Dependent Mixtures of Geometric Weights Priors

Spyridon J. Hatjispyros*, Christos Merkatas*, Theodoros Nicoleris**, Stephen G. Walker***

* Department of Mathematics, University of the Aegean, Karlovassi, Samos, GR-832 00, Greece.
** Department of Economics, National and Kapodistrian University of Athens, Athens, GR-105 59, Greece.
***Department of Mathematics, University of Texas at Austin, Austin, Texas 7812, USA.

Abstract

A new approach to the joint estimation of partially exchangeable observations is presented. This is achieved by constructing a model with pairwise dependence between random density functions, each of which is modeled as a mixture of geometric stick breaking processes. The claim is that mixture modeling with Pairwise Dependent Geometric Stick Breaking Process (PDGSBP) priors is sufficient for prediction and estimation purposes; that is, making the weights more exotic does not actually enlarge the support of the prior. Moreover, the corresponding Gibbs sampler for estimation is faster and easier to implement than the Dirichlet Process counterpart.

Keywords: Bayesian nonparametric inference; Mixture of Dirichlet process; Geometric stick breaking weights; Geometric Stick Breaking Mixtures; Dependent Process.

1. Introduction. In Bayesian nonparametric methods, the use of priors such as the Dirichlet process (Ferguson, 1973), is justified from the assumption that the observations are exchangeable, which means the distribution of \((X_1,\ldots,X_n)\) coincides with the distribution of \((X_{\pi(1)},\ldots,X_{\pi(n)})\), for all \(\pi \in S(n)\), where \(S(n)\) is the set of permutations of \(\{1,\ldots,n\}\). However, in real life applications, data are often partially exchangeable. For example, they may consist of observations sampled from \(m\) populations, or may be sampled from an experiment conducted in \(m\) different geographical areas. This means that the joint law is invariant under permutations within the \(m\) subgroups of observations \((X_{j,i})_{1 \leq i \leq n_j}, 1 \leq j \leq m\), so for all \(\pi_j \in S(n_j)\)

\[
((X_{1,i_1})_{1 \leq i_1 \leq n_1}, \ldots, (X_{m,i_m})_{1 \leq i_m \leq n_m}) \sim ((X_{1,\pi_1(i_1)})_{1 \leq i_1 \leq n_1}, \ldots, (X_{m,\pi_m(i_m)})_{1 \leq i_m \leq n_m}).
\] (1)

When the exchangeability assumption fails one needs to use non-exchangeable priors. There has been substantial research interest following the seminal work of MacEachern (1999) in the
construction of suitable dependent stochastic processes. Such then act as priors in Bayesian nonparametric models. These processes are distributions over a collection of measures indexed by values in some covariate space, such that the marginal distribution is described by a known nonparametric prior. The key idea is to induce dependence between a collection of random probability measures \( (\mathbb{P}_j)_{1 \leq j \leq m} \), where each \( \mathbb{P}_j \) comes from a Dirichlet process (DP) with concentration parameter \( c > 0 \) and base measure \( \mathbb{P}_0 \). Such random probability measures typically are used in mixture models to generate random density functions \( f(x) = \int_\Theta K(x|\theta) \mathbb{P}(d\theta) \); see Lo (1984).

There is a variety of ways that a DP can be extended to dependent DP. Most of them use the stick-breaking representation (Sethuraman, 1994), that is

\[
\mathbb{P}(\cdot) = \sum_{k=1}^{\infty} w_k \delta_{\theta_k}(\cdot),
\]

where \( (\theta_k)_{k \geq 1} \) are independent and identically distributed from \( \mathbb{P}_0 \) and \( (w_k)_{k \geq 1} \) is a stick-breaking process; so if \( (v_k)_{k \geq 1} \) are independent and identically distributed from \( \text{Be}(1, c) \), a beta distribution with mean \( (1 + c)^{-1} \), then \( w_1 = v_1 \) and for \( k > 1 \), \( w_k = v_k \prod_{l<k}(1 - v_l) \). Dependence is introduced through the weights and/or the atoms. A classical example of the use of dependent DP’s is the Bayesian nonparametric regression problem where a random probability measure \( \mathbb{P}_z \) is constructed for each covariate \( z \),

\[
\mathbb{P}_z(\cdot) = \sum_{k=1}^{\infty} w_k(z) \delta_{\theta_k(z)}(\cdot),
\]

where \( (w_k(z), \theta_k(z)) \) is a collection of processes indexed in \( z \)-space. Extensions to dependent DP models can be found in De Iorio et al. (2004), Griffin and Steel (2006), and Dunson and Park (2008).

Recently there has been growing interest for the use of simpler random probability measures which while simpler are yet sufficient for Bayesian nonparametric density estimation. The geometric stick breaking (GSB) random probability measure (Fuentes–García, et al. 2010) has been used for density estimation and has been shown to provide an efficient alternative to DP mixture models. Some recent papers extend this nonparametric prior to a dependent nonparametric prior. In the direction of covariate dependent processes, GSB processes have been seen to provide an adequate model to the traditional dependent DP model. For example, for Bayesian regression, Fuentes–García et al. (2009) propose a covariate dependent process based on random probability measures drawn from a GSB process. Mena et al. (2011) used GSB random probability measures in order to construct a purely atomic continuous time measure-valued process, useful for the analysis of time series data. In this case, the covariate \( z \geq 0 \) denotes the time that each observation is (discretely) recorded and conditionally on each observation is drawn from a time-dependent nonparametric mixture model based on GSB processes. However, to the best of our knowledge, random probability measures drawn from a
GSB process, for modeling related density functions when samples from each density function are available, has not been developed in the literature.

In this paper we will construct pairwise dependent random probability measures based on GSB processes. That is, we are going to model a finite collection of \( m \) random distribution functions \((G_j)_{1 \leq j \leq m}\), where each \( G_j \) is a GSB random probability measure, such that there is a unique common component for each pair \((G_j, G_{j'})\) with \( j \neq j' \). We are going to use these measures in the context of GSB mixture models, generating a collection of \( m \) GSB pairwise dependent random densities \((f_j(x))_{1 \leq j \leq m}\). Hence we obtain a set of random densities \((f_1, \ldots, f_m)\), where marginally each \( f_j \) is a random density function

\[
f_j(x) = \int_{\Theta} K(x|\theta) G_j(d\theta),
\]

thus generalizing the GSB priors to a multivariate setting for partially exchangeable observations.

In the problem considered here, these random density functions \((f_j)_{1 \leq j \leq m}\) are thought to be related or similar, e.g. perturbations of each other, and so we aim to share information between groups to improve estimation of each density, especially for those densities \( f_j \) for which the corresponding sample size \( n_j \) is small. In this direction, the main references include the work of Müller et al. (2014), Bulla et al. (2009), Kolossiatis et al. (2013) and Griffin et al. (2013); more rigorous results can be found in Lijoi et al. (2014A, 2014B). All these models have been proposed for the modeling of an arbitrary but finite number of random distribution functions, via a common part and an index specific idiosyncratic part so that for \( 0 < p_j < 1 \) we have \( P_j = p_j P_0 + (1 - p_j) P^*_j \), where \( P_0 \) is the common component to all other distributions and \( \{P^*_j : j = 1, \ldots, m\} \) are the idiosyncratic parts to each \( P_j \), and \( P_0, P^*_j \overset{iid}{\sim} \text{DP}(c, P_0) \). In Lijoi et al. (2014B) normalized random probability measures based on the \( \sigma \)-stable process are used for modeling dependent mixtures. Although similar (all models coincide only for the \( m = 2 \) case), these models are different from our model which is based on pairwise dependence of a sequence of random measures (Hatjispyros et al. 2011, 2016A).

We are going to provide evidence through numerical experiments that dependent GSB mixture models are an efficient alternative to pairwise dependent DP (PDDP) priors. First, we will randomize the existing PDDP model of Hatjispyros et al. (2011, 2016A), by imposing gamma priors on the concentration masses (leading to the more efficient rPDDP model). Then, for the objective comparison of execution times, we will conduct a-priori synchronized density estimation comparison studies between the randomized PDDP and the pairwise dependent GSB process (PDGSBP) models using synthetic and real data examples.

This paper is organized as follows. In Section 2 we will demonstrate the construction of pairwise dependent random densities, using a dependent model suggested by Hatjispyros et al. (2011). We also demonstrate how specific choices of latent random variables can recover the model of Hatjispyros et al. and the dependent GSB model introduced in this paper. These latent variables will form the basis of a Gibbs sampler for posterior inference, given in Section 3.
In Section 4 we resort to simulation. We provide comparison studies between the randomized version of the PDDP model and our newly introduced dependent GSB model, involving five cases of synthetic data and a real data set. Finally, Section 5 concludes with a summary and future work.

2. Preliminaries. We consider an infinite real valued process \( \{X_{ji} : 1 \leq j \leq m, i \geq 1\} \) defined over a probability space \((\Omega, \mathcal{F}, P)\), that is partially exchangeable as in \([\Pi]\). Let \(\mathcal{P}\) denote the set of probability measures over \(\mathbb{R}\); then de Finetti proved that there exists a probability distribution \(\Pi\) over \(\mathcal{P}^m\), which satisfies

\[
P\{X_{ji} \in A_{ji} : 1 \leq j \leq m, 1 \leq i \leq n_j\} = \int_{\mathcal{P}^m} P\{X_{ji} \in A_{ji} : 1 \leq j \leq m, 1 \leq i \leq n_j|\mathcal{Q}_j\} \Pi(d\mathcal{Q}_1, \ldots, d\mathcal{Q}_m)
\]

\[
= \int_{\mathcal{P}^m} \prod_{j=1}^m P\{X_{ji} \in A_{ji} : 1 \leq i \leq n_j|\mathcal{Q}_j\} \Pi(d\mathcal{Q}_1, \ldots, d\mathcal{Q}_m)
\]

\[
= \int_{\mathcal{P}^m} \prod_{j=1}^m \prod_{i=1}^{n_j} \mathcal{Q}_j(A_{ji}) \Pi(d\mathcal{Q}_1, \ldots, d\mathcal{Q}_m).
\]

The de Finetti measure \(\Pi\) represents a prior distribution over partially exchangeable observations.

