On Bayesian nonparametric modelling of two correlated distributions

M. Kolossiatis, J. E. Griffin and M. F. J. Steel

November 4, 2010

Abstract

In this paper, we consider the problem of modelling a pair of related distributions using Bayesian nonparametric methods. A representation of the distributions as weighted sums of distributions is derived through normalisation. This allows us to define several classes of nonparametric priors. The properties of these distributions are explored and efficient Markov chain Monte Carlo methods are developed. The methodology is illustrated on simulated data and an example concerning hospital efficiency measurement.

Keywords: Hospital efficiencies, Markov chain Monte Carlo, Normalised Random Measures, Pólya urn

1 Introduction

Data often arise under different conditions, either experimentally (different treatments) or observationally (different socio-economic groups). A standard statistical analysis of such data involves the comparison of the distributions under the different conditions. A parametric analysis might compare means, medians, variances or other summaries of the data. In this paper, we take a nonparametric approach and consider the simplest case where we have two distributions. A simple nonparametric analysis would separately fit distributions to data under the two conditions. However, it is likely that

the distributions under the two conditions will be related and a Bayesian hierarchical approach is a natural way to exploit such a relationship.

The problem of modelling a finite number, J, of related distributions has enjoyed substantial attention recently. Models are usually expressed in a mixture model framework where the data in group j, $Y_{j1}, Y_{j2}, \ldots, Y_{jN_j}$ can be expressed as

$$Y_{ji} \overset{ind.}{\sim} f(Y_{ji}; \theta_{ji}),$$

$$\theta_{ji} \overset{ind.}{\sim} F_j^*$$

where $f(\cdot; \theta)$ is a probability density function with parameters θ and $F_1^*, F_2^*, \ldots, F_J^*$ are discrete distributions with an infinite number of atoms. The problem then reduces to modelling the dependence between $F_1^*, F_2^*, \ldots, F_J^*$.

The Hierarchical Dirichlet process (HDP) (Teh et al., 2006) has become a central tool in many nonparametric models for allowing dependence between $F_1^*, F_2^*, \ldots, F_J^*$. The model assumes that $F_j^* \overset{iid}{\sim} \text{DP}(M, H)$ and $H \sim \text{DP}(M_0, H_0)$, where $\text{DP}(M, H)$ denotes a Dirichlet Process (DP) prior with mass parameter M and centring distribution H. The introduction of an unknown centring distribution H for $F_1^*, F_2^*, \ldots, F_J^*$ encourages *a posteriori* dependence between the distributions. This model is appropriate if the distributions can be considered exchangeable and defines a particular type of dependence between distributions, as discussed in Section 2.

Several alternative approaches to modelling correlated distributions have been proposed. The bivariate Dirichlet process (Walker and Muliere, 2003) introduces latent variables to encourage dependence between two conditionally independent distributions which are given Dirichlet process priors. The distributions could also be modelled explicitly as

$$F_j^* = \sum_{k=1}^{\infty} p_{jk} \delta_{\phi_k}$$

where δ_x denotes a Dirac measure at x, while $\phi_1, \phi_2, \phi_3, \ldots$ is an infinite sequence of iid random variables and $\sum_{k=1}^{\infty} p_{jk} = 1$ for $j = 1, 2, \ldots, J$. The construction of priors for $\{p_{jk}\}$ is technically challenging but some possible constructions are described by Ishwaran and Zarepour (2009) and Leisen and Lijoi (2010). A simpler approach to defining dependent random measures $F_1^*, F_2^*, \ldots, F_J^*$ arises from taking mixtures of distributions. Müller et al. (2004) discuss such a method by taking

$$F_j^* = wF_0 + (1 - w)F_j$$

where $0 \leq w \leq 1$ and $F_0, F_1, \ldots, F_J$ are distributions. Each distribution is a mixture of a common component, F_0, shared by all distributions, and F_j which is specific to the j-th distribution and will be termed an idiosyncratic distribution. The weight w controls the dependence between the distributions, with larger weights associated with greater dependence. This idea is extended to spatial problems by Rao and Teh (2009). This paper will build on this framework by allowing the weight to depend on j, allowing F_j^* to have a given marginal prior and introducing more efficient Markov chain Monte Carlo (MCMC) methods for posterior simulation.

The paper is organised as follows: Section 2 discusses some general ideas about modelling correlated distributions and possible approaches, Section 3 describes MCMC methods for fitting our models, Section 4 includes applications of the methods to simulated data and an economic example, and Section 5 provides a brief discussion. Proofs are grouped in the Appendix.

2 Modelling 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. A full hierarchical model could then be of the form:

$$\begin{aligned} Y_{ji} &\overset{ind.}{\sim} f(Y_{ji}; \theta_{ji}, \boldsymbol{\psi}),\ i = 1, 2, \ldots, N_j,\ j = 1, 2, \ldots, J \\ \theta_{ji} &\overset{ind.}{\sim} F_j^*,\ j = 1, 2, \ldots, J \\ F_1^*, F_2^*, \ldots, F_J^* &\sim h(\boldsymbol{\lambda}),\ j = 1, 2, \ldots, J \\ \boldsymbol{\psi} &\sim \pi(\boldsymbol{\psi}),\ \boldsymbol{\lambda} \sim \pi(\boldsymbol{\lambda}). \end{aligned} \tag{1}$$

In the above, Y_{ji} are data from J different groups of sizes $N_1, N_2, \ldots$ and N_J, while θ_{ji} are the parameters to be flexibly modelled using nonparametric, correlated distributions, which have prior h. In this paper we will focus on defining appropriate priors for the random distributions F_j^*. The model also includes $\boldsymbol{\psi}$ which are (potential) additional parameters in the distribution of Y_{ji} given θ_{ji} and $\boldsymbol{\lambda}$ which groups the parameters of the prior distribution of $F_j^*,\ j = 1, 2, \ldots, J$.

In the simplest case of two correlated random distributions, the following lemma gives a representation of the mixing distributions F_1^* and F_2^* as mixtures.

Lemma 1 *If F_1^* and F_2^* are discrete probability distributions, they can be represented as*

$$F_1^* = \varepsilon F_0 + (1-\varepsilon)\left[\varphi_1 F_1^{(0)} + (1-\varphi_1)F_1\right]$$
$$F_2^* = \varepsilon F_0 + (1-\varepsilon)\left[\varphi_2 F_2^{(0)} + (1-\varphi_2)F_2\right]$$

where $0 \leq \varepsilon, \varphi_1, \varphi_2 \leq 1$ and $F_1^{(0)}$ and $F_2^{(0)}$ share no atoms but have atoms in common with F_0, while F_1 and F_2 are discrete probability distributions which share no atoms with each other or with F_0, $F_1^{(0)}$ and $F_2^{(0)}$.

This representation is hard to work with since F_0 shares atoms with $F_1^{(0)}$ and $F_2^{(0)}$ and so must be modelled by correlated priors. In fact, the model of Müller et al. (2004) assumes that $\varphi_1 = \varphi_2 = 0$, which avoids modelling this correlation. This is not a terribly restrictive simplification since $F_1^{(0)}$ and $F_2^{(0)}$ can be approximated by placing points mass "close" to the points of F_0. At the other extreme, the Hierarchical Dirichlet process (Teh et al., 2006) assumes that $\varphi_1 = \varphi_2 = 1$ and all atoms are shared by all distributions. The most general model involving mixtures of random measures which share no atoms arises when

$$\begin{aligned} F_1^* &= \varepsilon_1 F_0 + (1-\varepsilon_1)F_1, \\ F_2^* &= \varepsilon_2 F_0 + (1-\varepsilon_2)F_2. \end{aligned} \tag{2}$$

The model of Müller et al. (2004) assumes that $\varepsilon_1 = \varepsilon_2$. In this paper, we will restrict attention to the model in (2). This model allows for a simple interpretation of F_0 as the common part shared by F_1^* and F_2^*, whereas F_1 and F_2 are idiosyncratic parts.

