firms's reach (for example technical or economical inefficiency or poor performance of the employees). Naturally, those disturbances are assumed to take only positive values, and as in Griffin and Steel (2004), it also assumed that these errors are constant over time. For the implications of relaxing the last assumption, see Fernández et al. (1997). The two sets of error terms are assumed to be independent of each other. For the formulation of a production frontier, one just needs to change the sign of the u_i's in (4.2.1). Finally, the firm efficiencies r_i are defined as the exponential of the negative of the corresponding inefficiencies:

$$r_i = \exp\{-u_i\}, \ i = 1, 2, \ldots, n.$$

The first set of error terms, v_{it}, are assumed to be symmetric, independent and identically distributed and are usually given a normal prior distribution with zero mean. Regarding the prior of the u_i's, many distributions have been proposed, such as the half-normal (Aigner et al., 1977), the exponential (Meeusen and van den Broeck, 1977), the truncated normal and the gamma distributions. The above priors were given in a parametric setting and, as discussed in Griffin and Steel (2004), all of them cause counter-intuitive problems. More specifically, the exponential and the half-normal distributions restrict the probability mass given to specific intervals for the efficiencies, whereas the gamma and the truncated normal can lead to identification problems, when the data suggest a gamma distribution that is practically indistinguishable from a normal distribution (which is the case for the data sets examined below).

On the other hand, Griffin and Steel (2004) propose a nonparametric prior for the inefficiencies, by assuming that they are identically distributed from a distribution F, where F is considered also random and it is assigned a DP prior. For the centering distribution of this DP, they use gamma distributions with fixed integer shape parameters, also known as Erlang distributions.

4.2.2 The models

In this section I employ my basic model (2.1.4), the model of Müller et al. (2004) and a model with a similar form to basic model, but with N-IGP priors (instead of DP priors) for the correlated distributions in the SF setting. More specifically, two panel data sets of sizes $T \cdot N_1$ and $T \cdot N_2$ will be considered, say $Y_{jit}, \ i = 1, 2, \ldots, N_j, \ j = 1, 2, \ t = 1, 2, \ldots, T$. In all models and for each data

set in those models I will use the same structure as in (4.2.1), together with a $\mathrm{N}(0, \sigma^2)$ distribution for the symmetric errors v_{jit}. Therefore, the top level of the hierarchical model will be

$$Y_{jit} \sim \mathrm{N}(\alpha + \boldsymbol{X}'_{jit}\boldsymbol{\beta} + u_{ji}, \sigma^2),\ i = 1, 2, \ldots, N_j,\ j = 1, 2,\ t = 1, 2, \ldots, T.$$

In the above, u_{ji} are the one-sided errors, which will be modelled using correlated distributions, say F_1^* and F_2^*. All the other parameters are as described before.

For the case of the Müller et al. (2004) model, it is assumed that those F_1^* and F_2^* have a common part, F_0, and idiosyncratic parts, F_1 and F_2, respectively. The weight assigned to F_0,ε is given a flexible beta prior, with some mass assigned to the special cases $\varepsilon = 0$ or $\varepsilon = 1$. F_0, F_1 and F_2 are assigned independent DP priors, with the same centering distribution, an exponential distribution:

$$Y_{jit} \overset{ind}{\sim} N(a + \boldsymbol{X}'_{jit} \cdot \boldsymbol{\beta} + u_{ji}, \sigma^2),\ i = 1, 2, \ldots, N_j,\ j = 1, 2,\ t = 1, 2, \ldots, T$$

$$u_{ji} \sim F_j^* = \varepsilon F_0 + (1 - \varepsilon)F_j,\ i = 1, 2, \ldots, N_j,\ j = 1, 2$$

$$F_0 \sim \mathrm{DP}(M_0, H),\ F_1 \sim \mathrm{DP}(M_1, H),\ F_2 \sim \mathrm{DP}(M_2, H),\ \text{where } H \equiv \mathrm{Exp}(\lambda) \text{ and independent}$$

$$\pi(\varepsilon) \ = \ \pi_0\delta_0(\varepsilon) + \pi_1\delta_1(\varepsilon) + (1 - \pi_0 - \pi_1)\mathrm{Be}(a_\varepsilon, b_\varepsilon) \tag{4.2.2}$$

$$f(a, \boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2},\ \lambda \sim \mathrm{Exp}(-\log(r^*)),\ M_0/\eta_0, M_1/\eta_0, M_2/\eta_0 \overset{iid}{\sim} \mathrm{InvBe}(\eta, \eta).$$

A noninformative prior for $(a, \boldsymbol{\beta}, \sigma^2)$ is also assumed, which however leads to a proper posterior distribution (as shown in Fernández et al. (1997)) and an inverted beta distributions (Zellner, 1971) for the precision parameters M_0, M_1 and M_2 (each divided by a hyperparameter η_0, which is also the prior median), as in Griffin and Steel (2004). In statistics literature, the inverted beta distribution is also called the gamma-gamma distribution (see, for example Bernardo and Smith, 1994, p. 120). The distribution H is set to be an exponential distribution with mean $1/\lambda$, and an exponential prior with mean $-1/\log(r^*)$ is adopter for λ.

In the case of Model (2.1.4), the same structure as above is adopted, with the difference that the weight ε has a beta prior with parameters the precision parameters of the DP priors of F_0 and F_1. We also have the same concentration parameters for the DP priors of F_1 and F_2, resulting also in the last two being identically distributed:

$$Y_{jit} \overset{ind}{\sim} N(a + \boldsymbol{X}'_{jit} \cdot \boldsymbol{\beta} + u_{ji}, \sigma^2),\ i = 1, 2, \ldots, N_j,\ j = 1, 2,\ t = 1, 2, \ldots, T$$

$$u_{ji} \sim F_j^* = \varepsilon F_0 + (1 - \varepsilon)F_j,\ i = 1, 2, \ldots, N_j,\ j = 1, 2$$

$$F_0 \sim \mathrm{DP}(M_0, H),\ F_1, F_2 \overset{iid}{\sim} \mathrm{DP}(M_1, H),\ \text{where } H \equiv \mathrm{Exp}(\lambda)$$

$$\varepsilon \sim \text{Be}(M_0, M_1) \tag{4.2.3}$$

$$f(a, \boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2},\ \lambda \sim \text{Exp}(-\log(r^*)),\ M_0/\eta_0, M_1/\eta_0 \overset{iid}{\sim} \text{InvBe}(\eta, \eta).$$

Finally, the model with N-IGP priors has the same structure as in the case of the DP priors, apart from the priors of F_0, F_1 and F_2 (N-IGPs instead of DPs) and a N-IG prior (instead of beta) distribution for ε:

$$Y_{jit} \overset{ind}{\sim} N(a + \boldsymbol{X}'_{jit} \cdot \boldsymbol{\beta} + u_{ji}, \sigma^2),\ i = 1, 2, \ldots, N_j,\ j = 1, 2,\ t = 1, 2, \ldots, T$$

$$u_{ji} \sim F_j^* = \varepsilon F_0 + (1 - \varepsilon) F_j,\ i = 1, 2, \ldots, N_j,\ j = 1, 2$$

$$F_0 \sim \text{N-IGP}(M_0, H),\ F_1, F_2 \overset{iid}{\sim} \text{N-IGP}(M_1, H),\ \text{where } H \equiv \text{Exp}(\lambda)$$

$$\varepsilon \sim \text{N-IG}(M_0, M_1) \tag{4.2.4}$$

$$f(a, \boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2},\ \lambda \sim \text{Exp}(-\log(r^*)),\ M_0/\eta_0, M_1/\eta_0, M_2/\eta_0 \overset{iid}{\sim} \text{InvBe}(\eta, \eta).$$

Notice also that the above models have obvious similarities with the second model used in Griffin and Steel (2004), a PDP-type model. In that model an extra set of covariates was used to account for firm characteristics (i.e. staff ratio and type of hospital: non-profit, for profit or government) and the corresponding parameters were used to borrow inference between the groups. They also used a DP prior for the distribution of the inefficiencies, with a separate distribution for each group. Links between their nonparametric distributions were provided by a parametric centering distribution and a common mass parameter. On the other hand, in the model proposed here, inference between the two groups is mainly borrowed through the common nonparametric part F_0 and also through the common weight ε.

4.2.3 Computational implementation

In order to simulate from the posterior distributions of the parameters in the above models we resort to MCMC algorithms. In the case of the last model, slice sampling ideas (together with the auxiliary variables needed in order to perform this slice sampler) as in Walker (2007), Kalli et al. (2008) and Griffin and Walker (2009) are also used. A mix-split step (similar to the one used in the previous models) is also feasible for Models (4.2.2) and (4.2.3). On the other hand, as explained below, integrating out the weight in the case of Model (4.2.4) was computationally problematic, so I did not include it in the MCMC algorithm.

The full conditional distributions of the parameters in the above models are as follows:

Model (4.2.2)

The posterior distribution of all parameters in the model is:

$$\begin{aligned} f(\mathbf{s},\mathbf{r},\boldsymbol{\phi},\varepsilon,\lambda,\sigma^2,\alpha,\boldsymbol{\beta},M_0,M_1,M_2|\mathbf{Y},X) &\propto \prod_{j,i,t} f(Y_{jit}|\boldsymbol{X}_{jit},\alpha,\boldsymbol{\beta},r_{ji},s_{ji},\boldsymbol{\phi},\sigma^2) f(\alpha,\boldsymbol{\beta},\sigma^2) f(\varepsilon) f(\lambda) \\ &\times f(M_0,M_1,M_2)\prod_{j,i} f(\phi_{ji}|\lambda)\prod_{j,i} f(r_{ji}|\varepsilon) f(\mathbf{s}|\mathbf{r},M_0,M_1,M_2) \end{aligned}$$

where $\boldsymbol{Y}$ is the vector of all data, X is the table of all covariate values and $\boldsymbol{s},\boldsymbol{r}$ and $\boldsymbol{\phi}$ are the indicators and discrete values of the inefficiencies u_{ji}. Let also K_j and n_j denote the number of clusters and the number of firms (not observations) assigned to $F_j,\ j=0,1,2$, respectively.
The full conditional distribution of each parameter is as follows:

- $\lambda|\cdots \text{Ga}(K_0+K_1+K_2+1, \sum_{j,i}\phi_{ji}-\log(r^*))$.
- $\sigma^{-2}|\cdots \text{Ga}((N_1+N_2)T/2, \sum_{j,i,t}(Y_{jit}-\alpha-\boldsymbol{X}'_{jit}\boldsymbol{\beta}-u_{ji})^2/2)$.
- $f(\varepsilon|\cdots)=\pi_0 1_{\left(\sum_{j,i} r_{ji}=0\right)}\delta_0(\varepsilon)+\pi_1 1_{\left(\sum_{j,i} r_{ji}=N_1+N_2\right)}\delta_1(\varepsilon)+(1-\pi_0-\pi_1)\text{Be}(a_\varepsilon+n_0, b_\varepsilon+n_1+n_2)$.
- For each $k=1,2,\ldots,K_j,\ j=0,1,2$ the full conditional distribution of ϕ_{jk} will be a truncated at 0 normal distribution with mean $\frac{\sum(Y^{(jk)}-\alpha-\boldsymbol{X}^{(jk)'}\boldsymbol{\beta})-\lambda\sigma^2}{n_{jk}T}$ and variance $\frac{\sigma^2}{n_{jk}T}$, where the superscript denotes the observations and the corresponding covariate vectors that correspond to the specific cluster ϕ_{jk} and n_{jk} is the number of firms allocated to the same cluster.
- $f(\boldsymbol{s},\boldsymbol{r}|\cdots)$: as in Griffin and Steel (2004).
- $f(M_0|\cdots)\propto \frac{M_0^{\eta+K_0-1}\Gamma(M_0)}{\Gamma(M_0+n_0)(M_0+\eta_0)^2\eta}$
  $f(M_1|\cdots)\propto \frac{M_1^{\eta+K_1-1}\Gamma(M_1)}{\Gamma(M_1+n_1)(M_1+\eta_0)^2\eta}$
  $f(M_2|\cdots)\propto \frac{M_2^{\eta+K_2-1}\Gamma(M_2)}{\Gamma(M_2+n_2)(M_2+\eta_0)^2\eta}$.
- $f(\alpha,\boldsymbol{\beta}|\cdots)$ is the usual linear regression model update (see, for example, equation (A.7) of Koop et al. (1997)).

In the above, the inefficiency terms u_{ji}, although not explicitly stated in the full parameter space, can easily be retrieved from the discrete values and the corresponding indicators. For example, if $r_{12}=0$ and $s_{12}=3$, then $u_{12}=\phi_{03}$. The reason for having u_{ji} in some of the steps above is merely for simplicity of the expressions. I also use the indicator function $1_{(\cdots)}$, which is 1 if the expression in the subscript is true, and 0 otherwise.
The full conditional distributions of the M's are not of known form, so we confront to RWMH updating steps for each of them.

Another interesting subject here is an identifiability issue regarding the intercept α: both α and the inefficiencies u_{ji} have an additive effect, therefore identification of the those should be provided by the prior of the latter. However, in the nonparametric setting these assumptions are not so strong, so the usual MCMC algorithm might be moving too slowly, regarding α. A solution to this is to use the reparametrisation $\alpha, z_{ji} = \alpha + u_{ji}$. The usual parametrisation (α, u_{ji}) is called the non-centred and (α, z_{ji}) is the centred one. As shown in Gelfand et al. (1995), the latter parametrisation rapidly adjusts the intercept. As a result, a mixed updating scheme for $(\alpha, \boldsymbol{\beta})$ is used: in each step of the MCMC chain we update α and $\boldsymbol{\beta}$ using the usual updating step, as mentioned above (i.e. under the non-centred parametrisation), and also conditional on the z_{ji}'s (i.e. under the centred parametrisation). In the latter case, $\boldsymbol{\beta}|\cdots$ can be easily calculated from the full conditional of the regression parameter in the linear regression model, whereas for α we use the fact that $\min(z_{ji}) - \alpha$ follows an exponential distribution with mean $((K_0 + K_1 + K_2)\lambda)^{-1}$.

As mentioned above, an additional mix-split step can be introduced in the algorithm, and performed in the same way as before (except, of course, from the acceptance probabilities, since the prior distribution of the ϕ's is different).

Finally, the predictive densities for F_1^* and F_2^*, as well as those of the component distributions F_0, F_1 and F_2 are calculated in a similar way as in Section 3.

Model (4.2.3)

The parameter space is the same as before, except for M_2, and the posterior distribution of all parameters is:

$$\begin{aligned} f(\mathbf{s}, \mathbf{r}, \boldsymbol{\phi}, \varepsilon, \lambda, \sigma^2, \alpha, \boldsymbol{\beta}, M_0, M_1 | \mathbf{Y}, X) \quad &\propto \quad \prod_{j,i,t} f(Y_{jit} | X_{jit}, \alpha, \boldsymbol{\beta}, r_{ji}, s_{ji}, \boldsymbol{\phi}, \sigma^2) f(\alpha, \boldsymbol{\beta}, \sigma^2) f(M_0, M_1) \\ &\times \quad f(\varepsilon | M_0, M_1) \prod_{j,i} f(\phi_{ji} | \lambda) f(\lambda) \prod_{j,i} f(r_{ji} | \varepsilon) f(\mathbf{s} | \mathbf{r}, M_0, M_1) \end{aligned}$$

using the same notation as before.
The full conditional distributions of most parameters will be the same as above. The only differences will be:

- $f(M_0|\cdots) \propto \varepsilon^{M_0} \frac{M_0^{\eta + K_0 - 1} \Gamma(M_0 + M_1)}{\Gamma(M_0 + n_0)(M_0 + \eta_0)^2 \eta}$ and
  $f(M_1|\cdots) \propto (1-\varepsilon)^{M_1} \frac{M_1^{\eta + K_1 + K_2 - 1} \Gamma(M_1)\Gamma(M_0 + M_1)}{\Gamma(M_1 + n_1)\Gamma(M_1 + n_2)(M_1 + \eta_0)^2 \eta}$.
- $\varepsilon|\cdots \sim \text{Be}(M_0 + n_0, M_1 + n_1 + n_2)$.
- The acceptance probabilities in the mix-split step.

As before, the full conditional distributions of M_0 and M_1 are not of known form, so we confront to RWMH updates.

Model (4.2.4)

Implementing Model (4.2.4) was quite different than the previous two models. The reason, as mentioned in Section 3.7, is that Matlab does not have good built-in functions for the incomplete gamma function appearing in the Pólya-urn representation of the N-IGP. Therefore, I used a method similar to the one proposed in (Griffin and Walker, 2009, and also described in Section 3.7) for simulating such models, using slice sampling ideas. The basic idea is that, using auxiliary variables, one can simulate from the posterior distribution of the parameters in the nonparametric mixture model without any truncation error, using slice sampling. The basic difference in this algorithm from the one described in Section 3.7 is that here there are $N_1 + N_2$ auxiliary variables $v_{ji}, i = 1, 2, \ldots, N_j, j = 1, 2$, each corresponding to firm (j, i), instead of only three $v_j,\ j = 0, 1, 2$ we had before (each corresponding to component distribution $F_j,\ j = 0, 1, 2$).

More specifically, the full likelihood $f(Y_{jit}|X_{jit}, \alpha, \boldsymbol{\beta}, \sigma^2, s_{ji}, r_{ji}, \phi_{ji})$ is extended to

$$\begin{aligned} f(Y_{jit}, v_{ji}, U_{ji}|X_{jit}, \alpha, \boldsymbol{\beta}, \sigma^2, s_{ji}, r_{ji}, J_{ji}) &= \prod_{j,i,t} N(Y_{jit}|X_{jit}, \alpha, \boldsymbol{\beta}, \sigma^2, s_{ji}, r_{ji}, \boldsymbol{\phi}) \\ &\times \prod_{j,i:r_{ji}=0} 1_{(U_{ji}<J_{0s_{ji}})} \prod_{i:r_{1i}=1} 1_{(U_{1i}<J_{1s_{1i}})} \prod_{i:r_{2i}=1} 1_{(U_{2i}<J_{2s_{2i}})} \\ &\times \prod_{j,i\ :\ r_{ji}=0} e^{-v_{ji}H_0} \prod_{j,i\ :\ r_{ji}=1} e^{-v_{ji}H_j} \end{aligned}$$

where r_{ji}, s_{ji} and ϕ_{ji} are as before, $U_{ji},\ j = 1, 2$ are auxiliary variables with a $U(0, \infty)$ prior distribution, $v_{ji},\ i = 1, 2, \ldots, N_j,\ j = 1, 2$ are additional exponentially-distributed auxiliary variables, used to make the algorithm more efficient (see Nieto-Barajas et al. (2004)) and $H_j = \sum_{m=1}^{\infty} J_{jm},\ j = 0, 1, 2,$ where $\boldsymbol{J}_j = (J_{j1}, J_{j2}, \ldots)$ are the (unnormalised) weights in the three component distributions $F_j,\ j = 0, 1, 2$. The *a priori* density of those weights is as in (3.7.4). The infinite sums appearing in H_j can be avoided by truncating these sums at a value K such that $J_{jk} < L,\ \forall\ k > K,\ j = 0, 1, 2$. The value L is arbitrary, and as Griffin and Walker (2009) suggest, setting this to be the minimum of all U_{ji} seems to work well in practise. Finally, we can then integrate over all those $J_{jk} \geq L$, therefore avoiding the infinite sums.

To sum up, the full parameters vector and the full conditionals are as follows:

$$\begin{aligned} f(\alpha, \boldsymbol{\beta}, \sigma^2, \boldsymbol{s}, \boldsymbol{r}, \boldsymbol{\phi}, \boldsymbol{U}, \boldsymbol{J}, v_{ji}, \lambda, \varepsilon, M_0, M_1|\boldsymbol{Y}, X) &\propto f(\alpha, \boldsymbol{\beta}, \sigma^2) f(M_0, M_1) f(\lambda) f(\varepsilon|M_0, M_1) \\ &\times f(\boldsymbol{\phi}|\lambda) f(\boldsymbol{J_0}|M_0) f(\boldsymbol{J_1}|M_1) f(\boldsymbol{J_2}|M_1) f(\boldsymbol{U}) f(\boldsymbol{r}|\varepsilon) \\ &\times f(\boldsymbol{Y}|X, \alpha, \boldsymbol{\beta}, \sigma^2, \boldsymbol{s}, \boldsymbol{r}, \boldsymbol{\phi}) f(\boldsymbol{s}|\boldsymbol{r}, \boldsymbol{U}, \boldsymbol{J}) \prod_{j,i} f(v_{ji}|\boldsymbol{s}, \boldsymbol{r}, \boldsymbol{J}) \end{aligned}$$

where the last product is proportional to $\prod_{j,i:r_{ji}=0} e^{-v_{ji}H_0} \prod_{j,i:r_{ji}=1} e^{-v_{ji}H_j}$.
Therefore, we have:

- $f(\varepsilon|\cdots) \propto \varepsilon^{n_0}(1-\varepsilon)^{n_1+n_2} \frac{K_{-1}\left(\sqrt{\frac{M_0^2}{\varepsilon}+\frac{M_1^2}{1-\varepsilon}}\right)}{\varepsilon^{3/2}(1-\varepsilon)^{3/2}\sqrt{\frac{M_0^2}{\varepsilon}+\frac{M_1^2}{1-\varepsilon}}}$
  $\Rightarrow f(\varepsilon|\cdots) \propto \varepsilon^{n_0-1}(1-\varepsilon)^{n_1+n_2-1} \frac{K_{-1}\left(\sqrt{\frac{M_0^2}{\varepsilon}+\frac{M_1^2}{1-\varepsilon}}\right)}{\sqrt{M_0^2(1-\varepsilon)+M_1^2\varepsilon}}$.
- $f(M_0|\cdots) \propto \frac{M_0^{\eta-1}}{(M_0+\eta_0)^{2\eta}} M_0 e^{M_0} \frac{K_{-1}\left(\sqrt{\frac{M_0^2}{\varepsilon}+\frac{M_1^2}{1-\varepsilon}}\right)}{\sqrt{M_0^2(1-\varepsilon)+M_1^2\varepsilon}} M_0^{K_0^*} e^{-M_0\left(\int_L^\infty q(x)dx+\int_0^L(1-e^{-v_0 x})q(x)dx\right)}$
  $\Rightarrow f(M_0|\cdots) \propto \frac{M_0^{\eta+K_0^*}}{(M_0+\eta_0)^{2\eta}} \frac{K_{-1}\left(\sqrt{\frac{M_0^2}{\varepsilon}+\frac{M_1^2}{1-\varepsilon}}\right)}{\sqrt{M_0^2(1-\varepsilon)+M_1^2\varepsilon}} e^{-M_0\left(\int_L^\infty q(x)dx+\int_0^L(1-e^{-v_0 x})q(x)dx-1\right)}$.
- $$f(M_1|\cdots) \propto \frac{M_1^{\eta-1}}{(M_1+\eta_0)^{2\eta}} M_1 e^{M_1} \frac{K_{-1}\left(\sqrt{\frac{M_0^2}{\varepsilon}+\frac{M_1^2}{1-\varepsilon}}\right)}{\sqrt{M_0^2(1-\varepsilon)+M_1^2\varepsilon}} M_1^{K_1^*+K_2^*} \times e^{-M_1\left(\int_L^\infty q(x)dx+\int_0^L(1-e^{-v_1 x})q(x)dx+\int_L^\infty q(x)dx+\int_0^L(1-e^{-v_2 x})q(x)dx\right)}$$
  $$\Rightarrow f(M_1|\cdots) \propto \frac{M_1^{\eta+K_1^*+K_2^*}}{(M_1+\eta_0)^{2\eta}} \frac{K_{-1}\left(\sqrt{\frac{M_0^2}{\varepsilon}+\frac{M_1^2}{1-\varepsilon}}\right)}{\sqrt{M_0^2(1-\varepsilon)+M_1^2\varepsilon}} \times e^{-M_1\left(\int_L^\infty q(x)dx+\int_0^L(1-e^{-v_1 x})q(x)dx+\int_L^\infty q(x)dx+\int_0^L(1-e^{-v_2 x})q(x)dx-1\right)}.$$
- For each pair (s_{ji}, r_{ji}), $i = 1, 2, \ldots, N_j$, $j = 1, 2$, we have:
  $P(s_{ji}=k, r_{ji}=l|\cdots) \propto \begin{cases} \varepsilon e^{-v_{ji}H_0} \prod_t N(Y_{jit}|X_{jit}, \alpha, \boldsymbol{\beta}, \sigma^2, \phi_{0k}) 1_{(U_{ji}<J_{0k})} & , k = 1, \ldots, K_0^*, l = 0 \\ (1-\varepsilon) e^{-v_{ji}H_j} \prod_t N(Y_{jit}|X_{jit}, \alpha, \boldsymbol{\beta}, \sigma^2, \phi_{jk}) 1_{(U_{ji}<J_{jk})}, & k = 1, \ldots, K_j^*, l = 1. \end{cases}$
- For each v_{ji}, $i = 1, 2, \ldots, N_j$, $j = 1, 2$, we have:
  $f(v_{ji}|\cdots) \propto \begin{cases} e^{-v_{ji}H_0} & , \text{ if } r_{ji} = 0 \\ e^{-v_{ji}H_j} & , \text{ if } r_{ji} = 1. \end{cases}$
- For all $i = 1, 2, \ldots, N_j$, $j = 1, 2$, $U_{ji}|\cdots \sim \begin{cases} U(0, J_{0,s_{ji}}) & , \text{ if } r_{ji} = 0 \\ U(0, J_{j,s_{ji}}) & , \text{ if } r_{ji} = 1. \end{cases}$
- For the jumps J_{jk} with at least one observation allocated to it (i.e. $n_{kl} > 0$), we have:
  $J_{jk}|\cdots \sim \text{Ga}(n_{jk} - 0.5, 1 + \sum_{i:r_{ji}=1} v_{ji})$, $j = 1, 2$ and
  $J_{0k}|\cdots \sim \text{Ga}(n_{0k} - 0.5, 1 + \sum_{i,j:r_{ji}=0} v_{ji})$.
  For the jumps J_{kl} with no observations allocated to them, see Griffin and Walker (2009) for the full conditional distributions and below for the updating method.

- The full conditional distributions of $\lambda, \sigma^2, \alpha, \boldsymbol{\beta}$ and $\boldsymbol{\phi}$, and the corresponding updating schemes are as in the previous models.

In the above, K_j and n_j are as defined in the previous models, K_{-1} is the modified Bessel function of the third type and $q(x) = w(x)/M$.

The full conditional distributions for ε, M_0 and M_1 are not of known form, so we confront to RWMH steps, proposing ε', M_0' and M_1', respectively, where $\text{logit}(\varepsilon') \sim N(\text{logit}(\varepsilon), \sigma_\varepsilon^2)$,
$\log(M_0') = \log(M_0) + \zeta_1,\ \zeta_1 \sim N(0, \sigma_{\zeta_1}^2)$ and $\log(M_1') = \log(M_1) + \zeta_2,\ \zeta_2 \sim N(0, \sigma_{\zeta_2}^2)$.

Regarding the v_{ji}'s, the infinite sum appearing in the above expressions precludes updating them directly from its full conditional distribution. On the other hand, MH steps can be used, together with the method in Griffin and Walker (2009) for the infinite sum.

In order to update the jumps with no observations allocated to them, we simulate a Poisson process with intensity $e^{-V_k x} w(x)$ on (L, ∞),where w is as in (3.7.4) and V_k is the sum of all v_{ji} that belong in $F_k,\ k = 0, 1, 2$.

Finally, similar to Griffin and Walker (2009), the predictive densities for F_1^* and F_2^* can be calculated using an additional pair of indicators for each F_j^*, i.e. $(s_{1,N_1+1}, r_{1,N_1+1})$ and $(s_{2,N_2+1}, r_{2,N_2+1})$.

Note: As mentioned above, I did not include a mix-split step in this algorithm, although this is theoretically possible. In practise, though, it was not possible to integrate the weight ε out of this step, again due to limitations of Matlab. Since not integrating ε out in extra mix-split step would result in very little improvement in the mixing of the chains, this step was omitted completely.