We start off by describing the PDDP model, with no auxiliary variables, using only the de Finetti measure \(\Pi\), marginal measures \(\mathcal{Q}_j\), then, we proceed to the definition of a randomized version of it, and to the specific details for the case of the GSB random measures.

A. In Hatjispyros et al. (2011), the following hierarchical model was introduced. For \(m\) subgroups of observations \(\{(x_{ji})_{1 \leq i \leq n_j} : 1 \leq j \leq m\}\),

\[
x_{ji} | \theta_{ji} \overset{\text{ind}}{\sim} K(\cdot | \theta_{ji})
\]

\[
\theta_{ji} | \mathcal{Q}_j \overset{\text{ind}}{\sim} \mathcal{Q}_j(\cdot)
\]

\[
\mathcal{Q}_j = \sum_{l=1}^m p_{jl} P_{jl}, \quad \sum_{l=1}^m p_{jl} = 1, \quad P_{jl} = P_{lj}
\]

\[
P_{jl} \overset{\text{ind}}{\sim} \mathcal{DP}(c, P_0), \quad 1 \leq j \leq l \leq m,
\]

for some kernel density \(K(\cdot | \cdot)\), concentration parameter \(c > 0\) and parametric central measure \(P_0\) for which \(E(P_{jl}(d\theta)) = P_0(d\theta)\). So, we have assumed that the random densities \(f_j(x)\) are dependent mixtures of the dependent random measures \(\mathcal{Q}_j\) via \(f_j(x|\mathcal{Q}_j) = \int_\Theta K(x|\theta)\mathcal{Q}_j(d\theta)\), or equivalently, dependent mixtures of the \(m\) independent mixtures \(g_{jl}(x|P_{jl}) = \int_\Theta K(x|\theta) P_{jl}(d\theta)\), \(l = 1, \ldots, m\). To introduce the rPDDP model, we randomize the PDDP model by sampling the \(P_{jl}\) measures from the independent Dirichlet processes \(\mathcal{DP}(c_{jl}, P_0)\) and then impose gamma priors on the concentration masses, i.e. \(P_{jl} \overset{\text{ind}}{\sim} \mathcal{DP}(c_{jl}, P_0), \quad c_{jl} \overset{\text{ind}}{\sim} \mathcal{G}(a_{jl}, b_{jl}), \quad 1 \leq j \leq l \leq m.\)
B. To develop a pairwise dependent geometric stick breaking version, we let the random density functions $f_j(x)$ generated via

$$f_j(x) := f_j(x|Q_j) = \sum_{l=1}^{m} p_{jl} g_{jl}(x|G_{jl}), \quad Q_j = \sum_{l=1}^{m} p_{jl} G_{jl}, \quad 1 \leq j \leq m. \quad (2)$$

The $g_{jl}(x) := g_{jl}(x|G_{jl}) = \int_{\Theta} K(x|\theta) G_{jl}(d\theta)$ random densities are now independent mixtures of GSB processes, satisfying $g_{jl} = g_{j},$ under the slightly altered definition

$$G_{jl} = \sum_{k=1}^{\infty} q_{jlk} \delta_{\theta_{jlk}} \quad \text{with} \quad q_{jlk} = \lambda_{jl}(1 - \lambda_{jl})^{k-1}, \quad \lambda_{jl} \sim h(\cdot|\xi_{jl}), \quad \theta_{jlk} \overset{iid}{\sim} G_0, \quad (3)$$

where $h$ is a parametric density supported over the interval $(0, 1)$ depending on some parameter $\xi_{jl} \in \Xi,$ and $G_0$ is the associated parametric central measure.

The independent GSB processes $\{G_{jl} : 1 \leq j, l \leq m\}$ form a matrix $G$ of random distributions with $G_{jl} = G_{lj}$. In matrix notation

$$Q = (p \otimes G) \mathbf{1}, \quad (4)$$

where $p = (p_{jl})$ is the matrix of random selection weights, and $p \otimes G$ is the Hadamard product of the two matrices defined as $(p \otimes G)_{jl} = p_{jl} G_{jl}.$ By letting $\mathbf{1}$ to denote the $m \times 1$ matrix of ones it is that the $j$th element of vector $Q$ is given by equation (2).

C. Following a univariate construction of geometric slice sets (Fuentes–García et al. 2010), we define the stochastic variables $N = (N_{ji})$ for $1 \leq i \leq n_j$ and $1 \leq j \leq m$, where $N_{ji}$ is an almost surely finite random variable of mass $f_N$ possibly depending on parameters, associated with the sequential slice set $S_{ji} = \{1, \ldots, N_{ji}\}$. Following Hatjispyros et al. (2011, 2016a) we introduce:

1. The GSB mixture selection variables $\delta = (\delta_{ji})$; for an observation $x_{ji}$ that comes from $f_j(x)$, $\delta_{ji}$ selects one of the mixtures $\{g_{jl}(x) : l = 1, \ldots, m\}$. Then the observation $x_{ji}$ came from mixture $g_{j\delta_{ji}}(x)$.

2. The GSB clustering variables $d = (d_{ji})$; for an observation $x_{ji}$ that comes from $f_j(x)$, given $\delta_{ji}$, $d_{ji}$ allocates the component of the GSB mixture $g_{j\delta_{ji}}(x)$ that $x_{ji}$ came from. Then the observation $x_{ji}$ came from component $K(x|\theta_{j\delta_{ji}d_{ji}})$.

In what follows, unless otherwise specified, the random densities $f_j(x)$ are mixtures of independent GSB mixtures.

**Proposition 1.** Suppose that the clustering variables $(d_{ji})$ conditionally on the slice variables $(N_{ji})$ are having the discrete uniform distribution over the sets $(S_{ji})$ that is $d_{ji}|N_{ji} \sim DU(S_{ji})$, and $P\{N_{ji} = r|\delta_{ji} = l\} = f_N(r|\lambda_{jl})$, then

$$f_j(x_{ji}, N_{ji} = r) = r^{-1} \sum_{l=1}^{m} p_{jl} f_N(r|\lambda_{jl}) \sum_{k=1}^{r} K(x_{ji}|\theta_{jlk}), \quad (5)$$
and
\[ f_j(x_{ji}, N_{ji} = r, d_{ji} = k|\delta_{ji} = l) = \frac{1}{r} f_N(r|\lambda_{jl}) \mathcal{I}(k \leq r) K(x_{ji}|\theta_{jlk}). \tag{6} \]

The proof is given in Appendix A.

The following proposition gives a multivariate analogue of equation (2) in Fuentes–García, et al. (2010):

**Proposition 2.** Given the random set \( S_{ji} \), the random functions in (2) become finite mixtures of a.s. finite equally weighted mixtures of the \( K(\cdot|\cdot) \) probability kernels, that is
\[ f_j(x_{ji}|N_{ji} = r) = \sum_{l=1}^{m} W(r|\lambda_{jl}) \sum_{k=1}^{r} r^{-1} K(x_{ji}|\theta_{jlk}), \tag{7} \]
where the probability weights \( \{W(r|\lambda_{jl}) : 1 \leq l \leq m\} \) are given by
\[ W(r|\lambda_{jl}) = \frac{p_{jl} f_N(r|\lambda_{jl})}{\sum_{l'=1}^{m} p_{jl'} f_N(r|\lambda_{jl'})}. \]

The proof is given in Appendix A.

Note that, the one–dimensional model introduced in Fuentes–García et al. (2010), under our notation attains the representation
\[ f_j(x_{ji}|N_{ji} = r, \delta_{ji} = l) = \sum_{k=1}^{r} r^{-1} K(x_{ji}|\theta_{jlk}). \]

2.1 The model. Marginalizing (6) with respect to the variable \((N_{ji}, d_{ji})\), we obtain
\[ f_j(x_{ji}|\delta_{ji} = l) = \sum_{k=1}^{\infty} \left( \sum_{r=k}^{\infty} r^{-1} f_N(r|\lambda_{jl}) \right) K(x_{ji}|\theta_{jlk}). \tag{8} \]