Bayesian inference in this model involves placing priors on the parameters F_0, F_1, F_2, ε_1 and ε_2. It is natural to assume that F_0, F_1 and F_2 are independent random probability measures since they share no atoms. However, we would often want to assume that ε_1 and ε_2 are correlated *a priori* since they will usually relate to distributions under similar conditions. The two correlated distributions can then be naturally embedded at an intermediate level of a larger hierarchical model. The intermediate levels of the hierarchical model in (1) will then be the following:

$$\begin{gathered} \theta_{ji} \overset{ind.}{\sim} F_j^*, \text{ where } F_j^* = \varepsilon_j F_0 + (1-\varepsilon_j)F_j,\ j = 1,2 \\ F_j \overset{ind.}{\sim} \mathrm{DP}(M_j, H(\boldsymbol{\lambda})),\ j = 0,1,2 \\ \varepsilon_1, \varepsilon_2 \sim \pi(\varepsilon_1, \varepsilon_2) \end{gathered} \tag{3}$$

$$M_j \overset{ind.}{\sim} \pi(M_j),\ j = 0, 1, 2,$$

where we have chosen Dirichlet Process (DP) priors for F_j. This paper will focus on DP priors, but other nonparametric priors can be considered (Griffin et al., 2010). We employ different concentration parameters for the three DPs, but use the same centring distribution, H, with parameter $\boldsymbol{\lambda}$. 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 $\boldsymbol{\lambda}$.

The hierarchical model of Müller et al. (2004) for $J = 2$ is a special case of model (3), where $\varepsilon_1 = \varepsilon_2 = \varepsilon$ and a certain prior distribution is given to the common weight, ε. However, the form of the model described in (3) has some attractive features which will be investigated in the following two subsections.

2.1 The Normalisation Model

The model for F_1^* and F_2^* in (3) can be constructed by normalising sums of Gamma processes. Let $\text{Ga}(a, b)$ denote a Gamma distribution with shape parameter a and mean a/b and use the notation $G \sim \Gamma\text{P}(M, H)$, where $M > 0$ and H is a distribution function, to represent that G follows a Gamma process for which $G(B) \sim \text{Ga}(MH(B), 1)$ for all measurable sets B. The model in (3) can then be obtained in the following way. Let G_0, G_1 and G_2 be independent and $G_i \sim \Gamma\text{P}(M_i, H)$ for $i = 0, 1, 2$ and define $G_1^* = G_0 + G_1$ and $G_2^* = G_0 + G_2$. Then, $G_1^* \sim \Gamma\text{P}(M_0 + M_1, H)$ and $G_2^* \sim \Gamma\text{P}(M_0 + M_2, H)$. Normalising G_1^* to give $F_1^* = \frac{G_1^*}{G_1^*(\Omega)}$ and, similarly, $F_2^* = \frac{G_2^*}{G_2^*(\Omega)}$ leads to $F_1^* \sim \text{DP}(M_0 + M_1, H)$ and $F_2^* \sim \text{DP}(M_0 + M_2, H)$ (Ferguson, 1973). Then,

$$\begin{aligned} F_1^*(B) &= \frac{G_1^*(B)}{G_1^*(\Omega)} = \frac{G_0(\Omega)}{G_0(\Omega) + G_1(\Omega)} \frac{G_0(B)}{G_0(\Omega)} + \frac{G_1(\Omega)}{G_0(\Omega) + G_1(\Omega)} \frac{G_1(B)}{G_1(\Omega)} \\ &= \varepsilon_1 F_0(B) + (1 - \varepsilon_1) F_1(B) \end{aligned}$$

where $\varepsilon_1 = \frac{G_0(\Omega)}{G_0(\Omega)+G_1(\Omega)}$, $F_0(B) = \frac{G_0(B)}{G_0(\Omega)}$, and $F_1(B) = \frac{G_1(B)}{G_1(\Omega)}$. It follows from the properties of Gamma processes that ε_1, F_0 and F_1 are independent and that $\varepsilon_1 \sim \text{Be}(M_0, M_1)$, $F_0 \sim \text{DP}(M_0, H)$ and $F_1 \sim \text{DP}(M_1, H)$. Similarly,

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

where $F_2 \sim \text{DP}(M_2, H)$ and $\varepsilon_2 \sim \text{Be}(M_0, M_2)$. So, for $M_1 = M_2$, F_1^* and F_2^* are identically DP-distributed, but not independent, due to the common part F_0. The

same holds for the two weights, which are both marginally beta-distributed, but are not independent. In fact, their joint density is

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

for $0<\varepsilon_1,\varepsilon_2<1$.

So, we can construct correlated distributions with DP marginals with parameters M and H by taking weighted sums of independent DPs with the same base distribution H. This idea could be extended to larger numbers of distributions or any process constructed by normalising a random measure with independent increments (James et al., 2005), as discussed by Griffin et al. (2010). However, the model with Dirichlet process marginals is the only one where the weights are independent of the component random distributions. This is due to the properties of the Gamma process and does not hold for any other infinitely divisible process.

2.2 The Single-ε Model

The model defined by normalisation leads to correlated weights ε_1 and ε_2. A simplified version of this model assumes a common weight ε and is closer to the model of Müller et al. (2004). The model is

$$\begin{gathered}\theta_{ji}\sim F_j^*,\ \text{ where } F_j^*\ =\ \varepsilon F_0\ +\ (1-\varepsilon)F_j,\ j=1,2\\ F_0\sim \text{DP}(M_0,H(\boldsymbol{\lambda})),\ F_1,F_2\overset{iid}{\sim}\text{DP}(M_1,H(\boldsymbol{\lambda}))\\ \varepsilon\sim \text{Be}(M_0,M_1)\\ M_0,M_1\overset{iid}{\sim}\text{Ga}(a_0,b_0),\ \boldsymbol{\lambda}\sim\pi(\boldsymbol{\lambda}).\end{gathered}\tag{4}$$

The simplification $\varepsilon_1=\varepsilon_2$ allows for more direct sharing of information between the two distributions (since the weights are now the same rather than just correlated). This sharing of information can be particularly useful in cases of few observations from one or both distributions. On the other hand, unless someone is particularly interested in inferring the weights ε_1 and ε_2, 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. This is a direct result of model fitting and Ockham's razor, as explained in Müller et al. (2004).

2.2.1 Properties of the Single-ε Model

The Single-ε model has some very nice properties, both theoretical and computational, many 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 Section 3.

First of all, the marginal distributions of F_1^* and F_2^* can be shown to be (see Appendix)

$$F_1^*, F_2^* \sim \text{DP}(M_0 + M_1, H). \tag{5}$$

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

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

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

and

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

The last expression is an interesting result, as it indicates that the correlation between the masses allocated to a set A by F_1^* and F_2^* does not depend on A or H.

Let $\boldsymbol{s}$ denote the vector of all allocation parameters, assigning each data point to a distinct value in the Dirichlet process. The exchangeable product partition formula (EPPF) of the Dirichlet process with mass parameter M has the well-known form

$$p(\boldsymbol{s}|M) = M^k \frac{\Gamma(M)}{\Gamma(M+n)} \prod_{i=1}^{K} \Gamma(n_i) \tag{6}$$

where there are K distinct values with data points allocated to them and n_i is the number of data points allocated to the i-th distinct value. Next, the EPPF and the Pólya-urn representations for model (4) are derived. In order to do this, it is useful to notice that the model for F_1^* and F_2^* in (4) is a mixture model, so we introduce two sets of indicators, r_{ji} and $s_{ji},\ i = 1, 2, \ldots, N_j,\ j = 1, 2$. 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, is drawn from the common part or the idiosyncratic part: $\theta_{ji} \sim F_0$ if $r_{ji} = 0$ and $\theta_{ji} \sim F_j$ if $r_{ji} = 1$ for $i = 1, 2, \ldots, N_j,\ j = 1, 2$. The indicators s_{ji} assign each θ_{ji} to one of the discrete values of the component distributions $F_j,\ j = 0, 1, 2$ (given the value of r_{ji}):

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

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 (for example K_1 is the number of distinct s_{1i}, for which $r_{1i} = 1$).