4.2.4 Hospital data

The data to be analysed are the panel data of 382 nonteaching hospitals in the U.S.A. for a period of $T = 5$ years, from 1987 to 1991, also analysed in Koop et al. (1997). The output consists of five different measurements: number of cases, number of impatient days, number of beds, number of outpatient visits and a case mix index (say $W^{(1)} - W^{(5)}$, respectively). We also consider a measure of capital stock C and an aggregate wage index P. In the vector of covariates we also include t and t^2, in order to capture any time trend. For Y's we have the logarithm of cost for each firm.

Next, the hospitals are separated in six categories, based on their ownership status (for-profit, non-profit or government) and their number of clinical workers per patient. The first characteristic is straightforward. The second one, called "staff ratio" for simplicity from now on, is a binary variable taking the value 1 if the average (over the years) of the ratio of the number of clinical workers over the number of patients for a specific hospital is higher than the median of those averages of all 382 hospitals, and 0 otherwise. Doing so, the following group sizes were produced:

	Non-profit	For-profit	Government
Staff ratio=0	141	34	22
Staff ratio=1	127	30	28

Table 4.3: Group sizes for the six groups of hospital firms based on ownership status and staff ratio.

Notice that Griffin and Steel (2004) used the same two characteristics (ownership status and ratio of workers per patient) in defining similar groups of the hospitals. However, the clusters obtained here were not exactly the same as the ones in that article, probably due to a different way for setting the "staff ratio" index (for example, using the mean instead of the median over all hospitals).

Finally, I applied the models of Section 4.2.2 in each pair of groups of hospitals with the same ownership status. As seen in Table 4.3, the non-profit groups were the largest ones and will be studied in more depth. Fortunately, those two groups were the same as in Griffin and Steel (2004). The effect of the small sample sizes for the for-profit and the government-owned hospitals will become apparent when they are analysed.

4.2.5 Results

I applied the three models (4.2.2)-(4.2.4) to the hospital data. In all three models a translog function of the covariates ($W^{(1)}$ to $W^{(5)}$, C and P) was used. Therefore, $\forall\ i = 1, 2, \ldots, N_j,\ j = 1, 2$ and $t = 1, 2, \ldots, T$,

$$\begin{aligned}
\boldsymbol{X}'_{jit}\boldsymbol{\beta} &= \sum_{k=1}^{5} \beta_k \log(W_{jit}^{(k)}) + \beta_6 \log(P_{jit}) + \sum_{k=1}^{5}\sum_{l=k}^{5} \xi_{kl} \log(W_{jit}^{(k)}) \log(W_{jit}^{(k)}) \\
&+ \sum_{k=1}^{5} \beta_{7+k} \log(W_{jit}^{(k)}) \log(P_{jit}) + \beta_{13} \log(C_{jit}) + \sum_{k=1}^{5} \beta_{13+k} \log(W_{jit}^{(k)}) \log(C_{jit}) \\
&+ \beta_{19} \log(P_{jit}) \log(C_{jit}) + \beta_{20} \left(\log(C_{jit})\right)^2 + \beta_{21} t + \beta_{22} t^2,
\end{aligned}$$

where $\xi_{kl},\ k \leq l$, provide the remaining elements of the vector of covariates $\boldsymbol{\beta}$.
I also set r^* at the value 0.8 (as in Griffin and Steel, 2004) and $\eta = \eta_0 = 1$. The value of r^* implies a prior median for the efficiency r of 0.8, whereas the values of η and η_0 imply a prior median value of 1 for both M_0 and M_1.
In the case of Model (4.2.2), we have $\pi_0 = \pi_1 = 0.1$ and $a_\varepsilon = b_\varepsilon = 1$.
In all cases the burn-in period was 40000 iterations, whereas the length of the chain used was also long enough (ranging from 80000 to 250000 iterations).

Non-profit hospitals

These models are first applied to the largest data sets, i.e the non-profit hospitals with low staff ratio (data set 1), which consists of 141 firms, and the non-profit hospitals with high staff ratio (data set 2), with 127 firms in it.

Let us first consider the basic proposed model, Model (4.2.3). The acceptance rate of the split steps in the mix-split step was around 24.0%, whereas for merge steps the corresponding rate was around 19.8%. The mean, median, and 95% credible interval for the posterior distribution of various quantities are shown in Table 4.4:

	M_0	M_1	y	x	σ^2	σ^{-2}	λ	β_{21}	β_{22}
Mean	6.10	1.248	0.815	7.35	0.0031	344.38	3.61	0.124	-0.0045
Median	5.58	0.865	0.869	6.80	0.0029	344.97	3.55	0.124	-0.0045
2.5th percentile	1.48	0.055	0.351	2.79	0.0025	196.09	2.20	0.111	-0.0066
97.5th percentile	13.75	4.59	0.992	14.95	0.0051	396.65	5.36	0.138	-0.0024

Table 4.4: Posterior means, medians and 95% credible intervals for various parameters in Model (4.2.3).

In the above, $y = \frac{M_0}{M_0+M_1}$, $x = M_0 + M_1$ and β_{21}, β_{22} are the parameters of the time trends t and t^2, respectively.

Next, consider the posterior distribution of the weight parameter ε. This is shown in Figure 4.12. As can be seen, there is a very small mode at 0 (corresponding to the case of F_1^* and F_2^* not having a common part at all) and two very big modes at 1 (the case of F_1^* and F_2^* coinciding) and around 0.88 (roughly speaking, F_1^* and F_2^* sharing around 88% of their posterior distribution).

Furthermore, inference for this parameter also reveals the importance of the additional mix-split step: when this step was not applied, the small mode around 0 was not captured, probably because it is far from the other two, much bigger modes. For comparison purposes, I show the trace plots for this parameter with and without the mix-split step in Figure 4.13. The other results presented are those with the mix-split step applied in the algorithm.

Figure 4.14 next shows the posterior densities of M_0 and M_1 and of the reparametrisation $y = \frac{M_1}{M_0+M_1}, x = M_0 + M_1$. The posterior distribution of M_0 is flatter than the one of M_1, which is peaked around 1, indicating that F_1 and F_2 are very far from their expected centering distribution. The posterior of x is a compromise between those two, i.e. less flat than the posterior of M_0, but also less peaked than that of M_1. Finally, the posterior for y is left-skewed with a mode very close to 1, in line with Figure 4.12.

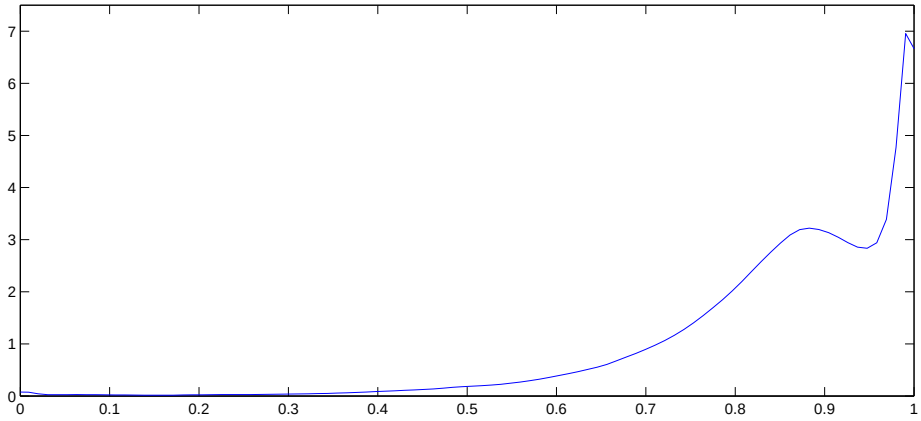


Figure 4.12: Posterior distribution of ε for Model (4.2.3) applied to the non-profit hospitals.

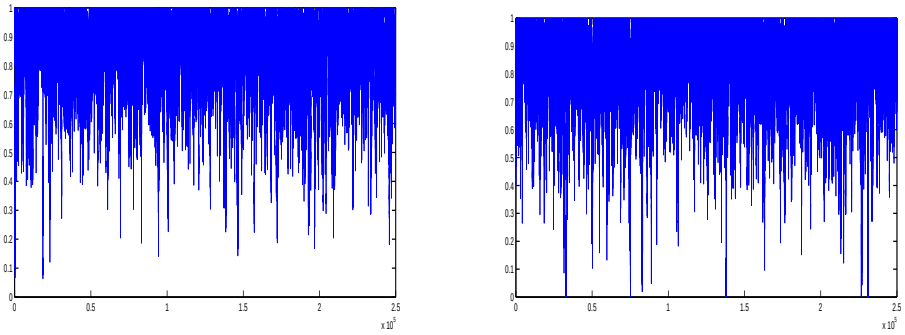


Figure 4.13: Trace plots for ε, with (right) and without (left) the mix-split step for Model (4.2.3) applied to the non-profit hospitals.

Another important aspect in this type of models is the predictive density of the efficiency of a new firm in each of the two groups F_1^*, F_2^*. These predictive densities were plotted, as well as the corresponding cumulative distribution functions (cdf) in Figure 4.15. For a better inspection of the difference in the two groups, the predictive densities of the efficiency of a firm in the common (F_0) and in the idiosyncratic (F_1, F_2) parts were also plotted (Figure 4.16).

The left graph in Figure 4.15 resembles the results of Griffin and Steel (2004). For F_1^*, we get a mode at 1, an antimode around 0.95, a small bump around 0.86 and two larger modes at 0.7 and 0.75 (the last two are slightly reversed in size here, compared to Griffin and Steel (2004)). The difference in this one is the mode around 0.67 of Griffin and Steel (2004), which is transposed left, around 0.6 and looks more like a bump. Regarding F_2^*, we have the same large modes at 0.7 and 0.75 and the bumps around 0.67 and 0.85. In this case, there is also a tiny mode around 0.47. The

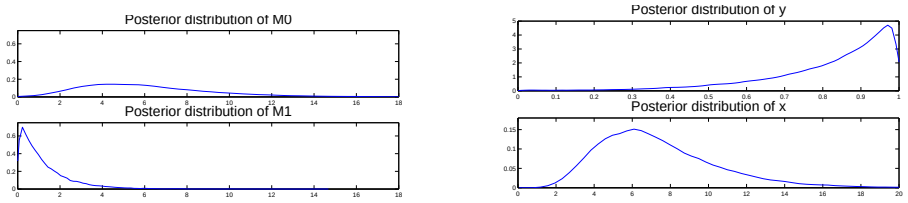


Figure 4.14: Posterior distributions of M_0 (top), M_1 (bottom)(left) and y (top), x (bottom)(right) for Model (4.2.3) and the non-profit hospitals.

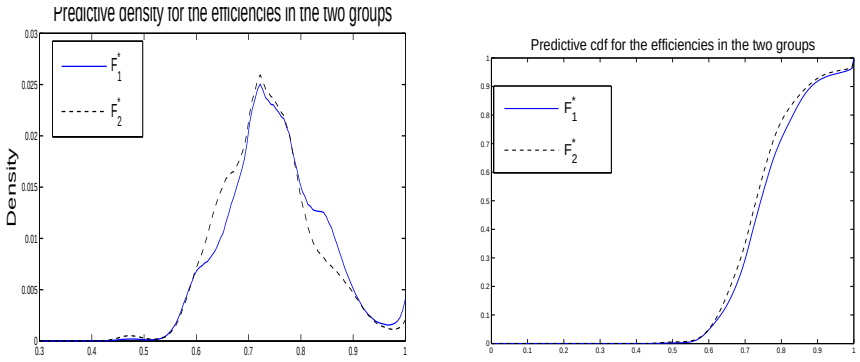


Figure 4.15: Predictive densities (left) and cumulative distributions (right) for the efficiency of firms in the low staff ratio (solid line) and the high staff ratio group (dashed line) for Model (4.2.3) applied to the non-profit hospitals.

main difference in this graph and the one of Griffin and Steel (2004) (although not a big difference, in any case) is the behaviour at the right end. In Griffin and Steel (2004) the mass of the predictive density is decreasing as the efficiency approaches 1, whereas here there is a small mode around 1, which also creates an antimode around 0.95. However, overall we might say that the results are very similar.

The right graph of Figure 4.15 clearly demonstrates that the first group (non-profit hospitals with low staff ratio) is more efficient than the second group (non-profit hospitals with high staff ratio). It is also interesting that this occurs in a rather specific way with an increase of probability of about 0.06 around 0.65, and this difference is preserved with small differentiations up to 0.99, where the two cdf's coincide.

Another interesting point here is that, comparing the predictive densities of F_1^* and F_2^*, it becomes clear that their main differences are in the intervals (0.6,0.7) (where F_2^* has more mass) and the interval (0.8,0.9) (where the opposite is true). In other words, it can be said that the mass of F_1^* in (0.8,0.9) has been moved to (0.6,0.7) for F_2^*. This difference is also clear from the predictive densities of the component distributions F_1, F_2 and F_0 in Figure 4.16. This graph is helpful in providing a

better insight as to where the characteristics of those predictives come from: the large mode at 1 and the bump around 0.86 in F_1^* are due to the idiosyncratic part F_1, whereas the other two around 0.7 and 0.75 and the small one around 0.6 come from the common part F_0. As for F_2^*, the small mode at 0.47 is due to its idiosyncratic part F_2, the mode at 1 and the bump at 0.85 are due to F_0, the mode around 0.75 is mostly (but not completely) due to F_0, whereas the bump around 0.67 is due to both F_2 and F_0. The last one might seem strange, since it is obvious that F_2 has much more mass around that point than F_0 does, but it makes sense if we take into account the weights of those two parts (i.e. ε).

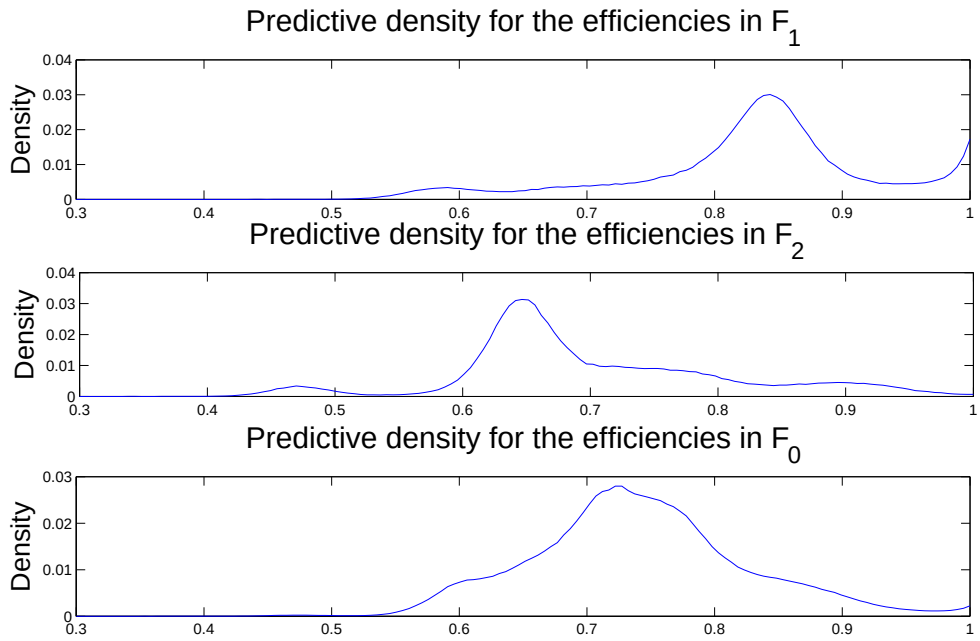


Figure 4.16: Predictive densities for the efficiency of a firm in F_1 (above), F_2 (centre) and F_0 (below) for Model (4.2.3) for the non-profit hospitals.

Finally, quartile plots, similar to the ones given in Figures 7 and 14 of Griffin and Steel (2004) are given in Figure 4.17 for all the firms in the first (left) and second group (right).
As in Griffin and Steel (2004), the last two plots represent the posterior probabilities of the efficiency of each firm falling in each quartile of the predictive efficiency distribution. More specifically, the size of the black shading for a specific firm represents the (posterior) probability that its efficiency falls in the lower quartile, i.e. in $(0, 0.25)$ of the predictive efficiency distribution. Dark gray shading corresponds to the second lower quartile $(0.25, 0.5)$, light gray to the third one $(0.5, 0.75)$ and the white one to the highest quartile $(0.75, 1)$.

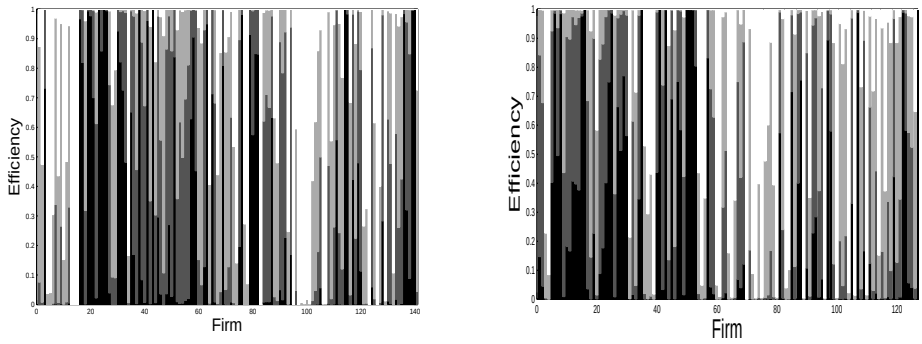


Figure 4.17: Quartile plots for the efficiencies of the firms in the F_1^* (left) and F_2^* (right) for Model (4.2.3) applied to the non-profit hospitals.

Next, Model (4.2.2) is applied to the same data sets. Overall, the results were similar to the results of the previous models. The acceptance rate for mixing steps was around 18.9%, whereas the same rate for split steps was 19.3%. Although mixing was not significantly improved, the results reported will correspond to the ones with this step included in the code.

Means, medians and 95% credible interval for the posterior distribution of the same parameters as before (except for x and y, since those two quantities do not have a natural interpretation here), as well as M_2, which was not in the previous model, are shown in Table 4.5:

It can be seen that the parameters $\lambda, \sigma^2, \beta_{21}$ and β_{22} are very close to the case of the previous

	M_0	M_1	M_2	σ^2	σ^{-2}	λ	β_{21}	β_{22}
Mean	5.70	6.81	5.97	0.0031	335.02	3.61	0.124	-0.0045
Median	5.08	1.39	0.972	0.0029	345.60	3.55	0.124	-0.0045
2.5th percentile	1.28	0.042	0.031	0.0025	198.19	2.20	0.111	-0.0066
97.5th percentile	13.79	20.95	16.77	0.0050	397.27	5.36	0.138	-0.0024

Table 4.5: Posterior means, medians and 95% credible intervals for various parameters in Model (4.2.2).

model. This is also true for M_0, although not as much as for the parameters above. On the other hand, the values of M_1 and M_2 are significantly larger than those of the (common) M_1 in Model (4.2.3). The reason for this is that the prior for the weight in the previous model, which involves M_0 and M_1, induces lower values for M_1.

Next, consider the posterior distribution of this weight ε for this model. This is shown in Figure 4.18 and we can see the same modes at 0.85 and 1. On the other hand, in this case there is not any mass at 0. The trace plot of this parameter are shown in Figure 4.19.

The posterior distributions of the three concentration parameters are shown in Figure 4.20. It

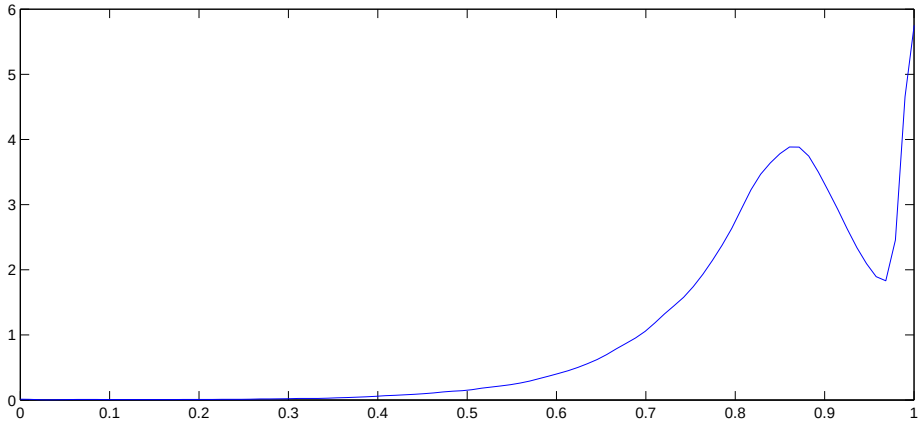


Figure 4.18: Posterior distribution of ε for Model (4.2.2) for the non-profit hospitals.

is interesting to notice the similarity of M_0 with before, as well as the similar shape of M_1 and M_2 with M_1 before.

The predictive densities and cumulative distribution functions of the efficiencies in the two groups are shown in Figure 4.21. The cdfs looks very similar to the ones of the previous model. This is also the case for the predictive densities, although one can spot some small differences from their graph. More specifically, there is higher difference between F_1^* and F_2^* at the mode around 0.6 and a much higher mode at 1 for F_1^*.

Finally, the predictive densities for the component distributions F_0, F_1 and F_2 are shown in Figure 4.22. Again, these distributions are very similar to the ones derived when Model (4.2.3) was applied to the same data. The only differences seem to be at the mass of all three predictives at 1, which here it is more than before, especially for F_2.

Next, Model (4.2.4) was applied to the same data sets. First, it is interesting to mention the difference in the behaviour of the algorithm used for this model, compared to the algorithms in the previous two models, since they are quite different. What I noticed is that the running time for this algorithm was significantly longer than the corresponding time for each of the other algorithms (around 18 times longer). This, of course, was also the case for the other two pairs of hospital groups used. As for the mixing of the chains, this is harder to answer, since I was not able to imitate an additional split/merge step for the last model. Additionally, the results of applying the last model to the specific data were more different to the results of the other two models than the results of the other two models, when compared to each other (as will be apparent later), so it is not so easy to compare the mixing. Regardless of this, the mixing of the algorithm for this model seemed good

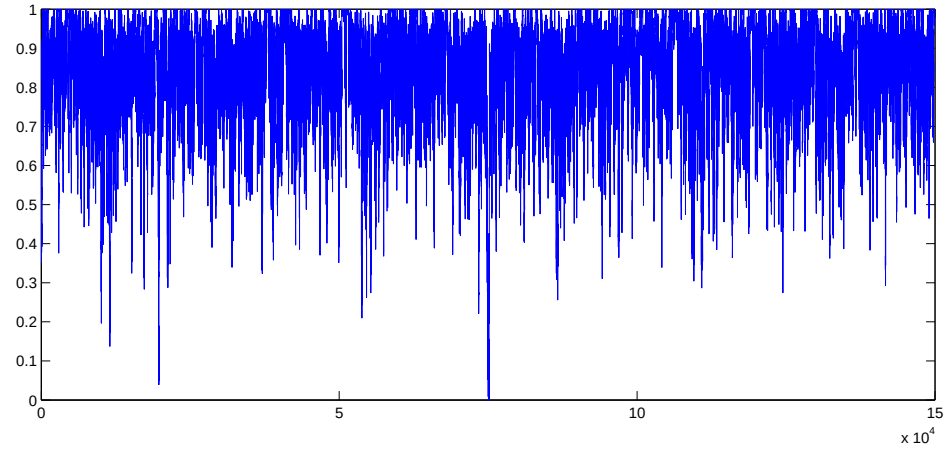


Figure 4.19: Trace plot for ε for Model (4.2.2) for the non-profit hospitals.

enough. On the other hand, this difference in the results should not affect the running time much.

Concentrating on the actual results of Model (4.2.4) applied to the two bigger groups of hospitals, the following mean and percentiles for some parameters were found:

	M_0	M_1	y	x	σ^2	σ^{-2}	λ	β_{21}	β_{22}
Mean	1.01	0.556	0.628	1.57	0.0026	387.72	3.67	0.125	-0.0045
Median	0.826	0.435	0.655	1.42	0.0026	387.37	3.62	0.126	-0.0045
2.5th perc	0.086	0.045	0.155	0.196	0.0024	354.55	2.43	0.113	-0.0064
97.5th perc	3.18	1.68	0.934	3.929	0.0028	422.51	5.21	0.138	-0.0027

Table 4.6: Posterior means, medians and 95% credible intervals for various parameters in Model (4.2.4).

In the above $x = M_0 + M_1$ and $y = \frac{M_0}{M_0+M_1}$ are as defined before. Although these parameters have a slightly different interpretation here, y can still be seen as the prior mean of the weight ε and x as a measure of variance in its prior distribution (although through a slightly more complicated relationship than in the case of the beta prior distribution in Model (4.2.3)).

Again, we see that there is a consensus with the previous models regarding the posterior mean, median and 95% credible intervals for the parameters $\lambda, \sigma^2, \beta_{21}$ and β_{22}. On the other hand, there are significantly different values for the concentration parameters M_0 and M_1 and for the reparametrisation x and y. The reason for those differences is, of course, the different prior distribution of ε, which involves those M's.

This different prior also influences significantly the posterior of ε, as seen in Figure 4.23. Here there is only one big mode around 0.8, instead of the two large modes before. Mixing of ε was good, as seen in Figure 4.24.

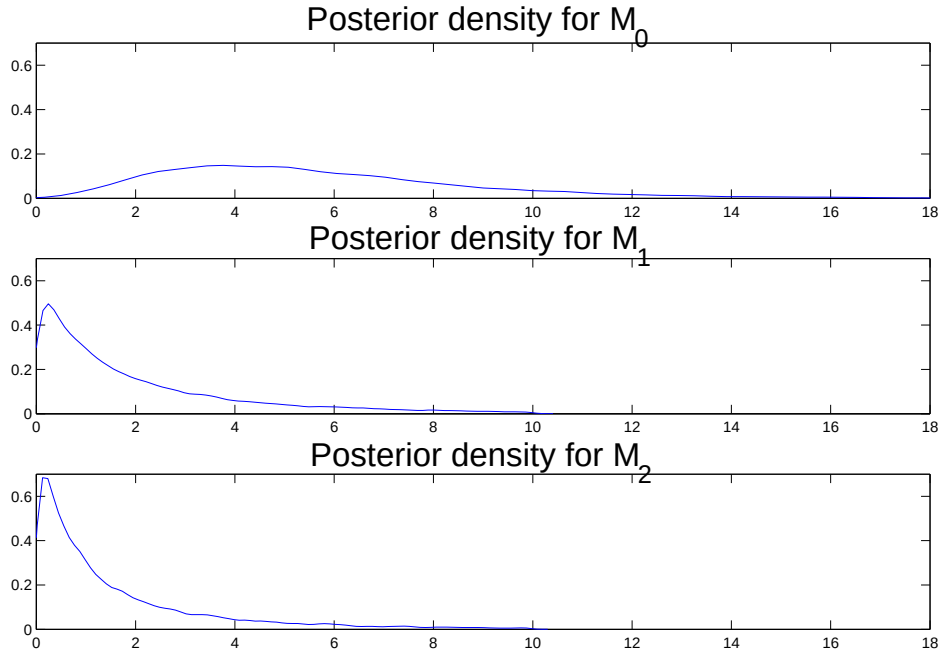


Figure 4.20: Posterior distributions for M_0, M_1 and M_2 for Model (4.2.2) for the non-profit hospitals.

Next we look at the posterior distributions of the two concentration parameters, as well as those of the reparametrisation x and y. These posteriors are shown in Figure 4.25. As mentioned before, those parameters exhibit a significantly different behaviour than in the previous two models, and the main reason for this is the prior $\varepsilon \sim \text{N-IG}(M_0, M_1)$.

