Comparing Distributions Using Dependent Normalized Random Measure Mixtures*

J.E. Griffin

University of Kent, Canterbury, UK

M. Kolossiatis

Cyprus University of Technology, Limassol, Cyprus

M.F.J. Steel

University of Warwick, Coventry, UK

December 15, 2010

Abstract

A methodology for the simultaneous Bayesian nonparametric modelling of several distributions is developed. Our approach uses normalized random measures with independent increments and builds dependence through the superposition of shared processes. The properties of the prior are described and the modelling possibilities of this framework are explored in some detail. Efficient slice sampling methods are developed for inference. Various posterior summaries are introduced which allow better understanding of the differences between distributions. The methods are illustrated on simulated data and examples from survival analysis and stochastic frontier analysis.

Keywords: Bayesian nonparametrics; Dependent distributions; Dirichlet process; Normalized Generalized Gamma process; Slice sampling; Utility function

*Address for correspondence: M. Steel, Department of Statistics, University of Warwick, Coventry CV4 7AL, UK. E-mail: M.F.Steel@stats.warwick.ac.uk.

1 Introduction

This paper considers the nonparametric modelling of data divided into different groups and the comparison of their distributions. For example, we may observe the results of different medical treatments or the performance of firms with different management structures. Statistical analysis will often concentrate on inference about the differences in the distributions. Analysis of Variance (ANOVA) concentrates on differences between means for different groups and links these to the effects of each factor. However, differences between groups may not be well modelled by restricting attention to location. For example, if there are distinct subpopulations within the observations then each group may contain different proportions of each subpopulation and a full summary of the differences would involve identifying parts of the support on which the two distributions place substantially different masses. We follow a full Bayesian analysis by firstly placing a prior on the distributions and secondly defining a decision problem which reports where the distributions are similar or substantially different.

We use a Bayesian nonparametric mixture model approach to understand the differences between the distributions. Let $F_1, F_2, \ldots, F_q$ be the distribution of observations for q different groups, then an infinite mixture model assumes that the density for the g-th group is

$$f_g = \int k(\cdot|\theta) dG_g(\theta)$$

where $k(\cdot|\theta)$ is a density parameterized by θ and G_g is a discrete random probability measure. Since the measure is discrete, it can be represented as

$$G_g = \sum_{i=1}^{\infty} w_{g,i} \delta_{\theta_{g,i}}$$

where δ_x is the Dirac delta function that places mass 1 at x and $\theta_{g,1}, \theta_{g,2}, \ldots$ and $w_{g,1}, w_{g,2}, \ldots$ are infinite sequences of random variables for which $\sum_{i=1}^{\infty} w_{g,i} = 1$ and $w_{g,i} > 0$ for all i. It follows that the mixture model can be written as

$$\sum_{i=1}^{\infty} w_{g,i} k(\cdot|\theta_{g,i}) \tag{1}$$

or, alternatively, the model can be represented hierarchically for an observation $y_{g,j}$ drawn from F_g as follows

$$y_{g,j} \sim k(\cdot|\theta_{g,s_{g,j}}), \qquad p(s_{g,j} = i) = w_{g,i}$$

where $s_{g,j}$ is an allocation variable indicating to which component distribution $k(\cdot|\theta)$ the j-th observation in group g is allocated. The groups will often be formed by all possible combinations of some categorical covariates and we will denote those covariates by z_g for the g-th group. This is a very general model and many previously proposed models fall within it. The ANOVA-DDP model of De Iorio et al. (2004) assumes that the density k is a $\mathrm{N}(\theta, \sigma^2)$, while $w_{g,i} = w_i$ and $\theta_{g,i} = z_g^T \beta_i$ where β_i is a vector of parameters. This allows the means of the different components to change with covariates.

A popular approach allows the weights to depend on covariates and sets $\theta_{g,i} = \theta_i$ so that the location of the components is fixed across each group. A finite mixture of normals model along these lines was proposed by Rodriguez et al. (2009) who allow the component weights to depend on covariates. Alternatively, the weights can be modelled through combinations of random variables, which encourages correlation between the random distributions. The Matrix Stick-Breaking process of Dunson et al. (2008) assumes that z_g is a two-dimensional vector and that $w_{g,1}, w_{g,2}, w_{g,3}, \ldots$ are derived using a Matrix Stick-Breaking construction where

$$w_{g,j} = V_{z_{g,1},1} V_{z_{g,2},2}$$

and $V_{1,1}, V_{2,1}, V_{3,1}, \ldots$ and $V_{1,2}, V_{2,2}, V_{3,2}, \ldots$ are infinite sequences of beta random variables. Müller et al. (2004) assume that

$$f_g = \psi \sum_{i=1}^{\infty} w^{\star}_{g,i} k(\cdot|\theta^{\star}_{g,i}) + (1-\psi) \sum_{i=1}^{\infty} w_i k(\cdot|\theta_i)$$

where $0 \leq \psi \leq 1$. The distribution of the g-th group is a mixture of a common component shared by all groups and an idiosyncratic component. The parameter ψ is the weight placed on the idiosyncratic component and so affects the correlation between distributions.

The Hierarchical Dirichlet process (Teh et al., 2006) assumes, in its simplest form, that

$$G_g \sim \mathrm{DP}(MG_0), \quad g = 1, \ldots, q, \qquad G_0 \sim \mathrm{DP}(M_0 H) \tag{2}$$

The distributions are exchangeable and this structure allows clusters to be shared by different groups (due to the discrete nature of the Dirichlet process at both levels). If the Hierarchical Dirichlet process is used as the mixture distribution in the mixture models then we have something of the form of (1). Teh et al. (2006) derive the stick-breaking construction for $w_{g,1}, w_{g,1}, w_{g,2}, \ldots$. The model can be extended to more levels of hierarchy in the standard way. This model assumes that distributions are exchangeable at some level. In contrast, this paper will mostly concentrate on the problem where groups are defined by covariates. There is normally no natural nesting in these settings, so that hierarchical models will then not be appropriate.

We propose to use a normalized superposition of random measures to induce dependence. This general framework leads to dependence structures that can be fairly easily controlled through the mass parameters of the underlying measures and extends naturally to any number of groups. In fact, we can use this framework to separately model the mass shared by any subset of the groups or we can use simpler settings, depending on the flexibility of the dependence structure we want to assume. We use shrinkage priors for the mass parameters to ensure consistent priors across different levels of model complexity. For posterior inference, we propose novel slice sampling Markov chain Monte Carlo (MCMC) methods, used in combination with a split-merge move. We also discuss ways of summarizing the differences between the nonparametric distributions for each group, based on decision theoretic ideas.

The paper is organized as follows. Section 2 describes the construction of random probability measures by normalization and our proposed framework for modelling dependence using normalized random measures, Section 3 describes efficient MCMC sampling methods for inference, Section 4 discusses a decision theoretic approach to comparing distributions, Section 5 analyzes simulated data and presents real data applications to stochastic frontier analysis and health, while Section 6 concludes.

2 Introducing Dependence in Normalized Random Measures

2.1 General Framework

Normalized Random Measures with Independent Increments (NRMIs) are a class of nonparametric priors for a random probability measure, G, constructed by normalizing a positive random measure with independent increments, $\tilde{G}(B)$, to give

$$G(B) = \frac{\tilde{G}(B)}{\tilde{G}(\Omega)}.$$

Throughout the paper we will use G to represent the normalized version of a random measure $\tilde{G}$. Generally, we will concentrate on random measures which only contain jumps and write

$$\tilde{G} = \sum_{i=1}^{\infty} J_i \delta_{\theta_i},$$

where θ_i are i.i.d. from some distribution H and $J_1, J_2, J_3, \ldots$ are jumps of a Lévy process with Lévy density $\zeta(x)$. The process is well-defined if $0 < \tilde{G}(\Omega) < \infty$ almost surely which happens if $\int \zeta(x)\, dx$ is infinite. The NRMI can be employed as the prior of the mixing measure G in an infinite mixture model $f(y) = \int k(y|\theta)\, dG(\theta)$ to define an NRMI mixture. This class of processes and their use in mixture models is studied in general by James et al. (2009). Several previously proposed processes fall within this class. The Dirichlet process (Ferguson, 1973) (DP) occurs if $\tilde{G}$ is a Gamma process, for which

$$\zeta(x) = Mx^{-1}\exp\{-x\}, \qquad M > 0.$$

The Normalized Generalized Gamma process (Lijoi et al., 2007) (NGG) is constructed by normalizing a Generalized Gamma process (Brix, 1999), for which

$$\zeta(x) = \frac{M}{\Gamma(1-a)}x^{-1-a}\exp\{-\lambda x\}, \qquad M > 0, \quad 0 < a < 1, \quad \lambda \geq 0. \tag{3}$$

This process tends to the Dirichlet Process as $a \to 0$ and $\lambda = 1$. The Normalized Inverse-Gaussian process (Lijoi et al., 2005) occurs if $a = 0.5$ and $\lambda = 1$. Another special case is the Normalized Stable Process of Kingman (1975), which corresponds to $\lambda = 0$.

Dependence between two distributions G_1 and G_2 can be introduced through the unnormalized random measures $\tilde{G}_1$ and $\tilde{G}_2$. Intuitively, it is clear that the dependence between G_1 and G_2 will grow as the dependence between $\tilde{G}_1$ and $\tilde{G}_2$ grows. A similar approach for constructing processes of random probability measures over time is discussed by Griffin (2009).

Suppose that we have q groups, then the random measures can be defined in the following way. Firstly, we can define p underlying random measures $\tilde{G}^\star_1, \tilde{G}^\star_2, \ldots, \tilde{G}^\star_p$ such that

$$\tilde{G}^\star_j = \sum_{i=1}^{\infty} J_{j,i}\delta_{\theta_{j,i}}, \quad j = 1, \ldots, p,$$

where $\theta_{j,i}$ are i.i.d. from some distribution H and $J_{j,1}, J_{j,2}, J_{j,3}, \ldots$ are jumps with Lévy density $\zeta^\star_j(x)$. Defining $\tilde{G}^\star = (\tilde{G}^\star_1, \tilde{G}^\star_2, \ldots, \tilde{G}^\star_p)^T$, the random measures in the vector $\tilde{G} = (\tilde{G}_1, \tilde{G}_2, \ldots, \tilde{G}_q)^T$ will be formed as

$$\tilde{G} = D\tilde{G}^\star,$$

where D is a $q \times p$-dimensional selection matrix. Then $\tilde{G}_j$ is a Lévy process and the Lévy density of $\tilde{G}_j$ is $\zeta_j(x) = D_{j\cdot}\zeta^\star(x)$ where $D_{j\cdot}$ is the j-th row of D and $\zeta^\star(x) = (\zeta^\star_1(x), \ldots, \zeta^\star_p(x))^T$. In particular, we take $\zeta^\star_h(x) = M_h\eta(x)$ so that $\zeta_j(x) = [D_{j\cdot}M]\eta(x)$ where $M = (M_1, \ldots, M_p)^T$. When we normalize, we obtain

$$G = WG^\star, \tag{4}$$

where $G = (G_1, \ldots, G_q)^T$, $G^\star = (G^\star_1, \ldots, G^\star_p)^T$ and W is a $q \times p$ matrix with elements

$$W_{ij} = \frac{D_{ij}\tilde{G}^\star_j(\Omega)}{\sum_{k=1}^{p} D_{ik}\tilde{G}^\star_k(\Omega)} \text{ and } G^\star_j = \frac{\tilde{G}^\star_j}{\tilde{G}^\star_j(\Omega)}.$$

Therefore, the distribution for each group is a mixture of $G_1^\star, G_2^\star, \ldots, G_p^\star$ where the weights for the i-th group are given by the i-th row of W. This process will be denoted generally as a Correlated Normalized Random Measure with Independent Increments, or CNRMI$(M, H, D; \eta)$. Often, we will choose a specific functional form for η so that the marginal processes $G_1, \ldots, G_q$ come from a known process (for example, a Dirichlet process). We will consider two possibilities: a Correlated Dirichlet Process CDP(M, H, D) where $\eta(x) = x^{-1}\exp\{-x\}$ and the marginal processes are DP and a Correlated Normalized Generalized Gamma Process CNGG$(M, H, D; a, \lambda)$ where $\eta(x) = x^{-1-a}\exp\{-\lambda x\}$ and the marginal processes are NGG. The mixture form for $G_1, G_2, \ldots, G_q$ is an important difference to the Hierarchical Dirichlet process, which is a framework that leads to all atoms being shared by all distributions and assumes that all distributions are *a priori* equally correlated.