Proposition 1 *The EPPF for model (4) is:*

$$p(\boldsymbol{s}, \boldsymbol{r}|\boldsymbol{M}) = \frac{\Gamma(M_0 + M_1)}{\Gamma(M_0 + M_1 + N)} M_0^{K_0} M_1^{K_1+K_2} \frac{\Gamma(M_1 + n_1 + n_2)\Gamma(M_1)}{\Gamma(M_1 + n_1)\Gamma(M_1 + n_2)} \prod_{j=0}^{2} \prod_{i=1}^{K_j} \Gamma(n_{j,i}) \tag{7}$$

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 in use, $n_{j,i}$ 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 $j \in \{0, 1, 2\}$.

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

Proposition 2 *Suppose that $\theta_{1,1}, \theta_{1,2}, \ldots, \theta_{1,N_1} \sim F_1^*$ and $\theta_{2,1}, \theta_{2,2}, \ldots, \theta_{2,N_2} \sim F_2^*$. The Pólya-urn representations for model (4) will be as follows: $\forall\ A \in \mathcal{F}$ and $j = 1, 2$*

$$P\left(\theta_{j,N_j+1} \in A|D\right) = w_0 \bar{F}_0(A) + w_j \bar{F}_j(A) + (1 - w_0 - w_j)H(A)$$

where $w_0 = \frac{n_0}{M_0+M_1+N}$, $w_j = \frac{M_1+N-n_0}{M_0+M_1+N} \frac{n_j}{M_1+n_j}$ and

$$\bar{F}_j = \frac{1}{n_j} \sum_{i=1}^{K_j} n_{ji} \delta_{\phi_{ji}}, \quad j = 0, 1, 2.$$

Here D denotes the set of all data and the rest is as in Proposition 1.

The distribution of a future observation θ_{j,N_j+1} is drawn from a mixture of the empirical distribution of the observations allocated to the common component ($\bar{F}_0$), the

empirical distribution of the observations in group j which are not allocated to the common component ($\bar{F}_j$) and the centring distribution H.

Instead of M_0 and M_1, one can also use the alternative parameterisation $x = M_0 + M_1$ and $y = \frac{M_0}{M_0+M_1}$. Based on the results of Theorem 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 and allows us to rewrite (7) as

$$p(\boldsymbol{s}, \boldsymbol{r}|x, y) = \kappa_1 \kappa_2 \kappa_3, \tag{8}$$

where

$$\kappa_1 = \frac{\Gamma(x)}{\Gamma(x+N)} x^{K_0+K_1+K_2} \prod_{j=0}^{2} \prod_{i=1}^{K_j} \Gamma(n_{j,i}),$$

$$\kappa_2 = y^{K_0}(1-y)^{K_1+K_2},$$

and

$$\kappa_3 = \frac{\Gamma(x(1-y)+n_1+n_2)\Gamma(x(1-y))}{\Gamma(x(1-y)+n_1)\Gamma(x(1-y)+n_2)}.$$

Note from (6) that the factor κ_1 relates to $p(\boldsymbol{s}|x)$ if we were to sample from a single distribution with a DP prior with precision parameter $x = M_0 + M_1$. The second part κ_2 can be seen as the contribution to the joint distribution of "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, κ_3 refers to "splitting" the data not allocated to the common part into the two idiosyncratic parts.

Likewise, under the assumption that the allocation of observations between the different components is in line with the prior so that on average $n_0 = Ny$ and $n_j = N_j(1-y)$, the weights in Proposition 2 can be written as $w_0 = y\frac{N}{x+N}$ and $w_j = (1-y)\frac{N_j}{x+N_j}$. Both weights are expressed as linear functions of the parameter y, measuring strength of dependence, multiplied by a term which gets larger as N_j gets larger or x gets smaller, which controls the contribution of the empirical distribution to the predictive.

Whereas the Single-ε model and the model of Müller et al. (2004) are very similar, they have some notable differences in their behaviour and their properties. The reason for that lies in the way these 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. In model (4) the random distributions F_1^* and F_2^* are DP-distributed, whereas this is not always true for the other model. In model (4) the expressions for the first two central moments and the correlation structure are very simple and easy to use. The corresponding quantities for the model of Müller et al. (2004) are easy to derive, but more complicated. The same holds for the Pólya-urn representations and the EPPF. Another nice feature of (4) is the nice intuitive form of expression (8).

3 Computational Methods

In this section we describe the MCMC methods used to fit a hierarchical mixture model with a Single-ϵ or Normalisation prior for a pair of correlated distributions. As in Müller et al. (2004), we assume that the sampling model $f(Y_{ji}; \mu_{ji}, S)$ is a normal distribution with mean μ_{ji} and variance S and that $\mu_{ji} \sim F_j^*$. The base distribution H follows a normal distribution $\mathrm{N}(m, B)$. 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, $\mathrm{IGa}(c, cC)$ (so that the prior mean of B is $\frac{cC}{c-1}$ (for $c > 1$) and the prior variance is $\frac{c^2C^2}{(c-1)^2(c-2)}$ (for ($c > 2$)), and the variance S is also given an inverse gamma distribution with parameters q and qR. The full Single-ε model can thus be written as

$$
\begin{aligned}
Y_{ji} &\sim \mathrm{N}(\mu_{ji}, S),\ i = 1, 2, \ldots, N_j,\ j = 1, 2 \\
\mu_{ji} &\sim F_j^*, \text{ where } F_j^* = \varepsilon F_0 + (1 - \varepsilon) F_j \\
F_0 &\sim \mathrm{DP}(M_0, H),\ F_j \overset{iid}{\sim} \mathrm{DP}(M_1, H), \text{ where } H \equiv \mathrm{N}(m, B) \\
\varepsilon &\sim \mathrm{Be}(M_0, M_1) \\
M_0, M_1 &\overset{iid}{\sim} \mathrm{Ga}(a_0, b_0),\ (m, B) \sim \mathrm{N}(m_0, A) \times \mathrm{IGa}(c, cC),\ S \sim \mathrm{IGa}(q, qR).
\end{aligned}
\tag{9}
$$

In order to simulate from the posterior distribution of model (9), we use a Pólya-urn scheme and use the fact that the sampling model and the centring distribution are conjugate. As in Müller et al. (2004), we use the indicators s_{ji}, r_{ji} and the discrete values ϕ_{ji} defined in Section 2.2.1. The posterior distribution involves $\boldsymbol{r}$, $\boldsymbol{s}$, $\{\phi_{ji}\}$, M_0, M_1, ε, m, B and S. The full conditional distributions of all parameters will be

the same as in Müller et al. (2004) with $M_2 = M_1$, except for the parameters ε, M_0 and M_1, on which we will focus in the next subsection.