The predictive densities and the corresponding cdfs of the efficiencies in the two groups are shown in Figure 4.26. The densities are significantly different from the corresponding ones derived in the previous models. For F_1^* we have a small mode at 0.6, two larger ones around 0.7 and 0.75 and another one around 0.85. There is also a significant amount of mass at 1. For F_2^* there is a tiny mode around 0.45, a larger one at 0.65, the two largest modes at 0.7 and 0.75 (i.e. the same as the largest modes as for F_1^*) and a smaller mode at 1. There is also a small bump around 0.86. The predictive cdfs, on the other hand, are not that different than the corresponding cdfs derived with the previous two models. More importantly, in all three models we have higher probabilities on higher (predictive) efficiencies in F_1^* than in F_2^* (as shown by the fact that the predictive cdf for F_1^* is below the cdf for F_2^*), indicating that the low staff, non-profit hospitals are more efficient than the high staff, non-profit hospitals. Finally, in order to get a better understanding of the predictive behaviour

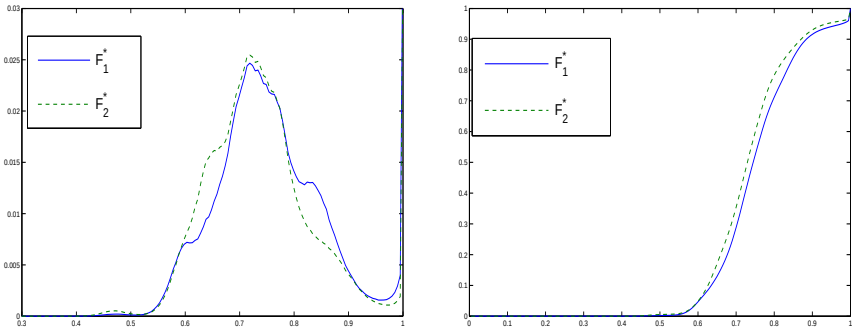


Figure 4.21: Predictive densities (left) and cumulative distributions (right) for the efficiency of firms in the low staff ratio (solid line) and the high staff ratio group (dashed line) for Model (4.2.2) applied to the non-profit hospitals.

of my model, I have plotted the predictive densities of the component distributions F_0, F_1 and F_2 (Figure 4.26). The predictive density for the first idiosyncratic part, F_1 has a big mode around 0.85 and some mass at 1. For the second idiosyncratic part, F_2, there is one mode at 0.65. For the common part, F_0, the predictive density exhibits two big clusters at 0.7 and 0.75 and a smaller mode at 0.6.

It is also interesting to compare the predictive densities for the two groups in this model and in the corresponding model with DP priors, i.e. Model (4.2.2). For the N-IGP model, there is an obvious enforcement of the modes at 0.6, 0.75 and 0.85 for F_1^* and 0.65 and 0.75 for F_2^*. This is in accordance with the intuition given in Lijoi et al. (2005) regarding the property of the N-IGP to better detect clusters and enforce the significant ones in a more sensible way.

Government-run hospitals

The three models above were also applied to the government-run hospitals with low staff ratio (data set 1, 22 firms) and the government-run hospitals with high staff ratio (data set 2, 28 firms). However, when I applied the first two models, the results were not as one would expect since, for example, I was not able to get similar predictive densities for F_1^* and F_2^* as in Griffin and Steel (2004). On the contrary, the predictives of the model applied here (Figure 4.28) assigned most of their mass at 1. On the other hand, by significantly reducing the effect of the mode at 1, these predictives were much closer to the results of Griffin and Steel (2004).

Next, we turn our attention to the posterior distribution of the weight ε. As seen in Figure 4.29, for Model (4.2.2) we have a big mode at 1 and a smaller mode at 0. For Model (4.2.3) we got the same modes, together with another big mode around 0.75. Both distributions are rather dispersed.

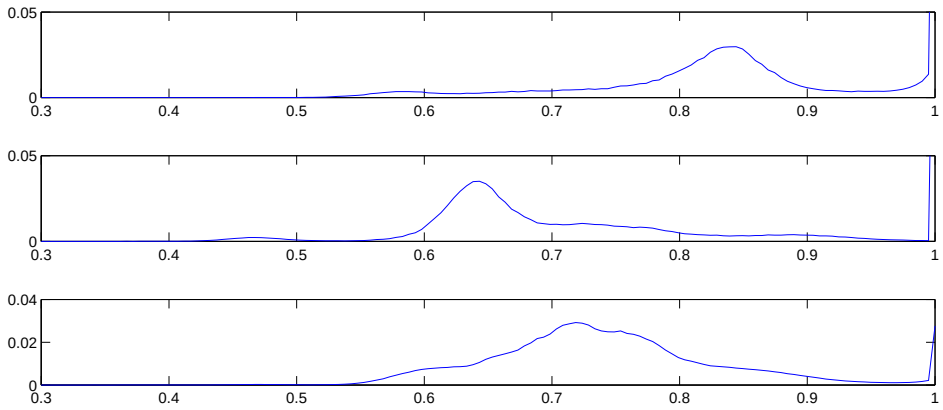


Figure 4.22: Predictive densities for the efficiency of a firm in F_1 (above), F_2 (centre) and F_0 (below) for Model (4.2.2) for the non-profit hospitals.

On the other hand, when Model (4.2.4) was applied to these data, the results were more reasonable. The posterior distribution and the trace plot for ε for this model are shown in Figure 4.30. This posterior distribution is unimodal around 0.7, and with lower variance than in the previous two models. From the trace plot, it seems that mixing is also good.

For comparison reasons, the prior distributions for the weight in each of the three models were also plotted (Figure 4.31).

The predictive densities of F_1^* and F_2^* are shown in Figure 4.32. From the left graph there, we see that the predictives for the efficiencies in the two groups are close to the corresponding distributions in Griffin and Steel (2004). On the other hand, there are also some differences such as the small mode at 0.55 for the first group and the reversed (in size) and shifted modes at 0.75 and 0.85 for the second group. The corresponding predictive densities for the component distributions F_1, F_2 and F_0 are shown in the right side of Figure 4.32. In any case, the above predictive densities, although not extremely close to the ones in Griffin and Steel (2004), are closer to them (and seem more sensible) than those of the other two models applied to the same data. As mentioned before, the obvious difference in the predictive densities of Model (4.2.2) against the corresponding distributions for Models (4.2.2) and (4.2.3) is the dominating mode at 1 for the latter models. On the other hand, notice that the difference between the efficiencies in F_1^* and F_2^* (shown, for example, by the distance between the two lines in each graph, taking also into account which line is higher) is the same in all graphs. Therefore, it can be said that all models perform similarly, in terms of assessing the relative predictive efficiency of the two groups of government-run hospitals (low or high staff ratio).

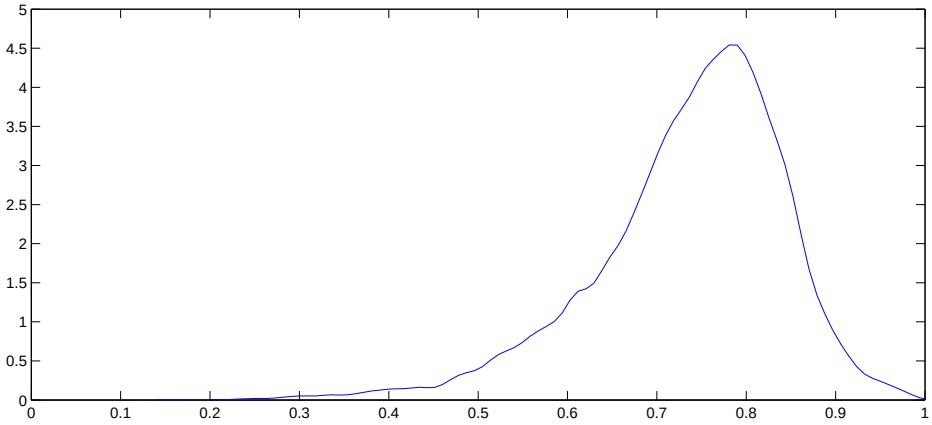


Figure 4.23: Posterior distribution of ε for Model (4.2.4) for the non-profit hospitals.

Finally, I applied the above three models to the for-profit hospitals with low and high staff ratio. In this case, however, the results were overdispersed for all three models (and not only for the first two, as for the government hospitals). The main reason for this is the small data sizes.

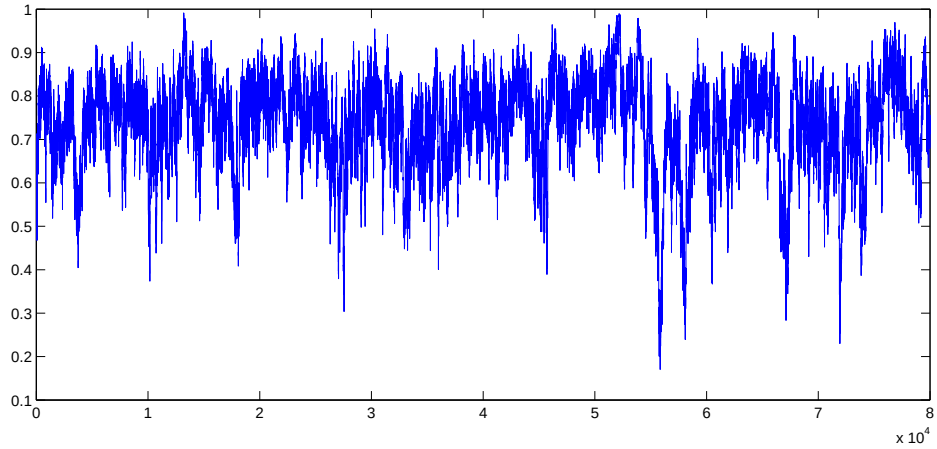


Figure 4.24: Trace plot for ε for Model (4.2.4) for the non-profit hospitals.

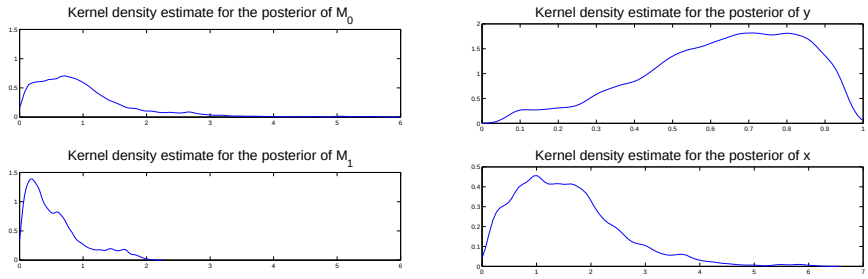


Figure 4.25: Posterior distributions of M_0 (top), M_1 (bottom)(left) and y (top), x (bottom)(right) for Model (4.2.4) for the non-profit hospitals.

4.3 Summary

In this chapter I applied some of the models introduced earlier to real data. First, my basic proposed model and the model of Müller et al. (2004) were applied to the daily stock returns of two firms. The extra mix-split step proposed in Section 3 was also applied and it was shown that mixing of the chains was improved when using this step. Using these models we get a better understanding of the similarities and differences of the underlying distributions of those data. Next, the same two models, together with the model with N-IGP priors (instead of DP priors) were applied to the stochastic frontier setting. For the first two models the corresponding algorithms were similar to the ones discussed in Section 3, and also used in the financial data above. For the last model, however, a similar approach could not be applied, due to limitations of the software used. As a result, the slice sampler of Griffin and Walker (2009) (as described in Section 3.7) was extended in this context.

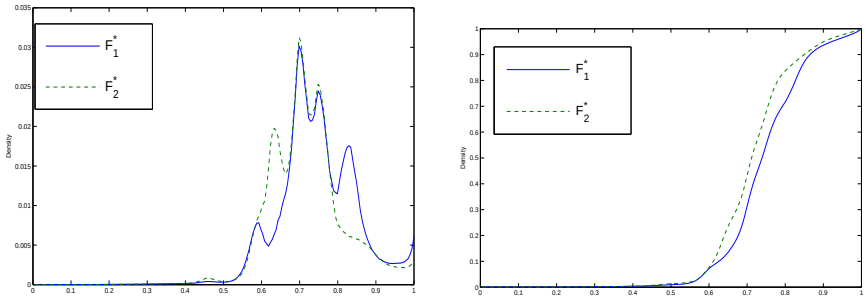


Figure 4.26: Predictive densities (left) and cumulative distributions (right) for the efficiency of firms in the low staff ratio (solid line) and the high staff ratio group (dashed line) for Model (4.2.4) applied to the non-profit hospitals.

Finally, the derived algorithms were applied to model hospital efficiency data, where each pair of data sets corresponded to the same ownership status and was separated according to a staff ratio index. The results for the first two models were quite similar to each other and rather different from the results for the last model.

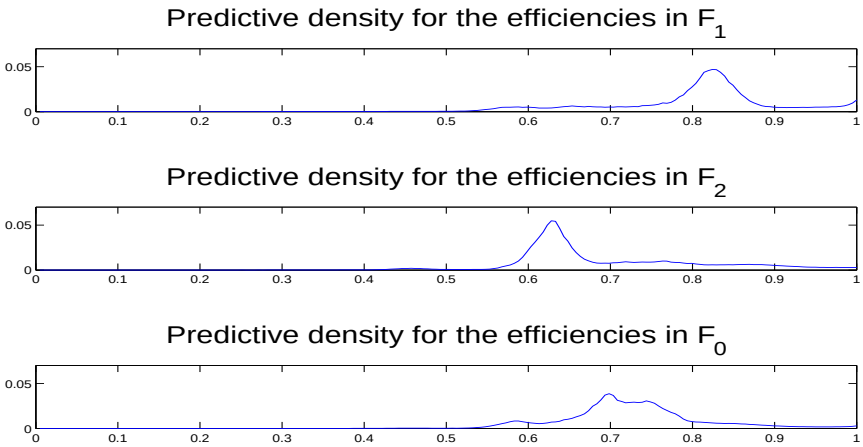


Figure 4.27: Predictive densities for the efficiency of a firm in F_1 (above), F_2 (centre) and F_0 (below) for Model (4.2.4) for the non-profit hospitals.

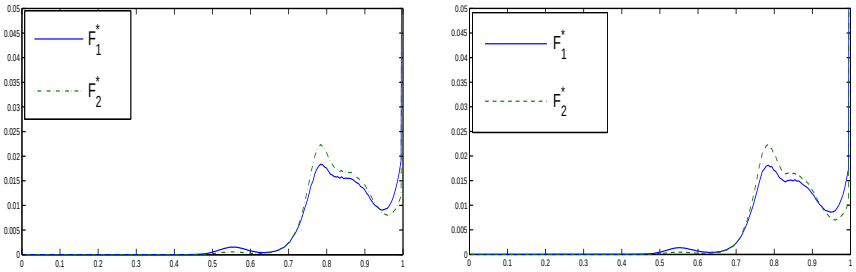


Figure 4.28: Predictive densities for the efficiency of firms in the low staff ratio (solid line) and the high staff ratio group (dashed line) for Models (4.2.3) (left) and (4.2.2) (right) applied to the government hospitals.

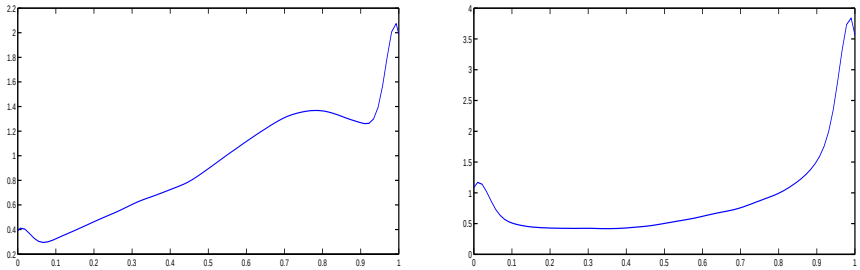


Figure 4.29: Posterior distribution of the weight ε for Models (4.2.3) (left) and (4.2.2) (right) applied to the government hospitals.

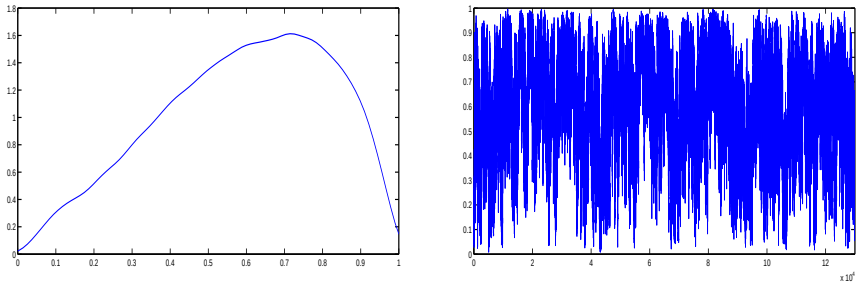


Figure 4.30: Posterior distribution (left) and trace plot (right) for ε for Model (4.2.4) for the government hospitals.

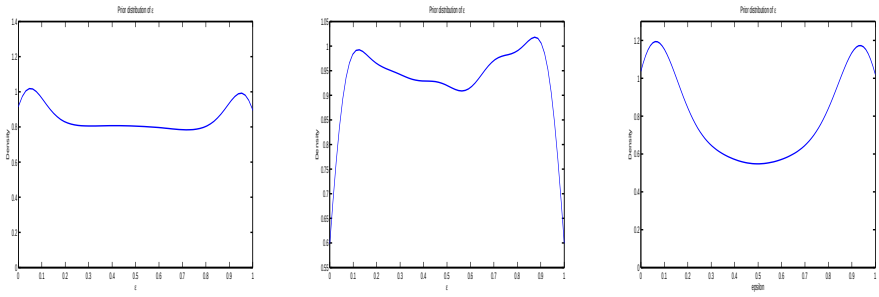


Figure 4.31: Prior distribution of ε in Model (4.2.2) (left), (4.2.3) (middle) and (4.2.4) (right).

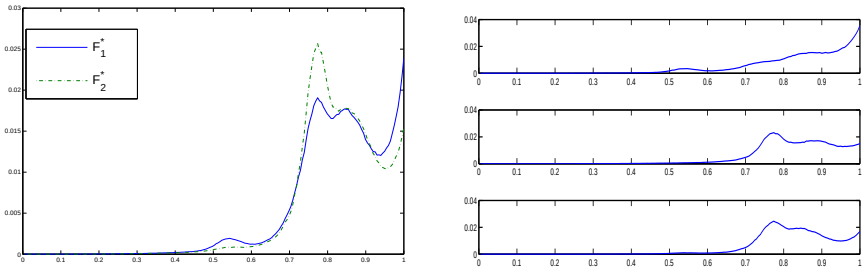


Figure 4.32: Predictive densities for the efficiency of firms in the low staff ratio (solid line) and the high staff ratio group (dashed line) (left) and predictive densities for F_1 (above), F_2 (centre) and F_0 (below) (right) for Model (4.2.4) applied to the government hospitals.

Chapter 5

Modelling Overdispersion With the Normalized Tempered Stable Distribution

As mentioned at the end of Chapter 3, a general formula for the moments and cross-moments of the N-IG distribution will be derived, as well as those of a more general class of distributions, called the normalised tempered stable distribution. This generalised distribution will be used in creating a novel distribution for modelling count data. The formulae for the models will then be used in the calculation of maximum likelihood estimates of the proposed model. It will be demonstrated that the latter is better than the simpler beta-binomial distribution in modelling overdispersed data and compare the two models, as well as other models proposed in the literature, using simulated and real data.

5.1 The Moments of the N-IG Distribution

Although only the one-dimensional N-IG distribution will be used in my proposed models for overdispersed data, I also have results for the moments of the more general n-dimensional N-IG distribution, by extending the results of Lijoi et al. (2005) and James et al. (2006):

Theorem 5.1.1. *Let*

$$\boldsymbol{W} = (W_1, W_2, \ldots, W_n) \sim \textit{N-IG}(a_1, a_2, \ldots, a_{n+1}),$$

where all $a_i \geq 0$ $i = 1, 2, \ldots, n+1$ *and at least one of them is strictly positive. Then,* $\forall\ N \in \mathbb{N},$ *and* $\forall\ i = 1, 2, \ldots, n,$

$$E(W_i^N) = \sum_{t=-2N+2}^{N-1} \sum_{l=0}^{N-1} c_N(t+2l+1, l) I_t \tag{5.1.1}$$

where

$$c_N(k,l) = \begin{pmatrix} N-1 \\ l \end{pmatrix} \frac{(-1)^{N+l+1} a_i^k e^a d_N(k)}{2^{N-1}\Gamma(N)k!},\ d_N(k) = \sum_{m=0}^{\lfloor \frac{k-1}{2} \rfloor} \begin{pmatrix} k \\ 2m+1 \end{pmatrix} \prod_{j=0}^{N-1} (2m+1-2j),$$

$[x]$ *is the floor function for* $x \in \mathbb{R},\ a = \sum_{i=1}^{n+1} a_i$ *and* $I_t = \int_1^\infty e^{-au} u^t du.$

Proof:

Since $\boldsymbol{W}$ is N-IG-distributed, each W_i can be written as $\frac{V_i}{V},\ i = 1, 2, \ldots, n,$ where $V = \sum_{j=1}^{n+1} V_i$ and each of those $V_i,\ i = 1, 2, \ldots, n+1$ is inverse-Gaussian distributed and independently of each other. Then:

$$\begin{aligned}
\mathrm{E}(W_i^N) &= \mathrm{E}\left(\frac{V_i^N}{V^N}\right) \\
&= \mathrm{E}\left(\frac{V_i^N}{\Gamma(N)} \int_0^\infty u^{N-1} e^{-uV} du\right) \\
&= \frac{1}{\Gamma(N)} \int_0^\infty u^{N-1} \mathrm{E}\left(V_i^N e^{-u(V_i+V_{-i})}\right) du, \text{ where } V_{-i} = V - V_i \qquad (5.1.2) \\
&= \frac{1}{\Gamma(N)} \int_0^\infty u^{N-1} \mathrm{E}\left(V_i^N e^{-uV_i}\right) \mathrm{E}(e^{-uV_{-i}}) du \\
&= \frac{(-1)^N}{\Gamma(N)} \int_0^\infty u^{N-1} \mathrm{E}\left(\frac{\partial^N}{\partial u^N}(e^{-uV_i})\right) \mathrm{E}(e^{-uV_{-i}}) du \\
&= \frac{(-1)^N}{\Gamma(N)} \int_0^\infty u^{N-1} \mathrm{E}(e^{-uV_{-i}}) \frac{\partial^N}{\partial u^N} \left(\mathrm{E}(e^{-uV_i})\right) du \qquad (5.1.3) \\
&= \frac{(-1)^N}{\Gamma(N)} \int_0^\infty u^{N-1} e^{-(a-a_i)(\sqrt{2u+1}-1)} \underbrace{\frac{\partial^N}{\partial u^N} \left(e^{-a_i(\sqrt{2u+1}-1)}\right)}_{(1)} du. \qquad (5.1.4)
\end{aligned}$$

In the above (5.1.2) is an application of the Fubini Theorem and (5.1.3) is an application of Theorem (16.8) in Billingsley (1995). Finally, for (5.1.4) the know formula for the moment generating function of the inverse-Gaussian distribution was used.

The difficult part in (5.1.4) is to calculate (1), i.e. the N-th derivative of the function $e^{-a_i(\sqrt{2u+1}-1)}$. This is possible using Meyer's formula, which is a variation of Faa di Bruno's formula (see, for example, Johnson (2002)):

$$\frac{\partial^N}{\partial x^N}(g \circ f)(x) = \frac{\partial^N}{\partial x^N}\left[g(f(x))\right] = \sum_{k=0}^{N} \frac{g^{(k)}(f(x))}{k!} \left\{ \frac{\partial^N}{\partial h^N}\left[f(x+h) - f(x)\right]^k \Big|_{h=0} \right\} \tag{5.1.5}$$

where $g^{(k)}$ is the $k-$th derivative of the function g.
In this case, $g(x) = e^x$ and $f(x) = -a_i(\sqrt{2x+1} - 1)$. So,
$f(x+h) - f(x) = a_i(\sqrt{2x+1} - \sqrt{2x+2h+1})$

$$\Rightarrow [f(x+h) - f(x)]^k = a_i^k(\sqrt{2x+1} - \sqrt{2x+2h+1})^k$$
$$= a_i^k \sum_{m=0}^{k} \begin{pmatrix} k \\ m \end{pmatrix} (-1)^m (\sqrt{2x+2h+1})^m (\sqrt{2x+1})^{k-m}$$

$$\Rightarrow \frac{\partial^N}{\partial h^N}[f(x+h) - f(x)]^k = a_i^k \sum_{m=0}^{k} \begin{pmatrix} k \\ m \end{pmatrix} (-1)^m (\sqrt{2x+1})^{k-m} \left[\frac{\partial^N}{\partial h^N}\left[(2x+2h+1)^{m/2}\right]\right].$$

The last derivative will be: $\frac{\partial^N}{\partial h^N}\left((2x+2h+1)^{m/2}\right) = (2x+2h+1)^{\frac{m-2N}{2}} \prod_{j=0}^{N-1}(m-2j)$.
At $h = 0$, we get: $\frac{\partial^N}{\partial h^N}\left((2x+2h+1)^{m/2}\right) = (2x+1)^{\frac{m-2N}{2}} \prod_{j=0}^{N-1}(m-2j)$.
So, using (5.1.5), we get:

$$\frac{\partial^N}{\partial u^N}\left(e^{-a_i(\sqrt{2u+1}-1)}\right) = -\frac{e^{-a_i(\sqrt{2u+1}-1)}}{(2u+1)^N} \sum_{k=1}^{N} \frac{a_i^k}{k!}(2u+1)^{k/2} \sum_{\substack{m=1 \\ m:\text{odd}}}^{k} \begin{pmatrix} k \\ m \end{pmatrix} \prod_{j=0}^{N-1}(m-2j).$$

In the above, the sum over k starts from 1, since for $k = 0$ the term is zero. As a result, the counter m can also be set to start from 1. Additionally, m can be set to take only odd values, since the terms in the sum over m will be zero for even values of m, since $m \leq N$.
By using a slightly different parametrisation of the counter m, we can substitute

$$\sum_{\substack{m=1 \\ m:\text{odd}}}^{k} \begin{pmatrix} k \\ m \end{pmatrix} \prod_{j=0}^{N-1}(m-2j) \text{ with } \sum_{m=0}^{\lfloor \frac{k-1}{2} \rfloor} \begin{pmatrix} k \\ 2m+1 \end{pmatrix} \prod_{j=0}^{N-1}(2m+1-2j).$$

So,

$$\frac{\partial^N}{\partial u^N}\left(e^{-a_i(\sqrt{2u+1}-1)}\right) = -\frac{e^{-a_i(\sqrt{2u+1}-1)}}{(2u+1)^N} \sum_{k=1}^{N} \frac{a_i^k}{k!}(2u+1)^{k/2} \sum_{m=0}^{\lfloor \frac{k-1}{2} \rfloor} \begin{pmatrix} k \\ 2m+1 \end{pmatrix} \prod_{j=0}^{N-1}(2m+1-2j).$$

Plugging the above expression into (5.1.4), we get:
$\mathrm{E}(W_i^N) = \frac{(-1)^{N+1}}{\Gamma(N)} \sum_{k=1}^{N} \frac{a_i^k}{k!} d_N(k) \int_0^\infty e^{-a(\sqrt{2u+1}-1)} u^{N-1}(2u+1)^{k/2-N} du$,
where $d_N(k) = \sum_{m=0}^{\lfloor \frac{k-1}{2} \rfloor} \begin{pmatrix} k \\ 2m+1 \end{pmatrix} \prod_{j=0}^{N-1}(2m+1-2j)$.

The integral in the last equation can be found as follows:

$$\begin{aligned}\int_0^\infty e^{-a(\sqrt{2u+1}-1)}u^{N-1}(2u+1)^{k/2-N}du &= \int_1^\infty e^{-a(y-1)}\frac{(y^2-1)^{N-1}}{2^{N-1}}y^{k-2N+1}dy,\\ &\qquad\qquad \text{using the substitution } y=\sqrt{2u+1}\\ &= \frac{1}{2^{N-1}}\int_1^\infty e^{-a(y-1)}\sum_{l=0}^{N-1}\binom{N-1}{l}(-1)^l y^{2N-2-2l}y^{k-2N+1}dy,\\ &\qquad\qquad \text{using the binomial theorem}\\ &= \frac{1}{2^{N-1}}\sum_{l=0}^{N-1}\binom{N-1}{l}(-1)^l e^a\int_1^\infty e^{-ay}y^{k-2l-1}dy.\end{aligned}$$

So, we have:

$$\mathrm{E}(W_i^N)=\sum_{k=1}^{N}\frac{(-1)^{N+1}}{\Gamma(N)2^{N-1}}\frac{a_i^k}{k!}e^a d_N(k)\sum_{l=0}^{N-1}\binom{N-1}{l}(-1)^l I^*_{k,l}$$

where $I^*_{k,l}=\int_1^\infty y^{k-2l-1}e^{-ay}dy$

The above expression is, of course, a valid one. However, it can be simplified even more by noting that the integrals $I^*_{k,l}$ depend on k and l only through $k-2l$. This will reduce the number of integrals to be calculated from N^2 to $3N-2$.