If we use $G_1, G_2, \ldots, G_q$ as mixing measures for q mixture models, the distribution of an observation, y, in the i-th group is now given by

$$f_i(y) = \int k(y|\theta)\, dG_i(\theta).$$

Then we can write

$$f_i = \frac{\tilde{f}_i}{\tilde{F}_i(\Omega)},$$

where $\tilde{f}_i(y) = \int k(y|\theta) d\tilde{G}_i(\theta)$ and $\tilde{F}_i(A) = \int_A \tilde{f}_i(y)\, dy$. Now, $\tilde{F}_i$ expresses an unnormalized distribution in terms of basis functions (where the kernel $k(\cdot)$ are the basis functions) and so F_i is a normalized basis function model.

A natural measure of the dependence between two distributions is the correlation between $G_i(B)$ and $G_j(B)$ where B is a measurable set. Using the construction in this paper, this correlation does not depend on B and so can be used as a single measure of dependence between distributions, which we denote by Corr(G_i, G_j). The following results present an expression for the correlation, using a particular form of the framework described above for $q = 2$, $p = 3$ and $D = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$. This is a simple, yet illustrative example.

Theorem 1 *Suppose that $\tilde{G}_1 = \tilde{G}_1^\star + \tilde{G}_2^\star$ and $\tilde{G}_2 = \tilde{G}_1^\star + \tilde{G}_3^\star$ where the Lévy measure of $\tilde{G}_k^\star$ is $M_k\eta(x)$. Define*

$$L_\eta(v) = \int_0^\infty (1 - \exp\{-vx\})\eta(x)\, dx.$$

The covariance of G_1 and G_2 is

$$\mathit{Cov}(G_1(B), G_2(B)) = H(B)(1 - H(B))M_1 \int_0^\infty \int_0^\infty \beta(v_1, v_2; M_1, M_2, M_3)\, dv_1 dv_2,$$

where

$$\beta(v_1, v_2; M_1, M_2, M_3) = -L''_\eta(v_1+v_2)\exp\left\{-M_1 L_\eta(v_1+v_2) - M_2 L_\eta(v_1) - M_3 L_\eta(v_2)\right\}.$$

Proof: See Appendix

Similarly, expressions can be derived for Var$(G_1(B))$ and Var$(G_2(B))$ and so

$$\rho = \text{Corr}(G_1, G_2) = \frac{M_1 \int_0^\infty \int_0^\infty \beta(v_1, v_2; M_1, M_2, M_3)\, dv_1 dv_2}{\sqrt{(M_1+M_2)(M_1+M_3)\beta^*(M_1+M_2)\beta^*(M_1+M_3)}},$$

where

$$\beta^*(M) = \int_0^\infty \int_0^\infty -L''_\eta(v_1+v_2)\exp\left\{-M L_\eta(v_1+v_2)\right\}\, dv_1 dv_2.$$

In the special case where $M_1 = M\rho^\star$ and $M_2 = M_3 = M(1-\rho^\star)$ for $0 < \rho^\star < 1$, we obtain

$$\rho = \rho^\star\left[1+\epsilon\right],$$

where

$$\epsilon = \frac{\int_0^\infty \int_0^\infty -L''_\eta(v_1+v_2)\exp\left\{-M L_\eta(v_1+v_2)\right\}\gamma(v_1,v_2)\, dv_1 dv_2}{\beta^*(M)},$$

with

$$\gamma(v_1, v_2) = \exp\left\{-M(1-\rho^\star)\left[L_\eta(v_1) + L_\eta(v_2) - L_\eta(v_1+v_2)\right]\right\} - 1.$$

Therefore, the correlation between G_1 and G_2, ρ, can be well-approximated by $\rho^\star$ if $\gamma(v_1, v_2)$ is close to zero for all $\rho^\star$ which will be the case for many forms of processes. It is important to point out we do not necessarily advocate adopting the restricted parametrization for M_1, M_2 and M_3 in the special case used above, but it is a useful device to better understand the properties of our models, as illustrated in the following examples:

Dirichlet process marginals (CDP)

Here

$$L_\eta(v) = \log(1+v), \qquad L''_\eta(v) = -\frac{1}{(1+v)^2},$$

so that

$$\gamma(v_1, v_2) = \exp\left\{-M(1-\rho^\star)\log\left(1 + \frac{v_1 v_2}{1+v_1+v_2}\right)\right\} - 1.$$

Figure 1 plots the correlation between G_1 and G_2 as a function of $\rho^\star$ and illustrates that $\rho^\star$ closely approximates the correlation, especially for larger M.

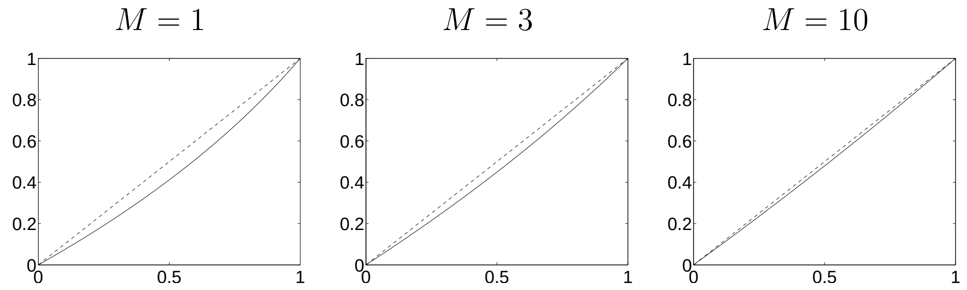

Figure 1: Plot of the actual correlation, ρ, (solid line) and $\rho^\star$ (dashed line) against $\rho^\star$ for the CDP.

Normalized Generalized Gamma process marginals (CNGG)

In this case

$$L_\eta(v) = \frac{1}{a}\left((v+\lambda)^a - \lambda^a\right), \qquad L''_\eta(v) = (a-1)(v+\lambda)^{a-2},$$

which implies that

$$\gamma(v_1, v_2) = \exp\left\{-M(1-\rho^\star)\frac{1}{a}\left[(v_1+\lambda)^a + (v_2+\lambda)^a - (v_1+v_2+\lambda)^a - \lambda^a\right]\right\} - 1.$$

Figure 2 shows the relationship between $\rho^\star$ and the actual correlation, ρ, for CNGG processes with different choices of the parameters. The correlation is close to $\rho^\star$ for each choice of the hyperparameters with the largest differences for the smaller values of a and λ. For general M_1, M_2 and M_3, this results suggests that increasing M_1 relative to M_2 and M_3 leads to a larger correlation between G_1 and G_2.

When $q > 2$, we can always write a pair of unnormalized distribution $\tilde{G}_j$ and $\tilde{G}_k$, where $j \neq k$, as

$$\tilde{G}_j = \tilde{G}^{(c)} + \tilde{G}^{(j)}$$
$$\tilde{G}_k = \tilde{G}^{(c)} + \tilde{G}^{(k)},$$

where, using $\mathrm{I}(\cdot)$ to denote the indicator function, the Lévy measure of $\tilde{G}^{(c)}$ is given by $\left[\sum_{m=1}^{p} \mathrm{I}(D_{jm}=1, D_{km}=1)M_m\right]\eta(x)$, $\tilde{G}^{(j)}$ has Lévy measure $\left[\sum_{m=1}^{p} \mathrm{I}(D_{jm}=1, D_{km}=0)M_m\right]\eta(x)$ and $\tilde{G}^{(k)}$ has Lévy measure $\left[\sum_{m=1}^{p} \mathrm{I}(D_{jm}=0, D_{km}=1)M_m\right]\eta(x)$. This suggest using the general approximation

$$\mathrm{Corr}(G_j, G_k) \approx \frac{M^{(c)}}{\sqrt{M^{(c)}+M^{(j)}}\sqrt{M^{(c)}+M^{(k)}}}, \qquad (5)$$

where $M^{(c)} = \sum_{m=1}^{p} \mathrm{I}(D_{jm}=1, D_{km}=1)M_m$, $M^{(j)} = \sum_{m=1}^{p} \mathrm{I}(D_{jm}=1, D_{km}=0)M_m$, and $M^{(k)} = \sum_{m=1}^{p} \mathrm{I}(D_{jm}=0, D_{km}=1)M_m$. Therefore, the correlation between

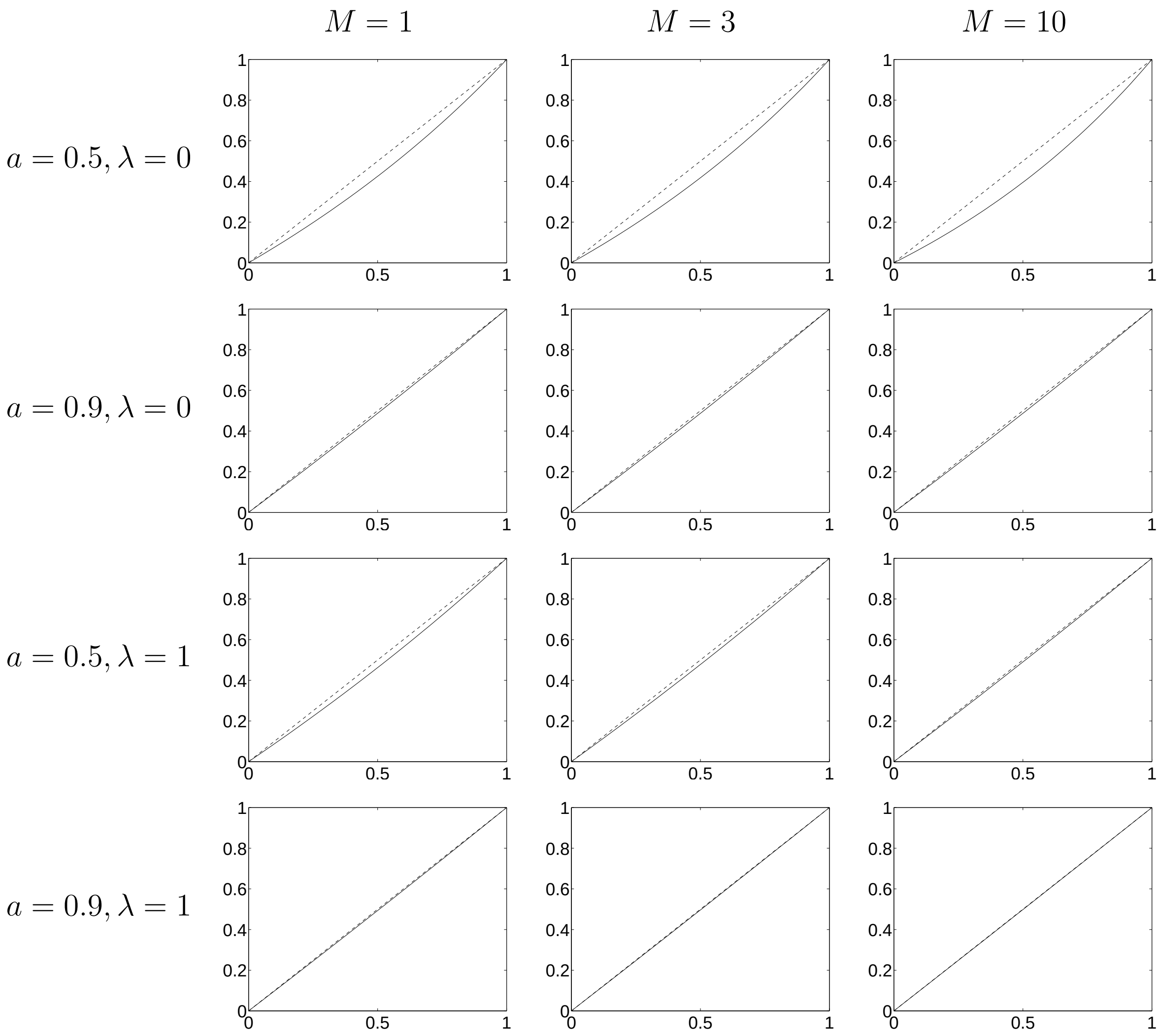

Figure 2: Plot of the actual correlation, ρ, (solid line) and $\rho^\star$ (dashed line) against $\rho^\star$ for the CNGG.