3.1 MCMC sampler for the Single-ε model

The full conditionals of M_0, M_1 and ε are:

- $\varepsilon|\cdots \sim \text{Be}(M_0 + N - \sum_{j,i} r_{ji}, M_1 + \sum_{j,i} r_{ji})$ which can be simulated directly.
- $f(M_0|\cdots) \propto M_0^{a_0+K_0-1} e^{-M_0[b_0 - \log(\varepsilon)]} \frac{\Gamma(M_0+M_1)}{\Gamma(M_0+n_0)}$ and $f(M_1|\cdots) \propto M_1^{a_0+K_1+K_2-1} e^{-M_1[b_0-\log(1-\varepsilon)]} \frac{\Gamma(M_1)\Gamma(M_0+M_1)}{\Gamma(M_1+n_1)\Gamma(M_1+n_2)}$. We can use Random Walk Metropolis-Hastings (RWMH) steps for these parameters.

The marginal posterior distribution of ε is often bimodal which can cause slow mixing for the algorithms described so far. To combat this problem, we introduce an additional split/merge step.

3.1.1 The Split/Merge Step

The split/merge step allows faster movement between the modes of the marginal distribution of ε. 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 calculate the Metropolis-Hastings acceptance probability. 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, which depends 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, in use (i.e. the number of distinct s_{ji} within each of F_0, F_1 and F_2, according to the corresponding r_{ji}'s), 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.

To simplify expressions, we will write the multinomial Beta function as $B(\boldsymbol{a}) = \frac{\prod \Gamma(a_i)}{\Gamma(\sum a_i)}$ and define

$$
\begin{aligned}
e(m_1, m_2) = &\exp\left\{-\frac{1}{2}\left[\frac{(m_1+m_2)m^2 - 2m(\sum Y_1' + \sum Y_2') - \frac{B}{S}(\sum Y_1' + \sum Y_2')^2}{(m_1+m_2)B+S}\right]\right\} \\
&\times \exp\left\{\frac{1}{2}\left[\frac{m_1 m^2 - 2m\sum Y_1' - \frac{B}{S}(\sum Y_1')^2}{m_1 B + S}\right]\right\} \\
&\times \exp\left\{\frac{1}{2}\left[\frac{m_2 m^2 - 2m\sum Y_2' - \frac{B}{S}(\sum Y_2')^2}{m_2 B + S}\right]\right\}
\end{aligned}
$$

and

$$
d(m_1, m_2) = \sqrt{\frac{S[(m_1+m_2)B+S]}{(m_1 B+S)(m_2 B+S)}}.
$$

The sums appearing in $e(m_1, m_2)$ are taken over the Y_{1i} or Y_{2i} associated with the clusters chosen to be split or merged. The algorithm for the split/merge step and the corresponding acceptance probabilities $\alpha(\mathbf{c}, \mathbf{c}')$, where $\mathbf{c} = (\mathbf{r}, \mathbf{s})$ is the current and $\mathbf{c}' = (\mathbf{r}', \mathbf{s}')$ is the proposed complete vector of indicators for model (9), are as follows:

Split/Merge Method:

1. Choose split or merge, each with probability 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_1 or F_2).
   (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.

The acceptance probabilities $\alpha(\mathbf{c}, \mathbf{c}')$ will be as follows:

i. If we propose to move a cluster from F_0 to $F_k,\ k = 1, 2$, say the cluster corresponding to the d-th discrete value in F_0, ϕ_{0d},
$\alpha(\mathbf{c}, \mathbf{c}') = \min\left\{1, \frac{M_1}{M_0}\frac{B(M_1+n_1+n_2+m_{0k}, M_1+n_k)}{B(M_1+n_1+n_2, M_1+n_{3-k}+m_{0k})}\frac{K_0}{(K_k+2)(K_{3-k}+1)-1}\right\}$.

ii. 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, ϕ_{0d}, the acceptance probability will be:
$\alpha(\mathbf{c}, \mathbf{c}') = \min\left\{1, \frac{M_1^2}{M_0}a\frac{K_0}{(K_1+2)(K_2+2)-1}\frac{d(m_{01}, m_{02})}{e(m_{01}, m_{02})}\right\}$
where
$a = \frac{B(M_1+n_1+n_2+m_{01}+m_{02}, M_1+n_1, M_1+n_2)}{B(M_1+n_1+n_2, M_1+n_1+m_{01}, M_1+n_2+m_{02})}B(m_{01}, m_{02})$.

We accept the split with the corresponding probability above. 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.

(b) Otherwise, if only $K_k = 0,\ k = 1, 2$, we propose to move a cluster from the other idiosyncratic part to the common one. In other words, we propose merging a cluster from F_{3-k} with an empty cluster from $F_k,\ k = 1, 2$.
In this case, we uniformly choose a cluster from the other idiosyncratic part (corresponding to, say, $\phi_{3-k,d}$) and move it to the common part with probability $\alpha(\mathbf{c}, \mathbf{c}') = \min\left\{1, \frac{M_0}{M_1}\frac{K_{3-k}}{K_0+1}\right\}$.
If the step is rejected, we do nothing.

(c) 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 (including the empty cluster) is chosen with probability $1/(K_1 + 1)$. We similarly choose a cluster from F_2 or an empty cluster. If two empty clusters are chosen, we repeat the above draw, since this merging is prohibited (in order to have a reversible MCMC algorithm). The acceptance probability in this case will be:

i. If we propose to transfer the selected cluster from F_k to $F_0,\ k = 1, 2$,
$\alpha(\mathbf{c}, \mathbf{c}') = \min\left\{1, \frac{M_0}{M_1}\frac{B(M_1+n_1+n_2-m_k, M_1+n_k)}{B(M_1+n_1+n_2, M_1+n_k-m_k)}\frac{(K_1+1)(K_2+1)-1}{K_0+1}\right\}$.

ii. If two existing clusters are chosen, corresponding to, say, (ϕ_{1d}, ϕ_{2d}), the acceptance probability is:
$\alpha(\mathbf{c}, \mathbf{c}') = \min\left\{1, \frac{M_0}{M_1^2} a \frac{(K_1+1)(K_2+1)-1}{K_0+1} \frac{e(m_1,m_2)}{d(m_1,m_2)}\right\}$
where
$a = \frac{B(M_1+n_1+n_2-m_1-m_2, M_1+n_1, M_1+n_2)}{B(M_1+n_1+n_2, M_1+n_1-m_1, M_1+n_2-m_2)} \frac{1}{B(m_1,m_2)}$.

If the proposed step is accepted, we perform the merging.

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 Markov chain. This is 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^*.

3.2 MCMC sampler for the Normalisation model

Now, we consider the Normalisation model described in Subsection 2.1 with $M_1 = M_2$. It is useful to write

$$\varepsilon_j = \frac{\gamma_0}{\gamma_0 + \gamma_j}, \qquad j = 1, 2$$

where γ_0, γ_1 and γ_2 are mutually independent with $\gamma_0 \sim \text{Ga}(M_0, 1), \gamma_1 \sim \text{Ga}(M_1, 1)$ and $\gamma_2 \sim \text{Ga}(M_1, 1)$. It is computationally more convenient here to work with the parametrisation γ_0, γ_1 and γ_2 instead of $\varepsilon_1 = \frac{\gamma_0}{\gamma_0+\gamma_1}$ and $\varepsilon_2 = \frac{\gamma_0}{\gamma_0+\gamma_2}$. The parameters that need to be updated differently to the algorithm for the model in Section 3.1 are γ_0, γ_1, γ_2, M_0 and M_1. We also describe the necessary expressions to use the split-merge move. The joint full conditional distribution for the γ's will be:

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

In order to simulate from the above distribution, we use the identity

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

and introduce latent variables U_{ki} for $k = 1, 2$ and $i = 1, 2, \ldots, N_k$ and define $\boldsymbol{U}$ to be the set of U_{ki}, so that

$$f(\gamma_0, \gamma_1, \gamma_2, \boldsymbol{U}|\cdots) \propto \gamma_0^{M_0-1} e^{-\gamma_0} \gamma_1^{M_1-1} e^{-\gamma_1} \gamma_2^{M_1-1} e^{-\gamma_2} \prod_{k=1}^{2} \prod_{i=1}^{N_k} \gamma_0^{1-r_{ki}} \gamma_k^{r_{ki}} e^{-(\gamma_0+\gamma_k)U_{ki}}.$$