By clustering $I^*_{k,l}$ according to $t=k-2l-1\in\{-2N+2,-2N+3,\ldots,-1,0,1,2,\ldots,N-1\}$ (and also replacing k by $t+1+2l$), and denoting the clustered integrals by $I_t=\int_1^\infty e^{-au}u^t du$, we get the desired formula. □

Another interesting formula than can be derived using the same basic method as above is the one for the cross moments of each pair $W_i, W_j,\ i\neq j$, where both W_i and W_j are components of the same N-IG-distributed $\boldsymbol{W}$:

Theorem 5.1.2. *Let*

$$\boldsymbol{W}=(W_1,W_2,\ldots,W_n)\sim N\text{-}IG(a_1,a_2,\ldots,a_{n+1}),$$

where all $a_i\geq 0\ i=1,2,\ldots,n+1$ *and at least one of them is strictly positive. Then,*
$\forall\ N_1,N_2\in\mathbb{N},$ *and* $\forall\ i,j\in\{1,2,\ldots,n\},\ i\neq j,$

$$E(W_i^{N_1}W_j^{N_2})=\sum_{t=-2N_1-2N_2+3}^{N_1+N_2-1}\ \sum_{l=1}^{N_2}\ \sum_{m=0}^{N_1+N_2-1}c_{N_1,N_2}(t+1+2m-l,l,m)\ I_t \qquad (5.1.6)$$

where $c_{N_1,N_2}(k,l,m)=\frac{(-1)^{N_1+N_2+m}e^a}{\Gamma(N_1+N_2)2^{N_1+N_2-1}}\frac{a_i^k a_j^l}{k!l!}d_{N_1}(k)d_{N_2}(l)\binom{N_1+N_2-1}{m},\ a=\sum_{i=1}^{n+1}a_i$

$d_N(k)=\sum_{m=0}^{\lfloor\frac{k-1}{2}\rfloor}\binom{k}{2m+1}\prod_{j=0}^{N-1}(2m+1-2j)$ *and* $I_t=\int_1^\infty e^{-au}u^t du.$

Proof:

For simplicity, in the following we have substituted $N_1 + N_2$ by N. We also used the fact that $W_i = \frac{V_i}{V}, W_j = \frac{V_j}{V}$, $V = \sum_{k=1}^{n+1} V_k$ for some independent inverse-Gaussian-distributed V_k, $k = 1, 2, \ldots, n+1$.

$$\begin{aligned}
\mathrm{E}(W_i^{N_1} W_j^{N_2}) &= \mathrm{E}\left(\frac{V_i^{N_1} V_j^{N_2}}{V^N}\right) \\
&= \mathrm{E}\left(\frac{V_i^{N_1} V_j^{N_2}}{\Gamma(N)} \int_0^\infty u^{N-1} e^{-uV} du\right) \\
&= \frac{1}{\Gamma(N)} \int_0^\infty u^{N-1} \mathrm{E}\left(V_i^{N_1} V_j^{N_2} e^{-u(V_i+V_j+V_{-i,j})}\right) du \text{ where } V_{-i,j} = V - V_i - V_j \\
&= \frac{1}{\Gamma(N)} \int_0^\infty u^{N-1} \mathrm{E}\left(V_i^{N_1} e^{-uV_i}\right) \mathrm{E}\left(V_j^{N_2} e^{-uV_j}\right) \mathrm{E}(e^{-uV_{-i,j}}) du \\
&= \frac{(-1)^N}{\Gamma(N)} \int_0^\infty u^{N-1} \mathrm{E}\left(\frac{\partial^{N_1}}{\partial u^{N_1}}(e^{-uV_i})\right) \mathrm{E}\left(\frac{\partial^{N_2}}{\partial u^{N_2}}(e^{-uV_j})\right) \prod_{k\neq i,j} \mathrm{E}(e^{-uV_k}) du \\
&= \frac{(-1)^N}{\Gamma(N)} \int_0^\infty u^{N-1} \frac{\partial^{N_1}}{\partial u^{N_1}} \left(\mathrm{E}(e^{-uV_i})\right) \frac{\partial^{N_2}}{\partial u^{N_2}} \left(\mathrm{E}(e^{-uV_j})\right) \prod_{k\neq i,j} \mathrm{E}(e^{-uV_k}) du \\
&= \frac{(-1)^N}{\Gamma(N)} \int_0^\infty u^{N-1} \frac{\partial^{N_1}}{\partial u^{N_1}} \left(e^{-a_i(\sqrt{2u+1}-1)}\right) \frac{\partial^{N_2}}{\partial u^{N_2}} \left(e^{-a_j(\sqrt{2u+1}-1)}\right) e^{-(a-a_i-a_j)(\sqrt{2u+1}-1)} du.
\end{aligned}$$

Again, using Meyer's formula and the transformation $y = \sqrt{2u+1}$ in the derived integral, we get:

$$\mathrm{E}(W_i^{N_1} W_j^{N_2}) = \sum_{k=1}^{N_1} \sum_{l=1}^{N_2} \sum_{m=0}^{N_1+N_2-1} c_N(k,l,m) I^*_{k,l,m}$$

where $c_{N_1,N_2}(k,l,m) = \frac{(-1)^{N_1+N_2+m} e^a}{\Gamma(N_1+N_2) 2^{N_1+N_2-1}} \frac{a_i^k a_j^l}{k!l!} d_{N_1}(k) d_{N_2}(l) \begin{pmatrix} N_1+N_2-1 \\ m \end{pmatrix}$,

$I^*_{k,l,m} = \int_1^\infty e^{-ay} y^{k+l-2m-1} dy$,

$d_{N_1}(k) = \sum_{q=0}^{\lfloor \frac{k-1}{2} \rfloor} \begin{pmatrix} k \\ 2q+1 \end{pmatrix} \prod_{j=0}^{N_1-1} (2q+1-2j)$, and

$d_{N_2}(l) = \sum_{q=0}^{\lfloor \frac{k-1}{2} \rfloor} \begin{pmatrix} k \\ 2q+1 \end{pmatrix} \prod_{j=0}^{N_2-1} (2q+1-2j)$.

Again, by clustering the integrals $I^*_{k,l,m}$ according to $t = k + l - 2m - 1 \in \{-2N+3, -2N+4, \ldots, -1, 0, 1, 2, \ldots, N-1\}$, we arrive at expression (5.1.6). □

Notice also that, using the above method, the general expression for the moments of products of powers of more that two quantities, for example $\mathrm{E}(W_i^{N_1} W_j^{N_2} W_k^{N_3})$, can also be calculated.

5.1.1 Some results

Using Theorems 5.1.1 and 5.1.2, some basic moment results are derived:

Corollary 5.1.1. *Let* $\boldsymbol{W} = (W_1, W_2, \ldots, W_n)$ *be distributed as in Theorem 5.1.2 and* $i, j \in \{1, 2, \ldots, n\}$, $i \neq j$. *Then, the following hold:*

1. $E(W_i) = \frac{a_i}{a}$.
2. $E(W_i^2) = \left(\frac{a_i}{a}\right)^2 + a_i(a - a_i)e^a\Gamma(-2, a)$.
3. $Var(W_i) = a_i(a - a_i)e^a\Gamma(-2, a)$.
4. $E(W_i^3) = \frac{a_i}{16a}[6 - 10a + (a - a_i)(a - 2a_i)(a - 1) + \frac{a_i(2a_i+3a)(a+1)}{a^2}$
   $+ a(a - a_i)(a^2 - 2aa_i - 12)e^aE_i(-a)]$, *where* $E_i(a) = -\int_{-a}^{\infty} \frac{e^{-t}}{t}dt$.
5. $E(W_i - E(W_i))^3 = -\frac{a_i}{8a^3}(2a_i^2 - 3a_ia + a^2)[3 - a + a^2(a^2 - 12)e^a\Gamma(-2, a)]$.
6. *The skewness coefficient of* W_i,

   $$Sk(W_i) := \frac{E\left(W_i - E(W_i)\right)^3}{\left(Var(W_i)\right)^{3/2}}$$

   is given by:

   $$Sk(W_i) = -\frac{(2a_i^2 - 3a_ia + a^2)\left(3 - a + a^2(a^2 - 12)e^a\Gamma(-2, a)\right)}{8e^{3a/2}\sqrt{a_i}a^{7/2}\left[(a - a_i)\Gamma(-2, a)\right]^{3/2}}.$$

7. $E(W_iW_j) = \frac{a_ia_j(a+1)}{2a^2} + \frac{a_ia_je^aE_i(-a)}{2}$.
8. $Cov(W_i, W_j) = -a_ia_je^a\Gamma(-2, a)$.
9. $Corr(W_i, W_j) = -\sqrt{\frac{a_i}{a-a_i}\frac{a_j}{a-a_j}}$.

The above results are appealing, as the expressions for the mean and variance of any element of the vector $\boldsymbol{W}$, as well as the correlation of any two elements in $\boldsymbol{W}$ are really simple ones. It is also nice that the effect of the rest of the elements in $\boldsymbol{W}$ in the above results is expressed only through the total sum of the parameters, a. Alternatively, if we denote the ratios $\frac{a_i}{a}$ by p_i, we get the simple formulae:

1. $\mathrm{E}(W_i) = p_i$.
2. $\mathrm{Var}(W_i) = p_i(1 - p_i)a^2e^a\Gamma(-2, a)$.
3. $\mathrm{Corr}(W_i, W_j) = -\sqrt{\frac{p_i}{1-p_i}\frac{p_j}{1-p_j}}$.

The last three results are also shown in Lijoi et al. (2005).

Proof:

Note: the calculation of the integrals I_t can be found in the subsection directly after this proof.

1. For the first one, we use Theorem 5.1.1 for $N = 1$. Therefore, $l = 0, t = 0, c_1(1, 0) = a_i e^a,\ I_0 = \frac{e^{-a}}{a},$ so $\mathrm{E}(W_i) = \frac{a_i}{a}$.

2. For the second one, we again use Theorem 5.1.1, but now for $N = 2$. So, $l \in \{0, 1\}, t \in \{-2, -1, 0, 1\}$ and $c_2(1, 0) = a_i e^a/2,\ c_2(1, 1) = -a_i e^a/2,\ c_2(2, 0) = a_i^2 e^a/2,\ c_2(2, 1) = -a_i^2 e^a/2, I_{-2} = aE_i(-a) + e^{-a},\ I_{-1} = -E_i(-a),\ I_0 = \frac{e^{-a}}{a}$ and $I_1 = e^{-a}\left(\frac{1}{a} + \frac{1}{a^2}\right)$.
   Putting all these together, we get the desired result.

3. The third part is straightforward from the two results above and the elementary formula:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}^2(X).$$

4. The next result is easy to verify, using Theorem 5.1.1 for $N = 3$ and some tedious algebra.

5. The third central moment of W_i follows immediately from the above and some algebra.

6. For the skewness of W_i, all we need is to use the definition of skewness, and some algebraic calculations.

7. The expectation of the product of W_i and W_j can be shown using Theorem 5.1.2 and $N_1 = N_2 = 1$. Therefore, $l = 1,\ m \in \{0, 1\}$ and $t \in \{-1, 0, 1\}$, so $c_{1,1}(1, 1, 0) = \frac{a_i a_j e^a}{2},\ c_{1,1}(1, 1, 1) = \frac{-a_i a_j e^a}{2}, I_{-1} = -E_i(-a),\ I_0 = \frac{e^{-a}}{a}$ and $I_1 = e^{-a}\left(\frac{1}{a} + \frac{1}{a^2}\right)$ and we arrive at the desired result.

8. The covariance of W_i and W_j can be directly derived using the above results and the known formula:

$$\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y).$$

9. Finally, the correlation between W_i and W_j follows from the above and the formula:

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{E}(X)\mathrm{E}(Y)}}.$$ □

5.1.2 Calculating the integrals I_t

Consider the integral $I_t = \int_1^\infty e^{-au} u^t du$.

For $t = 0$,

$$I_0 = \int_1^\infty e^{-au} du = \frac{e^{-a}}{a}.$$

For $t = -1$,

$$I_{-1} = \int_1^\infty e^{-au} u^{-1} du = -E_i(-a)$$

where $E_i(a) = -\int_{-a}^\infty \frac{e^{-t}}{t} dt$ is the known exponential-integral function. It is also trivial to show the following relationship between $E_i(a)$ and the incomplete gamma function $\Gamma(a, x) = \int_x^\infty e^{-t} t^{a-1} dt$:

$$E_i(-a) = \frac{e^{-a}}{a^2}(1 - a) - 2\Gamma(-2, a). \tag{5.1.7}$$

For $t \leq -2$,

$$I_t = (-1)^t \frac{a^{-t-1} E_i(-a)}{(-t-1)!} - e^{-a} \sum_{k=0}^{-t-2} \frac{a^k}{t(t+1)\dots(t+k)}.$$

For $t \geq 1$,

$$I_t = e^{-a} \sum_{k=0}^{t} \frac{t!}{k!} a^{k-t-1}.$$

The above results were taken from Gradshteyn and Ryzhik (1994). Alternatively, see Abramowitz and Stegun (1964).

5.1.3 The one-dimensional N-IG distribution

In modelling overdispersed count data, the one-dimensional N-IG distribution will be used. So, it would be useful to just state its density function, general moment results and some special moments for this case:

Let $X \sim$ N-IG(a_1, a_2), where $a_1, a_2 > 0$. Then, as stated in Lijoi et al. (2005) , its probability density function is

$$f_X(x) = \frac{e^{a_1+a_2} a_1 a_2 K_{-1}(\sqrt{A_2(x)})}{\pi x^{3/2}(1-x)^{3/2}\sqrt{A_2(x)}}, \ 0 < x < 1 \tag{5.1.8}$$

where $A_2(x) = \frac{a_1^2}{x} + \frac{a_2^2}{1-x}$ and K is the modified Bessel function of the third type.

Corollary 5.1.2. *Let X be distributed as in (5.1.8). It holds that $\forall\ N \in \mathbb{N}$,*

$$E(X^N) = \sum_{t=-2N+2}^{N-1} \sum_{l=0}^{N-1} c_N(t + 2l + 1, l) I_t \tag{5.1.9}$$

where

$$c_N(k, l) = \begin{pmatrix} N-1 \\ l \end{pmatrix} \frac{(-1)^{N+l+1} a_1^k e^{a_1+a_2} d_N(k)}{2^{N-1}\Gamma(N)k!}, \ d_N(k) = \sum_{m=0}^{\lfloor \frac{k-1}{2} \rfloor} \begin{pmatrix} k \\ 2m+1 \end{pmatrix} \prod_{j=0}^{N-1} (2m + 1 - 2j)$$

and $I_t = \int_1^\infty e^{-(a_1+a_2)u} u^t du$.

Proof:

The above result is straightforward, as a special case of Theorem 5.1.1. □

Another interesting result that can again be used in the calculation of MLEs of the parameters in a N-IG-binomial model is the following:

Corollary 5.1.3. *Let X be distributed as in (5.1.8). It holds that* $\forall\ N_1, N_2 \in \mathbb{N}$,

$$E\left(X^{N_1}(1-X)^{N_2}\right) = \sum_{t=-2N_1-2N_2+3}^{N_1+N_2-1} \sum_{l=1}^{N_2} \sum_{m=0}^{N_1+N_2-1} c_{N_1,N_2}(t+1+2m-l,l,m) I_t \tag{5.1.10}$$

where $c_{N_1,N_2}(k,l,m) = \frac{(-1)^{N_1+N_2+m} e^{a_1+a_2}}{\Gamma(N_1+N_2) 2^{N_1+N_2-1}} \frac{a_1^k a_2^l}{k!l!} d_{N_1}(k) d_{N_2}(l) \begin{pmatrix} N_1+N_2-1 \\ m \end{pmatrix}$,

$d_N(k) = \sum_{m=0}^{\lfloor \frac{k-1}{2} \rfloor} \begin{pmatrix} k \\ 2m+1 \end{pmatrix} \prod_{j=0}^{N-1}(2m+1-2j)$ *and* $I_t = \int_1^\infty e^{-(a_1+a_2)u} u^t du$.

Proof:

It follows from Theorem 5.1.2, for $X = W_i$ and $W_j = 1 - X$. □

In this case, even if we did not have the more general Theorem 5.1.2, one could derive a slightly different expression for $\mathrm{E}\left(X^{N_1}(1-X)^{N_2}\right)$, using the binomial theorem for expressing $(1-X)^{N_2}$ as a polynomial of X, and then using Theorem 5.1.1, or Corollary 5.1.2. On the other hand, in this way the derived expression has more summation terms than in Corollary 5.1.2.

Some moment results:

Let $X \sim$ N-IG(a_1, a_2). Using the results of Section 5.1.2, we have:

1. $\mathrm{E}(X) = \frac{a_1}{a_1+a_2}$.
2. $\mathrm{E}(X^2) = \left(\frac{a_i}{a_1+a_2}\right)^2 + a_1 a_2 e^{a_1+a_2} \Gamma(-2, a_1+a_2)$.
3. $\mathrm{Var}(X) = a_1 a_2 e^{a_1+a_2} \Gamma(-2, a_1+a_2)$.
4. $\mathrm{E}(X^3) = \frac{E_i(-(a_1+a_2)) e^{a_1+a_2}}{16} \left[a_1(a_1+a_2)^3 + 6a_1^2(a_1+a_2) + 2a_1^3(a_1+a_2) - 12a_1(a_1+a_2) + 12a_1^2\right]$
$+\frac{a_1}{64(a_1+a_2)^3} \left[(a_1+a_2)^5 - 2(a_1+a_2)^4 - 18(a_1+a_2)^3 + 12a_1(a_1+a_2)^3 + 4a_1^2(a_1+a_2)^3\right.$
$\left.+24(a_1+a_2)^2 - 8a_1^2(a_1+a_2)^2 + 24a_1(a_1+a_2)^2 + 24a_1(a_1+a_2) + 16a_1^2 + 16a_1^2(a_1+a_2)\right]$.

5.2 A More General Class of Distributions

A more general class of distributions is the tempered stable distribution, which was introduced by Tweedie (1984):

Definition 14. *Let $\kappa > 0$, $\delta > 0$ and $\gamma > 0$. A random variable X follows a tempered stable distribution with parameters κ, δ and γ if its Lévy density is*

$$u(x) = \delta 2^{\kappa} \frac{\kappa}{\Gamma(1-\kappa)} x^{-1-\kappa} \exp\left\{-\frac{1}{2}\gamma^{1/\kappa} x\right\}.$$

We will write $X \sim TS(\kappa, \delta, \gamma)$.

In general, the probability density function is not available analytically. On the other hand, due to its relationship to the positive stable distribution (see Feller (1971)), it can be expressed through the following series representation:

$$p(x|\kappa, \delta, \gamma) = c \sum_{k=1}^{\infty} (-1)^{(k-1)} \sin(k\pi\kappa) \frac{\Gamma(k\kappa+1)}{k!} 2^{k\kappa+1} (x\delta^{-1/\kappa})^{(-k\kappa-1)} \exp\left\{-\frac{1}{2}\gamma^{1/\kappa} x\right\}$$

where $c = \frac{1}{2\pi}\delta^{-1/\kappa} \exp\{\delta\gamma\}$.
The expectation of X is $2\kappa\delta\gamma^{(\kappa-1)/\kappa}$ and its variance is $4\kappa(1-\kappa)\delta\gamma^{(\kappa-2)/\kappa}$. The moment generating function will be important for our derivations and is given by

$$\mathrm{E}(\exp\{tx\}) = \exp\{\delta\gamma - \delta(\gamma^{1/\kappa} - 2t)^{\kappa}\}. \tag{5.2.11}$$

The tempered stable distribution is infinitely divisible and self-decomposable. It has previously been applied to several problems. Barndorff-Nielsen and Shephard (2001) model stock prices using scale mixture of normal distribution where the mixing distribution is tempered stable and Palmer et al. (2008) apply it to the modelling of cell generation times.

There are two important subclasses. A TS $\left(\kappa, \frac{\nu}{\kappa\psi^{2\kappa}}, \psi^{2\kappa}\right)$ will limit in probability as $\kappa \to 0$ to a gamma distribution with probability density function

$$p(x) = \frac{(\psi^2/2)^{\nu}}{\Gamma(\nu)} x^{\nu-1} \exp\left\{-\frac{1}{2}\psi^2 x\right\}$$

and the inverse-Gaussian distribution arises when $\kappa = \frac{1}{2}$, with derived probability density function

$$p(x) = \frac{\delta}{\sqrt{2\pi}} \exp\{\delta\gamma\} x^{-3/2} \exp\left\{-\frac{1}{2}(\delta^2 x^{-1} + \gamma^2 x)\right\}.$$

In other words, due to the extra parameter κ, the TS distribution is a general class of distributions, covering the gamma and the inverse-Gaussian distributions. Because of its infinite divisibility, we will also be able to construct a new distribution by normalising it, in a similar fashion to how the Dirichlet and the normalised inverse-Gaussian distributions can be seen as normalised gamma and inverse-Gaussian distributions. As a result, the derived distribution will have the Dirichlet and the N-IG distributions as special cases. In the following of this section it is assumed, without loss of generality, that $\gamma = 1$ (since γ is a scale parameter, as seen from (5.2.11)).

Definition 15. *Let* $0 < \kappa < 1$ *and* $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_{n+1})$ *be a vector of positive numbers. If* $V_1, V_2, \ldots, V_{n+1}$ *are independent tempered stable random variables with* $V_i \sim TS(\kappa, \frac{\nu_i}{\kappa}, 1)$ *and*

$$W_i = \frac{V_i}{V_1 + V_2 + \ldots V_{n+1}}$$

then $\boldsymbol{W} = (W_1, W_2, \ldots, W_n)$ *follows a multivariate normalised tempered stable distribution with parameters* $\boldsymbol{\nu}$ *and* κ *which we write as* $MNTS(\nu_1, \nu_2, \ldots, \nu_{n+1}; \kappa)$.

As mentioned above, both the Dirichlet and the normalised inverse-Gaussian distribution are two special cases of this distribution. The Dirichlet distribution arises as $\kappa \to 0$ and the normalised inverse-Gaussian distribution arises if $\kappa = 1/2$:

$$MNTS(\nu_1, \nu_2, \ldots, \nu_{n+1}; \kappa) \xrightarrow{\kappa \to 0} \text{Dir}(\nu_1, \nu_2, \ldots, \nu_{n+1})$$

in probability, and

$$MNTS(\nu_1, \nu_2, \ldots, \nu_{n+1}; 1/2) \equiv \text{N-IG}(2\nu_1, 2\nu_2, \ldots, 2\nu_{n+1}).$$

All the moments and cross-moments of the n-dimensional MNTS distribution exist (since the distribution is defined on the n-th dimensional unit simplex) and they can be calculated using the following theorems:

Theorem 5.2.1. *Suppose that* $\boldsymbol{W} = (W_1, W_2, \ldots, W_n) \sim MNTS(\nu_1, \nu_2, \ldots, \nu_{n+1}; \kappa)$ *and let* $N, N_1, N_2 \in \mathbb{N}$. *It follows that:*

1. $E\left(W_i^N\right) = \sum_{l=1}^{N} \sum_{j=0}^{N-1} b_N(l, j)\, \Gamma\left(l - j/\kappa, \frac{S}{\kappa}\right)$
2. $E\left(W_i^{N_1} W_j^{N_2}\right) = \sum_{l=1}^{N_1} \sum_{m=1}^{N_2} \sum_{t=0}^{N_1+N_2-1} c_{N_1,N_2}(l, m, t)\, \Gamma\left(l + m - t/\kappa, \frac{S}{\kappa}\right)$

where

$$b_N(l, j) = \begin{pmatrix} N-1 \\ j \end{pmatrix} \frac{(-1)^{N+j}\ (S/\kappa)^{j/\kappa}\ \exp\left\{\frac{S}{\kappa}\right\}\ \ d_N(\kappa, l)}{\Gamma(N)\ l! \kappa} \mu_i^l,$$

$$c_{N_1,N_2}(l, m, t) = \begin{pmatrix} N_1 + N_2 - 1 \\ t \end{pmatrix} \frac{(-1)^{N_1+N_2+t}\ (S/\kappa)^{t/\kappa}\ \exp\left\{\frac{S}{\kappa}\right\}\ d_{N_1}(\kappa, l)\ d_{N_2}(\kappa, m)}{\Gamma(N_1 + N_2)\ l!\ m!\ \kappa} \mu_i^l \mu_j^m,$$

$d_N(\kappa, l) = \sum_{i=1}^{l} \begin{pmatrix} l \\ i \end{pmatrix} (-1)^i \prod_{c=0}^{N-1} (\kappa i - c)$, $S = \sum_{i=1}^{n+1} \nu_i$, $\mu_i = \frac{\nu_i}{S}$ *and* $\Gamma(a, x) = \int_x^\infty t^{a-1}\ \exp\{-t\} dt$ *is the incomplete gamma function.*

Proof:

Again, the same procedure as in the previous theorems is followed, which is based on the method of Lijoi et al. (2005) and James et al. (2006).

We start by noting that, since $\boldsymbol{W} \sim \text{MNTS}(\nu_1, \nu_2, \ldots, \nu_{n+1}; \kappa)$, then there exist independent $V_i \sim \text{TS}(\kappa, \frac{\nu_i}{\kappa}, 1),\ i = 1, 2, \ldots, n+1$ such that $W_i = \frac{V_i}{V},\ i = 1, 2, \ldots, n$ and $V = \sum_{i=1}^{n+1} V_i$. So,

$$\begin{aligned}
\text{E}\left(W_i^N\right) &= \text{E}\left(\frac{V_i^N}{V^N}\right) \\
&= \text{E}\left(\frac{V_i^N}{\Gamma(N)} \int_0^\infty u^{N-1} e^{-uV} du\right) \\
&= \frac{1}{\Gamma(N)} \int_0^\infty u^{N-1} \prod_{j \neq i} \text{E}\left(e^{-uV_j}\right)\ \text{E}\left(V_i^N e^{-uV_i}\right)\ du \qquad (5.2.12) \\
&= \frac{(-1)^N}{\Gamma(N)} \int_0^\infty u^{N-1} \prod_{j \neq i} \text{E}\left(e^{-uV_j}\right)\ \text{E}\left(\frac{\partial^N}{\partial u^N} e^{-uV_i}\right)\ du \\
&= \frac{(-1)^N}{\Gamma(N)} \int_0^\infty u^{N-1} \prod_{j \neq i} \text{E}\left(e^{-uV_j}\right) \frac{\partial^N}{\partial u^N}\ \text{E}\left(e^{-uV_i}\right)\ du \qquad (5.2.13) \\
&= \frac{(-1)^N}{\Gamma(N)} \exp\left\{\frac{S}{\kappa}\right\} \int_0^\infty u^{N-1}\ \exp\left\{-\sum_{j \neq i} \frac{\nu_j}{\kappa}(1+2u)^\kappa\right\} \underbrace{\frac{\partial^N}{\partial u^N} \exp\left\{-\frac{\nu_i}{\kappa}(1+2u)^\kappa\right\}}_{(1)} du. \qquad (5.2.14)
\end{aligned}$$

Again, (5.2.12) is an application of the Fubini Theorem and (5.2.13) can be seen as an application of Theorem (16.8) of Billingsley (1995). Finally, for (5.2.14) I used the formula for the moment generating function of the tempered stable distribution (5.2.11).