The quantity inside the parentheses on the right-hand side of the previous equation is \( f_j(d_{ji}|\delta_{ji} = l) \). Following Fuentes–García, et al. (2010), we substitute \( f_N(r|\lambda_{jl}) \) with the negative binomial distribution \( NB(r|2, \lambda_{jl}) \), i.e.
\[ f_N(r|\lambda_{jl}) = r \lambda_{jl}^2 (1 - \lambda_{jl})^{r-1} \mathcal{I}(r \geq 1), \tag{9} \]
so equation (8) becomes
\[ f_j(x_{ji}|\delta_{ji} = l) = \sum_{k=1}^{\infty} q_{jlk} K(x_{ji}|\theta_{jlk}) \] with \( q_{jlk} = \lambda_{jl} (1 - \lambda_{jl})^{k-1} \),
and the \( f_j \) random densities take the form of a finite mixture of GSB mixtures
\[ f_j(x_{ji}) = \sum_{l=1}^{m} p_{jl} \sum_{k=1}^{\infty} q_{jlk} K(x_{ji}|\theta_{jlk}). \]
We denote the set of observations along the \( m \) groups as \( \mathbf{x} = (x_{ji}) \) and with \( \mathbf{x}_j \) the set of observations in the \( j \)th group. The three sets of latent variables in the \( j \)th group will be denoted as \( \mathbf{N}_j \) for the slice variables, \( \mathbf{d}_j \) for the clustering variables, and finally \( \mathbf{\delta}_j \) for the set of GSB mixture allocation variables. From now on, we are going to leave the auxiliary variables unspecified; especially for \( \delta_{ji} \) we use the notation \( \delta_{ji} = (\delta_{ji}^1, \ldots, \delta_{ji}^m) \in \{e_1, \ldots, e_m\} \) with \( \mathbb{P}\{\delta_{ji} = e_i\} = p_{ji} \), where \( e_i \) denotes the usual basis vector having its only nonzero component equal to 1 at position \( l \). Hence, for a sample of size \( n_1 \) from \( f_1 \), a sample of size \( n_2 \) from \( f_2 \), etc., a sample of size \( n_m \) from \( f_m \) we can write the full likelihood as a multiple product:

\[
f(\mathbf{x}, \mathbf{N}, \mathbf{d} | \mathbf{\delta}) = \prod_{j=1}^{m} f(\mathbf{x}_j, \mathbf{N}_j, \mathbf{d}_j | \mathbf{\delta}_j)
\]

\[
= \prod_{j=1}^{m} \prod_{i=1}^{n_j} \mathcal{I}(d_{ji} \leq N_{ji}) \prod_{l=1}^{m} \left\{ \lambda_{ji}^2 (1 - \lambda_{ji})^{N_{ji} - 1} K(x_{ji} | \theta_j d_{ji}) \right\}^{\delta_{ji}}.
\]

In a hierarchical fashion, using the auxiliary variables, we have for \( j = 1, \ldots, m \) and \( i = 1, \ldots, n_j \),

\[
x_{ji}, N_{ji}, d_{ji}, \delta_{ji}, (\theta_j \delta_{ji})_{1 \leq r \leq m}, \lambda_{ji} \delta_{ji} \sim \prod_{r=1}^{m} \left\{ \lambda_{jr}^2 (1 - \lambda_{jr})^{N_{jr} - 1} K(x_{ji} | \theta_j d_{ji}) \right\}^{\delta_{ji}} \mathcal{I}(N_{ji} \geq d_{ji})
\]

\[
d_{ji} \mid N_{ji} \sim \mathcal{DU}(S_{ji}), \quad \mathbb{P}\{\delta_{ji} = e_i\} = p_{ji}
\]

\[
q_{jik} = \lambda_{ji}(1 - \lambda_{ji})^{k-1}, \quad \theta_{jik} \sim G_0, \quad k \in \mathbb{N}.
\]

2.2 The PDGSBP covariance and correlation. In this sub-section we find the covariance and the correlation between \( f_j(x) \) and \( f_i(x) \). First we provide the following lemma.

**Lemma 1.** Let \( g_G(x) = \int_\Theta K(x|\theta)G(d\theta) \) be a random density, with \( G = \lambda \sum_{j=1}^{\infty} (1 - \lambda)^{j-1} \delta_{\theta_j} \) and \( \theta_j \sim G_0 \), then

\[
\mathbb{E}[g_G(x)^2] = \left( \frac{1}{2 - \lambda} \right) \left\{ \lambda \int_\Theta K(x|\theta)^2 G_0(d\theta) + 2(1 - \lambda) \left( \int_\Theta K(x|\theta)G_0(d\theta) \right)^2 \right\}.
\]

The proof is given in Appendix A.

**Proposition 3.** It is that

\[
\text{Cov}(f_j(x), f_i(x)) = p_{ji} p_{ij} \text{Var} \left( \int_\Theta K(x|\theta)G_{ji}(d\theta) \right), \tag{10}
\]

with

\[
\text{Var} \left( \int_\Theta K(x|\theta)G_{ji}(d\theta) \right) = \frac{\lambda_{ji}}{2 - \lambda_{ji}} \text{Var}(K(x|\theta)). \tag{11}
\]

The proof is given in Appendix A.
Suppose now that \((f_j^P(x))_{1 \leq j \leq m} \) and \((f_i^G(x))_{1 \leq j \leq m} \) are two collections of \(m\) DP and \(m\) GSB pairwise dependent random densities respectively, i.e. \(f_j^P(x) = \sum_{i=1}^{m} p_{ji} g_{ji}^P(x) \) with \(g_{ji}^P(x) = g_{ji}(x|\mathbb{P}_{ji}) \), and \(f_i^G(x) = \sum_{j=1}^{m} p_{ij} g_{ij}^G(x) \) with \(g_{ij}^G(x) = g_{ij}(x|\mathbb{G}_{ji}) \). Then we have the following proposition:

**Proposition 4.** For given parameters \((\lambda_{ji}), (c_{ji}), \) and matrix of selection probabilities \((p_{ji})\) it is that

1. The PDGSBP and rPDDP correlations are given by

\[
\text{Corr}(f_j^G(x), f_i^G(x)) = \frac{\lambda_{ji} p_{ji} p_{ij}}{2 - \lambda_{ji}} \left( \sum_{i=1}^{m} \sum_{r=1}^{m} \frac{p_{ji}^2 p_{ir}^2 \lambda_{ji} \lambda_{ir}}{(2 - \lambda_{ji})(2 - \lambda_{ir})} \right)^{-1/2},
\]

and

\[
\text{Corr}(f_j^P(x), f_i^P(x)) = \frac{p_{ji} p_{ij}}{1 + c_{ji}} \left( \sum_{i=1}^{m} \sum_{r=1}^{m} \frac{p_{ji}^2 p_{ir}^2}{(1 + c_{ji})(1 + c_{ir})} \right)^{-1/2}.
\]

2. When \(\lambda_{ji} = \lambda\) and \(c_{ji} = c\) for all \(1 \leq j \leq i \leq m\), the expressions for the rPDDP and PDGSBP correlations simplify to

\[
\text{Corr}(f_j^G(x), f_i^G(x)) = \text{Corr}(f_j^P(x), f_i^P(x)) = p_{ji} p_{ij} \left( \sum_{i=1}^{m} \sum_{r=1}^{m} p_{ji}^2 p_{ir}^2 \right)^{-1/2}.
\]

The proof is given in Appendix A.

It is clear that, irrespective of the model, the random densities \(f_j(x)\) and \(f_i(x)\) are positively correlated whenever \(p_{ji} = p_{ij} = 1\). Similarly, the random densities \(f_j(x)\) and \(f_i(x)\) are independent (have no common part) whenever \(p_{ji} = p_{ij} = 0\). Another, less obvious feature, upon synchronization, is the ability of controlling the correlation among the models. For example, suppose that for \(m = 2\), the random densities \(f_1(x)\) and \(f_2(x)\) are dependent, and that \(\lambda_{ji} = (1 + c_{ji})^{-1}\); then consider the expression

\[
D_{12} := \lambda_{12}^2 p_{12}^2 p_{21}^2 \left\{ \text{Corr}(f_1^G(x), f_2^G(x))^{-2} - \text{Corr}(f_1^P(x), f_2^P(x))^{-2} \right\}.
\]

Since correlations are positive, \(D_{12} \geq 0\) whenever \(\text{Corr}(f_1^G(x), f_2^G(x)) \leq \text{Corr}(f_1^P(x), f_2^P(x))\), and that \(D_{12} < 0\) whenever \(\text{Corr}(f_1^G(x), f_2^G(x)) > \text{Corr}(f_1^P(x), f_2^P(x))\). Then, it is not difficult to see that

\[
D_{12} = \left( p_{12}^2 \lambda_{12} + r_1 p_{11}^2 \lambda_{11} \right) \left( p_{21}^2 \lambda_{12} + r_2 p_{22}^2 \lambda_{22} \right) - \left( p_{12}^2 \lambda_{12} + p_{11}^2 \lambda_{11} \right) \left( p_{21}^2 \lambda_{12} + p_{22}^2 \lambda_{22} \right)
\]

with \(r_k = (2 - \lambda_{12})(2 - \lambda_{kk}), k = 1, 2\). We have the following cases:

1. \(\lambda_{12} > \max\{\lambda_{11}, \lambda_{22}\} \Leftrightarrow r_1 < 1, r_2 < 1 \Leftrightarrow \text{Corr}(f_1^G(x), f_2^G(x)) > \text{Corr}(f_1^P(x), f_2^P(x))\).
2. \( \lambda_{12} < \min\{\lambda_{11}, \lambda_{22}\} \Leftrightarrow r_1 > 1, r_2 > 1 \Leftrightarrow \text{Corr}(f_1^{D}(x), f_2^{D}(x)) < \text{Corr}(f_1^{P}(x), f_2^{P}(x)). \)

3. \( \lambda_{12} = \lambda_{11} = \lambda_{22} \Leftrightarrow r_1 = r_2 = 1 \Leftrightarrow \text{Corr}(f_1^{D}(x), f_2^{D}(x)) = \text{Corr}(f_1^{P}(x), f_2^{P}(x)). \)

3. **The PDGSBP Gibbs sampler.** In this section we will describe the PDGSBP Gibbs sampler for estimating the model. The details for the sampling algorithm of the PDDP model can be found in Hatjispyros et al. (2011, 2016A). At each iteration we will sample the variables,

\[
\theta_{jlk}, 1 \leq j \leq l \leq m, 1 \leq k \leq N^*,
\]

\[
d_{ji}, N_{ji}, \delta_{ji}, 1 \leq j \leq m, 1 \leq i \leq n_j,
\]

\[
p_{jl}, 1 \leq j \leq m, 1 \leq l \leq m,
\]

with \( N^* = \max_{j,i} N_{ji} \) being almost surely finite.