Integrating across $\boldsymbol{U}$ leads to the correct distribution. Therefore, consider the augmented vector of parameters $(\gamma_0, \gamma_1, \gamma_2, \boldsymbol{U})$. In this case, the full conditional distributions are of known form:

$U_{ki}|\cdots \sim \text{Exp}(\gamma_0 + \gamma_k), \quad k = 1, 2, \quad i = 1, 2, \ldots, N_k,$

$\gamma_0|\cdots \sim \text{Ga}(M_0 + N_1 + N_2 - \sum_{k=1}^{2}\sum_{i=1}^{N_k} r_{ki}, 1 + \sum_{k=1}^{2}\sum_{i=1}^{N_k} U_{ki}),$

$\gamma_k|\cdots \sim \text{Ga}(M_1 + \sum_{i=1}^{N_k} r_{ki}, 1 + \sum_{i=1}^{N_k} U_{ki}),\ k = 1, 2.$

Here, $\text{Exp}(\theta)$ denotes the exponential distribution with mean $1/\theta$.

The full conditionals of M_0 and M_1, due to the different prior of the weights (actually, the priors of the γ's), will be:

$f(M_0|\cdots) \propto M_0^{a_0+K_0-1} e^{-b_0 M_0} \gamma_0^{M_0} \frac{1}{\Gamma(M_0+n_0)}$ and

$f(M_1|\cdots) \propto M_1^{a_0+K_1+K_2-1} e^{-b_0 M_1} \gamma_1^{M_1} \gamma_2^{M_1} \frac{1}{\Gamma(M_1+n_1)\Gamma(M_1+n_2)}$

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.

Despite the fact that now there are $N_1 + N_2$ auxiliary variables, 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. Also notice that, since the size of these auxiliary variables is equal to the data size, there will not be any additional problems of varying dimensionality of the parameter space.

Additionally, an extra split/merge step, similar to the one presented before (the differences will be in the acceptance probabilities) can also be incorporated in this algorithm and help improve mixing of the chains. The corresponding acceptance probabilities will now be:

2b, i) $\alpha(\mathbf{c}, \mathbf{c}') = \min\left\{1, \frac{M_1}{M_0} \cdot \left(\frac{1+\sum_{j=1}^{2}\sum_{i=1}^{N_j} U_{ji}}{1+\sum_{i=1}^{N_k} U_{ki}}\right)^{m_{0(3-k)}} \frac{K_0}{(K_{3-k}+1)(K_k+2)-1}\right\},\ k = 1, 2,$

2b, ii) $\alpha(\mathbf{c}, \mathbf{c}') = \min\left\{1, \frac{M_1^2}{M_0} \left(\frac{1+\sum_{j=1}^{2}\sum_{i=1}^{N_j} U_{ji}}{1+\sum_{i=1}^{N_1} U_{1i}}\right)^{m_{01}} \left(\frac{1+\sum_{j=1}^{2}\sum_{i=1}^{N_j} U_{ji}}{1+\sum_{i=1}^{N_2} U_{2i}}\right)^{m_{02}} a\right\}$, where

$a = B(m_{01}, m_{02}) \frac{d(m_{01},m_{02})}{e(m_{01},m_{02})} \frac{K_0}{(K_1+2)(K_2+2)-1}$,

3b) $\alpha(\mathbf{c}, \mathbf{c}') = \min\left\{1, \frac{M_0}{M_1} \cdot \left(\frac{1+\sum_{i=1}^{N_{3-k}} U_{3-k,i}}{1+\sum_{j=1}^{2}\sum_{i=1}^{N_j} U_{ji}}\right)^{m_{3-k}} \cdot \frac{K_{3-k}}{K_0+1}\right\},\ k = 1, 2,$

3c, i) $\alpha(\mathbf{c},\mathbf{c}') = \min\left\{1, \frac{M_0}{M_1}\cdot\left(\frac{1+\sum_{i=1}^{N_k}U_{ki}}{1+\sum_{j=1}^{2}\sum_{i=1}^{N_j}U_{ji}}\right)^{m_k}\cdot\frac{(K_1+1)(K_2+1)-1}{K_0+1}\right\},\ k=1,2,$

3c, ii) $\alpha(\mathbf{c},\mathbf{c}') = \min\left\{1, \frac{M_0}{M_1^2}\left(\frac{1+\sum_{i=1}^{N_1}U_{1i}}{1+\sum_{j=1}^{2}\sum_{i=1}^{N_j}U_{ji}}\right)^{m_1}\left(\frac{1+\sum_{i=1}^{N_2}U_{2i}}{1+\sum_{j=1}^{2}\sum_{i=1}^{N_j}U_{ji}}\right)^{m_2} a\right\}$, where

$a = \frac{1}{B(m_1,m_2)}\frac{e(m_1,m_2)}{d(m_1,m_2)}\frac{(K_1+1)(K_2+1)-1}{K_0+1}$.

4 Applications

4.1 Simulated data

The models developed in this paper were applied to a simulated data set with two groups which each contained 200 observations. The data in group 1 were generated from the distribution $0.5\mathrm{N}(1,1)+0.5\mathrm{N}(-10,1)$ and the data in group 2 were generated from the distribution $0.7\mathrm{N}(1,1)+0.3\mathrm{N}(8,1)$. We apply the Single-$\varepsilon$ model to the data

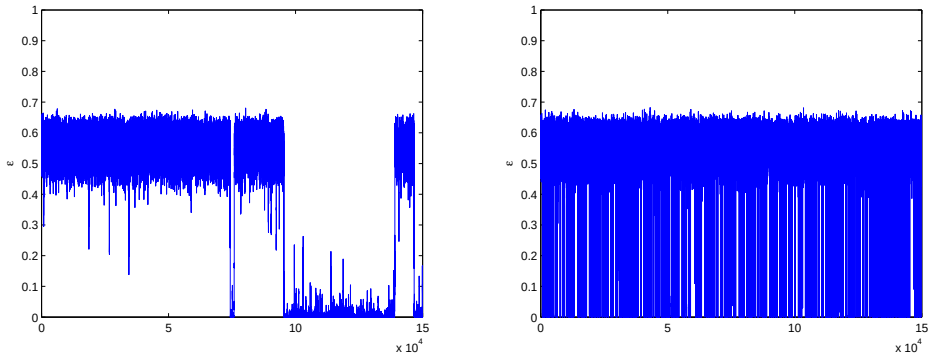

Figure 1: Trace plot for the posterior of ε with (right) and without (left) the extra split/merge step for model (4) for the simulated data set.

taking $f(Y_{ji};\theta_{ji},S)$ to be a $\mathrm{N}(\theta_{ji},S)$, together with the rest of the prior distributions in (9) (with $a_0 = b_0 = 0.5, m_0 = 0, A = 10, c = 2.1, C = 2, q = 0.01$ and $R = 0.0001$) and use the MCMC sampler with and without the split/merge step. The trace plots for the weight ε are shown in Figure 1. The posterior distribution of ε is bimodal, with modes at 0 and 0.5 (the minimum of 0.7 and 0.5, as discussed earlier). The trace plots also illustrate a possible mixing problem in the algorithm without the split/merge step. The split/merge step improves mixing of the chain by increasing the frequency of the jumps between the two modes of ε. The move has a 4.5% acceptance rate of split steps and 4.6% acceptance of merge steps. The mode at 0 is quite large in

this case. However, the mode could be much smaller with other data and the algorithm without split/merge moves might not visit the mode at zero in a reasonable number of iterations. As a result, we used the algorithm with the additional split/merge step in all subsequent analyses.