As before, the difficult part is expression (1) in (5.2.14), i.e. the N-th derivative of the function $\exp\left\{-\frac{\nu_i}{\kappa}(1+2u)^\kappa\right\}$. Calculating this integral is again feasible using Meyer's formula, and the final result is:

$$\begin{aligned}
\frac{\partial^N}{\partial u^N} \exp\left\{-\frac{\nu_i}{\kappa}(1+2u)^\kappa\right\} &= 2^N \sum_{l=1}^N \frac{\exp\left\{-\frac{\nu_i}{\kappa}(1+2u)^\kappa\right\}}{l!} \frac{\nu_i^l}{\kappa^l} \sum_{j=1}^l \binom{l}{j} (-1)^j (1+2u)^{\kappa l - N} \prod_{c=0}^{N-1} (\kappa j - c) \\
&= 2^N \sum_{l=1}^N \frac{\exp\left\{-\frac{\nu_i}{\kappa}(1+2u)^\kappa\right\}}{l!} \frac{\nu_i^l}{\kappa^l} (1+2u)^{\kappa l - N} d_N(\kappa, l) \qquad (5.2.15)
\end{aligned}$$

where

$$d_N(\kappa, l) = \sum_{j=1}^l \binom{l}{j} (-1)^j \prod_{c=0}^{N-1} (\kappa j - c) = \sum_{j=1}^l \binom{l}{j} (-1)^j \frac{\Gamma(\kappa j + 1)}{\Gamma(\kappa j - N + 1)}.$$

It is now straightforward to verify that

$$\text{E}\left(W_i^N\right) = \frac{(-1)^N 2^N \exp\left\{\frac{S}{\kappa}\right\}}{\Gamma(N)} \sum_{l=1}^N \frac{\frac{\nu_i^l}{\kappa^l} d_N(\kappa, l)}{l!} \int_0^\infty u^{N-1}\ \exp\left\{-\frac{S}{\kappa}(1+2u)^\kappa\right\} (1+2u)^{\kappa l - N} du.$$

The integral in the last expression can be simplified using the substitution $y = (1+2u)^\kappa$ and the binomial theorem:

$$\int_0^\infty u^{N-1} \exp\left\{-\frac{S}{\kappa}(1+2u)^\kappa\right\}(1+2u)^{\kappa l-N}du = \frac{1}{2^N\kappa\frac{S^l}{\kappa^l}}\sum_{m=0}^{N-1}\left(\begin{array}{c} N-1 \\ m \end{array}\right)(-1)^m\frac{S^{m/\kappa}}{\kappa^{m/\kappa}}\Gamma\left(l-m/\kappa,\frac{S}{\kappa}\right)$$

and therefore

$$\mathrm{E}\left(W_i^N\right) = \frac{(-1)^N\exp\left\{\frac{S}{\kappa}\right\}}{\Gamma(N)\ \kappa}\sum_{l=1}^{N}\frac{\nu_i^l d_N(\kappa,l)}{l!\ S^l}\sum_{m=0}^{N-1}\left(\begin{array}{c} N-1 \\ m \end{array}\right)(-1)^m\left(\frac{S}{\kappa}\right)^{m/\kappa}\Gamma\left(l-m/\kappa,\frac{S}{\kappa}\right).$$

For the cross-moments, we have:

$$\begin{aligned}
\mathrm{E}\left(W_i^{N_1}W_j^{N_2}\right) =& \mathrm{E}\left(\frac{V_i^{N_1}V_j^{N_2}}{V^N}\right),\ \text{where } N = N_1+N_2\\
=& \mathrm{E}\left(\frac{V_i^{N_1}V_j^{N_2}}{\Gamma(N)}\int_0^\infty u^{N-1}\ e^{-uV}du\right)\\
=& \frac{(-1)^N}{\Gamma(N)}\int_0^\infty u^{N-1}\ \mathrm{E}\left(\frac{\partial^{N_1}}{\partial u^{N_1}}e^{-uV_i}\right)\ \mathrm{E}\left(\frac{\partial^{N_2}}{\partial u^{N_2}}e^{-uV_j}\right)\prod_{t\neq i,j}\mathrm{E}\left(e^{-uV_t}\right)du\\
=& \frac{(-1)^N}{\Gamma(N)}\int_0^\infty u^{N-1}\ \frac{\partial^{N_1}}{\partial u^{N_1}}\mathrm{E}\left(e^{-uV_i}\right)\ \frac{\partial^{N_2}}{\partial u^{N_2}}\mathrm{E}\left(e^{-uV_j}\right)\ \prod_{t\neq i,j}\mathrm{E}\left(e^{-uV_t}\right)du\\
=& \frac{(-1)^N}{\Gamma(N)}\exp\left\{\frac{S}{\kappa}\right\}\int_0^\infty\ u^{N-1}\ \frac{\partial^{N_1}}{\partial u^{N_1}}\left(e^{-\frac{\nu_i}{\kappa}(1+2u)^\kappa}\right)\ \frac{\partial^{N_2}}{\partial u^{N_2}}\left(\exp\left\{-\frac{\nu_j}{\kappa}(1+2u)^\kappa\right\}\right)\\
&\times\exp\left\{-\sum_{t\neq i,j}\frac{\nu_t}{\kappa}(1+2u)^\kappa\right\}du.
\end{aligned}$$

Using the result for the $N-$th derivative of the function $\exp\left\{-\frac{\nu_i}{\kappa}(1+2u)^\kappa\right\}$ from above, we find:

$$\mathrm{E}\left(W_i^{N_1}W_j^{N_2}\right) = \frac{(-1)^N}{\Gamma(N)}\exp\left\{\frac{S}{\kappa}\right\}2^N\sum_{l=1}^{N_1}\sum_{m=1}^{N_2}\frac{d_{N_1}(\kappa,l)d_{N_2}(\kappa,m)}{l!m!}\frac{\nu_i^l}{\kappa^l}\frac{\nu_j^m}{\kappa^m}I^*_{l,m}(\kappa)$$

where

$$I^*_{l,m}(\kappa) = \int_0^\infty u^{N-1}\ \exp\left\{-\frac{S}{\kappa}(1+2u)^\kappa\right\}(1+2u)^{\kappa(l+m)-N}\ du.$$

Using the same substitution as above, $y = (1+2u)^\kappa$, in $I^*_{l,m}(\kappa)$, together with the binomial theorem, we find:

$$\mathrm{E}\left(W_i^{N_1}W_j^{N_2}\right) = \sum_{l=1}^{N_1}\sum_{m=1}^{N_2}\sum_{t=0}^{N-1}c_{N_1,N_2}(l,m,t)\Gamma\left(l+m-t/\kappa,\frac{S}{\kappa}\right) \tag{5.2.16}$$

where

$$c_{N_1,N_2}(l,m,t) = \left(\begin{array}{c} N_1+N_2-1 \\ t \end{array}\right)\frac{(-1)^{N_1+N_2+t}\ (S/\kappa)^{t/\kappa}\ \exp\left\{\frac{S}{\kappa}\right\}\ d_{N_1}(\kappa,l)\ d_{N_2}(\kappa,m)}{\Gamma(N_1+N_2)\ l!\ m!\ \kappa}\mu_i^l\mu_j^m.\quad\square$$

Notice that in this, more general case, one cannot cluster the incomplete gamma functions, as was done in the N-IG distribution case. Such a clustering can only be done for rational values of κ, for example 1/2 or 1/3. For such cases, the clustering will be made according to the first argument of the incomplete gamma function, which is then a function of the indexes l and j or l, m and t (and since the second argument is the same for all terms). This procedure will require a sum over $u = l - j/\kappa$ or $u = l + m - t/\kappa$ in the first and second expression, respectively. This new sum can then replace one of the existing sums in each expression. On the other hand, one might find this clustering appealing, both in terms or algebraic simplicity, as well as for computational convenience.

Algorithms for the calculation of the incomplete gamma function are described by Zhang and Jin (1996) and implementations in Fortran and Matlab are available. Unfortunately, the corresponding command in Matlab proved to be inefficient for negative arguments. On the other hand, there is a built-in command for the incomplete gamma function in Mathematica, which I used in my calculations. This command is exact, even for negative arguments (since Mathematica is a symbolic language, and therefore exact).

As in the case of the N-IG distribution, the effect of the components of $\boldsymbol{W}$ which are not included in the moment calculations is only through the sum of their corresponding parameters, $S - \nu_i$ (for the simple moments) or $S - \nu_i - \nu_j$ (for the cross-moments).

It is also worth mentioning that the function $d_N(\kappa, l)$ defined above is related to the generalized Stirling numbers, or generalized factorial coefficients, $G(n, k, \sigma)$ (see, for example, Charalambides and Singh (1988) and Charalambides (2005)), through the simple formula

$$d_N(\kappa, l) = (-1)^N l! G(N, l, k).$$

Finally, the ideas could be easily extended to the case of deriving the moments of products of powers of more that two quantities, such as $\mathrm{E}\left(W_i^{N_1} W_j^{N_2} W_k^{N_3}\right)$, in which case the procedure and the final formula will be similar to the above (in this case, with four sums).

5.2.1 Some basic moment results

Corollary 5.2.1. *If* $\boldsymbol{W} = (W_1, W_2, \ldots, W_n) \sim MNTS(\nu_1, \nu_2, \ldots, \nu_{n+1}; \kappa)$ *then*

$$E(W_i) = \frac{\nu_i}{S} = \mu_i$$

$$Var(W_i) = (1-\kappa)\mu_i(1-\mu_i)\left[1 - \left(\frac{S}{\kappa}\right)^{1/\kappa} \exp\left\{\frac{S}{\kappa}\right\} \Gamma\left(1 - 1/\kappa, \frac{S}{\kappa}\right)\right]$$

$$Cov(W_i, W_j) = \mu_i\mu_j \left[\kappa + \kappa\frac{S}{\kappa} - \kappa \exp\left\{\frac{S}{\kappa}\right\} \left(\frac{S}{\kappa}\right)^{1/\kappa} \Gamma\left(2 - 1/\kappa, \frac{S}{\kappa}\right) - 1 \right]$$
$$= \mu_i\mu_j(1-\kappa) \left[\exp\left\{\frac{S}{\kappa}\right\} \left(\frac{S}{\kappa}\right)^{1/\kappa} \Gamma\left(1 - 1/\kappa, \frac{S}{\kappa}\right) - 1 \right]$$

$$Corr(W_i, W_j) = -\sqrt{\frac{\mu_i}{1-\mu_i}\frac{\mu_j}{1-\mu_j}}.$$

Proof:

For the first moment we apply the first formula of Theorem 5.2.1 for $N = 1$:

$$b_1(1,0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \frac{(-1)^1 \exp\left\{\frac{S}{\kappa}\right\} \nu_i^1}{\Gamma(1)1!\kappa^{1+0/\kappa}\ S^{1-0/\kappa}}(-\kappa), \text{ and } d_1(\kappa,1) = -\kappa \Rightarrow b_1(1,0) = \exp\left\{\frac{S}{\kappa}\right\}\frac{\nu_i}{S}.$$

The results follows from noting that $\Gamma\left(1 - 0/\kappa, \frac{S}{\kappa}\right) = \Gamma\left(1, \frac{S}{\kappa}\right) = \int_{\frac{S}{\kappa}}^{\infty} \exp\{-t\}dt = \exp\left\{-\frac{S}{\kappa}\right\}$. The second moment is

$$\mathrm{E}\left(W_i^2\right) = -(1-\kappa)\mu_i(1-\mu_i)\frac{S^{1/\kappa}}{\kappa^{1/\kappa}} \exp\left\{\frac{S}{\kappa}\right\} \Gamma\left(1 - 1/\kappa, \frac{S}{\kappa}\right) + \mu_i(1 - \kappa + \mu_i\kappa)$$

since $d_2(\kappa,1) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} (-1)^1(\kappa - 0)(\kappa - 1) = \kappa(1-\kappa)$ and $d_2(\kappa,2) = 2\kappa^2$, which implies that $b_2(1,0) = (1-\kappa)\exp\left\{\frac{S}{\kappa}\right\}\mu_i$, $b_2(1,1) = -(1-\kappa)\exp\left\{\frac{S}{\kappa}\right\}\mu_i\frac{S^{1/\kappa}}{\kappa^{1/\kappa}}$, $b_2(2,0) = \kappa\exp\left\{\frac{S}{\kappa}\right\}\mu_i^2$ and $b_2(2,1) = -\kappa\exp\left\{\frac{S}{\kappa}\right\}\mu_i^2\frac{S^{1/\kappa}}{\kappa^{1/\kappa}}$. The results follows from the fact that $\Gamma\left(1 - 0/\kappa, \frac{S}{\kappa}\right) = \exp\left\{-\frac{S}{\kappa}\right\}$ and $\Gamma\left(2 - 0/\kappa, \frac{S}{\kappa}\right) = \exp\left\{-\frac{S}{\kappa}\right\}\left(1 + \frac{S}{\kappa}\right)$. The cross-moment $E(W_iW_j)$ can be calculated using the second part of Theorem 5.2.1 for $N_1 = N_2 = 1$.

We only have to calculate $c_{1,1}(1,1,0)$ and $c_{1,1}(1,1,1)$. Noting that $d_1(\kappa,1) = -\kappa$ it follows that

$$c_{1,1}(1,1,0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \frac{(-1)^2 \exp\left\{\frac{S}{\kappa}\right\}(-\kappa)^2\frac{\nu_i^1}{\kappa^1}\frac{\nu_j^1}{\kappa}}{\Gamma(2)1!1!\kappa\frac{S^{1+1-0/\kappa}}{\kappa^{1+1-0/\kappa}}} = \exp\left\{\frac{S}{\kappa}\right\}\kappa\ \frac{\nu_i\nu_j}{S^2} = \exp\left\{\frac{S}{\kappa}\right\}\kappa\mu_i\mu_j$$

$$c_{1,1}(1,1,1) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \frac{(-1)^3 \exp\left\{\frac{S}{\kappa}\right\}\nu_i^1\nu_j^1}{\Gamma(2)1!1!\kappa\frac{S^{1+1-1/\kappa}}{\kappa^{1+1-1/\kappa}}} = -\exp\left\{\frac{S}{\kappa}\right\}\kappa\frac{S^{1/\kappa}}{\kappa^{1/\kappa}}\frac{\nu_i\nu_j}{S^2} = -\exp\left\{\frac{S}{\kappa}\right\}\kappa\mu_i\mu_j\frac{S^{1/\kappa}}{\kappa^{1/\kappa}}.$$

The fact that $\Gamma\left(1 + 1 - 0/\kappa, \frac{S}{\kappa}\right) = \Gamma\left(2, \frac{S}{\kappa}\right) = \exp\left\{-\frac{S}{\kappa}\right\}\left(1 + \frac{S}{\kappa}\right)$ implies that

$$\mathrm{E}\left(W_iW_j\right) = \kappa\mu_i\mu_j \left[1 + \frac{S}{\kappa} - \exp\left\{\frac{S}{\kappa}\right\}\left(\frac{S}{\kappa}\right)^{1/\kappa} \Gamma\left(2 - 1/\kappa, \frac{S}{\kappa}\right)\right].$$

Subtracting $\mathrm{E}(W_i)\mathrm{E}(W_j)$ from the above, we derive the formula for the covariance of W_i and W_j. Finally, by dividing the covariance of W_i and W_j by the square root of the product of their variances, we get the desired result. □

The results shown in Corollary 5.2.1 are quite pleasant and the expectation and correlation structure are the same as in the simpler Dirichlet and N-IG distributions. The expectation of W_i does not depend on κ and the variance only depends on $\boldsymbol{\nu}$ through the sum S, (similar to Dirichlet and the N-IG distributions). In fact, the form of the variance generalizes the form for the Dirichlet distribution since the variance only depends on $\boldsymbol{\nu}$ through the mean $\mu_i = \frac{\nu_i}{\sum_{i=1}^{n+1} \nu_i}$ and $S = \sum_{i=1}^{n+1} \nu_i$. Therefore we can write $\text{Var}(W_i) = \alpha(\kappa, S)\mu_i(1-\mu_i)$ for some function α. Finally, the expression for the correlation does not depend on κ and it is particularly simple.

5.2.2 The normalised tempered stable distribution

The univariate MNTS, which will be called the normalized tempered stable distribution (with parameters, say, ν_1, ν_2 and κ, $\text{NTS}(\nu_1, \nu_2; \kappa)$) is an important special case and the one that will used in the next subsection. As $\kappa \to 0$, the distribution tends to a beta distribution with parameters ν_1 and ν_2, whereas for $k = 1/2$ we get the univariate N-IG distribution with parameters $2\nu_1$ and $2\nu_2$.

The general form of the moments of this distribution can be easily derived from Theorem 5.2.1. From the same theorem the general form of the expectation of $\text{E}\big(X^{N_1}(1-X)^{N_2}\big)$ can also be derived, where $X \sim \text{NTS}(\nu_1, \nu_2; \kappa)$, by setting $W_i = X$ and $W_j = 1 - X$ in the second part of the theorem. The formula for the cross-moments will be particularly useful in the maximum likelihood estimation of the parameters of the models in Section 5.3.2.
The two central moments are:

$$\text{E}(X) = \frac{\nu_1}{\nu_1 + \nu_2} =: \mu$$

$$\text{Var}(X) = (1-\kappa)\mu(1-\mu)\left[1 - \left(\frac{\nu_1+\nu_2}{\kappa}\right)^{1/\kappa} \exp\left\{\frac{\nu_1+\nu_2}{\kappa}\right\} \Gamma\left(1 - 1/\kappa, \frac{\nu_1+\nu_2}{\kappa}\right)\right].$$

As κ increases, the tempered stable distribution becomes heavier tailed and this carries over to the NTS distribution. For example, Figure 5.1 shows how the variance changes with κ (left), how the kurtosis changes with κ (middle) and the relationship between the two, for a $\text{NTS}(\nu_1, \nu_2; \kappa)$ with $\nu_1 = \nu_2$ distribution. Kurtosis is defined as the standardised fourth central moment:

$$\text{Kurt}(X) = \frac{E(X - E(X))^4}{(Var(X))^2}.$$

It is clear from Figure 5.1 that the variance decreases as κ increases. The shape of the variance as a function of κ is the same for other values of the first moment $\mu = \frac{\nu_1}{\nu_1+\nu_2}$, however the values of the variance become smaller as μ moves further from $1/2$.

From the shape of the graph of kurtosis plotted against κ, one can see the point made above that as κ increases, the tails of the underlying TS distributions become heavier. Notice especially the

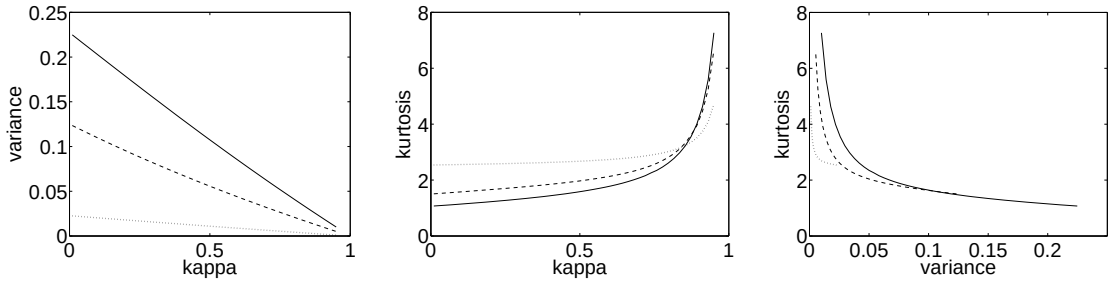


Figure 5.1: The Variance and kurtosis of NTS distribution with mean 0.5: (a) shows κ versus the variance, (b) shows κ versus the kurtosis and (c) shows variance versus kurtosis. In each graph: $S = 0.1$ (solid line), $S = 1$ (dashed line) and $S = 10$ (dotted line).

dramatic increase in kurtosis for values of κ greater than 0.8. The shape of this graph is preserved for all values of the other parameters, ν_1 and ν_2, whereas it is interesting to see that the minimum kurtosis is not always achieved at the limiting case $\kappa \to 0$, although the value at this limit is very close to the overall minimum. For $\mu \to 0$, the value of κ that gives the minimum value of kurtosis tends to 0.2, whereas for not very small values of μ, the case $\kappa \simeq 0$ seems to provide the smallest kurtosis. There is also symmetry around $\mu = 1/2$, in the sense of that for μ and $1 - \mu$, we get the same graph. The values of kurtosis increase as μ moves away from $1/2$, whereas for large values of κ, kurtosis decreases as $S = \nu_1 + \nu_2$ increases and for small values of κ, kurtosis increases as S increases.

In the right graph in Figure 5.1 we see the relationship between the variance and the kurtosis for $\mu = 0.5$. The shape again is the same for other values of the parameters ν_1 and ν_2 and the graph is exactly the same for μ and $1 - \mu$. The beta distribution corresponds to the point at the right end of the graph (i.e. for largest variance).

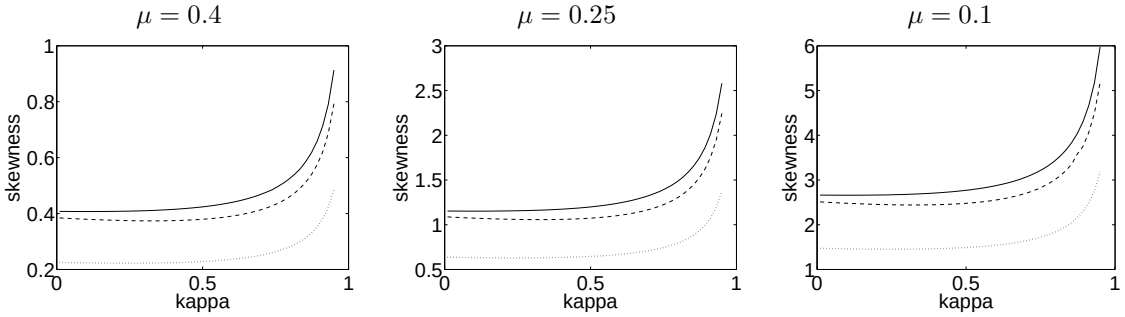


Figure 5.2: Skewness vs κ for various values of the mean for some MNTS distributions. In each graph: $S = 0.1$ (solid line), $S = 1$ (dashed line) and $S = 10$ (dotted line).

Let skewness of a distribution denote its standardised third central moment, i.e.

$$\text{Skew}(X) = \frac{E(X - E(X))^3}{(Var(X))^{3/2}}.$$

If $\mu = 1/2$ the skewness is zero for all values of κ. Figure 5.2 shows the skewness against κ for various values of μ. I only plot the skewness $\text{Skew}(\mu, S, \kappa)$ for $\mu < 0.5$ since $Skew(\mu, S, \kappa) = -Skew(1 - \mu, S, \kappa)$ (which follows from the construction of the distribution). As the value of μ moves away from 1/2, the values of skewness also increase in absolute terms. On the other hand, when the value of $S = \nu_1 + \nu_2$ increases, skewness is decreased in absolute value. Finally, note that, as in the case of the kurtosis, the minimum skewness (maximum, for $\mu > 1/2$) is not achieved for $\kappa \simeq 0$, but usually for some value between 0 and 0.6.

In Figure 5.3 I plotted the relationship between skewness and variance (left) and kurtosis vs skewness (right). For both graphs I used distributions with $\mu < 0.5$. The beta distribution is again at the right end of this graph, whereas the minimum skewness is not necessarily at the same point. The graph of kurtosis versus skewness is the most intriguing ones. The reason for this is the little curl of the curve at its endpoint where we have the smallest values of kurtosis, and it is caused by the fact that the minimum for skewness is not achieved at the limiting case $\kappa \simeq 0$, as happens for the kurtosis for not very small values of μ. In this case, the endpoint of the graph for which we have minimum kurtosis corresponds to the beta distribution. For cases of very small values for the mean, the same endpoint corresponds to the beta distribution, but to neither minimum skewness, nor the minimum kurtosis and the graph shows curliness for both those quantities (rather than only for the skewness, as in Figure 5.3).

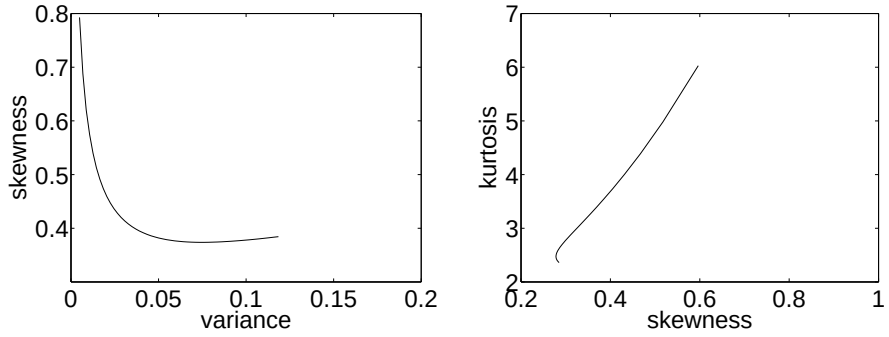


Figure 5.3: Skewness vs variance and kurtosis vs skewness for some MNTS distributions.

Finally, it is worth mentioning that for any mean and variance, one can find a NTS distribution that has exactly those central moments. This is not a surprise, since there are three parameters in this distribution, and only two parameters are needed to match these two conditions. On the other hand, the extra parameter can result in a better fit of the distribution of some data, for example by

matching also the skewness or the kurtosis. This is something one cannot do with some competing distributions, for example the beta distribution, since there are only two parameters there, and therefore by fixing the mean and the variance, all the other moments will also be fixed. In fact, for fixed mean and variance, the skewness coefficient increases (in absolute terms) as the value of κ increases (and, of course, $\nu_1 \neq \nu_2$), so NTS distributions can be particularly useful in cases where we believe that the skewness of the data is greater (in absolute terms) than the one implied by the beta distribution (i.e. for $\kappa \simeq 0$).

Having stated the characteristics and the differences of my proposed distribution from the beta distribution, I will use it as the distribution of the probabilities of success for binomial data. The derived model will be compared with the widely used beta-binomial (BB) model, as well as with other models proposed in the literature, using both simulated and real-world data.

5.3 Modelling Overdispersed Count Data

5.3.1 A brief literature review

In many experiments we observe data as the number of observations with some property out of a sample. The binomial distribution is a natural model for these type of data. However, the data are often found to be overdispersed which is often explained and modelled through differences in the binomial success probability p for different units. It is assumed that x_i successes (*i.e.* observations possessing a certain property) are observed from n_i observations and that $x_i \sim \text{Bi}(n_i, p_i)$ where p_i follows some distribution. The likelihood is formed through the cross-moments of this distribution (see equation (5.3.17)) and the beta is a natural choice for the distribution of the p_i, since these cross-moments are simply calculated. This leads to the beta-binomial (BB) model. However, this distribution may be misspecified leading to biased estimates of the parameters of the model. Other possible choices of distribution include the logistic-normal-binomial model (Williams, 1982) and the probit-normal-binomial model (Ochi and Prentice, 1984). Several authors have considered alternative specifications. Altham (1978) and Kupper and Haseman (1978) propose a two-parameter distribution, the correlated-binomial, which allows for direct interpretation and assignment of the correlation between any two of the underlying Bernoulli observations of a binomial random variable through one of its two parameters. Paul (1985) proposes a three-parameters generalisation of the beta-binomial distribution, the beta-correlated binomial distribution, with a modified version of the latter in Paul (1987). Brooks et al. (1997) use finite mixture distributions to provide a flexible specification. However, the introduction of a mixture distribution leads to more complicated inference and harder interpretation of parameters. Kuk (2004) suggests the q-power distribution which mod-

els the joint success probabilities of all orders by a power family of completely monotone functions which extends the folded logistic class of George and Bowman (1995). Pang and Kuk (2005) define a shared response model by allowing each response to be independent of all other with probability π or taking a value Z with probability $1-\pi$. Therefore more than one observation can take the common value Z. They argue that this is more interpretable than the q-power distribution of Kuk (2004). Rodríguez-Avi et al. (2007) use a generalized beta distribution as the mixing distribution.