G_j and G_k increases as the value of $M^{(c)}$ increases relative to $M^{(j)}$ and $M^{(k)}$. Generally, increasing M_h leads to increased correlations between all distributions with a 1 in the h-th column of D.

2.2 Modelling of Groups

In the simple case with 2 groups, there are naturally three underlying random measures $\tilde{G}^\star_j$ in our model, one modelling the common mass shared between the groups and two for the idiosyncratic components. In cases with more groups, we need to make modelling decisions, more fully explored in this subsection. The most flexible models in our class are generated by allocating a separate random measure for modelling the mass shared by each nonempty subset of group distributions. The most complete model for q groups in the CRNMI class with a given M, H and η can thus be defined by taking $p = 2^q - 1$ and letting the i-th column of D be the binary representation of i for $1 \leq i \leq 2^q - 1$. For example if $q = 3$, then

$$D = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

where $\tilde{G}_1^\star$, $\tilde{G}_2^\star$ and $\tilde{G}_4^\star$ are idiosyncratic components, $\tilde{G}_3^\star$, $\tilde{G}_5^\star$ and $\tilde{G}_6^\star$ are shared by two groups and $\tilde{G}_7^\star$ is shared by all three groups. This will be called the saturated model. The levels of correlation between the distributions can be accommodated by choosing appropriate values of $M_1, \ldots, M_p$ and using (5). Clearly this model becomes increasingly complicated as q increases. More parsimonious models can be constructed by removing columns of D from the saturated model (which is equivalent to setting some M_h to zero). A version of the model introduced by Müller et al. (2004) would use the $q \times (q+1)$-dimensional matrix

$$D = \left(\mathbf{1}_q \quad I_q\right),$$

where $\mathbf{1}_q$ is a q-dimensional vector of ones (representing the single common component) and I_q is the $q \times q$-dimensional identity matrix. Alternatively, if the distributions relate to observations at different times then a simple model could be defined using

$$D = \left(\mathbf{1}_q \quad I_q \quad F\right),$$

where F is a $q \times (q-1)$-dimensional matrix for which

$$F_{ij} = \begin{cases} 1 & \text{if } j = i \text{ or } j = i - 1 \\ 0 & \text{otherwise} \end{cases}.$$

The model then includes a common underlying measure (in the first column), idiosyncratic underlying measures (in the next q columns) and underlying random measures shared by consecutive distributions (in the next $q - 1$ columns). More specific form of problem-specific dependence could also be modelled. Suppose that we take observations from three distributions where we think that distribution 1 and 2 are more related to each other than to distribution 3. A suitable model would be

$$D = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

where the inclusion of the final column allows extra dependence between the first two distributions.

In practice, we may not have prior information that leads us to consider models simpler than the saturated model. We suggest using regularization to avoid overfitting (since the number of underlying processes, p, grows quickly with q). A standard approach would be Bayesian variable selection on the columns of D for the saturated model. This is equivalent to setting $M_h = 0$ in the Lévy measure of the underlying measure $\tilde{G}^{\star}_{h}$. We will take an alternative approach and define a prior for M_h which encourages substantial shrinkage towards zero (this is similar to the shrinkage prior approach to regression as described by Scott and Polson (2011) and Griffin and Brown (2010)). The prior for $M_1, M_2, \ldots, M_p$ is chosen in the following way. The values of $M_1, M_2, \ldots, M_p$ control the dependence between distributions and can be chosen to represent prior beliefs. The additive effect of the M_j's is useful here. Suppose that we have one distribution with M chosen to take the value M^*. Moving to two distributions in the saturated model suggests that $M_1 + M_3 = M^*$ and $M_2 + M_3 = M^*$, if we assume that G_1 and G_2 are exchangeable, so that we have $M_1 = M_2$. If we are indifferent between an observation being allocated to a shared cluster or an idiosyncratic cluster then $M_1 = M_2 = M_3$. Repeated use of this argument allows extension to any value of q and suggests that $M_1, M_2, \ldots, M_p$ are independent and $M_i \sim \text{Ga}(M^*/2^{q-1}, \beta)$. Figure 3 shows the prior distribution induced on ρ, the correlation coefficient between G_1 and G_2 when $q = 2$, for various values of M^*. All priors are centred around $1/2$ with the variability decreasing as M^* increases. We will use $M^* = 1$ in our applications. This prior is relatively flat for correlations larger than 0.1 and has larger mass close to zero. This will lead to some shrinkage of small correlations.

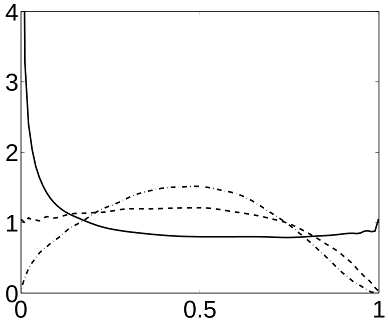

Figure 3: The prior on $\rho = \text{Corr}(G_1, G_2)$ with $M_i \sim \text{Ga}(M^*/2, \beta)$ where $M^* = 1$ (solid line), $M^* = 2$ (dashed line) and $M^* = 3$ (dot-dashed line) and $\beta = 1$.

3 Computational Methods

This section describes an MCMC sampler for fitting the general mixture model

$$y_{g,i} \sim k(y_{g,i}|\theta_{g,i}), \qquad i = 1, 2, \dots, n_g$$

$$\theta_{g,i} \sim G_g$$

$$G_1, G_2, \dots, G_q \sim \text{CNRMI}(M, H, D; \eta)$$

where M is given the prior described in Section 2 and H and η potentially have hyperparameters which also have priors.

Several slice sampling algorithms for normalized random measure mixture models were introduced by Griffin and Walker (2010). We will extend their "Slice 1" algorithm. For a single normalized random measure mixture the posterior is proportional to

$$p(J)p(\theta)\prod_{i=1}^{n} w_{s_i} k\left(y_i|\theta_{s_i}\right)$$

where $w_i = J_i/\sum_{l=1}^{\infty} J_l$, $J = (J_1, J_2, J_3, \dots)$ and $\theta = (\theta_1, \theta_2, \theta_3, \dots)$. They demonstrate that the following posterior with additional auxiliary variables $u_1, u_2, \dots, u_n$ and $v_1, v_2, \dots, v_n$ and integrating over all jumps smaller than $L = \min\{u_i\}$ is a much simpler form for computational purposes:

$$p(J_1, J_2, \dots, J_K)p(\theta)\prod_{i=1}^{n} \mathrm{I}(u_i < J_{s_i}) \exp\left\{-v_i \sum_{l=1}^{K} J_l\right\} \mathrm{E}\left[\exp\left\{-v_i \sum_{l=K+1}^{\infty} J_l\right\}\right] k\left(y_i|\theta_{s_i}\right)$$

where $J_1 > J_2 > J_3 > ... > J_K > L$, $u_1, u_2, \dots, u_n > 0$ and $v_1, v_2, \dots, v_n > 0$. The expectation can be evaluated using the Lévy-Khintchine formula and so

$$\mathrm{E}\left[\exp\left\{-v \sum_{i=K+1}^{\infty} J_i\right\}\right] = \exp\left\{-M \int_0^L (1 - \exp\{-vx\})\eta(x)\,dx\right\}.$$

The integral in the exponential is sometime available in terms of special function (this is the case for the Dirichlet process) or can be evaluated using standard quadrature methods.

The likelihood for a mixture model using the weights in Section 2 can be expressed in a suitable form for computation by introducing latent variables $\{s_{j,i}\}_{j=1:q,i=1:n_j}$ which are allocation variables for mixture components while $\{r_{j,i}\}_{j=1:q,i=1:n_j}$ allocates each observation to one of p underlying random measures $\tilde{G}_1^\star, \tilde{G}_2^\star, \dots, \tilde{G}_p^\star$. Thus, the observation $y_{g,i}$ is assumed to be drawn from $k(\cdot|\theta_{r_{g,i},s_{g,i}})$. Using auxiliary variables $u_{j,1},\ldots,u_{j,n_j}$ and $v_{j,1},\ldots,v_{j,n_j}$ for group j, the likelihood can now be expressed as

$$p(\theta)\prod_{i=1}^{q}V_i^{n_i-1}\prod_{i=1}^{p}p(J_{i,1},J_{i,2},\ldots,J_{i,K_i})\prod_{j=1}^{q}\prod_{i=1}^{n_j}\mathrm{I}\left(u_{j,i}<J_{r_{j,i},s_{j,i}}\right)\exp\left\{-V^TDJ^{(+)}\right\}$$
$$\times\,\mathrm{E}\left[\exp\left\{-V^TDJ^{(\infty)}\right\}\right], \tag{6}$$

where $J_{i,1}>J_{i,2}>J_{i,3}>\ldots$ for all i, $\theta=\{\theta_{i,j}\}_{i=1:p,j=1,2,3,\ldots}$ and $J_{i,1},\ldots,J_{i,K_i}$ are all jumps in process $\tilde{G}_i^\star$ which are larger than L. V is a q-dimensional vector where $V_j=\sum_{i=1}^{n_j}v_{j,i}$, $J^{(+)}$ is a p-dimensional vector with $J_i^{(+)}=\sum_{l=1}^{K_i}J_{i,l}$ and $J^{(\infty)}$ is a p-dimensional vector where $J_i^{(\infty)}=\sum_{l=K_i+1}^{\infty}J_{i,l}$. Integrating out $u_{j,i}$ from (6), the likelihood can be expressed as

$$p(\theta)\prod_{i=1}^{q}V_i^{n_i-1}\prod_{i=1}^{p}p(J_{i,1},J_{i,2},\ldots,J_{i,K_i})\prod_{j=1}^{p}\prod_{i=1}^{K_j}J_{j,i}^{n_{j,i}}\exp\left\{-V^TDJ^{(+)}\right\}$$
$$\times\,\mathrm{E}\left[\exp\left\{-V^TDJ^{(\infty)}\right\}\right],$$

where $n_{j,i}=\#\{(l,k)|s_{l,k}=i$ and $r_{l,k}=j,\ 1\leq k\leq n_l,\ 1\leq l\leq q\}$ is the size of the cluster of observations associated with $\theta_{j,i}$.

Each expectation in the product can be evaluated using the Lévy-Khintchine formula and so

$$\mathrm{E}\left[\exp\left\{-V^TDJ^{(\infty)}\right\}\right]=\exp\left\{-\mathbf{1}_p^T\tilde{M}\tilde{E}\right\},$$

where $\tilde{M}$ is a $p\times p$-diagonal matrix with $\tilde{M}_{hh}=M_h$ and, defining $D_{\cdot i}$ as the i-th column of D, $\tilde{E}$ is the p-dimensional vector with i-th element

$$\tilde{E}_i=\int_0^L\left(1-\exp\left\{-V^TD_{\cdot i}x\right\}\right)\eta(x)\,dx.$$

Therefore the posterior retains a lot of the linearity introduced in the model. The chain can be initialized in the following way. Choose a starting truncation point L and generate p different Poisson processes where the number of jumps, K_j, in the j-th process is simulated from a Poisson distribution with mean $M_j\int_L^\infty\eta(x)\,\mathrm{d}x$. The jumps of the j-th process are simulated by first drawing K_j random numbers $\{\xi_{j,k}\}_{k=1}^{K_j}$ from a uniform distribution between 0 and $\int_L^\infty\eta(x)\,\mathrm{d}x$ and ordered so that $\xi_{j,1}<\xi_{j,2}<\ldots<\xi_{j,K_j}$ and then setting $J_{j,k}=Q^{-1}(\xi_{j,k})$ for $k=1,2,\ldots,K_j$ where Q^{-1} is the inverse of $Q(x)=\int_x^\infty\eta(y)\,dy$. The locations $\theta_{j,1},\theta_{j,2},\ldots,\theta_{j,K_j}$ are taken to be i.i.d. from H and the latent variables $r_{j,i}$ and $s_{j,i}$ can be simulated from the discrete distributions

$$p(r_{j,i}=k)=\frac{D_{jk}M_k}{\sum_{l=1}^{p}D_{jl}M_l},\qquad k=1,2,\ldots,p$$

and

$$p(s_{j,i} = k) \propto k\left(y_i|\theta_{r_{j,i},k}\right) J_{r_{j,i},k}, \qquad 1 \le k \le K_{r_{j,i}}.$$

The slice latent variables can be taken as $u_{j,i} \sim \mathrm{U}(L, J_{r_{j,i},s_{j,i}})$.