1. For the locations of the random measures for \( k = 1, \ldots, d^* \) where \( d^* = \max_{j,i} d_{ji} \), it is that

\[
f(\theta_{jlk} | \cdots) \propto f(\theta_{jlk}) \begin{cases} \prod_{i=1}^{n_j} K(x_{ji} | \theta_{jlk})^I(\delta_{ji} = e_i, d_{ji} = k) \prod_{i=1}^{n_l} K(x_{li} | \theta_{jlk})^I(\delta_{ji} = e_j, d_{ji} = k) & l > j, \\ \prod_{i=1}^{n_j} K(x_{ji} | \theta_{jlk})^I(\delta_{ji} = e_j, d_{ji} = k) & l = j. \end{cases}
\]

If \( N^* > d^* \) we sample additional locations \( \theta_{ji,d^*+1}, \ldots, \theta_{ji,N^*} \) independently from the prior.

2. Here we sample the allocation variables \( d_{ji} \) and the mixture component indicator variables \( \delta_{ji} \) as a block. For \( j = 1, \ldots, m \) and \( i = 1, \ldots, n_j \), we have

\[
P(d_{ji} = k, \delta_{ji} = e_i | N_{ji} = r, \cdots) \propto p_{jl} K(x_{ji} | \theta_{jlk}) I(l \leq m) I(k \leq r).
\]

3. The slice variables \( N_{ji} \) have full conditional distributions given by

\[
P(N_{ji} = r | \delta_{ji} = e_i, d_{ji} = l, \cdots) \propto (1 - \lambda_{ji})^r I(r \geq l),
\]

which are truncated geometric distributions over the set \( \{l, l+1, \ldots\} \).

4. The full conditional for \( j = 1, \ldots, m \) for the selection probabilities \( p_j = (p_{j1}, \ldots, p_{jm}) \), under a Dirichlet prior \( f(p_j | a_j) \propto \prod_{i=1}^{m} p_j^{a_j-1} \), with hyperparameter \( a_j = (a_{j1}, \ldots, a_{jm}) \), is Dirichlet

\[
f(p_j | \cdots) \propto \prod_{i=1}^{m} p_j^{a_j + \sum_{i=1}^{m} l I(\delta_{ji} = e_i) - 1}.
\]

5. Here we update the geometric probabilities \( (\lambda_{jl}) \) of the GSB measures. For \( 1 \leq j \leq l \leq m \), it is that

\[
f(\lambda_{jl} | \cdots) \propto f(\lambda_{jl}) \begin{cases} \prod_{i=1}^{n_j} \{\lambda_{jl}^2 (1 - \lambda_{jl})^{N_{jl} - 1}\}^I(\delta_{jl} = e_i) \prod_{i=1}^{n_l} \{\lambda_{jl}^2 (1 - \lambda_{jl})^{N_{jl} - 1}\}^I(\delta_{jl} = e_j) & l > j, \\ \prod_{i=1}^{n_j} \{\lambda_{jl}^2 (1 - \lambda_{jl})^{N_{jl} - 1}\}^I(\delta_{jl} = e_j) & l = j. \end{cases}
\]
To complete the model, we assign priors to the geometric probabilities. For a fair comparison of the execution time between the two models, we apply
\[ \lambda_{jl} = (1 + c_{jl})^{-1} \]
transformed priors. So, by placing gamma priors \( c_{jl} \sim G(a_{jl}, b_{jl}) \) over the concentration masses \( c_{jl} \) of the PDDP model, we have
\[
f(\lambda_{jl}) = T G(\lambda_{jl} | a_{jl}, b_{jl}) \propto \lambda_{jl}^{-a_{jl} - 1} e^{-b_{jl}/\lambda_{jl}} (1 - \lambda_{jl})^{a_{jl} - 1} I(0 < \lambda_{jl} < 1).
\]
In the Appendix, we give the full conditionals for \( \lambda_{jl} \)’s, their corresponding embedded Gibbs sampling schemes, and the sampling algorithm for the concentration masses.

3.1 The complexity of the rPDDP and PDGSBP samplers. The main difference between the two samplers in terms of execution time, comes from the blocked sampling of the clustering and the mixture indicator variables \( \delta_{ji} \) and \( \delta_{ji} \).

The rPDDP model: The state space of the variable \((d_{ji}, \delta_{ji})\) conditionally on the slice variable \( u_{ji} \) is \( (d_{ji}, \delta_{ji})(\Omega) = \bigcup_{l=1}^{m} (A_{w_{jl}}(u_{ji}) \times \{e_l\}) \), where \( A_{w_{jl}}(u_{ji}) = \{ r \in \mathbb{N} : u_{ji} < w_{jlr} \} \) is the a.s. finite slice set corresponding to the observation \( x_{ji} \) (Walker, 2007). At each iteration of the Gibbs sampler, we have \( m(m+1)/2 \) vectors of stick-breaking weights \( w_{jl} \), each of length \( N_{jl}^{*} \); where \( N_{jl}^{*} \sim 1 + \text{Poisson}(-c_{jl} \log u_{jl}^{*}) \) with \( c_{jl} \) being the concentration parameter of the Dirichlet process \( P_{jl} \) and \( u_{jl}^{*} \) being the minimum of the slice variables in densities \( f_j \) and \( f_l \). Algorithm 1 gives the blocked sampling procedure of the clustering and mixture indicator variables. An illustration of the effect of the slice variable \( u_{ji} \) is given in Figure 1(a).

Algorithm 1: rPDDP

1: procedure SAMPLE \((d_{ji}, \delta_{ji})\)
2: for random densities \( f_j \), \( j = 1 \) to \( m \) do
3: for each data point \( x_{ji} \in f_j \ i = 1 \) to \( n_j \) do
4: for each mixture component \( K(x_{ji}|\theta_{jl}) \), \( l = 1 \) to \( m \) do
5: Construct slice sets \( A_{w_{jl}}(u_{ji}) \)
6: end for
7: Sample \((d_{ji} = k, \delta_{ji} = r|\cdots) \propto K(x_{ji}|\theta_{jrk}) I((k, r) \in \bigcup_{l=1}^{m} (A_{w_{jl}}(u_{ji}) \times \{e_l\})) \)
8: end for
9: end for
10: end procedure

Since the weights forming the stick-breaking representation are not in an ordered form, the construction of the slice sets in step 5 of Algorithm 1 requires a complete search in the array where the weights are stored. This operation is done in \( O(N_{jl}^{*}) \) time. For the sampling of the \( d_{ji} \) and \( \delta_{ji} \) variables in step 6, the choice of their value is an element from the union \( \bigcup_{l=1}^{m} (A_{w_{jl}}(u_{ji}) \times \{e_l\}) \). This means that the rPDDP algorithm for each \( j \), must create \( m \) slice sets which require \( N_{jl}^{*} \) comparisons each. The worst case scenario is that the sampled \((d_{ji}, \delta_{ji})\) is
the last element of $\bigcup_{l=1}^{m} \left( A_{wji}(u_{ji}) \times \{e_l\} \right)$. Thus, the DP based procedure of sampling $(d_{ji}, \delta_{ji})$ is of order
\[
\mathcal{O} \left( m^2 n_j N^*_j \sum_{l=1}^{m} |A_{wji}(u_{ji})| \right) = \mathcal{O} \left( N^*_j \sum_{l=1}^{m} |A_{wji}(u_{ji})| \right).
\]

The **PDGSBP model**: The state space of the variable $(d_{ji}, \delta_{ji})$ conditionally on the slice variable $N_{ji}$ is $(d_{ji}, \delta_{ji})(\Omega) = \bigcup_{l=1}^{m} \left( S_{ji} \times \{e_l\} \right)$. In the GSB case, the slice variable has a different rôle. It indicates at which random point the search for the appropriate $d_{ji}$ will stop. In Figure 1(b) we illustrate this argument. In Algorithm 2 the worst case scenario is that the sampled $(d_{ji}, \delta_{ji})$ will be the last element of $\bigcup_{l=1}^{m} \left( S_{ji} \times \{e_l\} \right)$. Thus, the GSB based procedure of sampling $(d_{ji}, \delta_{ji})$ is of order $\mathcal{O} \left( m^2 n_j N_{ji} \right) = \mathcal{O} \left( N_{ji} \right)$.

**Algorithm 2 : PDGSBP**

1: procedure SAMPLE $(d_{ji}, \delta_{ji})$
2: for random densities $f_j$, $j = 1$ to $m$ do
3: for each data point $x_{ji} \in f_j$ $i = 1$ to $n_j$ do
4: for each mixture component $K(x_{ji} | \theta_{jl})$, $l = 1$ to $m$ do
5: Sample $(d_{ji} = k, \delta_{ji} = r | \cdots) \propto K(x_{ji} | \theta_{jrk}) \mathcal{I}(k \leq N_{ji}) \mathcal{I}(r \leq m)$
6: end for
7: end for
8: end for
9: end procedure

(a) Stick-breaking weights for some $N^*_j = 20$. The red dashed line represents the slice variable $u_{ji} = 0.05$. The algorithm must check all the $N^*_j$ values to accept those that they satisfy $u_{ji} < w_{jlk}$. After a complete search, the slice set is $A_{wji}(u_{ji}) = \{1, 2, 3, 5, 7, 8\}$.

(b) Geometric stick-breaking weights for $N^*_j = 20$. The red dashed line represents the slice variable $N_{ji} = 6$. The slice set is simply $S_{ji} = \{1, 2, 3, 4, 5, 6\}$.