5.3.2 The NTS-binomial distribution

My proposed model is the NTS-binomial model, where it is assumed that we have data coming from a binomial distribution with probabilities of success p_i that follow a NTS distribution. In other word, the NTS distribution is used as the mixing distribution in a binomial model.

Definition 16. *A random variable X is said to follow a NTS-binomial distribution if and only if:*

$$X \sim Bin(n, p)$$

$$p \sim NTS(\nu_1, \nu_2; \kappa)$$

for parameters $\nu_1, \nu_2 > 0,\ 0 < \kappa < 1$ and $n \in \mathbb{N}$.

Of course, in the limit $\kappa \to 0$, the above model tends to the BB model.

We can interpret the mixing distribution $\text{NTS}(\nu_1, \nu_2; \kappa)$ as representing the heterogeneity in the probability of success across the different observed groups. The response can be written as the sum of n Bernoulli random variables, $X = \sum_{i=1}^{n} Z_i$ where $Z_1, Z_2, \ldots, Z_n$ are i.i.d. Bernoulli with success probability P where $P \sim \text{NTS}(\nu_1, \nu_2; \kappa)$. The intra-group correlation is defined as $\text{Corr}(Z_i, Z_j)$ which in my model has the form

$$\rho = (1-\kappa)\left[1 - \left(\frac{S}{\kappa}\right)^{1/\kappa} \exp\left\{\frac{S}{\kappa}\right\} \Gamma\left(1 - 1/\kappa, \frac{S}{\kappa}\right)\right].$$

where $S = \nu_1 + \nu_2$.
In these mixture models the variance can be written as

$$\text{Var}(X) = n^2 \text{Var}(P) + n\text{E}\left(P(1-P)\right)$$

where $P \sim \text{NTS}(\nu_1, \nu_2; \kappa)$. The first term of this sum can be interpreted as the variance due to differences between individuals in the sample (between-subject variance), whereas the second term represents the intra-subject variability. In my model these have a simple form

$$\text{Var}(P) = (1-\kappa)\mu(1-\mu)\left[1 - \left(\frac{S}{\kappa}\right)^{1/\kappa} \exp\left\{\frac{S}{\kappa}\right\} \Gamma\left(1 - 1/\kappa, \frac{S}{\kappa}\right)\right]$$

$$\begin{aligned}\mathrm{E}(P(1-P)) &= \mu(1-\mu)\left\{1-(1-\kappa)\left[1-\left(\frac{S}{\kappa}\right)^{1/\kappa}\exp\left\{\frac{S}{\kappa}\right\}\Gamma\left(1-1/\kappa,\frac{S}{\kappa}\right)\right]\right\}\\ &= \mu(1-\mu)(1-\rho).\end{aligned}$$

where $\mu = \mathrm{E}(P) = \frac{\nu_1}{\nu_1+\nu_2}$.

Estimation of the parameters in a NTS-binomial model can be easily performed using maximum likelihood estimation (MLE) methods. The formulae for the moments derived in the previous sections will be particularly useful for those methods:

Let

$$x_i \overset{ind}{\sim} \mathrm{Bin}(n_i, p_i),\ i = 1, 2, \ldots, N$$

$$p_i \overset{iid}{\sim} \mathrm{NTS}(\nu_1, \nu_2; \kappa),\ i = 1, 2, \ldots, N.$$

Then, the likelihood of the data $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, after integrating out the parameters p_i, is:

$$\begin{aligned} f(\mathbf{x}) &= \int_0^1 \cdots \int_0^1 f(\mathbf{x}|p_1, \ldots, p_N) f(p_1, \ldots, p_N) dp_1 \ldots dp_N \\ &= \prod_{i=1}^N \int_0^1 f(x_i|p_i) f(p_i) dp_i \\ &= \prod_{i=1}^N \int_0^1 \left(\begin{array}{c} n_i \\ x_i \end{array}\right) p_i^{x_i}(1-p_i)^{n_i-x_i} f(p_i) dp_i \\ &= \prod_{i=1}^N \left(\begin{array}{c} n_i \\ x_i \end{array}\right) \mathrm{E}_{p_i}\left(p_i^{x_i}(1-p_i)^{n_i-x_i}\right) \end{aligned} \tag{5.3.17}$$

where the expectation in the last line is taken with respect to p_i and can be calculated using the formulae proven before.

An alternative model would consider the same number of trials, n, for each x_i. The results in this case would be almost the same, as we just need to replace n_i with n in all the above.

If we now have data $x_1, x_2, \ldots, x_N$, the MLE of the parameters ν_1, ν_2 and κ, will be a triplet of numbers, denoted by $\hat{\nu}_1, \hat{\nu}_2$ and $\hat{\kappa}$, that maximise (5.3.17). Needless to say, these estimates will have to satisfy the natural requirements for the parameters, i.e. $\hat{\nu}_1, \hat{\nu}_2 > 0$ and $0 < \hat{\kappa} < 1$.

The standard errors of these estimates can be estimated using the asymptotic result:

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \to N(\mathbf{0}, (I(\boldsymbol{\theta}))^{-1})$$

where $\boldsymbol{\theta}$ is the parameter of interest, $\hat{\boldsymbol{\theta}}$ is its MLE and $I(\boldsymbol{\theta})$ is the Fisher information matrix. In this case, the vectors are of dimension 3 (since $\boldsymbol{\theta} = (\nu_1, \nu_2, \kappa)$) and I will be a 3×3 matrix, whose $(j,k)-$th element $I_{j,k}$ is given by $-\mathrm{E}\left(\frac{\partial^2 logL}{\partial u_j \partial u_k}; \boldsymbol{x}\right)$, where $\boldsymbol{x}$ is the data set, $logL$ is the log-likelihood,

$u_1 = \nu_1,\ u_2 = \nu_2$ and $u_3 = \kappa$. By using simple differentiation and the fact that the data are independent, it can be shown that:

$$\begin{aligned}
\frac{\partial^2 logL}{\partial u_j \partial u_k} &= \frac{\partial}{\partial u_j}\left(\frac{\partial logL}{\partial u_k}\right) \\
&= \frac{\partial}{\partial u_j}\left(\frac{\partial}{\partial u_k}\left(\sum_{i=1}^{N}\log(\mathrm{E}_i)\right)\right) \\
&= \sum_{i=1}^{N}\frac{\partial}{\partial u_j}\left(\frac{\partial}{\partial u_k}\log(\mathrm{E}_i)\right) \\
&= \sum_{i=1}^{N}\frac{\partial}{\partial u_j}\left(\frac{\frac{\partial \mathrm{E}_i}{\partial u_k}}{\mathrm{E}_i}\right) \\
&= \sum_{i=1}^{N}\left[\frac{\frac{\partial^2 \mathrm{E}_i}{\partial u_j \partial u_k}}{\mathrm{E}_i} - \frac{\frac{\partial \mathrm{E}_i}{\partial u_j}\frac{\partial \mathrm{E}_i}{\partial u_k}}{\mathrm{E}_i^2}\right]
\end{aligned}$$

where E_i is (apart from the binomial coefficient, which is dropped during the differentiation, as it does not involve any of the parameters) the likelihood of the data, having integrated the p_i's out, $\mathrm{E}_i = \mathrm{E}(p_i^{x_i}(1-p_i)^{n_i - x_i})$.
So, $I_{jk} = -\sum_{i=1}^{N}\mathrm{E}\left[\frac{\frac{\partial^2 \mathrm{E}_i}{\partial u_j \partial u_k}}{\mathrm{E}_i} - \frac{\frac{\partial \mathrm{E}_i}{\partial u_j}\frac{\partial \mathrm{E}_i}{\partial u_k}}{\mathrm{E}_i^2}\right]$, where the expectation is taken with respect to each x_i, whose distribution is E_i, defined on $\{0, 1, 2, \ldots, n_i\}$. Finally, I is evaluated at the MLE of the parameters $\nu_i,\ \nu_2$ and κ and its inverse is calculated.

However, when I followed this procedure, the method failed: in some cases the matrix was not positive semi-definite and/or the values at the diagonal of its inverse (i.e. where asymptotic variances should be) were negative. So, I thought that the problem might be that the regularity condition

$$\int \frac{\partial^2}{\partial\theta_i \partial\theta_j} f(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x} = 0$$

might not hold (if we were allowed to change the differentiation with the integration, i.e. to integrate and then differentiate, it would of course have been trivial).
As a result, we used the actual definition of the Fisher information matrix, i.e.

$$I_{i,j} = \mathrm{E}\left(\frac{\partial logL}{\partial u_i}\frac{\partial logL}{\partial u_j}; \boldsymbol{x}\right).$$

Following the same procedure as before, we have:

$$\begin{aligned}\frac{\partial logL}{\partial u_i}\frac{\partial logL}{\partial u_j} &= \sum_{k=1}^{N}\sum_{l=1}^{N}\frac{\partial \log(\mathrm{E}_k)}{\partial u_i}\frac{\partial \log(\mathrm{E}_l)}{\partial u_j} \\ &= \sum_{k=1}^{N}\sum_{l=1}^{N}\frac{\frac{\partial \mathrm{E}_k}{\partial u_i}}{\mathrm{E}_k}\frac{\frac{\partial \mathrm{E}_l}{\partial u_j}}{\mathrm{E}_l} \\ &= \sum_{k,l=1;\ l\neq k}^{N}\frac{\frac{\partial \mathrm{E}_k}{\partial u_i}}{\mathrm{E}_k}\frac{\frac{\partial \mathrm{E}_l}{\partial u_j}}{\mathrm{E}_l}+\sum_{k=1}^{N}\frac{\frac{\partial \mathrm{E}_k}{\partial u_i}\frac{\partial \mathrm{E}_k}{\partial u_j}}{\mathrm{E}_k^2}.\end{aligned}$$

The reason for splitting the double sum in this way in the last equation is that for $k = l$ the expectation will be taken with respect to the respective x_k, whereas when $k \neq l$, the expectation will be taken with respect to both x_k and x_l. So, the $(i, j)-$th element of I will be

$$\sum_{k,l=1;\ l\neq k}^{N}\mathrm{E}\left(\frac{\frac{\partial \mathrm{E}_k}{\partial u_i}}{\mathrm{E}_k}\frac{\frac{\partial \mathrm{E}_l}{\partial u_j}}{\mathrm{E}_l}\right)+\sum_{k=1}^{N}\mathrm{E}\left(\frac{\frac{\partial \mathrm{E}_k}{\partial u_i}\frac{\partial \mathrm{E}_k}{\partial u_j}}{\mathrm{E}_k^2}\right)$$

where the expectation is taken with respect to the corresponding x_k or (x_k, x_l), with distribution(s) E_k or $\mathrm{E}_k E_l$. Finally, we evaluate I at the MLE of the parameters ν_i, ν_2 and κ and calculate its inverse.

The above formulae for the standard errors were given for general n_i. The case where all the data have the same number of Bernoulli trials, say n, follows immediately.

5.3.3 Simulated data

Based on the observation that the NTS distributions exhibit greater (in absolute terms) skewness than the beta distribution, for the same mean and variance, I first considered two data sets, which were obviously highly skewed:

In both cases the actual data were $x_i \sim \mathrm{Bin}(6, p_i),\ i = 1, 2, \ldots, N$.

1. This data set has size 660. I simulated 660 p_i's where: 20 values are uniformly distributed on (0,0.1), 50 values are uniformly distributed on (0.1,0,2), 90 values are uniformly distributed on (0.2,0.3), 120 values are uniformly distributed on (0.3,0.4), 150 values are uniformly distributed on (0.4,0.5), 110 values are uniformly distributed on (0.5,0.6), 60 values are uniformly distributed on (0.6,0.7), 30 values are uniformly distributed on (0.7,0.8), 20 values are uniformly distributed on (0.8,0.9) and 10 values are uniformly distributed on (0.9,1). We choose $n_i = 6\ \forall\ i = 1, 2, \ldots, 660$. The p_i's have a mean of 0.4384, a variance of 0.0374 and a skewness of 0.2880. This skewness is higher than the skewness implied by the beta distribution for the same mean and variance (which is 0.1693), so one would expect that the MLE for κ would be different than zero.

2. I followed the same procedure as (1), now with 770 data, but now with different proportions in the intervals: 20 in (0,0.1), 90 in (0.1,0. 2), 120 in (0.2,0.3), 150 in (0.3,0.4), 160 in (0.4,0.5), 110 in (0.5,0.6), 60 in (0.6,0.7), 30 in (0.7,0.8), 20 in (0.8,0.9) and 10 in (0.9,1). The mean of the p_i's was 0.413, the variance was 0.0381 and the skewness was 0.4026 (higher than the 0.25 for the corresponding beta distribution).

The distribution of the underlying probabilities p_i described above are shown in Figure 5.4. The

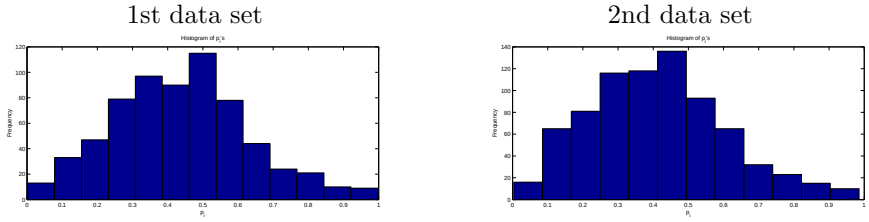


Figure 5.4: Histogram of p_i used in the first two simulated data sets.

maximum likelihood estimates of the parameters in the NTS-binomial and the beta-binomial model are shown in Table 5.1. As one would expect, these values were quite different and the more general NTS-binomial model provided a higher likelihood of the data than the BB model. On the other hand, when using the Akaike Information Criterion, $AIC = -2logL + 2k$ (where k is the number of parameters in a model), the more parsimonious beta-binomial model was preferred in both cases.

	NTS-binomial			beta-binomial	
Data	$\hat{\kappa}$	$\hat{\nu}_1$	$\hat{\nu}_2$	$\hat{\nu}_1$	$\hat{\nu}_2$
1	0.50	0.91	1.16	2.56	3.24
2	0.55	0.695	1.049	2.295	3.464

Table 5.1: Maximum likelihood estimates of the NTS-binomial and beta-binomial models for the first two simulated data sets.

I then considered 22 more simulated data sets, created using a more systematic method. More specifically, we used:

$$x_i \overset{ind}{\sim} Bin(n, p_i),\ i = 1, 2, \ldots, N$$

$$p_i \overset{iid}{\sim} NTS(\nu_1, \nu_2; \kappa),\ i = 1, 2, \ldots, N$$

and:

- For data sets 1-3: $n = 6, \nu_1 = 1, \nu_2 = 1/2, \kappa = 1/3$ and $N = 500, 1000$ and 1500.
- For data sets 4-6: $n = 6, \nu_1 = 4, \nu_2 = 1, \kappa = 3/5$ and $N = 500, 1000$ and 2000.
- For data sets 7-9: $n = 6, \nu_1 = 1, \nu_2 = 4, \kappa = 3/5$ and $N = 500, 1000$ and 2000.

- For data sets 10-12: $n = 6, \nu_1 = 2, \nu_2 = 4, \kappa = 4/5$ and $N = 500, 1000$ and 2000.
- For data sets 13-15: $n = 12, \nu_1 = 1, \nu_2 = 1/2, \kappa = 1/3$ and $N = 500, 1000$ and 1500.
- For data sets 16-18: $n = 12, \nu_1 = 1, \nu_2 = 4, \kappa = 3/5$ and $N = 500, 1000$ and 1500.
- For data sets 19-21: $n = 12, \nu_1 = 2, \nu_2 = 4, \kappa = 4/5$ and $N = 500, 1000$ and 1500.
- For data set 22: using beta-binomial data (i.e. with $\kappa \simeq 0$), with $n = 6, \nu_1 = 1, \nu_2 = 1$ and $N = 500$.

The data sets with $n = 12$ were used for comparison purposes when more trials were used in the binomial step.

Results:

In the above data sets, the order of the data sets is in accordance to the data size, with the smaller number of data corresponding to the first data set index in each triplet. For example, data sets 1-3 correspond to $n = 6, \nu_1 = 1, \nu_2 = 1/2, \kappa = 1/3$ and data set 1 has $N = 500$, data set 2 has $N = 1000$ and data set 3 has $N = 1500$. Similarly, data sets 4-6 correspond to $n = 6, \nu_1 = 4, \nu_2 = 1$ and $\kappa = 3/5$, with data size of 500 for data set 4, data size of 1000 for data set 5 and data size of 2000 for data set 6 etc.

The MLEs of the parameters when a NTS-binomial and a beta-binomial distribution were fitted to each data set are shown in Table 5.2. The numbers in parentheses are the standard errors of these estimates, calculated as in Section 5.3.2.

The first thing to notice is that for the last data set, the algorithm correctly "discovers" that the data come from a beta-binomial data, as the MLE for κ is very close to 0. Unfortunately, due to the fact that the value of κ is very close to 0, it was not possible to get the standard errors of the estimates in this case, using the Mathematica routine that I used for all the rest.

A second observation is that as the number of data increases, the MLEs are closer to the values of ν_1, ν_2 and κ that created the data, as one would expect. As for the standard errors of the estimates, they seem to be decreasing with N, but not in all cases. On the other hand, if one takes into account the relative value of the estimated parameter (e.g. calculate (standard error)/(value of estimator)), this holds in most of the cases. As for the inconsistent cases (in terms of decreasing standard errors of the estimates as N increases), for example data sets 10 and 11, we observe that they occur in cases of different set of maximum likelihood estimates for our parameters. This can be due to the fact that we do not have the actual NTS data, but only the binomial data with probability of success the NTS data, adding an extra level of randomness in the data. Another relevant, general observation is that, as the value of $\hat{\kappa}$ increases, the standard errors of all MLEs (not only the one for $\hat{\kappa}$) seem to increase. This is probably due to the fact that, skewness increases rapidly for values of κ larger

		NTS-binomial			beta-binomial	
Data	Data size	$\hat{\kappa}$	$\hat{\nu}_1$	$\hat{\nu}_2$	$\hat{\nu}_1$	$\hat{\nu}_2$
1	500	0.33 (1.22)	0.93 (3.35)	0.46 (1.71)	1.90 (0.20)	0.95 (0.038)
2	1000	0.26 (1.01)	1.27 (3.10)	0.62 (1.53)	2.42 (0.21)	1.16 (0.047)
3	1500	0.33 (0.74)	1.05 (2.19)	0.52 (1.11)	2.04 (0.13)	0.97 (0.026)
4	500	0.68 (0.77)	2.07 (6.94)	0.54 (1.83)	9.40 (4.16)	2.43 (0.89)
5	1000	0.58 (0.68)	3.55 (7.33)	0.84 (1.76)	10.60 (3.79)	2.52 (0.76)
6	2000	0.59 (0.71)	3.88 (8.49)	0.94 (2.08)	11.68 (3.17)	2.82 (0.66)
7	500	0.78 (0.92)	0.55 (3.04)	2.32 (12.67)	4.94 (2.60)	16.70 (12.35)
8	1000	0.54 (0.075)	1.51 (0.97)	5.94 (3.73)	2.84 (0.89)	11.32 (4.17)
9	2000	0.44 (0.000004)	1.49 (0.59)	5.97 (2.59)	3.00 (0.71)	12.03 (3.30)
10	500	0.62 (0.033)	2.25 (1.46)	4.87 (3.23)	21.10 (2.91)	43.51 (7.29)
11	1000	0.91 (0.31)	0.48 (2.56)	0.99 (5.21)	10.36 (14.23)	21.23 (9.66)
12	2000	0.92 (0.071)	2.05 (0.56)	4.12 (1.14)	31.89 (27.43)	64.03 (57.19)
13	500	0.32 (1.12)	0.94 (3.14)	0.46 (1.60)	1.90 (0.24)	0.93 (0.043)
14	1000	0.43 (0.75)	0.78 (2.10)	0.37 (1.05)	2.08 (0.20)	1.01 (0.038)
15	1500	0.38 (0.67)	0.85 (1.88)	0.42 (0.98)	2.05 (0.16)	1.02 (0.029)
16	500	0.60 (0.029)	0.99 (1.31)	3.75 (6.35)	3.07 (1.38)	11.68 (6.12)
17	1000	0.60 (0.00002)	1.09 (1.04)	4.53 (5.17)	3.59 (2.65)	15.36 (14.32)
18	1500	0.60 (0.0023)	1.08 (0.96)	4.30 (5.19)	3.30 (0.98)	13.18 (4.50)
19	500	0.93 (1.03)	1.39 (21.16)	2.84 (42.39)	29.16 (109.76)	59.34 (240.28)
20	1000	0.82 (1.28)	1.72 (13.59)	3.48 (27.01)	12.10 (12.53)	24.54 (29.86)
21	1500	0.85 (0.78)	1.36 (8.23)	2.73 (16.03)	12.42 (10.56)	24.85 (24.77)
22	500	0.001	1.04	1.00	1.05 (0.040)	1.00 (0.031)

Table 5.2: Maximum likelihood estimates of the NTS-binomial and beta-binomial models for the simulated data sets.

than 0.8, for example for data set 10, therefore increasing the overall uncertainty in our prediction. Next, one can see that the MLEs for ν_1 and ν_2 are, generally, very different in the cases of the NTS-binomial and the beta-binomial models. The difference is larger when the value of κ is larger. On the other hand, the parameters in the two models are not directly comparable, and it makes more sense to compare the estimates of some moments of the distributions. It is not shown here, but in most cases the first three central moments of the assumed distribution for p_i are well approximated by the first three central moments of the fitted distributions (see also Table 5.3 for the estimates of the mean). This is mostly true for the data sets that seem to be a good enough sample from the underlying distribution and the MLEs are close to the hypothetical values. As one would expect, the mean is approximated better than variance and skewness and variance is better approximated than skewness.

Finally, it is interesting to spot the differences between the cases with $n = 6$ and $n = 12$. The $NTS(1, 0.5; 0.33)$ case produces very similar, and in general terms good results. The $NTS(2, 4; 0.8)$ is generally troublesome for both cases. Finally, the more differences appear in the case of the underlying $p_i \overset{iid}{\sim} NTS(1, 4; 0.6)$. In general (and well demonstrated by the $NTS(1, 4; 0.6)$ data),

it seems that in the case of $n = 12$ we are able to get the "correct" values of the parameters, as well as matching first three central moments more often than when $n = 6$. On the other hand, the (relative-i.e. divided by the estimated value of the parameter) standard errors of our estimated parameters are generally smaller when less Bernoulli trials are used.

The most awkward aspect of the results in Table 5.2 is the standard errors. These results are probably due to the specific parametrisation used, where there is strong interaction between the three parameters. Alternatively, one can look at the reparametrisation (μ, α, κ), where

$$\mu = \frac{\nu_1}{\nu_1 + \nu_2} \text{ and } \alpha = (1 - \kappa)\left[1 - e^{\frac{\nu_1+\nu_2}{\kappa}}\left(\frac{\nu_1 + \nu_2}{\kappa}\right)^{1/\kappa}\Gamma(1 - 1/\kappa, \frac{\nu_1 + \nu_2}{\kappa})\right],$$

i.e. μ is the mean and α (multiplied by $\mu(1 - \mu)$) is the variance of a $NTS(\nu_1, \nu_2; \kappa)$ - distributed random variable.

The MLEs and the corresponding standard errors for these parameters are shown in Table 5.3. This

Data	Data size	$\hat{\kappa}$	$\hat{\mu}$	$\hat{\alpha}$	κ	μ	α
1	500	0.33 (1.76)	0.67 (0.26)	0.26 (0.24)			
2	1000	0.26 (1.45)	0.67 (0.18)	0.26 (0.17)	0.33	0.67	0.25
3	1500	0.33 (1.07)	0.67 (0.15)	0.24 (0.14)			
4	500	0.68 (0.99)	.79 (0.32)	0.079 (0.058)			
5	1000	0.58 (1.17)	0.81 (0.26)	0.072 (0.048)	0.60	0.80	0.062
6	2000	0.59 (0.90)	0.81 (0.20)	0.065 (0.034)			
7	500	0.78 (0.87)	0.19 (0.39)	0.051 (0.044)			
8	1000	0.55 (0.13)	0.20 (0.051)	0.052 (0.0067)	0.60	0.20	0.062
9	2000	0.44 (0.062)	0.20 (0.044)	0.063 (0.0074)			
10	500	0.62 (0.11)	0.31 (0.054)	0.044 (0.010)			
11	1000	0.91 (0.42)	0.33 (0.40)	0.031 (0.040)	0.80	0.33	0.026
12	2000	0.92 (1.53)	0.33 (1.16)	0.011 (0.058)			
13	500	0.32 (1.63)	0.67 (0.24)	0.26 (0.23)			
14	1000	0.43 (1.07)	0.68 (0.20)	0.24 (0.15)	0.33	0.67	0.25
15	1500	0.38 (0.96)	0.67 (0.16)	0.26 (0.13)			
16	500	0.60 (0.17)	0.21 (0.18)	0.064 (0.025)			
17	1000	0.60 (0.10)	0.19 (0.11)	0.065 (0.015)	0.60	0.20	0.062
18	1500	0.60 (0.078)	0.20 (0.089)	0.058 (0.012)			
19	500	0.93 (1.38)	0.33 (1.11)	0.012 (0.060)			
20	1000	0.82 (1.76)	0.33 (0.57)	0.027 (0.072)	0.80	0.33	0.026
21	1500	0.85 (1.07)	0.33 (0.42)	0.026 (0.051)			
22	500	0.001	0.51	0.33	0	0.50	0.33

Table 5.3: Maximum likelihood estimates for κ, μ, and α and the underlying values of these parameters for the simulated data sets.

table suggests that there is consistency in the MLEs for μ and α (and therefore for the estimated first two moments), using different data sizes. The only exception seems to be the maximum likelihood estimates for μ and α for the data sets 10-12 and for α for data set 19. Notice, however, that those

data sets correspond to underlying $p_i \overset{iid}{\sim} NTS(2, 4; 0.8)$, which did not give good and consistent estimates for ν_1 and ν_2, either. As for the standard errors of the estimates, there also seems to be a consistency here, as well as a decrease with the data size, in most of the cases. Again, data sets 10-12 produce the most irrational (in the sense of decreasing standard errors) results. Overall, the standard errors for the MLEs of those two parameters are much smaller than the standard errors for the MLEs of ν_1 and ν_2. The standard errors for $\hat{\kappa}$ are comparable to those produced in the previous parametrisation, but much more consistent in this case.

Finally, for comparison purposes, I plotted the kernel density estimates of the underlying p_i for both the NTS-binomial and beta-binomial models fitted at the MLEs of the parameters for selected data sets.

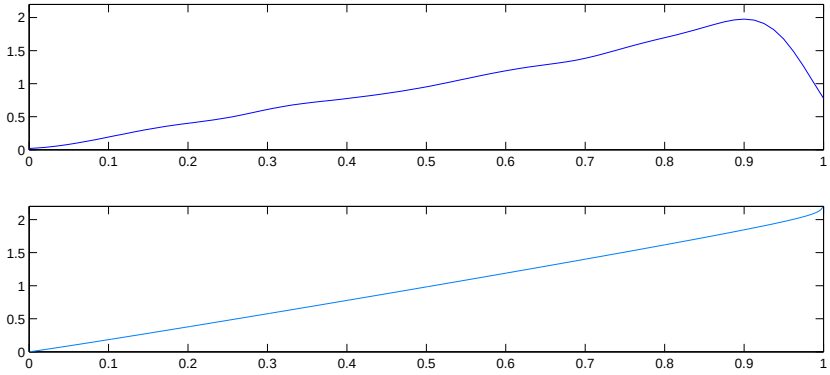


Figure 5.5: Kernel density estimates for the underlying p_i in a NTS-binomial (top) and beta-binomial model (bottom) for simulated data set 3.