In the following steps, we define $J_j^* = \{J_{j,i}|n_{j,i} \neq 0\}$, *i.e.* the jumps in the j-th component process which have observations allocated to them. The steps of the Gibbs sampler are as follows:

Step 1: Split-Merge move

The problem of multi-modality of the posterior distribution in these models and a computational solution, a split-merge move, are described in Kolossiatis et al. (2010). In our model, it is useful to link the underlying measures to their corresponding columns in the D matrix. For example, in the saturated model with $q = 3$ described at the start of Subsection 2.2, the underlying random measure $\tilde{G}_1^\star$ will be referred to as the "underlying random measure $(0, 0, 1)$". The split-merge move is performed in the following way. A split move is selected with probability 1/2, otherwise a merge move is proposed. An underlying random measure $\mathbf{e}$, a column of D, is selected at random from those underlying random measures which have observation allocated to them and a non-empty mixture component, $i^\star$, from $\mathbf{e}$ is selected uniformly at random. If the split move is selected, the members of the cluster are divided according to their group membership into two clusters $\mathbf{e}_1$ and $\mathbf{e}_2$. For example, in the saturated model with $q = 3$, if we choose $\mathbf{e} = (1, 1, 0)$ the cluster would be split into a cluster in the underlying measure $(1, 0, 0)$ and a cluster in the underlying measure $(0, 1, 0)$. In this case, there is only one possible split. However, if we choose $\mathbf{e} = (1, 1, 1)$, there are three possible splits: clusters in $(1, 0, 0)$ and $(0, 1, 1)$, clusters in $(0, 1, 0)$ and $(1, 0, 1)$, or clusters in $(0, 0, 1)$ and $(1, 1, 0)$. The particular split is chosen uniformly at random from all possible splits. The merge move performs the opposite operation. For this move, a set of allowable underlying measures is defined, $\mathcal{C} = \{\mathbf{e}^\star|\mathbf{e}_j^\star = 0 \text{ for all } j \text{ for which } \mathbf{e}_j = 1\}$, and an underlying measure $\mathbf{e}^\dagger$ is chosen uniformly at random from $\mathcal{C}$ (this happens regardless of whether there are any non-empty clusters allocated to that measure). One of the non-empty clusters, $j^\star$, in $\mathbf{e}^\dagger$ (if any exist) or a "null" cluster is chosen at random with equal probability. A new cluster is then formed in the underlying random measure $\mathbf{e}^{comb}$ where $\mathbf{e}_i^{comb} = 1$ if $\mathbf{e}_i^\star = 1$ or $\mathbf{e}_i^\dagger = 1$ and $\mathbf{e}_i^{comb} = 0$ if $\mathbf{e}_i^\star = 0$ and $\mathbf{e}_i^\dagger = 0$. If a "null" cluster were selected then the cluster $i^\star$ is moved from the underlying measure $\mathbf{e}^\star$ to $\mathbf{e}^{comb}$. Otherwise, clusters $i^\star$ and $j^\star$ are combined to define a new cluster in $\mathbf{e}^{comb}$. Let (s, r) and (s', r') be the values of the latent allocation variables before and after making the move respectively. We assume that $M_j \sim \text{Ga}(a_j, b_j)$. The acceptance probability is calculated integrating out the jumps and for the split move has the form

$$\max\left\{1, \frac{p(y|s',r')p(s',r')}{p(y|s,r)p(s,r)} \frac{D_1 K^*(\mathbf{e}) S(\mathbf{e})}{2D_1' K^{*\prime}(\mathbf{e}_1) M(\mathbf{e}_1)(K^{*\prime}(\mathbf{e}_2)+1)}\right\}$$

where D_1 and D_1' are the number of underlying random measures with observations allocated to them before and after the move respectively, $K^*(\mathbf{e})$ and $K^{*\prime}(\mathbf{e})$ are the number of non-empty clusters in the random measure $\mathbf{e}$ before and after the move, $M(\mathbf{e})$ is the size of $\mathcal{C}$ and $S(\mathbf{e})$ is the number of pairs of underlying measures that can be formed by splitting $\mathbf{e}$. In addition, we can write

$$p(s,r) = \prod_{j=1}^{p} \frac{\Gamma(a_j + K_j^*)}{(b_j + \tilde{A}_j)^{a_j + K_j^*}} \prod_{\{i|n_{j,i}\neq 0\}} \int J_{j,i}^{n_{j,i}} \exp\left\{-J_{j,i} V^T D_{\cdot i}\right\} \eta(J_{j,i})\, dJ_{j,i},$$

where for $j = 1, 2, \ldots, p$ we have $K_j^* = \#\{k : n_{j,k} > 0\}$ and

$$\tilde{A}_j = \int_0^\infty \left(1 - \exp\left\{-V^T D_{\cdot j} x\right\}\right) \eta(x)\, dx.$$

The move is completed by sampling M from its full conditional distribution and then sampling u, K and J.

Step 2: Updating V

Defining $\tilde{A} = (\tilde{A}_1, \ldots, \tilde{A}_p)^T$, the full conditional distribution of V_j is proportional to

$$V_j^{n_j - 1} \prod_{l=1}^{K_j} \int J_{j,l}^{n_{j,l}} \exp\left\{-J_{j,l} V^T D_{\cdot j}\right\} dJ_{j,l} \exp\left\{-\mathbf{1}_p^T \tilde{M} \tilde{A}\right\}, \qquad V_j > 0.$$

The parameter can be updated using a Metropolis–Hastings random walk on the log-scale. We also found it useful to update $V^\star = \sum_{j=1}^{p} V_j$ conditional on $B = (b_1, \ldots, b_p)$ where $b_j = V_j / V^\star$. The full conditional distribution of $V^\star > 0$ is

$$V^{\star n-1} \prod_{j=1}^{p} \prod_{l=1}^{K_j} \int J_{j,l}^{n_{j,l}} \exp\left\{-J_{j,l} V^\star B^T D_{\cdot j}\right\} dJ_{j,l} \exp\left\{-\mathbf{1}_p^T \tilde{M} \tilde{A}\right\}.$$

The parameter can be updated using a Metropolis–Hastings random walk on the log-scale. If $V^{\star\prime}$ is accepted then each V_j is updated to $V_j V^{\star\prime} / V^\star$.

Step 3: Updating M

The full conditional distribution of M_j is proportional to

$$p(M_j)\, M_j^{K_j} \exp\left\{-M_j \int_0^\infty (1-\exp\{-V_j J\})\eta(J)\, dJ\right\}$$

and if $p(M_j) \sim \text{Ga}(a_j, b_j)$ then the full conditional distribution is $\text{Ga}(a_j + K_j, b_j + \int_0^\infty (1 - \exp\{-V_j J\})\eta(J)\, dJ)$.

Step 4: Updating u, K and J

This set of full conditional distributions can be updated using the efficient slice sampling method of Kalli et al. (2011) by integrating out $u = \{u_{j,i}\}_{j=1:q,i=1:n_j}$ when updating the jumps. The update is described for NRMI mixtures by Griffin and Walker (2010) and can be simply extended to our model. The elements of $J_1^*, J_2^*, \ldots, J_p^*$ are simulated first followed by the elements of u (which only depends on J_k through the elements of J_k^*) and finally the other $J_{k,l} > L$. The full conditional distribution of the element $J_{k,l} \in J_k^*$ is proportional to

$$J_{k,l}^{n_{k,l}} \exp\left\{-J_{k,l} V^T D_{\cdot k}\right\} \eta(J_{k,l}), \qquad J_{k,l} > 0.$$

The full conditional of $u_{j,i}$ is $\text{U}\left(0, J_{r_{j,i},s_{j,i}}\right)$ and this allows us to calculate $L = \min\{u_{j,i}\}$. Finally, the jumps for which $J_{k,l} > L$ and $n_{k,l} = 0$ can be simulated as realizations of k inhomogeneous Poisson processes with intensities $M_k \exp\{-V^T D_{\cdot k} x\}\eta(x)$ on (L, ∞) and associating a θ drawn from H with each point of the realisation. Details of simulating from these Poisson processes are given in Griffin and Walker (2010).

Step 5: Updating θ

The elements of θ are independent under their joint full conditional distribution, and the density of θ_k is proportional to

$$h(\theta_{l,k}) \prod_{\{(j,i)|s_{j,i}=k \text{ and } r_{j,i}=l\}} k(y_{j,i}|\theta_{l,k}),$$

where $h(\cdot)$ is the density of H. This is a familiar form used in samplers for many infinite mixture models, such as DP mixtures.

Step 6: Updating s and r

The latent variables $s_{j,i}$ and $r_{j,i}$ can be updated jointly and drawn from their full conditional distribution

$$p(s_{j,i} = k \text{ and } r_{j,i} = l) \propto D_{jl}\, \mathrm{I}(J_{l,k} > u_{j,i})\, k(y_{j,i}|\theta_{l,k}),$$

where $\{(l,k) : J_{l,k} > u_{j,i}\}$ is a finite set.

3.1 Specific examples

3.1.1 Dirichlet process marginals (CDP)

The DP has the Lévy density with

$$\eta(x) = x^{-1} \exp\{-x\}.$$

Then the full conditional distribution of $J_{j,i}$ is Ga$(n_{j,i}, (1 + V^T D_{\cdot j}))$ and

$$\int J_{j,i}^{n_{j,i}} \exp\left\{-J_{j,i} V^T D_{\cdot j}\right\} \eta(J_{j,i})\, dJ_{j,i} = \frac{\Gamma(n_{j,i})}{(1 + V^T D_{\cdot j})^{n_{j,i}}}.$$

3.1.2 Normalized Generalized Gamma process marginals (CNGG)

The Lévy density with

$$\eta(x) = \frac{1}{\Gamma(1-a)} x^{-1-a} \exp\{-\lambda x\}$$

leads to

$$\int J_{j,i}^{n_{j,i}} \exp\left\{-J_{j,i} V^T D_{\cdot j}\right\} \eta(J_{j,i})\, dJ_{j,i} = \frac{1}{\Gamma(1-a)} \frac{\Gamma(n_{j,i} - a)}{(\lambda + V^T D_{\cdot j})^{(n_{j,i} - a)}}.$$

4 Comparing Distributions

Once we have a posterior distribution on the distributions $G_1, G_2, \ldots, G_q$, it is useful to have some graphical summaries which help us to understand the differences between distributions. Most simply, we can write

$$G_i = \bar{G} + \Pi_i,$$

where $\bar{G} = \frac{1}{q} \sum_{j=1}^{q} G_j$ is a "grand mean" distribution and $\Pi_i = G_i - \bar{G}$ is a signed measure which gives measure zero to Ω and which represents the difference of each distribution from the grand mean. Their densities will be represented by $\bar{g}$ and π_i, respectively. This idea is similar to the modelling of continuous responses in a one-way ANOVA model. Analogies to higher order ANOVA models are also possible. Suppose that the groups are defined by two covariates (x_1 and x_2) and the distribution for the i-th level of x_1 ($i = 1, \ldots, n$) and the j-th level of x_2 ($j = 1, \ldots, m$) is represented as $G_{i,j}$. Then we can decompose

$$G_{i,j} = \bar{G} + \Pi_{i\cdot} + \Pi_{\cdot j} + \Gamma_{i,j}, \tag{7}$$

where $\bar{G} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} G_{i,j}$, $\Pi_{i\cdot} = \frac{1}{m}\sum_{j=1}^{m}(G_{i,j} - \bar{G})$, $\Pi_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n}(G_{i,j} - \bar{G})$. Here $\bar{G}$ is a probability measure and $\Pi_{i\cdot}$, $\Pi_{\cdot j}$ and $\Gamma_{i,j}$ are signed measures that put measure 0 on Ω (and their densities will be $\bar{g}$, $\pi_{i\cdot}$, $\pi_{\cdot j}$ and $\gamma_{i,j}$, respectively). This separates the effect of level i of x_1 averaged over all levels of x_2, denoted by $\Pi_{i\cdot}$, the average effect of level j of x_2 ($\Pi_{\cdot j}$) and the interaction effects of combinations of levels i and j of both variables ($\Gamma_{i,j}$), giving us a very useful decomposition of the differences between the distributions.