Figure 1: A visualization of the effect of the $u_{ji}$ and $N_{ji}$ slice variables are given in Figures 1(a) and 1(b) respectively.
4. Illustrations. In this section we illustrate the efficiency of the PDGSBP model. For the choice of a normal kernel (unless otherwise specified) \( K(x|\theta) = \mathcal{N}(x|\theta) \) where \( \theta = (\mu, \tau^{-1}) \) and \( \tau = \sigma^{-2} \) is the precision. The prior over the means and precisions of the PDGSBP \((G_0)\) and the rPDDP model \((P_0)\) is the independent normal-gamma measure, given by

\[
P_0(d\mu, d\tau) = G_0(d\mu, d\tau) = \mathcal{N}(\mu | \mu_0, \tau_0^{-1}) \mathcal{G}(\tau | \epsilon_1, \epsilon_2) d\mu d\tau.
\]

Attempting a noninformative prior specification (unless otherwise specified), we took \( \mu_0 = 0 \) and \( \tau_0 = \epsilon_1 = \epsilon_2 = 10^{-3} \). For the concentration masses of the rPDDP model, a-priori, we set \( a_{jl} \sim \mathcal{G}(\alpha_{jl}, \beta_{jl}) \). For an objective evaluation of the execution time, of the two algorithms under different scenarios, we choose a synchronized prior specification, namely, for the geometric probabilities, we set \( \lambda_{jl} \sim T\mathcal{G}(a_{jl}, b_{jl}) - \) the transformed gamma density given in equation (14). In the appendix B, we show that such prior specifications are valid for \( a_{jl} > 1 \). In all our numerical examples, we took \( a_{jl} = b_{jl} = 1.1 \). For our numerical experiments (unless otherwise specified), the hyperparameters \((\alpha_{jl})\) of the Dirichlet priors over the matrix of the selection probabilities \( p = (p_{jl}) \) has been set to \( \alpha_{jl} = 1 \).

In all cases, we measure the similarity between probability distributions with the Hellinger distance. So for example, \( H(\mathcal{G}(f, \hat{f})) \) and \( H(\mathcal{D}(f, \hat{f})) \), will denote the Hellinger distance between the true density \( f \) and the predictive density \( \hat{f} \) of the PDGSBP and rPDDP algorithms, respectively. The Gibbs samplers run for \( 11 \times 10^4 \) iterations leaving the first \( 10^4 \) samples as a burn-in period.

4.1 Time execution efficiency of the PDGSBP model.

**Nested normal mixtures with a unimodal common and idiosyncratic part:** Here, we choose to include all pairwise and idiosyncratic dependences in the form of unimodal equally weighted normal mixture components. The mixture components are well separated with unit variance. We define each data model \( \mathcal{M}_m = \{f_j^{(m)} : 1 \leq j \leq m\} \) of dimension \( m \in \{2, 3, 4\} \), based on a \( 4 \times 10 \) matrix \( M = (M_{jk}) \), with entries in the set \( \{0, 1\} \), having at most two ones in each column and exactly four ones in each row. When there is exactly one entry of one, the column defines an idiosyncratic part. The appearance of exactly two ones in a column defines a common component. We let the matrix \( M \) given by

\[
M = \begin{bmatrix}
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1
\end{bmatrix},
\]

and for \( m \in \{2, 3, 4\} \), we define

\[
\mathcal{M}_m : f_j^{(m)}(x) \propto \sum_{k=5-m}^{2(m+1)} M_{jk} \mathcal{N}(x|10(k - 6), 1), \quad 1 \leq j \leq m,
\]

We are taking independently samples of sizes \( n_j^{(2)} = 60 \) from the \( f_j^{(2)}\)'s, \( n_j^{(3)} = 120 \) from the \( f_j^{(3)}\)'s, and, \( n_j^{(4)} = 200 \) from the \( f_j^{(4)}\)'s. In all cases, the PDGSBP and the rPDDP density estimations are of the same quality.
In Figures 2(a)–(d) we give the histograms of the data sets for the specific case $m = 4$, which are overlaid with the kernel density estimations (KDE’s) based on the predictive samples of the $f_j^{(4)}$’s coming from the PDGSBP (solid line) and the rPDDP (dashed line) models. The differences between the two models are nearly indistinguishable. The Hellinger distances between the true and the estimated densities for the case $m = 4$ are given in table 1.

In Table 2 we summarize the mean execution times (MET’s) per $10^3$ iterations in seconds. The PDGSBP sampler is about three times faster than the rPDDP sampler. The corresponding MET ratios for $m = 2, 3$ and 4 are 2.96, 3.04 and 3.37 respectively. We can see that the PDGSBP Gibbs sampler gives slightly faster execution times with increasing $m$. This will become more clear in our next simulated data example, where the average sample size per mode is being kept constant.

![Figure 2](image-url)  
Figure 2: Histograms of data sets coming for the case $m = 4$. The superimposed KDE’s are based on the predictive samples obtained from the PDGSBP and the rPDDP models.

| $i$ | $\mathcal{H}_G(f_i^{(4)}, \hat{f}_i^{(4)})$ | $\mathcal{H}_D(f_i^{(4)}, \hat{f}_i^{(4)})$ |
|-----|--------------------------------|--------------------------------|
| 1   | 0.17                           | 0.17                           |
| 2   | 0.19                           | 0.18                           |
| 3   | 0.22                           | 0.22                           |
| 4   | 0.20                           | 0.20                           |

Table 1: Hellinger distances for the case $m = 4$.  

---
Sparse $m$–scalable data set models: In this example, we attempt to create $m$-scalable normal mixture data sets of the lowest possible sample size. To this respect, we sample independently $m$ groups of data sets from the densities

$$f_j^{(m)}(x) \propto N(x|(j-1)\xi, 1)\mathbb{I}(1 \leq j < m) + \sum_{k=1}^{m-1} N(x|(k-1)\xi, 1)\mathbb{I}(j = m),$$

with sample sizes $n_j^{(m)} = n\{\mathbb{I}(1 \leq j < m) + (m-1)\mathbb{I}(j = m)\}$. We have chosen $\xi = 10$ and an average sample size per mode of $n = 20$, for $m \in \{2, \ldots, 10\}$.

In Figure 3 we depict the average execution times as functions of the dimension $m$. We can see how fast the two MET-curves diverge with increasing $m$. In Figure 4(a)–(j), for the specific case $m = 10$, we give the histograms of the data sets, overlaid with the KDE’s based on the predictive samples of the $f_j^{(10)}$’s coming from the PDGSBP (solid line) and the rPDDP (dashed line) models. We can see that the PDGSBP and the rPDDP density estimations are of the same quality.

The Hellinger distances between the true and the estimated densities for the specific case $m = 10$ are given in Table 3. The large values of the Hellinger distances $\mathcal{H}_G(f_{10}^{(10)}, \hat{f}_{10}^{(10)}) \approx \mathcal{H}_D(f_{10}^{(10)}, \hat{f}_{10}^{(10)}) \approx 0.22$, are caused by the enlargement of the variances of the underrepresented modes due to the small sample size.

![Figure 3: Mean execution times for the two models, based on the sparse $m$-scalable data sets.](image)
Figure 4: Histograms of sparse $m$-scalable data sets for the case $m = 10$. The superimposed KDE’s are based on the predictive samples of the PDGSBP and the rPDDP models.

| $i$ | $1$ | $2$ | $3$ | $4$ | $5$ | $6$ | $7$ | $8$ | $9$ | $10$ |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| $\mathcal{H}_G(f^{(10)}_i, \hat{f}^{(10)}_i)$ | 0.08 | 0.10 | 0.09 | 0.14 | 0.14 | 0.13 | 0.14 | 0.09 | 0.11 | 0.22 |
| $\mathcal{H}_D(f^{(10)}_i, \hat{f}^{(10)}_i)$ | 0.09 | 0.11 | 0.10 | 0.15 | 0.12 | 0.10 | 0.14 | 0.09 | 0.09 | 0.22 |

Table 3: Hellinger distances between true and estimated densities for the case $m = 10$ of the sparse scalable data example.

4.2 Normal and gamma mixture models that are not well separated.

The normal mixture example: We will first consider a normal model for $m = 2$, first appeared in Lijoi et. al (2014B). The data models for $f_1$ and $f_2$ are 7-mixtures. Their common part is a 4-mixture that is weighted differently between the two mixtures. More specifically, we sample two data sets of sample size $n_1 = n_2 = 200$, independently from

$$(f_1, f_2) = \left( \frac{1}{7} g_{11} + \frac{1}{7} g_{12}, \frac{4}{7} g_{21} + \frac{3}{7} g_{22} \right),$$

with

$$g_{11} = \frac{2}{7} \mathcal{N}(-8, 0.25^2) + \frac{3}{7} \mathcal{N}(1, 0.5^2) + \frac{2}{7} \mathcal{N}(10, 1),$$
$$g_{12} = \frac{1}{7} \mathcal{N}(-10, 0.5^2) + \frac{3}{7} \mathcal{N}(-3, 0.75^2) + \frac{1}{7} \mathcal{N}(3, 0.25^2) + \frac{2}{7} \mathcal{N}(7, 0.25^2),$$
$$g_{21} = \frac{2}{8} \mathcal{N}(-10, 0.5^2) + \frac{3}{8} \mathcal{N}(-3, 0.75^2) + \frac{2}{8} \mathcal{N}(3, 0.25^2) + \frac{1}{8} \mathcal{N}(7, 0.25^2),$$
$$g_{22} = \frac{1}{3} \mathcal{N}(-6, 0.5^2) + \frac{1}{3} \mathcal{N}(-1, 0.25^2) + \frac{1}{3} \mathcal{N}(5, 0.5^2).$$
For this case, a-priori we took \((\mu_0, \tau_0, \epsilon_1, \epsilon_2) = (0, 10^{-3}, 1, 10^{-2})\).

In Figure 5(a)–(b) we give the histograms of the data sets, with the predictive densities of the PDGSBP and rPDDP models superimposed in black solid and black dashed curves, respectively. We can see that the PDGSBP and the rPDDP density estimations are of the same quality. In Table 4, we give the Hellinger distance between the true and the estimated densities.