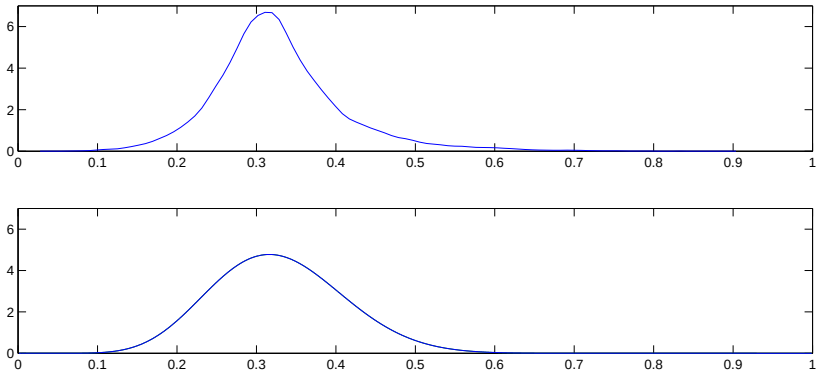


Figure 5.6: Kernel density estimates for the underlying p_i in a NTS-binomial (top) and beta-binomial model (bottom) for simulated data set 11.

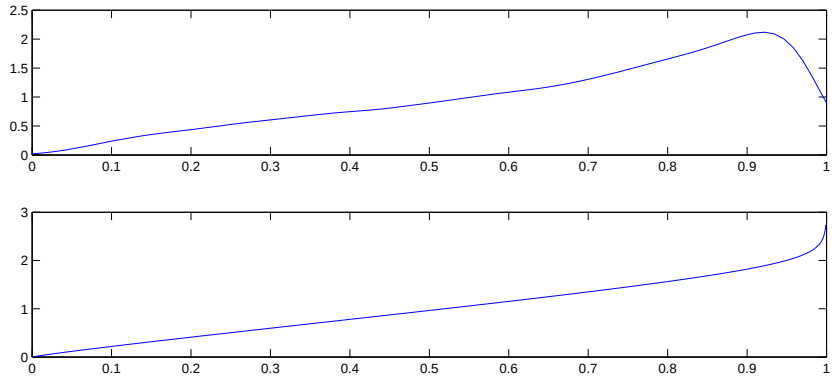


Figure 5.7: Kernel density estimates for the underlying p_i in a NTS-binomial (top) and beta-binomial model (bottom) for simulated data set 16.

5.3.4 An application to mice fetal mortality data

In this section I analyse some data of fetal mortality in mouse litters, used also in Brooks et al. (1997). Some small mistakes in the tables of the data there were spotted in Garren et al. (2001). The E1, E2 data sets were introduced in Brooks et al. (1997) and were created by pooling smaller data sets used by James and Smith (1982), whereas HS1, HS2 and HS3 were taken from Haseman and Soares (1976) and the AVSS data set was in Aeschbacher and Stalder (1977). In each data set, the data were more disperse than under a binomial distribution. Brooks et al. (1997) show that finite mixture models fit the data better than the standard beta-binomial model in all data sets except AVSS. Here, I fit the NTS-binomial and (its special case when $\kappa = 1/2$) the N-IG-binomial models, as alternatives. Several other models have been applied to these data including: the shared response model of Pang and Kuk (2005) and the q-power distribution of Kuk (2004). For parameter correspondence, note that the beta distribution with parameters (μ, θ) in Brooks et al. (1997) corresponds to $lim_{\kappa \to 0} NTS(\frac{\mu}{\theta}, \frac{1-\mu}{\theta}; \kappa)$.

E1 data:

E1 consists of 205 data points, and under the beta-binomial model,the maximum log-likelihood value is -283.70 at $\hat{\nu}_1 = 1.219$ and $\hat{\nu}_2 = 12.306$. For the NTS-binomial model, the maximum log-likelihood is -280.69 for $\hat{\nu}_1 = 0.165$, $\hat{\nu}_2 = 1.688$ and $\hat{\kappa} = 0.738$ (i.e. a model not so close to the beta-binomial one).

E2 data:

Here, there are 211 data points.

For unrestricted κ, the maximum log-likelihood value is -341.46 for $\hat{\nu}_1 = 0.097$, $\hat{\nu}_2 = 0.785$ and $\hat{\kappa} =$

0.743 On the other hand, for the beta-binomial case, the maximum log-likelihood is -344.88, and MLEs of $\hat{\nu}_1 = 1.001$ and $\hat{\nu}_2 = 7.900$.

HS1 data:

There are 524 data in this data set. For the beta-binomial case, the maximum log-likelihood is -777.79, obtained at $\hat{\nu}_1 = 1.217$ and $\hat{\nu}_2 = 12.291$. For the NTS-binomial model, the value of maximum log-likelihood is significantly higher (-772.67), for $\hat{\nu}_1 = 0.177,\ \hat{\nu}_2 = 1.780$ and $\hat{\kappa} = 0.727$.

HS2 data:

This is the largest data set, with 1328 observations. The maximum likelihood estimates for the general model are $\hat{\nu}_1 = 0.285,\ \hat{\nu}_2 = 2.346$ and $\hat{\kappa} = 0.832$, whereas for $k \simeq 0$ we get $\hat{\nu}_1 = 2.400$ and $\hat{\nu}_2 = 19.701$. The difference in log-likelihood is quite large: -1646.38 for the former and -1657,30 for the latter. This could be due to the fact that we have many observations, as well as the fact that the proposed value of κ is far from 0, indicating a substantial difference in the two models.

HS3 data:

This data set has 554 binomial data, and the difference in the maximum log-likelihood for the NTS-binomial and the beta-binomial models is again quite large. For the simpler model the MLEs are $\hat{\nu}_1 = 0.944,\ \hat{\nu}_2 = 12.305$ and maximum log-likelihood= -701.54. For the more general case, $\hat{\nu}_1 = 0.041,\ \hat{\nu}_2 = 0.516$ and $\hat{\kappa} = 0.804$, giving log-likelihood= -685.62.

AVSS data:

This data set consists of only 127 observations.

In this case, the maximum likelihood for the NTS-binomial model is achieved at the limiting case $k \to 0$, i.e. it coincides with the beta-binomial model. The MLE for the other two parameters are $\hat{\nu}_1 = 1.095$ and $\hat{\nu}_2 = 14.778$, and the corresponding log-likelihood is -168.93.

		NTS-binomial			beta-binomial	
Data	N	$\hat{\kappa}$	$\hat{\nu}_1$	$\hat{\nu}_2$	$\hat{\nu}_1$	$\hat{\nu}_2$
E1	205	0.738 (0.384)	0.165 (0.066)	1.688 (0.023)	1.219 (1.549)	12.306 (20.892)
E2	211	0.743 (0.044)	0.097 (0.168)	0.785 (1.569)	1.001 (0.875)	7.900 (9.815)
HS1	524	0.727 (0.548)	0.177 (0.713)	1.780 (7.169)	1.217 (0.961)	12.291 (13.058)
HS2	1328	0.832 (0.326)	0.285 (0.778)	2.346 (6.414)	2.400 (0.158)	19.701 (2.688)
HS3	554	0.804 (0.274)	0.041 (0.129)	0.516 (1.560)	0.944 (0.732)	12.305 (12.537)
AVSS	127	0	1.095	14.778	1.095 (2.082)	14.778 (36.812)

Table 5.4: Maximum likelihood estimates and standard errors of the estimates for the NTS-binomial and beta-binomial models for the six mice fetal mortality data sets.

The MLEs for all the above data sets and their related standard errors (shown in parentheses) are gathered and shown in Table 5.4. In all data sets except AVSS the estimate of κ is substantially different from $\kappa = 0$ (which corresponds to the beta-binomial case). The estimate is zero for the

AVSS data where the NTS-binomial model corresponds to the beta-binomial model. In the other data sets κ is estimated to be between 0.74 and 0.83 and the estimated mixing distributions are substantially different from a beta distribution. In fact the tails of the distribution are much heavier than those defined by a beta distribution with the same mean and variance. The estimated mixing distributions are shown in Figure 5.8 with the mixing distribution for the beta-binomial distribution (the AVSS data set is not included since the estimates for NTS-binomial and beta-binomial models imply the same mixing distribution). The graphs for the NTS-binomial were created using large enough samples from this distribution.

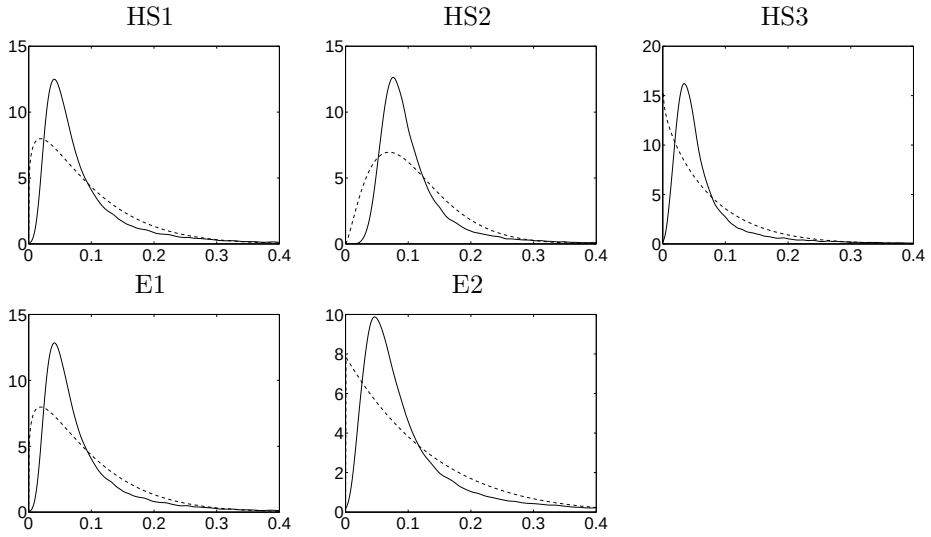


Figure 5.8: Density estimates for the mixing distribution for the NTS-binomial (solid line) and beta-binomial (dashed line) models evaluated at the maximum likelihood estimates for the mice fetal mortality data.

The first four central moments for the two distributions are shown in Table 5.5. The first two

		E1	E2	HS1	HS2	HS3
Expectation	beta	0.0901	0.1125	0.0901	0.1086	0.0713
	NTS	0.0888	0.1096	0.0902	0.1087	0.0734
Variance	beta	0.00565	0.01008	0.00565	0.00419	0.00464
	NTS	0.00636	0.01115	0.00650	0.00395	0.00707
Skewness	beta	1.41	1.42	1.41	1.00	1.65
	NTS	2.75	2.66	2.67	2.66	3.89
Kurtosis	beta	2.42	2.25	2.42	1.21	3.46
	NTS	13.26	11.99	12.62	13.57	22.93

Table 5.5: Estimates of the first four central moments of the mixing distributions for the beta-binomial and NTS-binomial distributions for five mice fetal mortality data sets.

moments are roughly equal with the exception of HS3 which has a larger variance for the NTS-

binomial model than the beta-binomial model. However, the third and especially fourth moments are larger for the NTS-binomial model, a result which is in accordance with the interpretation of κ given in Section 5.2.2.

As discussed above, the likelihood values for the NTS-binomial model are substantially better than those for the beta-binomial model. However, we would like to compare the NTS-binomial model to the other competing models: the finite mixture models of Brooks et al. (1997), the shared response model of Pang and Kuk (2005) and the q-power distribution of Kuk (2004). I also consider the normalised inverse-Gaussian distribution, which fixes $\kappa = 0.5$ in the NTS-binomial model. The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are used as measures of fit. If L is the maximum likelihood value, n is the data size and k is the number of parameters, the AIC $= -2\log L + 2k$ and the BIC $= -2\log L + k\log n$. Results for each data set are

Model	E1	E2	AVSS	HS1	HS2	HS3
beta-binomial	+8.3	+6.0	**341.9**	+9.3	+43.9	+32.8
NTS-binomial	+4.8	+1.1	+2.0	+1.2	+24.1	+3.4
N-IG-binomial	+5.1	+0.6	+0.5	+3.1	+38.9	+18.8
B-B/B mixture	+3.5	+0.4	+3.8	**1550.3**	**3274.7**	**1373.9**
2-d binom. mixture	+1.7	+0.6	+1.9	+20.6	+20.1	+4.6
3-d binom. mixture	+5.1	+1.9	+5.7	+1.1	+4.1	+1.8
Best binom. mixture	+5.1	+1.9	+9.7	+4.5	+1.7	+5.5
Shared response	**563.1**	**687.8**	+6.1	+20.3	+42.5	+8.9
q-power	+6.6	+8.4	+8.0	+6.5	+2.1	+0.5
Correlated-binomial	+26.2	+36.7	+1.2	+57.0	+67.3	+94.5
B-C-B	+7.8	+1.3	+2.0	+8.4	+35.7	+24.0

Table 5.6: AIC values for the competing models for each data set. The smallest value for each data set is shown in bold and other AIC values are shown as differences from that minimum.

given in Table 5.6 (for AIC) and Table 5.7 (for BIC). The best model has the smallest value of the information criterion. In both tables, the smallest value of AIC/BIC for each the data set is given in bold and the values for the other models are given as differences from the best model. In the tables, B-B/B mixture is the beta-binomial/binomial mixture (i.e. a mixture consisting of a beta-binomial part and a binomial part), 2-d/3-d correspond to mixtures of two or three binomials respectively, B-C-B is the beta-correlated-binomial model, and N-IG is the normalised inverse-Gaussian distribution. Finally, the best binomial mixture is a mixture of binomials with the number of components unknown. The number of components to be fitted is derived using the program C.A.MAN, using directional derivative methods (see Bohning et al. (1992)). I also found that the log likelihood value given by Pang and Kuk (2005) for E1 was not consistent with their estimates. Their value seems to correspond to the data set given by Brooks et al. (1997) rather than the corrected version given

by Garren et al. (2001). My proposed NTS-binomial model performs well for all six data sets, es-

Model	E1	E2	AVSS	HS1	HS2	HS3
beta-binomial	+8.3	+6.0	**347.6**	+6.2	+41.9	+32.3
NTS-binomial	+8.1	+4.5	+4.9	+2.4	+27.2	+7.2
N-IG-binomial	+5.1	+0.6	+0.5	**1561.9**	+36.8	+18.3
B-B/B mixture	+10.2	+4.6	+9.5	+5.4	+8.3	+8.2
2-d binom. mixture	+5.0	+4.0	+4.8	+21.7	+23.2	+8.4
3-d binom. mixture	+15.1	+12.0	+14.2	+10.8	+17.5	+14.3
Best binom. mixture	+15.1	+12.0	+23.9	+22.7	+25.6	+26.6
Shared response	**569.7**	**694.5**	+6.1	+17.2	+40.4	+8.4
q-power	+6.6	+7.9	+9.0	+3.4	**3287.2**	**1383.0**
Correlated-binomial	+26.2	+36.7	+1.2	+53.9	+65.2	+94.0
B-C-B	+11.1	+4.6	+4.9	+9.6	+38.8	+27.8

Table 5.7: BIC values for the competing models for each data set. The smallest value for each data set is shown in bold and other BIC values are shown as differences from that minimum.

pecially in terms of BIC. The N-IG model (which is a special case of my model) also performs very well for these data. For the E1 data, the best one in terms of both AIC and BIC is the shared response model, with the N-IG model very close to it. For E2, again the shared response model performs better than all the other regarding AIC and BIC, with the N-IG model again quite close (but with the B-B/B mixture even closer). For the AVSS data, the undisputed winner is the simple beta-binomial model, which seems to model the data sufficiently well. For HS1, the N-IG model is actually the best in terms of the BIC, whereas the B-B/B mixture performs best in terms of the AIC. The value of AIC for the N-IG model and the value of BIC for the B-B/B mixture are, as one would expect, close to the smallest value. My model is the second best in terms of BIC and third (but also very close to the second) in terms of the AIC. Regarding the HS2 data set, the differences between the criteria of the different models are larger than in the other data sets, due to the much larger data size. The best model in this case is (again) the mixture of beta-binomial and binomial in terms of AIC and the q-power model in terms of BIC. Finally, for the HS3 data set, the q-power distribution is the best in terms of the BIC, whereas the B-B/B mixture performs best in terms of the AIC. My model is second in terms of BIC and fourth in terms of AIC.

Another interesting point is the unimodality of the MLEs in the MNTS-binomial model. As seen in Figure 5.9, the profile log-likelihoods for the three parameters of this model for the HS1 data are unimodal, with the mode corresponding to the MLEs of them. Although for ν_1 and ν_2 the domain in which this is shown is limited ((0,0.3) for ν_1 and (0,3) for ν_2), compared to the actual domain of those two parameters (i.e. the positive real line), this is for illustration purposes. The unimodality holds for the whole positive real line for both ν_1 and ν_2. Moreover, the unimodality

holds for all the data sets considered here.

The reason that I think that this is a useful property for my model comes from a comparison with the corresponding concept for the beta-binomial/binomial mixture (see Figure 1 and Table 9 in Brooks et al. (1997)), in which the profile log-likelihood for the mixing parameter γ is bimodal for almost all data sets. The latter model performs well for most of the data sets, however the fact that it gives two possible modes for the parameters in the model is a drawback in terms of computations and/or theoretical properties. On the other hand, my model does not have this problem.

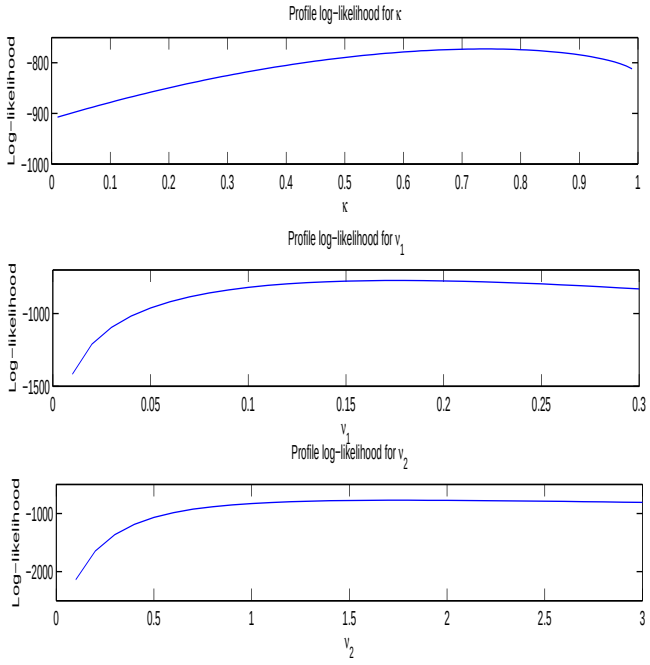


Figure 5.9: Profile log-likelihoods for the parameters in the MNTS-binomial model for the HS1 data.

5.4 Summary

In this chapter I move from the nonparametric context to the fully parametric setting. I propose a new, n-dimensional distribution defined on the unit simplex, called the multivariate normalised tempered stable (MNTS) distribution, which includes the Dirichlet, the beta and the normalised inverse-Gaussian distributions as special cases. The general formulae for the moments and cross-moments of the proposed distribution are also derived. These formulae are later used in the maximum

likelihood estimation of the parameters in a model for count data. More specifically, it is assumed that we have data from a binomial distribution, and the probabilities of success are assumed to follow the univariate MNTS distribution, called the normalised tempered stable (NTS) distribution. Because of the moment results of this new distribution, the binomial-NTS distributions will be particularly useful for modelling overdispersed data. This point is illustrated using both simulated and real data, more specifically mice fetal mortality data. The performance of my proposed model with other models proposed in the literature for the mice data is also compared, and it is shown that my model performs consistently well.

Chapter 6

Conclusions and Future Directions

6.1 Summary

This thesis covers models derived through normalisation methods for both parametric and Bayesian nonparametric inference. In the Bayesian nonparametric case, a class of models for grouped data that are assumed to follow dependent random probability measures (RPMs) is proposed. These dependent RPMs are constructed by normalising infinitely divisible random measures. In the parametric case a new distribution is proposed, the normalised tempered stable distribution. As the name suggests, this distribution is the distribution of a vector of normalised tempered stable-distributed random variables, i.e. each of them divided by the sum of all of them.

In Chapter 1 I begin with a general overview of Bayesian nonparametric methods. Most of these methods assume that the distributions of some data are also random. These random distributions are called random probability measures and the most widely used distribution of these RPMs is the Dirichlet process (DP). Properties of this process are presented, as well as the basic simulation methods for this type of models. A good alternative for the DP is the normalised inverse-Gaussian process (N-IGP), which is also defined and studied.
Next, various ways of introducing dependence in Bayesian nonparametric models are described. I focus on models for grouped data, and more specifically on a model presented in Müller et al. (2004). The normalisation method is also presented, which will be used in constructing my basic models.

In Chapter 2 a general class of models for two dependent distributions is presented, where it is assumed that each distribution consists of a weighted sum of a common component and an idiosyncratic component. Because of the common and the idiosyncratic components, these models can be naturally applied to grouped data. As shown, a model of this form can be constructed in

a systematic way by normalising specific infinitely divisible random measures. My basic proposed model, however, will be a slight simplification of this model. This model, apart from its intuitive appeal, also exhibits nice theoretical properties, for example very simple moment results. This model is also very similar to the model of Müller et al. (2004), although constructed using a different method, and a comparison of the two models is attempted. As mentioned above, the method of construction of my proposed model results in some nice properties that are not guaranteed in the model of Müller et al. (2004). On the other hand, the latter model is more general than my proposed model, since it is constructed in a more elaborate way.
Next, some generalisations of my basic proposed models and of the model of Müller et al. (2004) in three dimensions, i.e. for three dependent (random) distributions, are presented. The systematic way of construction of my models, i.e. the normalisation technique, results in a straightforward way of extending them in three, or even more, dimensions. This is not apparent for the Müller et al. (2004) model, however some options are presented.

Chapter 3 deals with the computational implementation of our models. As with most Bayesian nonparametric models, inference for the posterior distribution of the parameters in my models is achieved using Monte Carlo Markov Chain methods. For my basic model, the algorithm is very similar to the one described in Müller et al. (2004). On the other hand, when applying this algorithm to simulated data, I observed very slow mixing of the chains, due to the bimodality of the posterior distribution of a specific parameter. An additional step in the MCMC algorithm was therefore proposed, in which we either split a cluster from the common component to form clusters in the idiosyncratic components, or merge two clusters from the idiosyncratic components to form a common cluster in the common component. Using simulated data, it is demonstrated that this extra mix-split step improves mixing. This extra step is based on a generic idea, and can therefore be applied to a variety of models. In fact, this step is applied to the algorithm for the model of Müller et al. (2004), and again a better mixing is achieved.
Some ideas for simulating the models in higher dimensions are then presented. This area is still quite unexplored, although it should not be particularly difficult to extend the algorithms that are used for the simpler models.
Regarding the implementation of the model that directly emerges from the normalisation method, MCMC methods are again used, but now with a different parametrisation and slice sampling updating steps for some of the parameters. Although we could have mimicked the previous algorithms, the method used here is easy to code, straightforward, computationally inexpensive and does not require any sort of monitoring (as a RWMH step would require). An extra mix-split step was also included in this algorithm.

Finally, I present the algorithm for a model which has the same form as my basic model, but with N-IGP priors instead of DP priors. When I attempted to extend the algorithm for my basic model here, I came across some problems with the software used (Matlab), which made the algorithm unreliable. Specifically, the built-in commands in Matlab for calculating the incomplete gamma function $\Gamma(a, x)$ (that appears in the Pólya-urn representation of the N-IGP) do not give the correct values when a takes large negative values. As a result, a different simulation method was used. More specifically, I follow the method proposed in Griffin and Walker (2009), which makes use of slice sampling techniques for normalised random measures, and simulating from the full conditional distributions of all parameters of interest was made possible. When trying to implement a mix-split step, however, I ended up with an integral that did not have a known closed form. This was, actually, my motivation for deriving the general moment results for the N-IG distribution in Section 5.

In Chapter 4 various models presented above were applied to real-life data. At first my basic proposed model and the model of Müller et al. (2004) were applied to daily returns of two stocks. Using these models, we could get a better feeling about the common behaviour of the two stocks (corresponding to the common component in the two dependent distributions) and of the behaviour which is specific for each stock (as shown from the idiosyncratic components). Although the results are similar, it is also shown (using both predictive power and Bayes factors) that my model performs slightly better in terms of both criteria.
Next, my basic model, the model with N-IGP priors and the model of Müller et al. (2004) were embedded in a model for stochastic frontier analysis. The corresponding algorithms for all three models were derived and then applied to hospital cost frontier data, in order to examine their efficiencies, and in particular the differences between groups of hospitals. The hospitals in the same ownership status, but different clinical workers per patient ratios, were compared. The results are quite similar for my basic model and the model of Müller et al. (2004), and rather different for the model with the N-IGP priors, especially for small data sizes. Nevertheless, in all models we have the same results regarding which groups of hospitals (high or low staff ratio) are the most efficient.

In Chapter 5 a parametric model for count data is proposed. I start by deriving the general formulae for the moments and cross-moments for the N-IG distribution and derive some basic moments. Then, a novel, n-dimensional distribution defined on the unit simplex, the multivariate normalised tempered stable (MNTS) distribution is proposed. As its name indicates, this distribution can be derived as the distribution of a random vector of tempered stable-distributed random variables divided by their sum. This distribution has three parameters and includes many known distributions such as the Dirichlet and the N-IG distribution as special cases. We examine the theoretical characteristics of this distribution and derive the general moments and cross-moments for

this distribution, as well as some basic moments. A new model for discrete data is then constructed, where it is assumed that the probabilities of success of binomial data follow the normalised tempered stable (NTS) distribution, which is the univariate version of the MNTS distribution. Based on the moment results, the new model should be particularly successful in modelling overdispersed data, which is verified by applying and comparing the proposed model and the known beta-binomial model to simulated data. Notice that the formulae for general moments and cross-moments derived above are used in calculating the maximum likelihood estimates of the parameters in our model. Finally, the new model and the N-IG-binomial model are applied to mice fetal mortality data and the two models are compared to previously proposed models for the same data, using the AIC and the BIC. Although my proposed model is never the best model, it performs consistently well and has criterion values that are very close to the best model in each case.

6.2 Future Work

There are many possible extensions and future directions regarding the work presented in this thesis, in terms of both modelling and computational issues. Here, we will present some of these options.

The model with N-IGP priors seems to be an interesting alternative to the DP. This can be seen, for example, from the findings for the hospital data, and especially the government-run ones. It would be therefore interesting to further study this model and apply it to more data sets. Comparison with the corresponding model with DP priors can (and should) also be examined in a formal way, for example using Bayes factors. Simulation methods for this model can also be explored further, as well as constructing an additional mix-split step in these algorithms.

Apart from the proposed models and the model with N-IGP priors, one might consider models of the same structure, but with other normalised random measures as the prior distribution of the dependent random measures. Simulating from the posterior distributions of these models, however, might be challenging. A possible general algorithm to follow (or extend) is the slice sampling method for NRMs of Griffin and Walker (2009).

The slice sampler of Griffin and Walker (2009) performed well when used in this thesis, and could also be further studied, improved and applied to different data sets. This is also true for the extra mix-split step proposed in some models here. This extra step was shown to improve the mixing of the chains when used and therefore it would be interesting to embed it in more algorithms. Assessment of the contribution of this extra step in mixing can then be performed using different data sets.

The algorithms for the models for three correlated distributions are only described in general

terms in this thesis. Generalising the existing algorithms should not be extremely difficult, regardless of the significant increase in both notational complexity and computational burden. It is, therefore, an obvious area of further research. Developing the algorithms for posterior inference for these models will also help us better understand the specific models and their properties. Alternatively, other possible models for grouped data from three (or more) dependent distributions, together with the corresponding algorithms, may be considered.

Another possible area of future work would be a further and more thorough examination of the NTS distribution (or, more generally, the MNTS distribution) and the NTS-binomial model. The NTS distribution is a simple generalisation of many familiar distributions, with only one additional parameter, whereas the latter performed very well when applied to real data. As noted before, the NTS-binomial model can be particularly useful for modelling overdispersed count data, and it therefore seems worth applying it to data of this type, for example skewed data.

Finally, the NTS distribution and the NTS-binomial model can be naturally considered within the Bayesian context. A first, straightforward, and probably successful Bayesian model would be to model some count data using the NTS-binomial model, where we assign some prior distributions on the parameters ν_1, ν_2 and κ. As indicated by the examples in Chapter 5, this model should perform particularly well when the data are overdispersed.