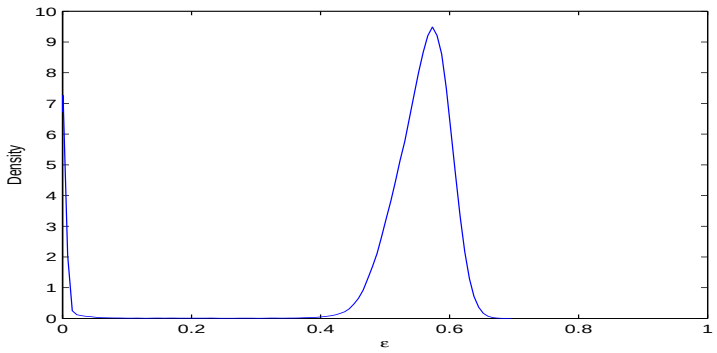

Figure 2: Posterior density of the weight ε for model (4) for the simulated data set.

The posterior density for the weight ε is shown in Figure 2, which puts most of its mass on values around 0.5 (which is the minimum of 0.7 and 0.5). For this value for ε, note that we can perfectly reproduce the distribution generating the data by taking F_0 to be a point mass at one, F_1 a point mass at -10, and F_2 to put weight 0.4 on a point mass at 1 and 0.6 on a point mass at 8. The predictive densities corresponding to the component distributions F_j (left) and the correlated distributions F_j^* (right) are shown in Figure 3. Indeed, we notice from the predictives corresponding to the components that F_0 is concentrated around the (correct) value 1, F_1 is concentrated around -10 and F_2 is bimodal, with about 40% of the mass around 1 and the rest around 8. The posterior of ε also has a smaller mode around 0. This value of ε corresponds to the case without a common part and explains the small second mode at 1 for F_1.

The predictives for group j (corresponding to $F_j^*, j = 1, 2$) closely match the distributions from which the data were generated. For comparison purposes, the latter distributions are also plotted on the same graph, using dashed lines.

The posterior mean, median and 95% credible intervals for the parameters in this model are shown in Table 1. The values of the concentration parameters M_0 and M_1 are quite small, indicating that F_0 and $F_j, j = 1, 2$ are quite far from their normal centring distribution. Indeed, it turns out from the inference on $K_j, j = 0, 1, 2$ that the number of clusters is quite small indeed, with only one for F_0 and F_1 in the median,

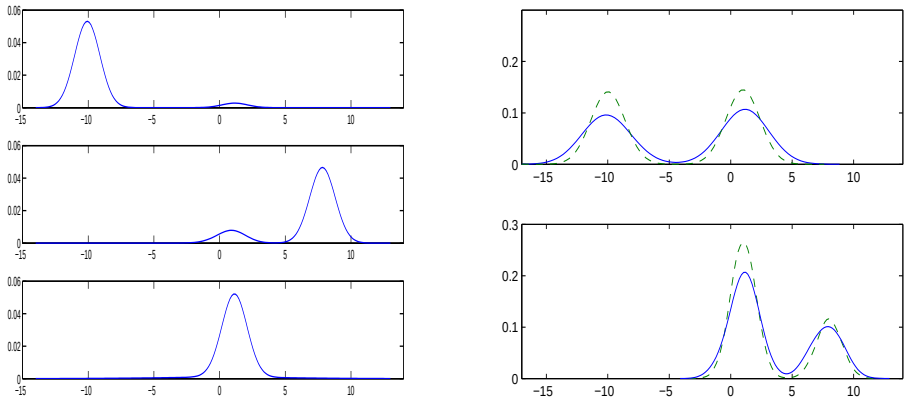

Figure 3: Predictive densities for the component distributions F_1 (top), F_2 (middle) and F_0 (bottom) (left-hand panels) and of F_1^* (top) and F_2^* (bottom) (right-hand panels) using the Single-ε model for the simulated data set. Dashed lines indicate the distributions that generated the data.

	M_0	M_1	K_0	K_1	K_2
Mean	0.199	0.221	1.215	1.319	2.022
Median	0.138	0.168	1	1	2
2.5th perc	0.010	0.013	-	-	-
97.5th perc	0.722	0.746	-	-	-

Table 1: Posterior mean, median and 95% credible intervals for the parameters in the Single-ε model for the simulated data set.

and two for F_2, as expected. So the inference corresponds quite accurately to the distribution that generated the data.

Results with the Normalisation model lead, as expected, to densities for ε_1 and ε_2 that concentrate most of their mass around 0.5 and 0.7, respectively, and the resulting predictive distributions are very similar to those found with the Single-ε model.

4.2 Hospital efficiency data

Stochastic frontier models were introduced by Aigner et al. (1977) and Meeusen and van den Broeck (1977) to model the efficiency of firms. We will consider a cost frontier for hospitals. The frontier corresponds to the minimum cost of producing a certain level of outputs, given specific input prices and represents the theoretical scenario

where a hospital is fully efficient. The observed cost is modelled by

$$Y_{ijt} = \alpha + X'_{ijt}\beta + u_{ij} + v_{ijt}, \qquad 1 \leq i \leq N_j,\ 1 \leq t \leq T,\ j = 1, 2 \tag{10}$$

where Y_{ijt} is the logarithm of cost and X_{ijt} is a vector of output levels and input prices for the i-th hospital in the j-th group in time period t, while $\alpha + X'_{ijt}\beta$ is the frontier. We have two types of error terms in (10). The first error, v_{ijt}, accounts for the uncertainty regarding the location of the frontier and is assumed to be normally distributed with mean 0 and variance σ^2. The second error term, u_{ij}, captures hospital-specific disturbances and represents the loss with respect to full efficiency. This inefficiency error can only take positive values, and is assumed to remain constant over time (the implications of relaxing the last assumption are discussed by Fernández et al., 1997). The two sets of error terms are taken to be independent of each other. The efficiency for firm i in group j is then defined as $\exp\{-u_{ij}\},\ i = 1, 2, \ldots, n, j = 1, 2$.

Understanding the effect of firm characteristics, such as management structure or regulatory framework, on the efficiency distribution is one important aim of stochastic frontier models. If the firm characteristics are discrete, the firms can be divided into groups and the problem then reduces to modelling the efficiency distribution for each group. Griffin and Steel (2004) describe a Product of Dirichlet Processes model for a Bayesian nonparametric analysis in this case. An alternative approach, based on the methods developed in this paper, uses the following model:

$$Y_{ijt} \overset{ind.}{\sim} \mathrm{N}(\alpha + X'_{ijt}\beta + u_{ij}, \sigma^2), \qquad 1 \leq i \leq N_j,\ 1 \leq t \leq T,\ j = 1, 2,$$

$$u_{ij} \sim F_j^* = \varepsilon F_0 + (1 - \varepsilon)F_j, \qquad 1 \leq i \leq N_j,\ j = 1, 2,$$

$$F_k \overset{ind.}{\sim} \mathrm{DP}(M_k, H), \qquad k = 0, 1, 2, \qquad H \sim \mathrm{Exp}(\lambda),$$

$$M_k/\eta_0 \overset{iid}{\sim} \mathrm{InvBe}(\eta, \eta), \qquad k = 0, 1, 2,$$

and

$$f(a, \beta, \sigma^2) \propto \sigma^{-2}, \quad \lambda \sim \mathrm{Exp}(-\log(r^*)).$$

The prior for λ (the inverse mean of H) is chosen so that prior predictive median efficiency is $r^\star$ (as in Griffin and Steel, 2004) and a noninformative prior for $(\alpha, \beta, \sigma^2)$ is assumed, which leads to a proper posterior distribution (as shown in Fernández et al., 1997). An inverted beta (gamma-gamma) distribution (Bernardo and Smith, 1994) for the precision parameters M_0, M_1 and M_2 (each divided by a hyperparameter η_0,

which is the prior median) was adopted, as in Griffin and Steel (2004). For the Single-ε and the Normalisation models we assume the same priors for ε and γ_0, γ_1 and γ_2 as before.