The summaries described so far allow us to understand and interpret the differences between distributions but we also want to say something meaningful about regions of the support where the distributions are particularly different. We will consider a pair of distributions, G_i and G_j, and find a partition $\mathcal{P}$ of Ω defining subsets $\mathcal{P}_k$ and an indicator vector d for which $d_k = -1$ if G_i places substantially more mass than G_j on $\mathcal{P}_k$, $d_k = 1$ if G_j places substantially more mass than G_i on $\mathcal{P}_k$ and $d_k = 0$ otherwise. The choice of $\mathcal{P}$ and d will be made by specifying a utility function and finding the partition that maximizes expected utility. The utility function is

$$U(\mathcal{P}, d) = \sum_{k=1}^{r} U^*(\mathcal{P}_k, d_k),$$

where $\mathcal{P}_1, \ldots, \mathcal{P}_r$ are the elements of $\mathcal{P}$ and

$$U^*(\mathcal{P}, d) = \begin{cases} G_i(\mathcal{P}) - G_j(\mathcal{P}) & , \quad d = -1 \\ \frac{\epsilon}{2}(G_i(\mathcal{P}) + G_j(\mathcal{P})) & , \quad d = 0 \\ G_j(\mathcal{P}) - G_i(\mathcal{P}) & , \quad d = 1, \end{cases}$$

where $0 < \epsilon < 2$ is chosen to determine the meaning of substantial difference. Increasing values of ϵ lead to a utility function that increasingly favours setting $d_k = 0$. To understand the choice of utility function, consider an element, $\mathcal{P}_k$ of a fixed partition, $\mathcal{P}$. Then, $d_k = 0$ if

$$\frac{|G_i(\mathcal{P}_k) - G_j(\mathcal{P}_k)|}{\frac{1}{2}(G_i(\mathcal{P}_k) + G_j(\mathcal{P}_k))} < \epsilon.$$

The left-hand side of the expression is the difference in the mass of the two distributions on $\mathcal{P}_k$ divided by the average mass and ϵ is then interpreted as a tolerance parameter which controls the size of that ratio which constitutes a substantial difference. The expression naturally scales the difference by the mean mass under the two distributions and larger absolute differences will be declared “similar” in areas with larger average mass.

As $U(\mathcal{P}, d)$ is additive over the elements in the partition, maximizing the utility over partitions is easily done by starting from a very fine partition $\tilde{\mathcal{P}}$ and maximizing U^* on each element. Then we simply join the elements of $\tilde{\mathcal{P}}$ to form the partition $\mathcal{P}$ that maximizes utility.

5 Illustrations

The methods developed in this paper are illustrated on simulated data, a survival analysis example and an example from efficiency measurement. In all cases, the model with NGG marginals with $\lambda = 1$ and unknown other hyperparameters was used. In practice, this is not a particularly restrictive choice. Writing $M = \check{M}/\lambda^a$ in (3) leads to a process where λ scales the jump sizes and so has no effect on the normalized process (we have also implemented inference with a prior on λ and indeed found that the posterior and prior were virtually identical). It is assumed that D corresponds to the saturated model with $p = 2^q - 1$ (even for the stochastic frontier example in Subsection 5.3, where $q = 6$ so $p = 63$). Throughout, the prior for a was a uniform distribution on $(0, 1)$ and the prior for M_i was Ga$(1/2^{q-1}, 1)$ which implies that the prior for each G_g is NGG with $M \sim$ Ga$(1, 1)$.

5.1 Simulated data

We use two examples to illustrate the flexibility of the model. The first example has two groups which both contain 50 data points. The data for the first group are generated from the mixture distribution

$$f_1(x) = \alpha_1 \mathrm{N}(0, 1) + (1 - \alpha_1)\mathrm{N}(-5, 1)$$

and in the second group from

$$f_2(x) = \alpha_2 \mathrm{N}(0, 1) + (1 - \alpha_2)\mathrm{N}(5, 1).$$

On average, $50\alpha_1$ points in group 1 and $50\alpha_2$ points in group 2 will come from a standard normal but the other points will come from normal distribution centred at -5 for group 1 and 5 for group 2. The model of Müller et al. (2004) can represent these distributions if $\alpha_1 = \alpha_2$ but that model will fit worse as the values of α_1 and α_2 become further apart. We first consider the choice $\alpha_1 = \alpha_2 = 0.5$.

The second example extends the first by defining a third group (so $q = 3$) with observations drawn according to the density

$$f_3(x) = \alpha_2 \mathrm{N}(0, 1) + (1 - \alpha_2)\mathrm{N}(5, 1).$$

The third group has the same distribution as the second group. In this case, we use $\alpha_1 = 0.5$ and $\alpha_2 = 0.9$. Each data set was fitted using the model with NGG marginals with unknown hyperparameters.

The model is

$$y_{g,j} \overset{ind.}{\sim} \mathrm{N}(\mu_{g,j}, \sigma^2_{g,j})$$

$$(\mu_{g,j}, \sigma^{-2}_{g,j}) \overset{ind.}{\sim} G_g$$

$$G_1, G_2, \ldots, G_q \sim \mathrm{CNGG}(M, H, D; a, \lambda),$$

with $H = \mathrm{N}(\mu|\bar{y}, \sigma^2/m_0)\mathrm{Ga}(\sigma^{-2}|1, 1)$ where $\bar{y}$ is the mean of all observations and $m_0 = 0.01$.

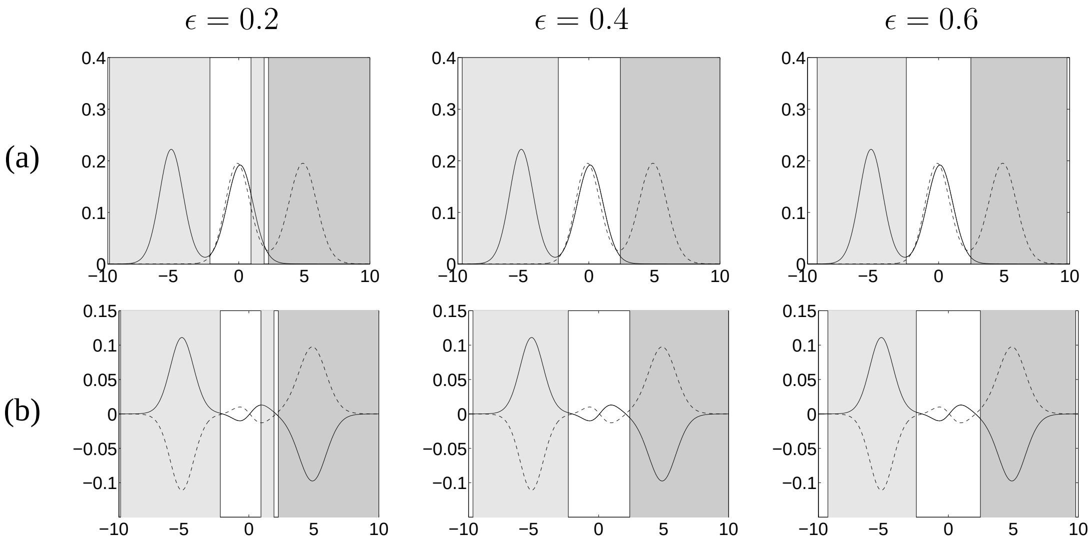

Figure 4: Example 1 ($\alpha_1 = \alpha_2 = 0.5$): (a) Posterior predictive density for the two groups (Group 1 is solid line and Group 2 is dashed line); (b) the difference π_1 (solid line) and π_2 (dashed line) indicating the area where Group 1 has substantially more mass than Group 2 (light grey) and vice versa (dark grey).

Some results of fitting the model to data in the first example are shown in Figure 4. The model estimates the densities well (shown in Row (a)). The graphs also show partitions of the support found using the approach in Section 4 for several values of the sensitivity parameter

ϵ. The results are reasonably robust to the choice of ϵ with $\epsilon > 0.2$ and they indicate that the distributions are similar between -2 and 2. This region seems slightly too small when $\epsilon = 0.2$, where the analysis reacts to the relatively small positive difference π_1 in between approximately 1 and 2. Row (b) shows the density of the differences π_1 and π_2. It is clear from the definition that $\pi_2 = -\pi_1$ when we have two groups and this is illustrated in the graphs which clearly show where the differences of the densities for the two groups are large.

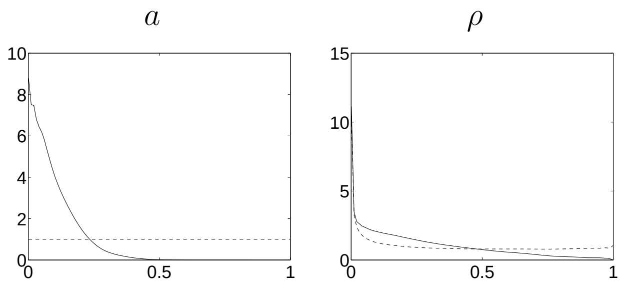

Figure 5: Example 1 ($\alpha_1 = \alpha_2 = 0.5$): Prior (dashed lines) and posterior (solid lines) densities of the parameter a and the correlation ρ for the NGG prior.

Figure 5 shows the posterior densities of the parameter a and the correlation ρ for the NGG prior. The data favour values of a smaller than 0.5. The posterior distribution of ρ (calculated using the result of Theorem 1) is not very different from the prior suggesting that the information in the data about correlation is not strong. The mass close to zero is in line with the fact that the distributions that generated both groups are quite different.

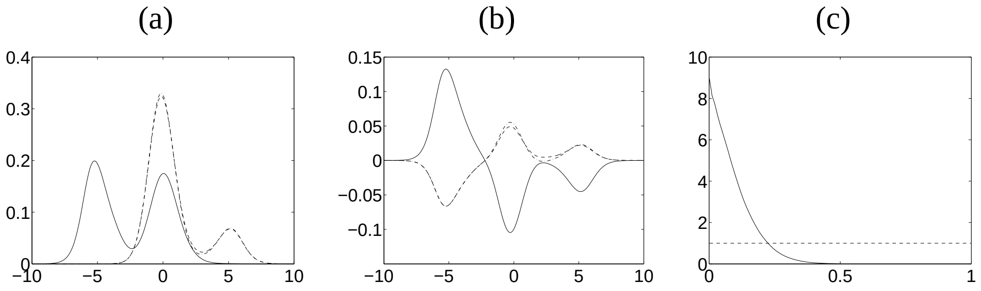

Figure 6: Example 2 ($\alpha_1 = 0.5, \alpha_2 = 0.9$): (a) Posterior predictive density for the three groups; (b) differences π_1, π_2 and π_3 (Group 1 (π_1) is solid line, Group 2 (π_2) is dashed line and Group 3 (π_3) is dot-dashed line. Results for Groups 2 and 3 are almost indistinguishable); (c) posterior distribution of a (prior overplotted as dashed line).

Figure 6 shows results of fitting the model to the second example with three groups. The density estimates clearly show the similarities between Groups 2 and 3 and the differences with respect to Group 1. The plots of π_1, π_2 and π_3 in panel (b) clearly illustrate the main differences. Group 1 places more mass than Groups 2 and 3 on values less than -2 whereas Groups 2 and 3 place more mass than Group 1 on values larger than -2. The posterior distribution of a is very similar to that shown in Figure 5.