![Figure 5: Density estimations of the 7-mixtures data sets, under the PDGSBP and the rPDDP models. The true densities have been superimposed in red.](image)

| \(i\) | \(H_G(f_i, \hat{f}_i)\) | \(H_D(f_i, \hat{f}_i)\) |
|------|-----------------|-----------------|
| 1    | 0.19            | 0.18            |
| 2    | 0.18            | 0.15            |

Table 4: Hellinger distance between the true and the estimated densities.

**The gamma mixture example:** In this example we took \(m = 2\). The data models for \(f_1\) and \(f_2\) are gamma 4-mixtures. The common part is a gamma 2-mixture, weighted identically among the two mixtures. More specifically, we sample two data sets of sample size \(n_1 = n_2 = 160\), independently from

\[
(f_1, f_2) = \left( \frac{2}{5} g_{11} + \frac{3}{5} g_{12}, \frac{7}{10} g_{12} + \frac{3}{10} g_{22} \right),
\]

with

\[
g_{11} = \frac{2}{3} G(2, 1.1) + \frac{1}{3} G(80, 2),
g_{12} = \frac{8}{14} G(10, 0.9) + \frac{6}{14} G(200, 8.1),
g_{22} = \frac{2}{3} G(105, 3) + \frac{1}{3} G(500, 10),
\]
Because we want to estimate the density of non-negative observations, we find it more appropriate to take the kernel to be a log-normal distribution (Hatjispyros et al. 2016B). That is \( K(x|\theta) = \mathcal{LN}(x|\theta) \) with \( \theta = (\mu, \sigma^2) \), is the log-normal density with mean \( \exp(\mu + \sigma^2/2) \). For this case, a-priori we set

\[
(\mu_0, \tau_0, \epsilon_1, \epsilon_2) = (\bar{S}, 0.5, 2, 0.01), \quad \bar{S} = \frac{1}{n_1 + n_2} \left( \sum_{j=1}^{n_1} \log x_{1j} + \sum_{j=1}^{n_2} \log x_{2j} \right).
\]

In Figure 6(a)-(b), we display the KDE’s based on the predictive samples of the two models. We can see that the PDGSBP and the rPDDP density estimations are of the same quality. In Table 5, we give the Hellinger distances.

![Figure 6](image_url)

Table 5: Hellinger distances for the gamma mixture data model.

| i | \( \mathcal{H}_G(f_i, \hat{f}_i) \) | \( \mathcal{H}_D(f_i, \hat{f}_i) \) |
|---|---|---|
| 1 | 0.13 | 0.11 |
| 2 | 0.19 | 0.18 |

4.3 Borrowing of strength of the PDGSBP model. In this example we consider three populations \( \{D_j^{(s)} : j = 1, 2, 3\} \), under three different scenarios \( s \in \{1, 2, 3\} \). The sample sizes are always the same, namely, \( n_1 = 200 \), \( n_2 = 50 \) and \( n_3 = 200 \) – the second population is
sampled only once. The three data sets $D_1^{(s)}$, $D_2^{(s)}$ and $D_3^{(s)}$, are sampled independently from the normal mixtures

$$(f_1^{(s)}, f_2^{(s)}, f_3^{(s)}) = ((1 - q^{(s)})f + q^{(s)}g_1, f, (1 - q^{(s)})f + q^{(s)}g_2),$$

where

$$f = \frac{3}{10} N(-10, 1) + \frac{2}{10} N(-6, 1) + \frac{2}{10} N(6, 1) + \frac{3}{10} N(10, 1),$$

$$g_1 = \frac{1}{2} N(-4, 1) + \frac{1}{2} N(4, 1),$$

$$g_2 = \frac{1}{2} N(-12, 1) + \frac{1}{2} N(12, 1).$$

More specifically, the three scenarios are:

1. For $s = 1$, we set, $q^{(1)} = 0$. This is the case where the three populations are coming from the same 4–mixture $f$. We depict the density estimations under the first scenario in Figures 7(a)–(c). This is the case where the small data set, benefits the most in terms of borrowing of strength.

2. For $s = 2$, we set, $q^{(2)} = 1/2$. The 2-mixtures $g_1$ and $g_2$ are the the idiosyncratic parts of the 6-mixtures $f_1^{(2)}$ and $f_3^{(2)}$, respectively. The density estimations under the second scenario are given in Figures 7(d)–(f). In this case, the strength of borrowing between the small data set and the two large data sets weakens.

3. For $s = 3$ we set $q^{(3)} = 1$. In this case the three populations have no common parts. The density estimations are given in Figures 7(g)–(i). This is the worst case scenario, where there is no borrowing of strength between the small and the two large data sets.

The Hellinger distances between the true and the estimated densities, for the three scenarios, are given in table 6. In the second column of the Table we can see how the Hellinger distance of the estimation $\hat{f}_2^{(s)}$ and the true density $f_2^{(s)}$ increases as the borrowing of strength weakens, it is that $\mathcal{H}_G(f_2^{(1)}, \hat{f}_2^{(1)}) < \mathcal{H}_G(f_2^{(2)}, \hat{f}_2^{(2)}) < \mathcal{H}_G(f_2^{(3)}, \hat{f}_2^{(3)}).$
Figure 7: Density estimation with the PDGSBP model (curves in black) under the three different scenarios. The true density has been superimposed in red.

| s   | $\mathcal{H}_G(f_1^{(s)}, \hat{f}_1^{(s)})$ | $\mathcal{H}_G(f_2^{(s)}, \hat{f}_2^{(s)})$ | $\mathcal{H}_G(f_3^{(s)}, \hat{f}_3^{(s)})$ |
|-----|------------------------------------------|------------------------------------------|------------------------------------------|
| 1   | 0.14                                     | 0.19                                     | 0.13                                     |
| 2   | 0.15                                     | 0.22                                     | 0.15                                     |
| 3   | 0.12                                     | 0.26                                     | 0.12                                     |

Table 6: Hellinger distances between the true and the estimated densities for the three scenario example.

4.4 Real data example. The data set is to be found at [http://lib.stat.cmu.edu/datasets/pbcseq](http://lib.stat.cmu.edu/datasets/pbcseq) and involves data from 310 individuals. We take the observation as SGOT (serum glutamic-oxaloacetic transaminase) level, just prior to liver transplant or death or the last observation recorded, under three conditions on the individual

1. The individual is dead without transplantation.
2. The individual had a transplant.
3. The individual is alive without transplantation.
We normalize the means of all three data sets to zero. Since it is reasonable to assume the densities for the observations are similar for the three categories (especially for the last two), we adopt the models proposed in this paper with \( m = 3 \). The number of transplanted individuals is small (sample size of 28) so it is reasonable to borrow strength for this density from the other two. In this example, we set the hyperparameters of the Dirichlet priors for the selection probabilities to

\[
\alpha_{jl} = \begin{cases} 
10, & \text{if } j = l = 1 \text{ or } j = l = 3 \\
1, & \text{otherwise.}
\end{cases}
\]

1. In Figure 8(a)–(c) we provide histograms of the real data sets and superimpose the KDE’s based on the predictive samples of the PDDP and PDGSBP samplers. The two models give nearly identical density estimations.

2. The estimated a-posteriori selection probabilities are given below

\[
\mathbb{E}_G(p \mid (x_{ji})) = \begin{pmatrix} 
0.61 & 0.23 & 0.16 \\
0.34 & 0.10 & 0.56 \\
0.08 & 0.12 & 0.80
\end{pmatrix}, \quad \mathbb{E}_D(p \mid (x_{ji})) = \begin{pmatrix} 
0.67 & 0.16 & 0.17 \\
0.29 & 0.15 & 0.56 \\
0.10 & 0.12 & 0.78
\end{pmatrix}.
\]

By comparing the second rows of the selection matrices, we conclude that the strength of borrowing is slightly larger in the case of PDGSBP model.

![Figure 8](image-url)  
**Figure 8:** Histograms of the real data sets with superimposed KDE curves based on the predictive samples of the PDGSBP and rPDDP models.

5. Discussion. In this paper we have generalized the GSB process to a multidimensional dependent stochastic process which can be used as a Bayesian nonparametric prior for density estimation in the case of partially exchangeable data sets. The resulting Gibbs sampler is as
accurate as its DP based counterpart, yet faster and far less complicated. The main reason for this is that the GSB sampled value of the allocation variable \(d_{ji}\) will be an element of the sequential slice set \(S_{ji} = \{1, \ldots, N_{ji}\}\). Thus, there is no need to search the arrays of the weights; we know the state space of the clustering variables in advance. On the other hand, the sampling of \(d_{ji}\) in the DP based algorithm will always have one more step; the creation of the slice sets.

For an objective comparison of the execution times of the two models, we have run the two samplers in an a-priori synchronized mode. This, involves the placing of \(G(a_{jl}, b_{jl})\) priors over the DP \(c_{jl}\) concentration masses, leading to a more efficient version of the PDDP model introduced in Hatjispyros et al. (2011, 2016A).

We have show that when the PDGSBP and PDDP models are synchronized, i.e. their parameters satisfy \(\lambda_{ji} = (1 + c_{ji})^{-1}\), the correlation between the models can be controlled by imposing further restrictions among the \(\lambda_{ji}\) parameters.

Finally, an interesting research path would be the generalization of the pairwise dependent \(Q_j\) measures to include all possible interactions, in the sense that

\[
Q_j(\cdot) = p_j G_j(\cdot) + \sum_{l=2}^{m} \sum_{\eta \in C_{j,l,m}} p_{j,\eta} G_{\eta(j)}(\cdot) \text{ with } p_j + \sum_{l=2}^{m} \sum_{\eta \in C_{j,l,m}} p_{j,\eta} = 1,
\]

where the \(G_j\) and the \(G_{\eta(j)}\)'s are independent GSB processes, \(C_{j,l,m} = \{(k_1, \ldots, k_{l-1}) : 1 \leq k_1 < \cdots < k_{l-1} \leq m, k_r \neq j, 1 \leq r \leq m - 1\}\) and \(\eta(j)\) is the ordered vector of the elements of the vector \(\eta\) and \(\{j\}\). Now the \(f_j\) densities will be a mixture of \(2^{m-1}\) GSB mixtures, and the total number of the independent GSB processes needed to model \((f_1, \ldots, f_m)\) will be \(2^m - 1\).