The data refer to 268 nonteaching hospitals in the U.S.A. for a period of $T = 5$ years, from 1987 to 1991, which are a subset of those analysed by Griffin and Steel (2004) and Koop et al. (1997). The same frontier as Koop et al. (1997) is used and the interested reader should consult that paper for its specification. In particular, we focussed on non-profit hospitals, which were divided into two categories according to the number of clinical workers per patient, which is termed "staff ratio": a binary variable taking the value 1 if the average (over the years) of the ratio of clinical workers per patient for a specific hospital is higher than the median of those averages of all 382 hospitals in the full sample, and 0 otherwise. This led to a sample of 141 hospitals with staff ratio of 0 (group 1) and 127 hospitals with staff ratio of 1 (group 2).

The models were fitted with $r^\star = 0.8$, $\eta = \eta_0 = 1$. The value of η_0 implies a prior median value of 1 for M_0, M_1 and M_2. The posterior distributions were simulated using the MCMC algorithm with the split/merge move.

We first considered the Single-ε model applied in this setting. The acceptance rate of the split steps in the split/merge step was around 24.0%, whereas for merge steps the corresponding rate was around 19.8%. The posterior distribution of the weight

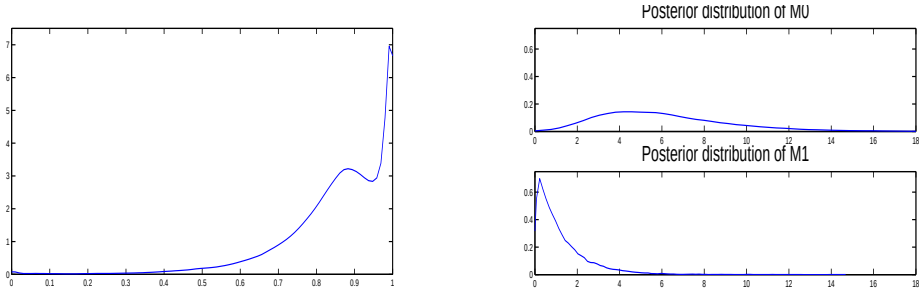

Figure 4: Posterior distribution of ε (left), M_0 (top-right) and M_1 (bottom-right) for the Single-ε model applied to the non-profit hospitals.

parameter ε is shown in the left panel of Figure 4. There is a very small mode at 0 (which corresponds to F_1^* and F_2^* not having a common part) and two larger 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 mass). The distribution illustrates the importance of the split-merge move. The mode at 0 is unlikely to be sampled without the split/merge

merge.

The right-hand side panels of Figure 4 show the posterior densities of M_0 and M_1. The posterior distribution of M_0 is flatter than the one of M_1, which is peaked below 1, indicating that F_1 and F_2 are very far from their expected centring distribution.

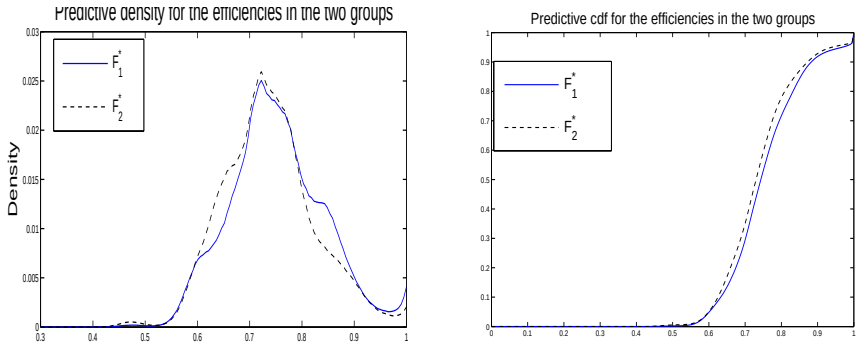

Figure 5: Predictive densities (left) and cumulative distributions (right) for the efficiency of firms in the low staff ratio (solid line) and the high staff ratio group (dashed line) for the Single-ε model applied to the non-profit hospitals.

The predictive density of the efficiency of a new firm in each of the two groups and the corresponding cumulative distribution functions (cdf) are plotted in Figure 5. The results resemble those of Griffin and Steel (2004). For group 1, there is a mode at 1, an antimode around 0.95, a "bump" around 0.86, a larger mode at 0.7 and a bump around 0.75. One difference is the mode around 0.67 of Griffin and Steel (2004), which is now transposed to the left, around 0.6, and looks more like a bump. For the second group, we have the same large mode at 0.7 and bumps around 0.67 and 0.75. In this case, there is also a tiny mode around 0.47. For this high staff ratio group, the main difference with the results in Griffin and Steel (2004) is the behaviour close to full efficiency, as 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. However, overall the results are very similar. The right graph of Figure 5 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 more or less preserved up to 0.9 or so, where the two cdf's start to coincide.

Another interesting point here is that, comparing the predictive densities corresponding to 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 some 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 6. 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 mode around 0.7 and the bumps around 0.75 and 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 bump around 0.75 is mostly (but not completely) due to F_0, whereas the bump around 0.67 is due to F_2

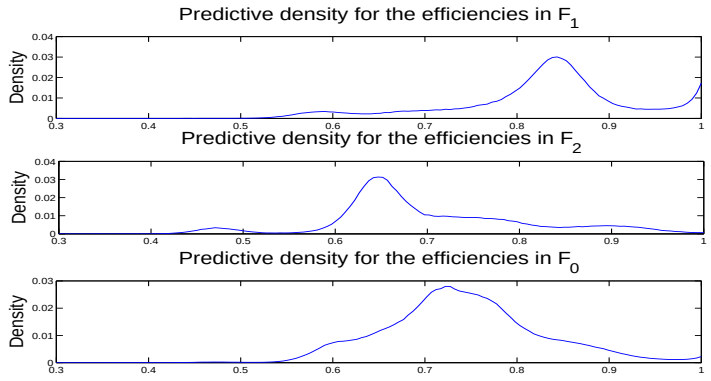

Figure 6: Predictive densities for the efficiency of a firm in F_1 (above), F_2 (centre) and F_0 (below) for the Single-ε model applied to the non-profit hospitals.

Next, we applied the Normalisation model on the same data. The acceptance rates were around 25.1% for the split steps and around 24.6% for the merge steps, and the results presented below are taken with this extra step. As a general comment, the results are very similar to the ones of the Single-ε model. The posterior distributions for the M's were similar, and so were the predictive distributions for F_1^*, F_2^* and the component distributions F_0, F_1, F_2, with the only difference worth mentioning being a larger mode at 1 for all of them. The only practically different posterior result is regarding the weights, since here we have two, instead of one in the previous model. The posterior distribution of the weights is shown in Figure 7. In both cases, the mode

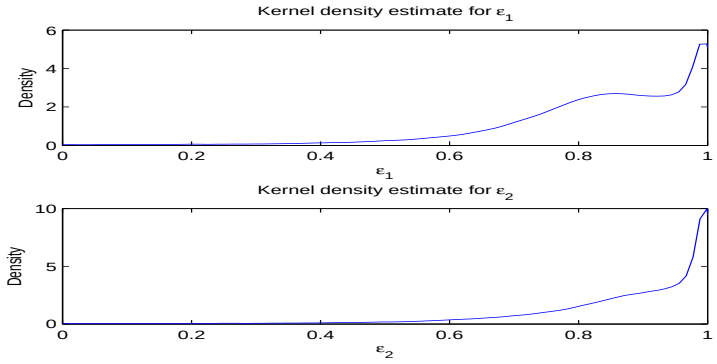

Figure 7: Posterior distribution of ε_1 (top) and ε_2 (bottom) for the Normalisation model.

at 0 is smaller than before. The largest mode at 1 is present for both ε_1 and ε_2, while for ε_1 we have another mode around 0.85.