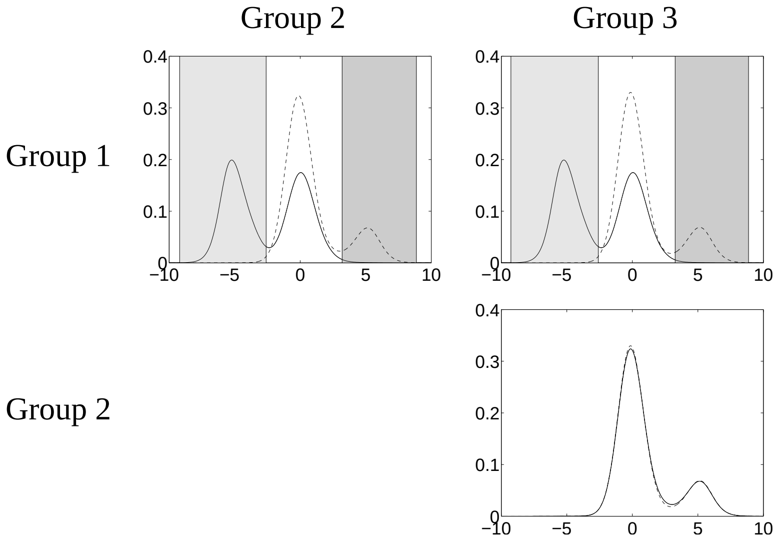

Figure 7: Example 2 ($\alpha_1 = 0.5, \alpha_2 = 0.9$): Posterior mean density for the group in the row (solid line) and column (dashed line) and comparison of the distributions with dark (light) grey areas indicating more mass for the group in the column (row).

Figure 7 shows the results of making pairwise comparisons for the three groups, using $\epsilon = 0.4$. The results follow from the discussion of the differences between the distributions. In the comparisons between Group 1 and Groups 2 and 3 there are two separate regions with important differences in the mass whereas the comparison between Group 2 and Group 3 shows no differences between the distributions (as we would expect).

5.2 Survival analysis

Doss and Huffer (2003) discuss modelling interval censored data in survival analysis using the DP as a prior for the distribution of the survival times. This application focuses on time to cosmetic deterioration of the breast of women with Stage 1 breast cancer who have undergone a lumpectomy under two treatments: radiation and radiation with chemotherapy. There are 46 subjects in the radiation only group and 48 subjects in the combination group. The data has been presented in Beadle et al. (1984). The indicator $d_{g,j} = 1$ if the j-th person in the g-th group suffers an event (in this case retraction of the breast) before the censoring time $T_{g,j}$ and $d_{g,j} = 0$ otherwise. If $d_{g,j} = 1$ then the observation is an interval $A_{g,j}$ in which the event occured. Doss and Huffer (2003) assign a Dirichlet process prior to the lifetime distribution for each group separately. Since the actual survival times are missing (due to the interval censoring), the posterior will then be a mixture of Dirichlet processes. Denoting the survival time of individual j in group g by $\tau_{g,j}$, we extend their approach to the model

$$\mathrm{I}\left(\tau_{g,j} \in A_{g,j}\right) \text{ if } d_{g,j} = 1 \text{ or } \mathrm{I}\left(\tau_{g,j} > T_{g,j}\right) \text{ if } d_{g,j} = 0$$

$$\tau_{g,j} \overset{ind.}{\sim} G_g$$

$$G_1, G_2, \ldots, G_q \sim \text{CNGG}(M, H, D; a, \lambda),$$

where H is an exponential distribution with mean $1/\xi$. The parameter ξ is given a vague Gamma prior with shape parameter 0.1 and mean 1.

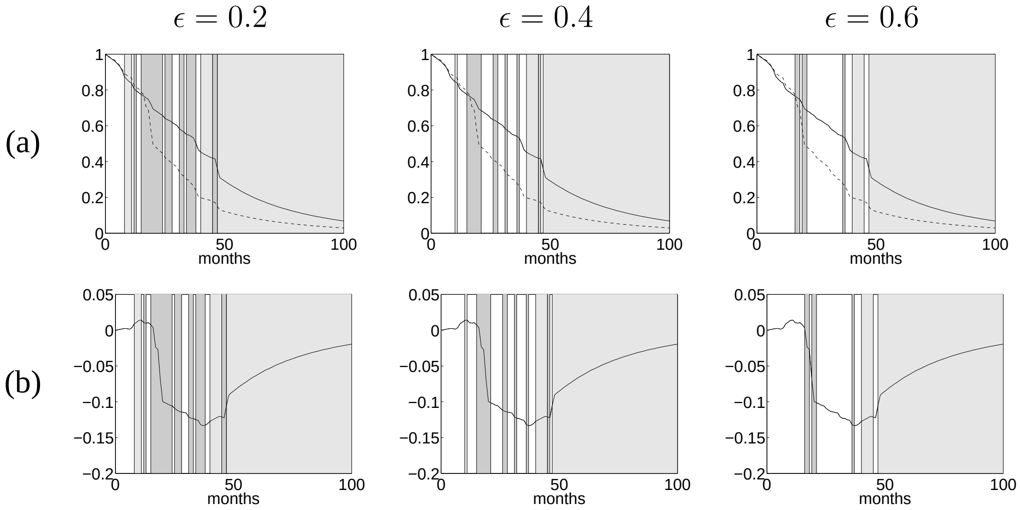

Figure 8: Survival analysis (combination group shown as dashed lines and radiation only group shown as solid lines): (a) the posterior mean survival functions for the two groups; (b) posterior mean for Π_1 where the radiation only group is coded as Group 1 (Dark (light) grey areas indicate more mass for Group 2 (1)).

Figure 8 displays results of the analysis of the clinical trial data. Row (a) shows that the survival function is similar for the two groups initially but the curves diverge around 16 months with the combination group associated with a much larger number of events. Row (b) shows the posterior mean of the difference between the survival functions for the groups. This also indicates that the mass is similar until 16 months but then the difference quickly becomes large until the survival functions converge again. The regions identified as similar change when moving from $\epsilon = 0.4$ to $\epsilon = 0.6$ with the latter having fewer, larger and more connected regions. The results with $\epsilon = 0.6$ more clearly highlight the larger differences in the survival functions, such as the sharp drop in the combination group around 16 months. Finally, for all values of ϵ the radiation only group places more mass than the combination group in the region beyond 45 months.

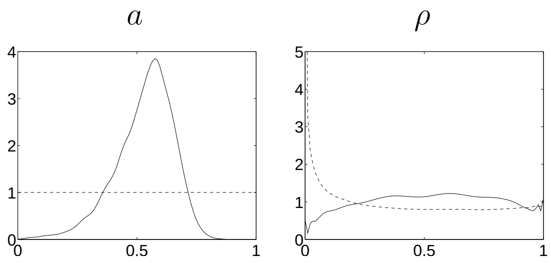

Figure 9: Survival analysis: Prior (dashed lines) and posterior (solid lines) densities of the parameter a and the correlation ρ.

The posterior distributions of a and ρ are shown in Figure 9, which indicates that the value $a = 0$ (the Dirichlet process case) is not well-supported by the data with a posterior median close to the Normalized Inverse-Gaussian process (where $a = 0.5$), but with substantial posterior uncertainty. The posterior distribution of the correlation parameter ρ indicates that the groups are different, but do share some common aspects.

5.3 Stochastic Frontier analysis

Stochastic frontier analysis is a popular method in econometrics for estimating the efficiency of firms. We will consider an application to the efficiency of US hospitals using data previously analyzed by Koop et al. (1997). It is assumed that all hospitals operate relative to a common cost frontier, which represents the minimum cost of performing the functions of that hospitals (including operations, patient care, etc.). It follows that inefficiency can be measured by how far a hospital operates above the optimal cost level given by the frontier. The costs are observed for the hospitals over a number of years. The model is written in terms of log cost, $C_{g,j,t}$, for the j-th hospital in the g-th group at the t-th time point

$$C_{g,j,t} = \alpha + x_{g,j,t}^T \beta + u_{g,j} + \varepsilon_{g,j,t},$$

where $x_{g,j,t}$ are variables used to define the frontier for j-th hospital in the g-th group at the t-th time point, $u_{g,j} > 0$ is the inefficiency for the j-th hospital in the g-th group and $\varepsilon_{i,j,t}$ are mutually independent, measurement errors which will be assumed to be normally distributed with mean 0 and variance σ^2. The model assumes that the efficiency of hospitals is fixed over the time period (a common assumption in the applied literature). The efficiency for the j-th hospital in the g-th group is defined to be $\exp\{-u_{g,j}\}$.

The main focus of this type of analysis is the distribution of the inefficiencies $u_{g,j}$ and estimation of the hospital efficiencies $\exp\{-u_{g,j}\}$. A Bayesian nonparametric analysis of the stochastic frontier model is described by Griffin and Steel (2004) who assume a DP prior for the inefficiency distribution and apply their methods to the data analyzed here. The model used here is

$$C_{g,j,t} \overset{ind.}{\sim} \mathrm{N}(\alpha + x_{g,j,t}^T \beta + u_{g,j}, \sigma^2)$$

$$u_{g,j} \overset{ind.}{\sim} G_g$$

$$G_1, G_2, \ldots, G_q \sim \mathrm{CNGG}(M, H, D; a, \lambda),$$

where α, β and σ^2 are given the priors described by Griffin and Steel (2004) and H is an exponential distribution with mean $1/\xi$, where ξ is given an exponential prior with mean $-1/\log r^\star$, so that $r^\star$ is the prior median efficiency. In this example $r^\star$ is chosen to take the value 0.8.

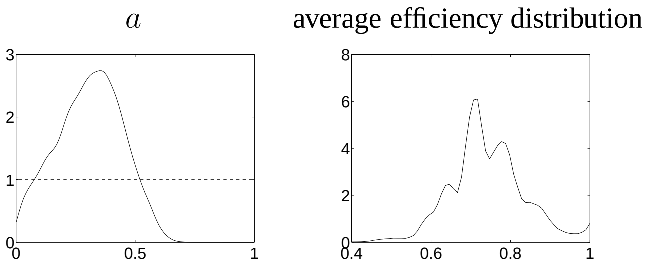

Figure 10: Stochastic Frontier Analysis: The posterior (solid line) and prior distributions (dashed line) of a and the posterior mean of the average efficiency distribution with the NGG prior.

The data also include information about the type of hospital and include two factors: the ownership status of the hospital (For-Profit, Non-Profit and Government) and a quality factor in terms of staff-patient ratio or SPR (Low or High). The definition of these factors is described in Koop et al. (1997). Figure 10 shows some posterior results of extending the model of Griffin and Steel (2004) using the prior developed in this paper. The posterior distribution of a has a mode at around 0.4. The posterior mean of the efficiency distribution averaged over all hospital types has three internal modes at roughly 0.65, 0.7 and 0.8 and a further mode at 1, which is quite in line with the results for the efficiency obtained in Griffin and Steel (2004) without using hospital type information. Figure 11 shows the posterior mean for the efficiency distribution within each group. For comparison, an analysis using a product of DP is provided by Griffin and Steel (2004). The prior developed in this paper leads to predictive distributions which vary substantially less between groups, illustrating the model's ability to effectively borrow information. This is particularly important in this application where group sizes are quite small, ranging from 20 to 141. All distributions are multi-modal with most distributions having modes at roughly 0.7 and 0.8 (and at 1). However, the sizes of the modes differ between the distributions.