Appendix A

Proof of Proposition 1. Starting from the \(N_{ji}\)-augmented random densities we have

\[
f_j(x_{ji}, N_{ji} = r) = \sum_{l=1}^{m} f_j(x_{ji}, N_{ji} = r, \delta_{ji} = l) = \sum_{l=1}^{m} p_{jl} f_j(x_{ji}, N_{ji} = r|\delta_{ji} = l)
\]

\[
= \sum_{l=1}^{m} p_{jl} \sum_{k=1}^{\infty} f_j(x_{ji}, N_{ji} = r, d_{ji} = k|\delta_{ji} = l)
\]

\[
= \sum_{l=1}^{m} p_{jl} f_j(N_{ji} = r|\delta_{ji} = l) \sum_{k=1}^{\infty} f_j(d_{ji} = k|N_{ji} = r) f_j(x_{ji}|d_{ji} = k, \delta_{ji} = l).
\]

Because \(f_j(N_{ji} = r|\delta_{ji} = l) = f_N(r|\lambda_{ji})\) and \(f_j(x_{ji}|d_{ji} = k, \delta_{ji} = l) = K(x_{ji}|\theta_{jlk})\), the last equation gives

\[
f_j(x_{ji}, N_{ji} = r) = \sum_{l=1}^{m} p_{jl} f_N(r|\lambda_{ji}) \sum_{k=1}^{\infty} \frac{1}{r} I(k \leq r) K(x_{ji}|\theta_{jlk})
\]

\[
= \frac{1}{r} \sum_{l=1}^{m} p_{jl} f_N(r|\lambda_{ji}) \sum_{k=1}^{r} K(x_{ji}|\theta_{jlk}).
\]
Augmenting further with the variables \(d_{ji}\) and \(\delta_{ji}\), yields

\[
f_j(x_{ji}, N_{ji} = r, d_{ji} = k, \delta_{ji} = l) = \frac{1}{r} p_{jl} f_N(r|\lambda_{ji}) I(k \leq r) K(x_{ji} | \theta_{jk}).
\]

Because \(P(\delta_{ji} = l) = p_{jl}\), the last equation leads to equation (6) and the proposition follows. \(\square\)

**Proof of Proposition 2.** Marginizing the joint of \(x_{ji}\) and \(N_{ji}\) with respect to \(x_{ji}\) we obtain

\[
f_j(N_{ji} = r) = \sum_{l=1}^{m} p_{jl} f_N(r|\lambda_{ji}).
\]

Then dividing equation (5) with the probability that \(N_{ji}\) equals \(r\) we obtain equation (7). \(\square\)

**Proof of Lemma 1.** Because \(g_{\mathcal{G}}(x) = \lambda \sum_{j=1}^{\infty} (1 - \lambda)^{j-1} K(x | \theta_j)\), we have

\[
\mathbb{E} \{g_{\mathcal{G}}(x)^2\} = \lambda^2 \mathbb{E} \left\{ \left( \sum_{j=1}^{\infty} (1 - \lambda)^{j-1} K(x | \theta_j) \right)^2 \right\}
\]

\[
= \lambda^2 \left\{ \sum_{j=1}^{\infty} (1 - \lambda)^{2j-2} \mathbb{E} [K(x | \theta_j)^2] + 2 \sum_{k=2}^{\infty} \sum_{j=1}^{k-1} (1 - \lambda)^{j+k-2} \mathbb{E}[K(x | \theta_j) K(x | \theta_k)] \right\}
\]

\[
= \lambda^2 \left\{ \sum_{j=1}^{\infty} (1 - \lambda)^{2j-2} \mathbb{E} [K(x | \theta_j)^2] + 2 \sum_{k=2}^{\infty} \sum_{j=1}^{k-1} (1 - \lambda)^{j+k-2} \mathbb{E}[K(x | \theta)]^2 \right\}
\]

\[
= \lambda^2 \left\{ \frac{1}{\lambda(2 - \lambda)} \mathbb{E} [K(x | \theta_j)^2] + 2 \frac{1 - \lambda}{\lambda^2(2 - \lambda)} \mathbb{E}[K(x | \theta)]^2 \right\},
\]

which gives the desired result. \(\square\)

**Proof of Proposition 3.** The random densities \(f_i(x) = \sum_{l=1}^{m} p_{il} g_{il}(x)\) and \(f_j(x) = \sum_{l=1}^{m} p_{jl} g_{jl}(x)\) depend to each other through the random measure \(\mathcal{G}_{ji}\), therefore

\[
\mathbb{E}[f_i(x)f_j(x)] = \mathbb{E}[\mathbb{E}(f_i(x)f_j(x)|\mathcal{G}_{ji})] = \mathbb{E}\{\mathbb{E}[f_i(x)|\mathcal{G}_{ji}] \mathbb{E}[f_j(x)|\mathcal{G}_{ji}]\}, \tag{15}
\]

and

\[
\mathbb{E}[f_i(x)|\mathcal{G}_{ji}] = \sum_{l \neq i} p_{il} \mathbb{E}[g_{il}(x)] + p_{ii} g_{ii}(x) = (1 - p_{ii}) \mathbb{E}[K(x | \theta)] + p_{ii} g_{ii}(x)
\]

\[
\mathbb{E}[f_j(x)|\mathcal{G}_{ji}] = \sum_{l \neq j} p_{jl} \mathbb{E}[g_{jl}(x)] + p_{jj} g_{jj}(x) = (1 - p_{jj}) \mathbb{E}[K(x | \theta)] + p_{jj} g_{jj}(x).
\]

Substituting back to equation (15), one obtains

\[
\mathbb{E}[f_i(x)f_j(x)] = (1 - p_{ii} p_{jj}) \mathbb{E}[K(x | \theta)]^2 + p_{ii} p_{jj} \mathbb{E} [g_{ji}(x)^2].
\]

Using lemma 1, the last equation becomes

\[
\mathbb{E}[f_i(x)f_j(x)] = \frac{\lambda_{ji} p_{jj} p_{ji}}{2 - \lambda_{ji}} \left\{ \mathbb{E}[K(x | \theta)^2] - \mathbb{E}[K(x | \theta)]^2 \right\} + \mathbb{E}[K(x | \theta)]^2,
\]

22
or that
\[ \text{Cov}(f_j(x), f_i(x)) = \frac{\lambda_{ji} p_{ji} p_{ij}}{2 - \lambda_{ji}} \text{Var}(K(x|\theta)). \]
The desired result, comes from the fact that
\[
\text{Var} \left( \int_{\Theta} K(x|\theta) G_{ji}(d\theta) \right) = \left\{ \frac{\lambda_{ji}}{2 - \lambda_{ji}} \mathbb{E}[K(x|\theta)]^2 + \frac{2(1 - \lambda_{ji}) \mathbb{E}[K(x|\theta)]^2}{2 - \lambda_{ji}} \right\} - \mathbb{E}[K(x|\theta)]^2
\]
\[= \frac{\lambda_{ji}}{2 - \lambda_{ji}} \left( \mathbb{E}[K(x|\theta)^2] - \mathbb{E}[K(x|\theta)]^2 \right). \]

**Proof of Proposition 4.**

(1.) From equation (11) and proposition 3, we have that
\[
\text{Var}(f_G^j(x)) = \text{Var}(\sum_{l=1}^{m} p_{jl} g_{jl}^j(x)) = \sum_{l=1}^{m} p_{jl}^2 \lambda_{ji} \text{Var}(K(x|\theta)).
\]
Normalizing the covariance in equation (10) with the associated standard deviations, yields
\[
\text{Corr}(f_G^j(x), f_G^i(x)) = \frac{\lambda_{ji} p_{ji} p_{ij}}{2 - \lambda_{ji}} \left( \sum_{l=1}^{m} \sum_{r=1}^{m} \frac{p_{jl}^2 p_{ir}^2 \lambda_{ji} \lambda_{ir}}{(2 - \lambda_{jl})(2 - \lambda_{ir})} \right)^{-1/2}.
\]

(16)

Similarly, from proposition 1 in Hatjispyros et al. (2011), it is that
\[
\text{Var}(f_D^j(x)) = \sum_{l=1}^{m} \frac{p_{jl}^2}{1 + c_{jl}} \text{Var}(K(x|\theta)),
\]
and
\[
\text{Corr}(f_D^j(x), f_D^i(x)) = \frac{p_{ji} p_{ij}}{1 + c_{ji}} \left( \sum_{l=1}^{m} \sum_{r=1}^{m} \frac{p_{jl}^2 p_{ir}^2 \lambda_{ji} \lambda_{ir}}{(1 + c_{jl})(1 + c_{ir})} \right)^{-1/2}.
\]

(17)

(2.) When \(\lambda_{ji} = \lambda\) and \(c_{ji} = c\) for all \(1 \leq j \leq i \leq m\), from equations (16) and (17), it is clear that
\[
\text{Corr}(f_G^j(x), f_G^i(x)) = \text{Corr}(f_D^j(x), f_D^i(x)) = p_{ji} p_{ij} \left( \sum_{l=1}^{m} \sum_{r=1}^{m} \frac{p_{jl}^2 p_{ir}^2}{(1 + c_{jl})(1 + c_{ir})} \right)^{-1/2}.
\]

**Appendix B**

1. Sampling of the concentrations masses for the rPDDP model.

In this case, the random densities \((f_j)\) are represented as finite mixtures of the DP mixtures \(g_{jl}(x|\mathbb{P}_{jl})\), where \(\mathbb{P}_{jl} \sim \mathcal{DP}(c_{jl}, P_0)\). We randomize the concentrations by letting \(c_{jl} \sim \mathcal{G}(a_{jl}, b_{jl})\).