Appendix A

Appendix

A.1 The Acceptance Probabilities For the Mix-Split Step in Section 3.3

The acceptance probability of the mix-split step described in Section 3.3 is

$$\alpha(\mathbf{c}, \mathbf{c}') = \min\left\{1, \frac{q(\mathbf{c}', \mathbf{c})}{q(\mathbf{c}, \mathbf{c}')} \frac{f(\mathbf{c}')}{f(\mathbf{c})}\right\}$$

where $q(a, b)$ is the transition probability from a value a to a value b and $f(\mathbf{c}) = f(\mathbf{c}|\cdots)$ is the posterior probability of $\mathbf{c}$, having integrated out the weight ε and all the discrete values ϕ_{ji}. The current state of all indicators s_{ji} and r_{ji} is denoted with $\mathbf{c}$ and the proposed one with $\mathbf{c}'$.
The second ratio of probabilities ($f(\mathbf{c}')/\ f(\mathbf{c})$) and can be deduced as follows:

$$\begin{aligned} f(\mathbf{c}) = f(\mathbf{c}|\cdots) &= \int\int f(\mathbf{c}, \varepsilon, \boldsymbol{\phi}) d\varepsilon d\boldsymbol{\phi} \\ &\propto \int\int f(\mathbf{Y}|\boldsymbol{\phi}, \mathbf{c}, S) f(\mathbf{c}, \varepsilon|\boldsymbol{\phi}, \cdots) f(\boldsymbol{\phi}|m, B) d\varepsilon d\boldsymbol{\phi} \\ &= \int\int f(\boldsymbol{Y}|\boldsymbol{\phi}, \boldsymbol{c}, S) f(\boldsymbol{c}, \varepsilon|\boldsymbol{M}) f(\boldsymbol{\phi}|m, B) d\boldsymbol{\phi} d\varepsilon \\ &= \int f(\boldsymbol{c}, \varepsilon|\boldsymbol{M}) d\varepsilon \int f(\boldsymbol{Y}|\boldsymbol{\phi}, \boldsymbol{c}, S) f(\boldsymbol{\phi}|m, B) d\boldsymbol{\phi} \\ &= \int f(\boldsymbol{\phi}|m, B) f(\boldsymbol{Y}|\boldsymbol{\phi}, \boldsymbol{c}, S) d\boldsymbol{\phi} \times f(\boldsymbol{c}|\boldsymbol{M}). \end{aligned} \tag{A.1.1}$$

The first part of (A.1.1) is the joint full conditional distribution of all ϕ_{ji}, and can be written as the product of the full conditionals of each of them. This is a convenient result, because most of the terms will cancel out when calculating the ratio $f(\boldsymbol{c}')/f(\boldsymbol{c})$. The terms that will be left will be those associated with clusters that are merged/split.

The second part of (A.1.1) is the distribution of $\boldsymbol{c}$, conditional on M_0, M_1. This is easily calculated by integrating out the weight from $f(\boldsymbol{c}, \varepsilon|\boldsymbol{M})$, and is given in equation (2.1.6).

So, by substituting the above results, and performing the simplifications mentioned above, we have:

Let m_{01} denote the number of data from the first data set allocated to the cluster to be split and m_{02} denote the number of data from the second data set allocated to the same cluster. Then:

Split proposal:

1. If $m_{01} = 0$, i.e. a cluster from F_0 is moved to F_2:
   $\frac{f(\boldsymbol{c}_{split})}{f(\boldsymbol{c})} = \frac{M_1}{M_0}\frac{\Gamma(M_1+n_1+n_2+m_{02})\Gamma(M_1+n_2)}{\Gamma(M_1+n_1+n_2)\Gamma(M_1+n_2+m_{02})}$.

2. If $m_{02} = 0$, i.e. a cluster from F_0 is moved to F_1:
   $\frac{f(\boldsymbol{c}_{split})}{f(\boldsymbol{c})} = \frac{M_1}{M_0}\frac{\Gamma(M_1+n_1+n_2+m_{01})\Gamma(M_1+n_1)}{\Gamma(M_1+n_1+n_2)\Gamma(M_1+n_1+m_{01})}$.

3. If $m_{01} > 0$, and $m_{02} > 0$, i.e. a cluster from F_0 is split to both F_1 and F_2:
   $\frac{f(\boldsymbol{c}_{split})}{f(\boldsymbol{c})} = \frac{M_1^2}{M_0}\frac{\Gamma(M_1+n_1+n_2+m_{01}+m_{02})\Gamma(M_1+n_1)\Gamma(M_1+n_2)}{\Gamma(M_1+n_1+n_2)\Gamma(M_1+n_1+m_{01})\Gamma(M_1+n_2+m_{02})}\frac{\Gamma(m_{01})\Gamma(m_{02})}{\Gamma(m_{01}+m_{02})}\sqrt{\frac{S[(m_{01}+m_{02})B+S]}{(m_{01}B+S)(m_{02}B+S)}} \times$
   $\exp\{\frac{(m_{01}+m_{02})m^2-2m(\sum Y_1'+\sum Y_2')-\frac{B}{S}(\sum Y_1'+\sum Y_2')^2}{2(m_{01}+m_{02})B+2S} - \frac{m_{01}m^2-2m\sum Y_1'-\frac{B}{S}(\sum Y_1')^2}{2m_{01}B+2S} - \frac{m_{02}m^2-2m\sum Y_2'-\frac{B}{S}(\sum Y_2')^2}{2m_{02}B+2S}\}$.

In all the above, it is assumed that $K_0 > 0$ (otherwise we do nothing) and $\sum Y_1',\ \sum Y_2'$ are taken over the data associated with the split cluster.

Merge proposal:

1. If exactly one of K_1 and K_2 is zero (note that then, exactly one of m_1 and m_2 is zero):
   $\frac{f(\boldsymbol{c}_{merge})}{f(\boldsymbol{c})} = \frac{M_0}{M_1}$.

2. If both K_1 and K_2 are positive:

   (a) If a cluster from F_2 is moved to F_0 (in this case, $m_1 = 0$):
   $\frac{f(\boldsymbol{c}_{merge})}{f(\boldsymbol{c})} = \frac{M_0}{M_1}\frac{\Gamma(M_1+n_1+n_2-m_2)\Gamma(M_1+n_2)}{\Gamma(M_1+n_1+n_2)\Gamma(M_1+n_2-m_2)}$.

   (b) If a cluster from F_1 is moved to F_0 (here, $m_2 = 0$):
   $\frac{f(\boldsymbol{c}_{merge})}{f(\boldsymbol{c})} = \frac{M_0}{M_1}\frac{\Gamma(M_1+n_1+n_2-m_1)\Gamma(M_1+n_1)}{\Gamma(M_1+n_1+n_2)\Gamma(M_1+n_1-m_1)}$.

   (c) If a cluster from F_2 and a cluster from F_1 are merged in F_0 :
   $\frac{f(\boldsymbol{c}_{merge})}{f(\boldsymbol{c})} = \frac{M_0}{M_1^2}\frac{\Gamma(M_1+n_1+n_2-m_1-m_2)\Gamma(M_1+n_1)\Gamma(M_1+n_2)}{\Gamma(M_1+n_1+n_2)\Gamma(M_1+n_1-m_1)\Gamma(M_1+n_2-m_2)}\frac{\Gamma(m_1+m_2)}{\Gamma(n_1)\Gamma(m_2)}\sqrt{\frac{(m_1B+S)(m_2B+S)}{S[(m_1+m_2)B+S]}}\exp$
   $\{\frac{-(m_1+m_2)m^2+2m(\sum Y_1'+\sum Y_2')+\frac{B}{S}(\sum Y_1'+\sum Y_2')^2}{2(m_1+m_2)B+2S} + \frac{m_1m^2-2m\sum Y_1'-\frac{B}{S}(\sum Y_1')^2}{2m_1B+2S} + \frac{m_2m^2-2m\sum Y_2'-\frac{B}{S}(\sum Y_2')^2}{2m_2B+2S}\}$.

In the above, it is assumed that at least K_1 or K_2 are non-zero (otherwise we do nothing),and $\sum Y_1',\ \sum Y_2'$ are taken over the data associated with the clusters to be merged.

The transition probabilities q can be calculated as follows:

1. Split step:
   $q(\boldsymbol{c}, \boldsymbol{c}_{split}) = P(\boldsymbol{c}_{split}|\boldsymbol{c}) = P(\boldsymbol{c}_{split}|\boldsymbol{c}, \text{split step})P(\text{split step}|\boldsymbol{c}) = \frac{1}{2K_0}$

$$q(\boldsymbol{c}_{split}, \boldsymbol{c}) = P(\boldsymbol{c}|\boldsymbol{c}_{split}, \text{merge})P(\text{merge}|\boldsymbol{c}_{split}) = \begin{cases} \frac{1}{2((K_1+2)(K_2+1)-1)} & , \quad \text{if } m_{01} > 0,\ m_{02} = 0 \\ \frac{1}{2((K_1+1)(K_2+2)-1)} & , \quad \text{if } m_{01} = 0,\ m_{02} > 0 \\ \frac{1}{2((K_1+2)(K_2+2)-1)} & , \quad \text{if } m_{01} > 0,\ m_{02} > 0. \end{cases}$$

Here, it is assumed that $K_0 > 0$. The values of K_1, K_2 do not matter.

2. Merge step:
   $q(\boldsymbol{c}, \boldsymbol{c}_{merge}) = P(\boldsymbol{c}_{merge}|\boldsymbol{c}, \text{merge step})P(\text{merge step}|\boldsymbol{c}) = \frac{1}{2((K_1+1)(K_2+1)-1)}$
   and $q(\boldsymbol{c}_{merge}, \boldsymbol{c}) = P(\boldsymbol{c}|\boldsymbol{c}_{merge}, \text{split step})P(\text{split step}|\boldsymbol{c}_{merge}) = \frac{1}{2(K_0+1)}$.
   Here, it is assumed that at least one of K_1 and K_2 is positive.

In the above, I have used that the probability of choosing to propose a merge or a split step is $1/2$ for each.
By combining the above, together with the expressions for the ratio $f(\boldsymbol{c}')/\ f(\boldsymbol{c})$, the acceptance probabilities for the proposed split or merge step can be calculated.

A.1.1 The acceptance probabilities for the alternative mix-split step

For the alternative mix-split step, i.e. when the values of K_0, K_1 and K_2 are taken into account in proposing a split or a merge step, the ratios $\frac{f(\mathbf{c}')}{f(\mathbf{c})z}$ will be the same as above. However, since we propose the mix and split steps differently, the ratios $\frac{q(\mathbf{c}',\mathbf{c}')}{q(\mathbf{c},\mathbf{c}')}$ will be different:
Let n_1 and n_2 denote the current (i.e. before the proposed mix or split step) number of data assigned in each idiosyncratic component distribution, F_1 and F_2, respectively. Then:

1. When $K_0 = 0$, we propose a merge step, and
   $q(\boldsymbol{c}_{merge}, \boldsymbol{c}) = P(\boldsymbol{c}|\boldsymbol{c}_{merge}, \text{split step})P(\text{split step}|\boldsymbol{c}_{merge}) = P(\text{split step}|\boldsymbol{c}_{merge}) = p_1$,
   since $P(\boldsymbol{c}|\boldsymbol{c}_{merge}, \text{split step}) = \frac{1}{K_0+1} = 1$, and
   $p_1 = \begin{cases} 1 & , \quad \text{if } K_1 = K_2 = 0 \\ 1/2 & , \quad \text{else.} \end{cases}$
   On the other hand,
   $q(\boldsymbol{c}, \boldsymbol{c}_{merge}) = P(\boldsymbol{c}_{merge}|\boldsymbol{c}, \text{merge step})P(\text{merge step}|\boldsymbol{c}) = \frac{1}{(K_1+1)(K_2+1)-1}$,
   since here $P(\text{merge step}|\boldsymbol{c}) = 1$.

2. When $K_1 = K_2 = 0$, we propose a split step, and
   $$q(\boldsymbol{c}_{split}, \boldsymbol{c}) = P(\boldsymbol{c}|\boldsymbol{c}_{split}, \text{merge step})P(\text{merge step}|\boldsymbol{c}_{split}) = \begin{cases} p_2 & , \quad \text{if } n_1 > 0,\ n_2 = 0 \\ p_2 & , \quad \text{if } n_1 = 0,\ n_2 > 0 \\ \frac{p_2}{3} & , \quad \text{if } n_1 > 0,\ n_2 > 0, \end{cases}$$

where $P(\text{merge step}|\boldsymbol{c}_{split}) = p_2 = \begin{cases} 1 & , \text{ if } K_0 = 1 \\ 1/2 & , \text{ else.} \end{cases}$

On the other hand,

$q(\boldsymbol{c}, \boldsymbol{c}_{split}) = P(\boldsymbol{c}_{split}|\boldsymbol{c}, \text{split step})P(\text{split step}|\boldsymbol{c}) = \frac{1}{K_0}$, since $P(\text{split step}|\boldsymbol{c}) = 1$.

3. In any other case, we propose a split or a merge step, each w.p. 1/2, and

$q(\boldsymbol{c}_{merge}, \boldsymbol{c}) = P(\boldsymbol{c}|\boldsymbol{c}_{merge}, \text{split step})P(\text{split step}|\boldsymbol{c}_{merge}) = \frac{p_3}{K_0+1}$,

where $p_3 = \begin{cases} 1 & , \text{ if } K_1 + K_2 = 1, \text{ or if both are equal to one and both are selected} \\ 1/2 & , \text{ else} \end{cases}$

$q(\boldsymbol{c}, \boldsymbol{c}_{merge}) = P(\boldsymbol{c}_{merge}|\boldsymbol{c}, \text{merge step})P(\text{merge step}|\boldsymbol{c}) = \frac{1}{2((K_1+1)(K_2+1)-1)}$

$q(\boldsymbol{c}_{split}, \boldsymbol{c}) = P(\boldsymbol{c}|\boldsymbol{c}_{split}, \text{merge step})P(\text{merge step}|\boldsymbol{c}_{split}) = \begin{cases} \frac{p_2}{(K_1+2)(K_2+1)-1} & , \text{ if } n_1 > 0,\ n_2 = 0 \\ \frac{p_2}{(K_1+1)(K_2+2)-1} & , \text{ if } n_1 = 0,\ n_2 > 0 \\ \frac{p_2}{(K_1+2)(K_2+2)-1} & , \text{ if } n_1 > 0,\ n_2 > 0, \end{cases}$

where p_2 is as above.

$q(\boldsymbol{c}, \boldsymbol{c}_{split}) = P(\boldsymbol{c}_{split}|\boldsymbol{c}, \text{split step})P(\text{split step}|\boldsymbol{c}) = \frac{1}{2K_0}$.

Bibliography

Abramowitz, M. and I. A. Stegun (1964). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, Volume 55 of *National Bureau of Standards Applied Mathematics Series*. For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C.

Aeschbacher, H. U., V. L. S. J. and R. Stalder (1977). The use of the beta-binomial distribution in dominant-lethal testing for weak mutagenic activity (part 1). *Mutation Research 44*.

Aigner, D., C. A. K. Lovell, and P. Schmidt (1977). Formulation and estimation of stochastic frontier production function models. *J. Econometrics 6*(1), 21–37.

Aldous, D. J. (1985). Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983*, Volume 1117 of *Lecture Notes in Math.*, pp. 1–198. Berlin: Springer.

Altham, P. M. E. (1978). Two generalizations of the binomial distribution. *J. Roy. Statist. Soc. Ser. C 27*(2), 162–167.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist. 2*, 1152–1174.

Barndorff-Nielsen, O. E. and N. Shephard (2001). Normal modified stable processes. *Teor. Ĭmovīr. Mat. Stat.* (65), 1–19.

Beal, M. J. and Z. Ghahramani (2003). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. In *Bayesian statistics, 7 (Tenerife, 2002)*, pp. 453–463. New York: Oxford Univ. Press.

Bernardo, J.-M. and A. F. M. Smith (1994). *Bayesian theory*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. Chichester: John Wiley & Sons Ltd.

Billingsley, P. (1995). *Probability and measure* (Third ed.). Wiley Series in Probability and Mathematical Statistics. New York: John Wiley & Sons Inc. A Wiley-Interscience Publication.

Blackwell, D. and J. B. MacQueen (1973). Ferguson distributions via pólya urn schemes. *Ann. Statist. 1*, 353–355.

Bohning, D., P. Schlattmann, and B. Lindsay (1992). Computer-assisted analysis of mixtures (c.a.man): Statistical algorithms. *Biometrics 48*(1), 283–303.

Brooks, S. P., B. J. T. Morgan, M. S. Ridout, and S. E. Pack (1997). Finite mixture models for proportions. *Biometrics 53*(3), 1097–1115.

Cifarelli, D. and E. Regazzini (1978). Problemi statistici non parametrici in condizioni di scambiabibilit parziale: impiego di medie associative. Technical report, Quaderni Istituto di Matematica Finanziaria dell'Universit di Torino.

Crauel, H. (2002). *Random probability measures on Polish spaces*, Volume 11 of *Stochastics Monographs*. London: Taylor & Francis.

Dalal, S. R. (1979). Dirichlet invariant processes and applications to nonparametric estimation of symmetric distribution functions. *Stochastic Process. Appl. 9*(1), 99–107.

Damien, P., J. Wakefield, and S. Walker (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *J. R. Stat. Soc. Ser. B Stat. Methodol. 61*(2), 331–344.

Dunson, D. B. and J.-H. Park (2008, June). Kernel stick-breaking processes. *Biometrika 95*(2), 307–323.

Escobar, M. D. (1988). *Estimating the Means of several Normal Populations by Nonparametric Estimation of the Distribution of the Means*. Ph. D. thesis, Yale University, Department of Statistics. Unpiblished PhD dissertation, Yale University, Department of Statistics.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *J. Amer. Statist. Assoc. 89*(425), 268–277.

Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc. 90*(430), 577–588.

Escobar, M. D. and M. West (1998). Computing nonparametric hierarchical models. In *Practical nonparametric and semiparametric Bayesian statistics*, Volume 133 of *Lecture Notes in Statist.*, pp. 1–22. New York: Springer.

Fearnhead, P. (2004). Particle filters for mixture models with an unknown number of components. *Stat. Comput. 14*(1), 11–21.

Feller, W. (1971). *An introduction to probability theory and its applications. Vol. II.* Second edition. New York: John Wiley & Sons Inc.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist. 1*, 209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist. 2*, 615–629.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent advances in statistics*, pp. 287–302. New York: Academic Press.

Fernández, C. and P. J. Green (2002). Modelling spatially correlated data via mixtures: a Bayesian approach. *J. R. Stat. Soc. Ser. B Stat. Methodol. 64*(4), 805–826.

Fernández, C., J. Osiewalski, and M. F. J. Steel (1997). On the use of panel data in stochastic frontier models with improper priors. *J. Econometrics 79*(1), 169–193.

Ferreira, J. T. and M. F. J. Steel (2007). A new class of skewed multivariate distributions with applications to regression analysis. *Statistica Sinica 17*, 505–529.

Garren, S. T., R. L. Smith, and W. W. Piegorsch (2001). Bootstrap goodness-of-fit test for the beta-binomial model. *J. Appl. Stat. 28*(5), 561–571.

Gelfand, A. E., S. K. Sahu, and B. P. Carlin (1995). Efficient parameterisations for normal linear mixed models. *Biometrika 82*(3), 479–488.

Gelfand, A. E. and A. F. M. Smith (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc. 85*(410), 398–409.

George, E. O. and D. Bowman (1995). A full likelihood procedure for analysing exchangeable binary data. *Biometrics 51*(2), 512–523.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc. 102*(477), 359–378.

Good, I. J. (1952). Rational decisions. *J. Roy. Statist. Soc. Ser. B. 14*, 107–114.

Gradshteyn, I. S. and I. M. Ryzhik (1994). *Table of integrals, series, and products* (Russian ed.). Boston, MA: Academic Press Inc. Translation edited and with a preface by Alan Jeffrey.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika 82*(4), 711–732.

Griffin, J. E. and M. F. J. Steel (2004). Semiparametric Bayesian inference for stochastic frontier models. *J. Econometrics 123*(1), 121–152.

Griffin, J. E. and M. F. J. Steel (2006). Order-based dependent Dirichlet processes. *J. Amer. Statist. Assoc. 101*(473), 179–194.

Griffin, J. E. and S. G. Walker (2009). Posterior simulation of normalised random measure mixtures. Technical report, Institue of Mathematics, Statistics and Actuarial Science, University of Kent.

Haseman, J. . and E. R. Soares (1976). The distribution of fetal death in control mice and its implications on statistical tests for dominant lethal effects. *Mutation Research 41*.

Ishwaran, H. and L. F. James (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc. 96*(453), 161–173.

Ishwaran, H. and M. Zarepour (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika 87*(2), 371–390.

Jain, S. and R. M. Neal (2005). Splitting and merging components of a nonconjugate dirichlet process mixture model. Technical report, Division of Biostatistics and Bioinformatics, University of California at San Diego; La Jolla CA 92093-0717, 2005.

James, D. A. and D. M. Smith (1982). Analysis of results from a collaborative study of the dominant lethal assay. *Mutation Research 97*.

James, L. F., A. Lijoi, and I. Pruenster (2005). Baysian inference via classes of normalized random measures.

James, L. F., A. Lijoi, and I. Prünster (2006). Conjugacy as a distinctive feature of the Dirichlet process. *Scand. J. Statist. 33*(1), 105–120.

Jefferys, W. H. and J. O. Berger (1992). Ockham's razor and bayesian analysis. *American Scientist 80*, 64–72.

Jeffreys, H. (1939). *Theory of Probability.* Oxford: Oxford University Press.

Johnson, W. P. (2002). The curious history of Faà di Bruno's formula. *Amer. Math. Monthly 109*(3), 217–234.

Kalli, M., J. E. Griffin, and S. G. Walker (2008). Slice sampling mixture models. Technical Report UKC/IMS/08/024, Institue of Mathematics, Statistics and Actuarial Science, University of Kent.

Koop, G., J. Osiewalski, and M. F. J. Steel (1997). Bayesian efficiency analysis through individual effects: Hospital cost frontiers. *Journal of Econometrics 76*(1-2), 77–105.

Kuk, A. Y. C. (2004). A litter-based approach to risk assessment in developmental toxicity studies via a power family of completely monotone functions. *J. Roy. Statist. Soc. Ser. C 53*(2), 369–386.

Kupper, L. L. and J. K. Haseman (1978). The use of a correlated binomial model for the analysis of certain toxicological experiments. *Biometrics 34*(1), 69–76.

Lavine, M. (1992). Some aspects of pólya tree distributions for statistical modelling. *Ann. Statist. 20*(3), 1222–1235.

Lavine, M. (1994). More aspects of pólya tree distributions for statistical modelling. *Ann. Statist. 22*(3), 1161–1176.

Lijoi, A., R. H. Mena, and I. Prünster (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *J. Amer. Statist. Assoc. 100*(472), 1278–1291.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist. 12*(1), 351–357.

MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Comm. Statist. Simulation Comput. 23*(3), 727–741.

MacEachern, S. N. (1998). Computational methods for mixture of Dirichlet process models. In *Practical nonparametric and semiparametric Bayesian statistics*, Volume 133 of *Lecture Notes in Statist.*, pp. 23–43. New York: Springer.

MacEachern, S. N. (1999). Dependent nonparametric process. ASA Proceeding of the Section on Bayesian Statistical Science. Alexandria, VA.

MacEachern, S. N., M. Clyde, and J. S. Liu (1999). Sequential importance sampling for nonparametric Bayes models: the next generation. *Canad. J. Statist. 27*(2), 251–267.

MacEachern, S. N. and P. Müller (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics 7*(2), 223–238.

Meeusen, W. and J. van den Broeck (1977, June). Efficiency estimation from cobb-douglas production functions with composed error. *International Economic Review 18*(2), 435–44.

Müller, P., F. Quintana, and G. Rosner (2004). A method for combining inference across related nonparametric Bayesian models. *J. R. Stat. Soc. Ser. B Stat. Methodol. 66*(3), 735–749.

Müller, P. and F. A. Quintana (2004). Nonparametric Bayesian data analysis. *Statist. Sci. 19*(1), 95–110.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist. 9*(2), 249–265.

Neal, R. M. (2003). Slice sampling. *Ann. Statist. 31*(3), 705–767. With discussions and a rejoinder by the author.

Newton, M. A. and A. E. Raftery (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *J. Roy. Statist. Soc. Ser. B 56*(1), 3–48. With discussion and a reply by the authors.

Nieto-Barajas, L. E., I. Prünster, and S. G. Walker (2004). Normalized random measures driven by increasing additive processes. *Ann. Statist. 32*(6), 2343–2360.

Ochi, Y. and R. L. Prentice (1984). Likelihood inference in a correlated probit regression model. *Biometrika 71*(3), 531–543.

Palmer, K. J., M. S. Ridout, and B. J. T. Morgan (2008). Modelling cell generation times by using the tempered stable distribution. *Journal Of The Royal Statistical Society Series C 57*(4), 379–397.

Pang, Z. and A. Y. C. Kuk (2005). A shared response model for clustered binary data in developmental toxicity studies. *Biometrics 61*(4), 1076–1084.

Papaspiliopoulos, O. and G. O. Roberts (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika 95*(1), 169–186.

Paul, S. R. (1985). A three-parameter generalisation of the binomial distribution. *Comm. Statist. Theory Methods 14*(6), 1497–1506.

Paul, S. R. (1987). On the beta-correlated binomial (BCB) distribution—a three-parameter generalization of the binomial distribution. *Comm. Statist. Theory Methods 16*(5), 1473–1478.

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, probability and game theory*, Volume 30 of *IMS Lecture Notes Monogr. Ser.*, pp. 245–267. Hayward, CA: Inst. Math. Statist.

Regazzini, E. (2001). Foundations of bayesian statistics and some theory of bayesian nonparametric methods. Lecture notes, Stanford University.

Richardson, S. and P. J. Green (1998). Corrigendum: "On Bayesian analysis of mixtures with an unknown number of components" [J. Roy. Statist. Soc. Ser. B **59** (1997), no. 4, 731–792]. *J. R. Stat. Soc. Ser. B Stat. Methodol. 60*(3), 661.

Roberts, G. O. and J. S. Rosenthal (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statist. Sci. 16*(4), 351–367.

Rodríguez-Avi, J., A. Conde-Sánchez, A. J. Sáez-Castillo, and M. J. Olmo-Jiménez (2007). A generalization of the beta-binomial distribution. *J. Roy. Statist. Soc. Ser. C 56*(1), 51–61.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica 4*(2), 639–650.

Sethuraman, J. and R. C. Tiwari (1982). Convergence of Dirichlet measures and the interpretation of their parameter. In *Statistical decision theory and related topics, III, Vol. 2 (West Lafayette, Ind., 1981)*, pp. 305–315. New York: Academic Press.

Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc. 101*(476), 1566–1581.

Tweedie, M. C. K. (1984). An index which distinguishes between some important exponential families. In *Statistics: applications and new directions (Calcutta, 1981)*, pp. 579–604. Calcutta: Indian Statist. Inst.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Comm. Statist. Simulation Comput. 36*(1-3), 45–54.

West, M., P. Müller, and M. D. Escobar (1994). Hierarchical priors and mixture models, with application in regression and density estimation. In *Aspects of uncertainty*, Wiley Ser. Probab. Math. Statist. Probab. Math. Statist., pp. 363–386. Chichester: Wiley.

Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *J. Roy. Statist. Soc. Ser. C 31*(2), 144–148.

Zellner, A. (1971). *An introduction to Bayesian inference in econometrics*. New York: John Wiley & Sons Inc. Wiley Series in Probability and Mathematical Statistics.

Zhang, S. and J. Jin (1996). *Computation of special functions*. A Wiley-Interscience Publication. New York: John Wiley & Sons Inc. With 1 IBM-PC floppy disk (3.5 inch; DD).