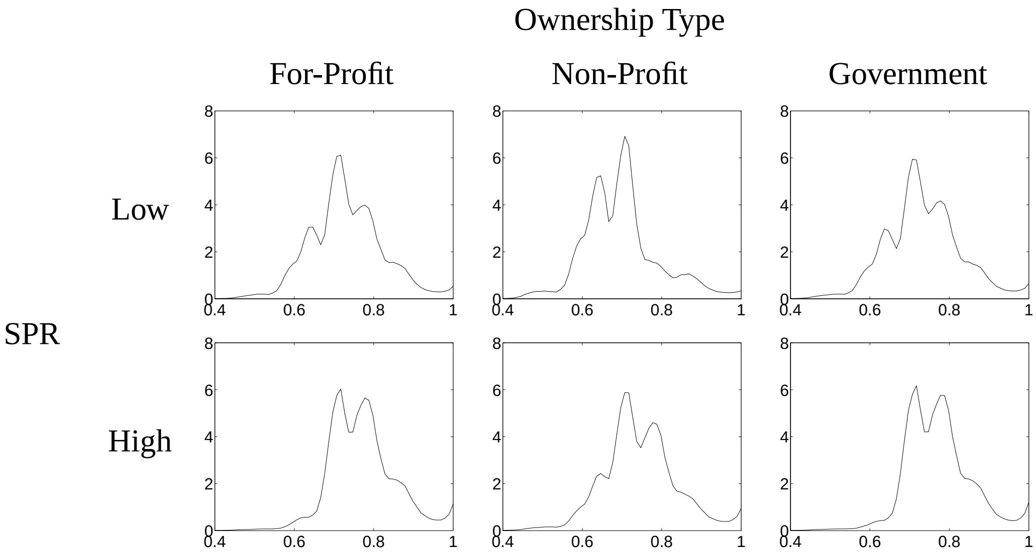

Figure 11: Stochastic Frontier Analysis: The posterior mean of the efficiency distribution for each hospital type with a NGG prior.

Figure 12 shows the decomposition of the estimated distribution defined in (7). These graphs more clearly show the differences and similarities between the distributions. The π's show the effect of one factor averaging over the other factors. Hospitals with High SPR tend have more mass at higher efficiency than Low SPR hospitals (suggesting that they tend to be more efficient). The effect of high SPR is to shift mass away from around 0.65 to around

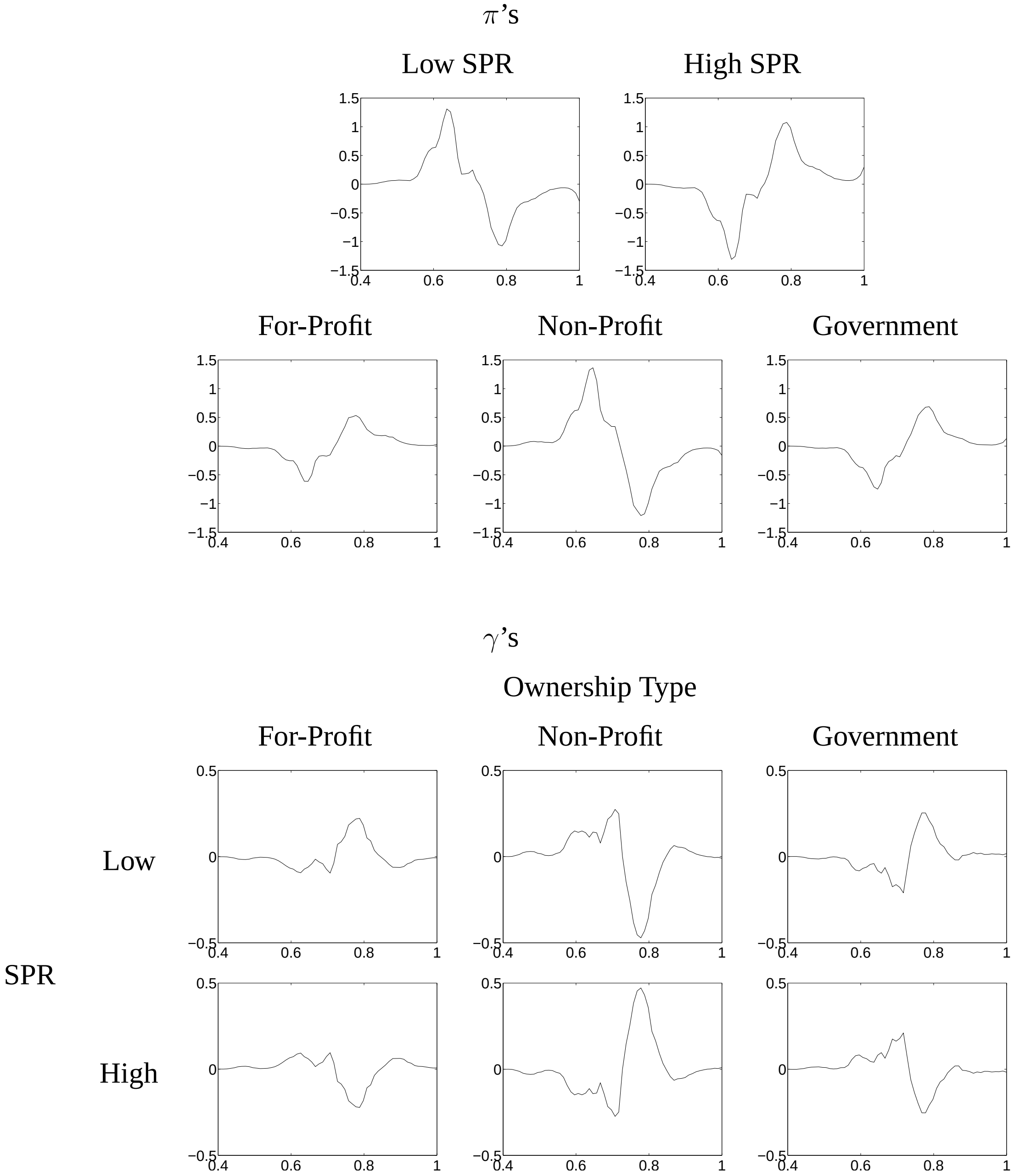

Figure 12: Stochastic Frontier Analysis: The posterior means of $\pi_{i\cdot}$, $\pi_{\cdot j}$ and $\gamma_{i,j}$ with NGG process marginals.

0.8. The For-Profit and Government hospitals have similar distributions and have more mass at higher level efficiency than Non-Profit hospitals, again mostly involving shifts from regions around 0.65 to those in the vicinity of 0.8. The densities γ relate to interaction terms which are most important for Non-Profit hospitals where Non-Profit hospitals with Low SPR tend to have particularly low mass at high levels of efficiency (around 0.8). Thus, the results clearly indicate which factors (or combinations of factors) lead to distributions that place more mass on higher levels of efficiency.

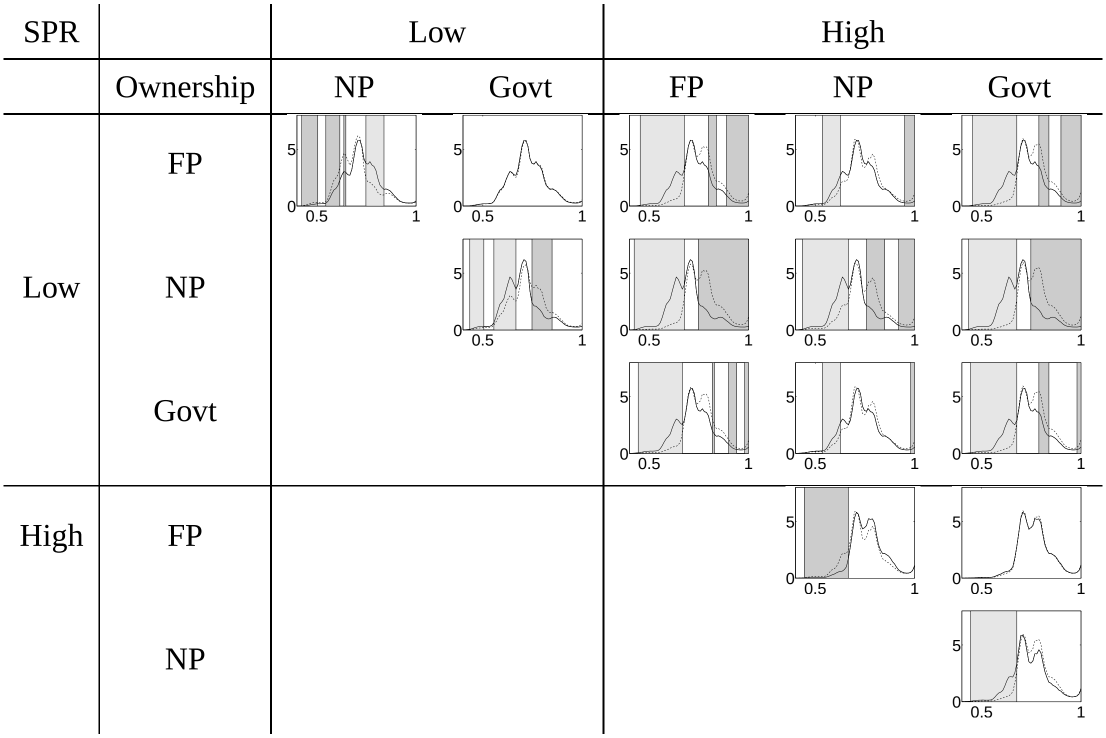

Figure 13: Stochastic Frontier Analysis: Graphs of pairwise comparisons of efficiency distributions according to ownership type (FP=For-Profit, NP=Non-Profit, Govt=Government) and Staff-Patient Ratio. The pairs are shown as the row (solid line) and column (dashed lines) with dark grey shading indicating higher mass in the column and light grey shading indicating higher mass in the row.

Figure 13 shows pairwise comparisons of the distributions which identify regions where the mass placed by the two corresponding distributions is substantially different, using $\epsilon = 0.4$. These indicate that there is a lack of evidence of a difference between the For-Profit and Government hospitals at both quality levels (in line with their very similar π's). There is also not much difference between the Non-Profit hospitals at high quality and the For-Profit and Government hospital at Low quality (the π's for both factors more or less balance each other out). The other combinations of factors lead to clear results where we can identify regions of the support where one distribution places more mass than the other and vice versa. Clearly, the For-Profit and Government hospitals with high quality are the most efficient combinations, placing more mass on higher efficiencies than other cases. Interestingly, the much more restrictive fully parametric model without interactions of Koop et al. (1997) leads to the very different (and counterintuitive) conclusions that For-Profit status and high SPR both reduce efficiencies.

6 Summary

This paper discusses a method for inferring differences between distributions associated with different groups of observations. A Bayesian nonparametric approach is taken and we introduce a novel form of priors, derived from Normalized Random Measures with Independent Increments. The prior allows the inclusion of information about partial exchangeability and so represents prior beliefs which could not be expressed using *e.g.* the Hierarchical Dirichlet process. This allows effective borrowing of strength between distributions without assuming exchangeability, and can easily and systematically accommodate widely varying levels of complexity in terms of dependence. Efficient, exact inference is possible using a slice sampling method, which extends the ideas of Griffin and Walker (2010). The prior is used with a new graphical method to compare pairs of distributions. The common support of any two distributions is partitioned and each element of the partition is characterized by obtaining more mass from either distribution or being allocated roughly similar mass by both distributions. This is an effective way of understanding the difference between two distributions. In particular, where the groups are defined by several covariates, we propose an informative ANOVA-type decomposition of the differences.

We analyze applications in survival analysis and stochastic frontiers with small numbers of observations, typical of real data applications in many fields. Despite this, the models perform very well and lead to sensible results. Interestingly, in both applications, models with Dirichlet process marginal processes are not well supported by the data and Normalized Generalized Gamma marginals are favoured. The posterior distribution of a in the survival example is centred around 0.5 which corresponds to the Normalized Inverse-Gaussian process.

We believe the methodology proposed in this paper is highly flexible, yet widely applicable to real data, and allows for quite informative inference on the (sources of the) differences between dependent distributions.

References

Beadle, G., S. Come, C. Henderson, B. Silver, and S. Hellman (1984). The effect of adjuvant chemotherapy on the cosmetic results after primary radiation treatment for early stage breast cancer. *Inter. J. Rad. Oncol., Biol. Phys. 10*, 2131–2137.

Brix, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Adv. in Appl. Probab. 31*, 929–953.

De Iorio, M., P. Müller, G. L. Rosner, and S. N. MacEachern (2004). An ANOVA Model for Dependent Random Measures. *J. Amer. Statist. Assoc. 99*, 205–215.

Doss, H. and F. W. Huffer (2003). Monte Carlo methods for Bayesian analysis of survival data using mixtures of Dirichlet process prior. *J. Comput. Graph. Statist. 12*, 282–307.