Finally, we applied the model of Müller et al. (2004) (with a roughly comparable prior for the weight) in the same context and on the same data, leading to very similar results as with the Single-ε model.

5 Discussion

This paper discusses the use of nonparametric mixture models for two correlated distributions. Several models are developed for representing the relationship between two nonparametric distributions, inspired by normalising random measures. We also develop and discuss efficient computational methods which use a novel split/merge move to improve mixing. We concentrate on a Dirichlet process-based framework which simplifies the derivation and the methodology leads to an effective borrowing of strength between the distributions. The modelling approach could be immediately extended to more distributions by extending the representation of Lemma 1. By making similar assumptions, the models proposed in this paper could then be extended to accommodate larger numbers of groups. This is a research direction that we are currently pursueing further.

References

D. Aigner, C. A. K. Lovell, and P. Schmidt. Formulation and estimation of stochastic frontier production function models. *J. Econometrics*, 6:21–37, 1977.

J.-M. Bernardo and A. F. M. Smith. *Bayesian theory*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester, 1994.

T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1:209–230, 1973.

C. Fernández, J. Osiewalski, and M. F. J. Steel. On the use of panel data in stochastic frontier models with improper priors. *J. Econometrics*, 79:169–193, 1997.

J. E. Griffin and M. F. J. Steel. Semiparametric Bayesian inference for stochastic frontier models. *J. Econometrics*, 123:121–152, 2004.

J. E. Griffin, M. Kolossiatis, and M. F. J. Steel. Comparing distributions with dependent normalized random measure mixtures, mimeo, 2010.

H. Ishwaran and M. Zarepour. Series representations for multivariate generalized gamma processes via a scale invariance principle. *Stat. Sinica*, 19:1665–1682, 2009.

L. F. James, A. Lijoi, and I. Pruenster. Bayesian inference via classes of normalized random measures, mimeo, 2005.

G. Koop, J. Osiewalski, and M. F. J. Steel. Bayesian efficiency analysis through individual effects: Hospital cost frontiers. *J. Econometrics*, 76:77–105, 1997.

F. Leisen and A. Lijoi. Vectors of two-parameter Poisson-Dirichlet processes, mimeo, 2010.

W. Meeusen and J. van den Broeck. Efficiency estimation from Cobb-Douglas production functions with composed error. *Int. Econ. Review*, 18:435–44, 1977.

P. Müller, F. Quintana, and G. Rosner. A method for combining inference across related nonparametric Bayesian models. *J. R. Stat. Soc. Ser. B*, 66:735–749, 2004.

V. Rao and Y. W. Teh. Spatial normalized gamma processes. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1554–1562. 2009.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.*, 101(476):1566–1581, 2006.

S. G. Walker and P. Muliere. A bivariate Dirichlet process. *Stat. Probab. Letters*, 64: 1–7, 2003.

A Appendix

A.1 Proof of Lemma 1

Let $F_1^* = \sum_{i=1}^{\infty} w_{1i}^\star \delta_{\theta_{1i}^*}$, $F_2^* = \sum_{i=1}^{\infty} w_{2i}^\star \delta_{\theta_{2i}^*}$, $\theta_1^\star = (\theta_{11}^\star, \theta_{12}^\star, \theta_{13}^\star, \dots)$ and $\theta_2^\star = (\theta_{21}^\star, \theta_{22}^\star, \theta_{23}^\star, \dots)$. Let $A = \theta_1^* \cap \theta_2^*$. Then we can write

$$F_k^* = \sum_{\theta_j \in A} w_{kj}^* \delta_{\theta_j} + \sum_{\theta_{kj}^\star \in \theta_k^\star - A} w_{kj}^* \delta_{\theta_{kj}^\star}, \ k = 1, 2.$$

Let $w_j^\dagger = \min\{w_{1j}^*, w_{2j}^*\}$ for $\theta_j \in A$, then

$$F_k^* = \sum_{\theta_j \in A} w_j^\dagger \delta_{\theta_j} + \sum_{\theta_j \in A} (w_{kj}^* - w_j^\dagger) \delta_{\theta_j} + \sum_{\theta_{kj}^\star \in \theta_k^\star - A} w_{kj}^* \delta_{\theta_{kj}^\star}, \ k = 1, 2$$

and it is clear that the result follows from taking $\varepsilon = \sum_{\theta_j \in A} w_j^\dagger$,

$$\varphi_k = \frac{\sum_{\theta_j \in A} (w_{kj}^* - w_j^\dagger)}{\sum_{\theta_j \in A} (w_{kj}^* - w_j^\dagger) + \sum_{\theta_{kj}^\star \in \theta_k^\star - A} w_{kj}^*}, \qquad F_0 = \frac{\sum_{\theta_j \in A} w_j^\dagger \delta_{\theta_j}}{\sum_{\theta_j \in A} w_j^\dagger}$$

$$F_k^{(0)} = \frac{\sum_{\theta_j \in A} (w_{kj}^* - w_j^\dagger) \delta_{\theta_j}}{\sum_{\theta_j \in A} (w_{1j}^* - w_j^\dagger)}, \qquad F_k = \frac{\sum_{\theta_{kj}^\star \in \theta_k^\star - A} w_{kj}^* \delta_{\theta_{kj}^\star}}{\sum_{\theta_{kj}^\star \in \theta_k^\star - A} w_{kj}}$$

for $k = 1, 2$. □

A.2 Proof of equation (5)

Let $A \in \mathcal{F}$ and let $\stackrel{d}{=}$ denote equality in distribution. For $F_1^*(A)$, we have that:

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

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

A.3 Proof of Theorem 1

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, first calculate the covariance between the two:

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

Then, 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. □

A.4 Proof of Proposition 1

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

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

Using the independence of s_{ji} in the three components (given the indicators r_{ji}) and applying expression (6) to each of them, the EPPF for model (4) can be derived. □

A.5 Proof of Proposition 2

To derive the Pólya-urn scheme, we first derive the Chinese restaurant representation. Let $\boldsymbol{c}_{ji} = (s_{ji}, r_{ji})$ and $\boldsymbol{c}$ be the set of all $\{\boldsymbol{c}_{ji}\}$. Suppose that the new observation falls in group k, then $\boldsymbol{c}_{k,new} = (s_{k,N_k+1}, r_{k,N_k+1})$. The conditional probability formula,

$$p(\mathbf{c}_{k,new}|\boldsymbol{c},M_0,M_1) = \frac{p(\mathbf{c}_{k,new},\boldsymbol{c}|M_0,M_1)}{p(\boldsymbol{c}|M_0,M_1)}$$

and equation (7) implies that

$$P(\mathbf{c}_{k,new}=(j,i)|\boldsymbol{c},M_0,M_1) = \begin{cases} \dfrac{M_0}{M_0+M_1+N}, & j=K_0+1, \quad i=0 \\ \dfrac{n_{0,j}}{M_0+M_1+N}, & 1\le j\le K_0, \quad i=0 \\ \dfrac{M_1}{M_1+n_k}\dfrac{M_1+N-n_0}{M_0+M_1+N}, & j=K_k+1, \quad i=1 \\ \dfrac{n_{k,j}}{M_1+n_k}\dfrac{M_1+N-n_0}{M_0+M_1+N}, & 1\le j\le K_k, \quad i=1. \end{cases}$$

The Pólya-urn scheme can then be derived by adding the corresponding probabilities. □