Following West (1992) we have the following two specific cases:

**A.** For \(j = l\), the posterior \(c_{jl}\)'s will be affected only by the size of the data set \(x_j\) and the number of unique clusters for which \(\delta_{ji} = e_j\). Letting
\[
\rho_{jj} = \#\{d_{jj} : \delta_{ji} = e_j, 1 \leq i \leq n_j\},
\]
we have

\[ \beta \sim \text{Be}(c_{jj} + 1, n_j) \]
\[ c_{jj} | \beta, \rho_{jj} \sim \pi_\beta \mathcal{G}(a_{jj} + \rho_{jj}, b_{jj} - \log(\beta)) + (1 - \pi_\beta) \mathcal{G}(a_{jj} + \rho_{jj} - 1, b_{jj} - \log(\beta)) \]

with the weights \( \pi_\beta \) satisfying \( \frac{\pi_\beta}{1 - \pi_\beta} = \frac{\alpha_{jj} + \rho_{jj} - 1}{\alpha_{jj} + \rho_{jj} - \log(\beta)} \).

**B.** For \( j \neq l \), the posterior \( c_{jl} \)’s will be affected by the size of the data sets \( x_j \) and \( x_l \) and the cumulative number of unique clusters \( d_{ji} \) for which \( \delta_{ji} = e_i \) and the unique clusters \( d_{li} \) for which \( \delta_{li} = e_j \). Letting

\[ \rho_{jl} = \#\{d_{ji} : \delta_{ji} = e_l, 1 \leq i \leq n_j\} + \#\{d_{li} : \delta_{li} = e_j, 1 \leq i \leq n_l\}, \]

it is that

\[ \beta \sim \text{Be}(c_{jl} + 1, n_j + n_l) \]
\[ c_{jl} | \beta, \rho_{jl} \sim \pi_\beta \mathcal{G}(a_{jl} + \rho_{jl}, b_{jl} - \log(\beta)) + (1 - \pi_\beta) \mathcal{G}(a_{jl} + \rho_{jl} - 1, b_{jl} - \log(\beta)), \]

with the weights \( \pi_\beta \) satisfying \( \frac{\pi_\beta}{1 - \pi_\beta} = \frac{\alpha_{jl} + \rho_{jl} - 1}{(\alpha_{jl} + \rho_{jl} - \log(\beta))}. \)

Bear in mind that \( \rho_{jl} = 0 \) is always a possibility, so that we impose \( a_{jl} > 1 \).

### 2. Sampling of the geometric probabilities for the PDGSBP model.

In this section we provide the full conditionals for the geometric probabilities \( \lambda_{jl} \) under beta conjugate and transformed gamma nonconjugate priors. We let

\[ S_{jl} = \sum_{i=1}^{n_j} \mathcal{I}(\delta_{ji} = e_l) \quad \text{and} \quad S'_{jl} = \sum_{i=1}^{n_j} \mathcal{I}(\delta_{ji} = e_l)(N_{ji} - 1). \]

**A.** For the choice of prior \( \lambda_{jl} \sim \text{Be}(a_{jl}, b_{jl}) \), for \( l = j \) it is that

\[ f(\lambda_{jj} | \ldots) = \text{Be}(\lambda_{jj} | a_{jj} + 2S_{jj}, b_{jj} + S'_{jj}), \]

also, for \( l \neq j \) we have

\[ f(\lambda_{jl} | \ldots) = \text{Be}(\lambda_{jl} | a_{jl} + 2(S_{jl} + S_{lj}), b_{jl} + S'_{jl} + S'_{lj}). \]

**B.** For the choice of prior \( \lambda_{jl} \sim T\mathcal{G}(a_{jl}, b_{jl}) \), for \( l = j \) it is that

\[ f(\lambda_{jj} | \ldots) \propto \lambda_{jj}^{2S_{jj} - a_{jj} - 1}(1 - \lambda_{jj})^{S'_{jj} + a_{jj} - 1}e^{-b_{jj}/\lambda_{jj}} \mathcal{I}(0 < \lambda_{jj} < 1). \]

To sample from this density, we include the positive auxiliary random variables \( \nu_1 \) and \( \nu_2 \) such that

\[ f(\lambda_{jj}, \nu_1, \nu_2 | \ldots) \propto \lambda_{jj}^{2S_{jj} - a_{jj} - 1} \mathcal{I}(\nu_1 < (1 - \lambda_{jj})^{S'_{jj} + a_{jj} - 1}) \mathcal{I}(\nu_2 < e^{-b_{jj}/\lambda_{jj}}) \mathcal{I}(0 < \lambda_{jj} < 1). \]
The full conditionals for \( \nu_1, \nu_2 \) are uniforms

\[
f(\nu_1 | \cdots) = U\left(\nu_1 | 0, (1 - \lambda_{jj}) S'_{jj} + a_{jj} - 1\right) \quad \text{and} \quad f(\nu_2 | \cdots) = U\left(\nu_2 | 0, e^{-b_{jj}/\lambda_{jj}}\right),
\]

and the new full conditional for \( \lambda_{jj} \) becomes

\[
f(\lambda_{jj} | \nu_1, \nu_2, \cdots) \propto \lambda_{jj}^{-\nu_1 - a_{jj} - 1} \begin{cases} \mathcal{I}\left(\frac{-b_{jj}}{\log \nu_2} < \lambda_{jj} < 1 - \nu_1^{1/L_{jj}}\right) & L_{jj} \geq 0 \\ \mathcal{I}\left(\max\left\{-\frac{b_{jj}}{\log \nu_2}, 1 - \nu_1^{1/L_{jj}}\right\} < \lambda_{jj} < 1\right) & L_{jj} < 0, \end{cases}
\]

where we have set \( L_{jj} = S'_{jj} + a_{jj} - 1 \). We can sample from this density using the inverse cumulative distribution function technique. Also, for \( l \neq j \) we apply the same embedded Gibbs sampling technique to the full conditional density

\[
f(\lambda_{jl} | \cdots) \propto \lambda_{jl}^{-\nu_1 - a_{jl} - 1} (1 - \lambda_{jl})^{S'_{jl} + S'_{lj} + a_{jl} - 1} e^{-b_{jl}/\lambda_{jl}} \mathcal{I}(0 < \lambda_{jl} < 1). \]

References.

Bulla, P., Muliere, P. and Walker, S.G. (2009). A Bayesian nonparametric estimator of a multivariate survival function. *Journal of Statistical Planning and Inference* 139, 3639–3648.

De Iorio, M., Müller, P., Rosner, G.L. and MacEachern, S.N. (2004). An ANOVA model for dependent random measures. *Journal of the American Statistical Association* 99, 205–215.

Dunson, D.B. and Park, J.H. (2008). Kernel stick-breaking processes. *Biometrika* 95, 307–323.

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

Fuentes–Garcia, R., Mena, R.H., Walker, S.G. (2009). A nonparametric dependent process for Bayesian regression *Statistics and Probability Letters* 79, 1112–1119.

Fuentes–Garcia, R., Mena, R.H., Walker, S.G. (2010). A new Bayesian nonparametric mixture model. *Comm.Statist.Simul.Comput* 39, 669–682.

Griffin, J.E. and Steel, M.F.J. (2006). Order–based dependent Dirichlet processes. *Journal of the American Statistical Association* 101, 179–194.

Griffin, J.E., Kolossiatis, M. and Steel, M.F.J. (2013). Comparing distributions by using dependent normalized ranom–measure mixtures. *Journal of the Royal Statistical Society, Series B* 75, 499–529.
Hatjispyros, S.J., Nicoleris, T. and Walker, S.G. (2011). Dependent mixtures of Dirichlet processes. *Computational Statistics and Data Analysis* **55**, 2011–2025.

Hatjispyros, S.J., Nicoleris, T. and Walker, S.G. (2016a). Dependent random density functions with common atoms and pairwise dependence. *Computational Statistics and Data Analysis* **101**, 236–249.

Hatjispyros, S.J., Nicoleris, T. and Walker, S.G. (2016b). Bayesian nonparametric density estimation under length bias. *Communications in Statistics* DOI: 10.1080/03610918.2016.1263735

Lijoi, A., Nipoti, B. and Prüenster, I. (2014a). Bayesian inference with dependent normalized completely random measures. *Bernoulli*, **20**, 1260–1291.

Lijoi, A., Nipoti, B. and Prüenster, I. (2014b). Dependent mixture models: clustering and borrowing information. *Computational Statistics and Data Analysis* **71**, 17–433.

Kolossiatis, M., Griffin, J.E. and Steel, M.F.J. (2013). On Bayesian nonparametric modelling of two correlated distributions. *Statistics and Computing* **23**, 1–15.

Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates I. Density estimates. *Annals of Statistics* **12**, 351–357.

MacEachern, S.N. (1999). Dependent nonparametric processes. In “Proceedings of the Section on Bayesian Statistical Science” pp. 50-55. American Statistical Association.

Müller, P., Quintana, F., and Rosner, G., (2004). A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society, Series B* **66**, 735–749.

Mena, R.H., Ruggiero, M. and Walker, S.G. (2011). Geometric stick–breaking processes for continuous–time Bayesian nonparametric modeling. *Journal of Statistical Planning and Inference* **141** (9), 3217–3230.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.

Walker, S.G. (2007). Sampling the Dirichlet mixture model with slices *Communications in Statistics* **36** 45–54.

West, M. (1992). Hyperparameter estimation in Dirichlet process mixture models. *Techni- cal report 92-A03*, Duke University, ISDS.