Dunson, D. B., Y. Xue, and L. Carin (2008). The matrix stick breaking process: Flexible Bayes meta analysis. *J. Amer. Statist. Assoc. 103*, 317–327.

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

Griffin, J. E. (2009). The Ornstein-Uhlenbeck Dirichlet Process and other time-varying nonparametric priors. Technical report, University of Warwick.

Griffin, J. E. and P. J. Brown (2010). Inference with Normal-Gamma prior distributions in regression problems. *Bayesian Analysis 5*, 171–188.

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

Griffin, J. E. and S. G. Walker (2010). Posterior simulation of Normalised Random Measure mixtures. *J. Comput. Graph. Statist.*, forthcoming.

James, L. F., A. Lijoi, and I. Prünster (2009). Posterior analysis for normalized random measures with independent increments. *Scand. J. Statist. 36*, 76–97.

Kalli, M., J. E. Griffin, and S. G. Walker (2011). Slice sampling mixture models. *Statistics and Computing 21*, forthcoming.

Kingman, J. (1975). Random discrete distributions. *J. R. Stat. Soc. Ser. B 37*, 1–22 (with discussion).

Kolossiatis, M., J. E. Griffin, and M. F. J. Steel (2010). On Bayesian nonparametric modelling of two correlated distributions. Technical report, University of Warwick.

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

Lijoi, A., R. H. Mena, and I. Prünster (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *J. Amer. Statist. Assoc. 100*, 1278–1291.

Lijoi, A., R. H. Mena, and I. Prünster (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *J. R. Stat. Soc. Ser. B 69*, 715–740.

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

Rodriguez, A., D. Dunson, and J. Taylor (2009). Bayesian hierarchically weighted finite mixture models for samples of distributions. *Biostatistics 10*, 155–171.

Scott, J. G. and N. G. Polson (2011). Shrink globally, act locally: Sparse Bayesian regularization and prediction. In *Bayesian Statistics 9*. Oxford University Press.

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

A Proof of Theorem 1

We know that $\mathrm{E}[G_1(B)] = \mathrm{E}[G_2(B)] = H(B)$. To calculate the covariance, we need

$$\begin{aligned}\mathrm{E}[G_1(B)G_2(B)] &= \mathrm{E}\left[\frac{\tilde{G}_1(B)}{\tilde{G}_1(\Omega)}\frac{\tilde{G}_2(B)}{\tilde{G}_2(\Omega)}\right]\\ &= \mathrm{E}\left[\frac{\left(\tilde{G}_1^\star(B)+\tilde{G}_2^\star(B)\right)\left(\tilde{G}_1^\star(B)+\tilde{G}_3^\star(B)\right)}{\left(\tilde{G}_1^\star(\Omega)+\tilde{G}_2^\star(\Omega)\right)\left(\tilde{G}_1^\star(\Omega)+\tilde{G}_3^\star(\Omega)\right)}\right]\\ &= \int_0^\infty\int_0^\infty \mathrm{E}\left[\gamma(v_1,v_2)\right]\,dv_1\,dv_2\end{aligned}$$

where

$$\begin{aligned}
\gamma(v_1, v_2) &= \left(\tilde{G}_1^\star(B) + \tilde{G}_2^\star(B)\right)\left(\tilde{G}_1^\star(B) + \tilde{G}_3^\star(B)\right) \\
&\quad \times \exp\left\{-v_1\left(\tilde{G}_1^\star(\Omega) + \tilde{G}_2^\star(\Omega)\right) - v_2\left(\tilde{G}_1^\star(\Omega) + \tilde{G}_3^\star(\Omega)\right)\right\} \\
&= \left(\tilde{G}_1^\star(B)^2 + \tilde{G}_1^\star(B)\tilde{G}_3^\star(B) + \tilde{G}_2^\star(B)\tilde{G}_1^\star(B) + \tilde{G}_2^\star(B)\tilde{G}_3^\star(B)\right) \\
&\quad \times \exp\left\{-(v_1+v_2)\tilde{G}_1^\star(\Omega) - v_1\tilde{G}_2^\star(\Omega) - v_2\tilde{G}_3^\star(\Omega)\right\}
\end{aligned}$$

The independence of the underlying processes $\tilde{G}_1^\star, \tilde{G}_2^\star$ and $\tilde{G}_3^\star$ and the independence of Lévy processes on disjoint sets gives

$$\begin{aligned}
&\mathrm{E}\left[\gamma(v_1, v_2)\right] \\
=&\mathrm{E}\left[\tilde{G}_1^\star(B)^2\exp\left\{-(v_1+v_2)\tilde{G}_1^\star(B)\right\}\right]\mathrm{E}\left[\exp\left\{-(v_1+v_2)\tilde{G}_1^\star(B^c)\right\}\right]\mathrm{E}\left[\exp\left\{-v_1\tilde{G}_2^\star(\Omega)\right\}\right] \\
&\times \mathrm{E}\left[\exp\left\{-v_2\tilde{G}_3^\star(\Omega)\right\}\right] + \mathrm{E}\left[\tilde{G}_1^\star(B)\exp\left\{-(v_1+v_2)\tilde{G}_1^\star(B)\right\}\right]\mathrm{E}\left[\tilde{G}_3^\star(B)\exp\left\{-v_2\tilde{G}_3^\star(B)\right\}\right] \\
&\times \mathrm{E}\left[\exp\left\{-(v_1+v_2)\tilde{G}_1^\star(B^c)\right\}\right]\mathrm{E}\left[\exp\left\{-v_1\tilde{G}_2^\star(\Omega)\right\}\right]\mathrm{E}\left[\exp\left\{-v_2\tilde{G}_3^\star(B^c)\right\}\right] \\
&+ \mathrm{E}\left[\tilde{G}_2^\star(B)\exp\left\{-v_1\tilde{G}_2^\star(B)\right\}\right]\mathrm{E}\left[\tilde{G}_1^\star(B)\exp\left\{-(v_1+v_2)\tilde{G}_1^\star(B)\right\}\right]\mathrm{E}\left[\exp\left\{-(v_1+v_2)\tilde{G}_1^\star(B^c)\right\}\right] \\
&\times \mathrm{E}\left[\exp\left\{-v_1\tilde{G}_2^\star(B^c)\right\}\right]\mathrm{E}\left[\exp\left\{-v_2\tilde{G}_3^\star(\Omega)\right\}\right] + \mathrm{E}\left[\tilde{G}_2^\star(B)\exp\left\{-v_1\tilde{G}_2^\star(B)\right\}\right] \\
&\times \mathrm{E}\left[\tilde{G}_3^\star(B)\exp\left\{-v_2\tilde{G}_3^\star(B)\right\}\right]\mathrm{E}\left[\exp\left\{-(v_1+v_2)\tilde{G}_1^\star(\Omega)\right\}\right]\mathrm{E}\left[\exp\left\{-v_1\tilde{G}_2^\star(B^c)\right\}\right] \\
&\times \mathrm{E}\left[\exp\left\{-v_2\tilde{G}_3^\star(B^c)\right\}\right]
\end{aligned}$$

The definition of $L_\eta(v)$ implies that

$$\mathrm{E}[\exp\{-v\tilde{G}_k^\star(B)\}] = \exp\left\{-H(B)M_kL_\eta(v)\right\}$$

and then

$$\begin{aligned}
\mathrm{E}\left[\tilde{G}_k^\star(B)\exp\{-v\tilde{G}_k^\star(B)\}\right] &= -\mathrm{E}\left[\frac{d}{dv}\exp\{-v\tilde{G}_k^\star(B)\}\right] = -\frac{d}{dv}\mathrm{E}\left[\exp\{-v\tilde{G}_k^\star(B)\}\right] \\
&= -\frac{d}{dv}\exp\left\{-H(B)M_kL_\eta(v)\right\} = H(B)M_kL'_\eta(v)\exp\left\{-H(B)M_kL_\eta(v)\right\}
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}\left[\left(\tilde{G}_k^\star(B)\right)^2\exp\{-v\tilde{G}_k^\star(B)\}\right] &= \mathrm{E}\left[\frac{d}{dv^2}\exp\{-v\tilde{G}_k^\star(B)\}\right] = \frac{d}{dv^2}\mathrm{E}\left[\exp\{-v\tilde{G}_k^\star(B)\}\right] \\
&= \left[H(B)^2M_k^2\left(L'_\eta(v)\right)^2 - H(B)M_kL''_\eta(v)\right]\exp\left\{-H(B)M_kL_\eta(v)\right\}.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\mathrm{E}\left[\gamma(v_1,v_2)\right] \\
&= \Big[H(B)^2 M_1^2 \left(L'_\eta(v_1+v_2)\right)^2 - H(B)M_1 L''_\eta(v_1+v_2)\Big] \exp\left\{-H(B)M_1L_\eta(v_1+v_2)\right\} \\
&\quad \times \exp\left\{-(1-H(B))M_1L_\eta(v_1+v_2)\right\} \exp\left\{-M_2L_\eta(v_1)\right\} \exp\left\{-M_3L_\eta(v_2)\right\} \\
&\quad + H(B)M_1L'_\eta(v_1+v_2) \exp\left\{-H(B)M_1L_\eta(v_1+v_2)\right\} H(B)M_3L'_\eta(v_2) \exp\left\{-H(B)M_3L_\eta(v_2)\right\} \\
&\quad \times \exp\left\{-(1-H(B))M_1L_\eta(v_1+v_2)\right\} \exp\left\{-M_2L_\eta(v_1)\right\} \exp\left\{-(1-H(B))M_3L_\eta(v_2)\right\} \\
&\quad + H(B)M_2L'_\eta(v_1) \exp\left\{-H(B)M_2L_\eta(v_1)\right\} H(B)M_1L'_\eta(v_1+v_2) \exp\left\{-H(B)M_1L_\eta(v_1+v_2)\right\} \\
&\quad \times \exp\left\{-(1-H(B))M_1L_\eta(v_1+v_2)\right\} \exp\left\{-(1-H(B))M_2L_\eta(v_1)\right\} \exp\left\{-M_3L_\eta(v_2)\right\} \\
&\quad + H(B)M_2L'_\eta(v_1) \exp\left\{-H(B)M_2L_\eta(v_1)\right\} H(B)M_3L'_\eta(v_2) \exp\left\{-H(B)M_3L_\eta(v_2)\right\} \\
&\quad \times \exp\left\{-M_1L_\eta(v_1+v_2)\right\} \exp\left\{-(1-H(B))M_2L_\eta(v_1)\right\} \exp\left\{-(1-H(B))M_3L_\eta(v_2)\right\} \\
&= \left[H(B)^2 \left(M_2L'_\eta(v_1) + M_1L'_\eta(v_1+v_2)\right)\left(M_3L'_\eta(v_2) + M_1L'_\eta(v_1+v_2)\right) - H(B)M_1L''_\eta(v_1+v_2)\right] \\
&\quad \times \exp\left\{-M_1L_\eta(v_1+v_2)\right\} \exp\left\{-M_2L_\eta(v_1)\right\} \exp\left\{-M_3L_\eta(v_2)\right\}.
\end{aligned}$$

Then

$$\mathrm{Cov}(G_1(B),G_2(B)) = H(B)^2\left[\int_0^\infty\int_0^\infty \alpha\gamma\, dv_1\, dv_2 - 1\right] - H(B)\int_0^\infty\int_0^\infty \beta\gamma\, dv_1\, dv_2$$

where

$$\alpha = \left(M_2L'_\eta(v_1) + M_1L'_\eta(v_1+v_2)\right)\left(M_3L'_\eta(v_2) + M_1L'_\eta(v_1+v_2)\right),$$

$$\beta = M_1L''_\eta(v_1+v_2) \quad \text{and}$$

$$\gamma = \exp\left\{-M_1L_\eta(v_1+v_2) - M_2L_\eta(v_1) - M_3L_\eta(v_2)\right\}.$$

The result follows from the fact that

$$\int_0^\infty\int_0^\infty \alpha\gamma dv_1 dv_2 = 1 + \int_0^\infty\int_0^\infty M_1L''_\eta(v_1+v_2)\exp\{-M_2L_\eta(v_1) - M_3L_\eta(v_2) - M_1L_\eta(v_1+v_2)\}\, dv_1 dv_2.$$