Bayesian Analysis (2014) **9**, Number 2, pp. 307–330

Local-Mass Preserving Prior Distributions for Nonparametric Bayesian Models

Juhee Lee * Steven N. MacEachern † Yiling Lu ‡ Gordon B. Mills ‡

Abstract. We address the problem of prior specification for models involving the two-parameter Poisson-Dirichlet process. These models are sometimes partially subjectively specified and are always partially (or fully) specified by a rule. We develop prior distributions based on local mass preservation. The robustness of posterior inference to an arbitrary choice of overdispersion under the proposed and current priors is investigated. Two examples are provided to demonstrate the properties of the proposed priors. We focus on the three major types of inference: clustering of the parameters of interest, estimation and prediction. The new priors are found to provide more stable inference about clustering than traditional priors while showing few drawbacks. Furthermore, it is shown that more stable clustering results in more stable inference for estimation and prediction. We recommend the local-mass preserving priors as a replacement for the traditional priors.

Keywords: nonparametric Bayes, Dirichlet process, two-parameter Poisson-Dirichlet process, local mass, prior misspecification, clustering

1 Introduction

Two main schools of thought lead to the use of Bayesian methods. The first is the subjective school which focuses on the elicitation of personal probabilities and is driven by the axioms of rational behavior, as described by Savage (1972). The second is the objective school which is driven by the complete class theorems of decision theory. While the former encourages careful elicitation of prior distributions, as in Lindley (1965) and De Finetti (1975), the latter opens the door to rule-based specification of the prior distribution, whether through the time-tested methods of Jeffreys (1998) or the more recent development of reference priors (Berger and Bernardo 1992) and the subsequent development of objective Bayesian methods.

Both schools are Bayesian, but they emphasize different aspects of inference. Subjective methods emphasize "knowledge" while objective methods emphasize "performance". When focused on performance, the user must specify a target of inference as well as desired properties for inference – thus the choice of the relative importance of sets of parameters when determining a reference prior (Bernardo 1979; Berger et al. 2009), the emphasis on consistency of estimators (e.g. Salinetti (2003); Kleijn and van der Vaart

*Department of Applied Mathematics and Statistics, University of California Santa Cruz, CA, U.S.A. juheelee@soe.ucsc.edu
†Department of Statistics, The Ohio State University, Columbus, Ohio, U.S.A.
‡Department of Systems Biology, UT MD Anderson Cancer Center, Houston, TX, U.S.A

 DOI:10.1214/13-BA857

(2006)), and the investigation of rates of convergence (e.g. Ghosal and van der Vaart (2001); Rousseau (2010)). These considerations, along with a reluctance to inject strong subjective beliefs into the prior distribution, lead to relatively diffuse priors. A generally agreed upon aspect of good performance is stability of inference as minor details of the rule used to specify the prior are varied.

Nonparametric Bayesian models are based on infinite dimensional objects, and so are inevitably specified, at least in part, with rule-based prior distributions. The most popular methods are those based on the Dirichlet process (Ferguson 1973) which is a special case of the two-parameter Poisson-Dirichlet process (Pitman 1996; Pitman and Yor 1997) and the many more recent directions, including Pólya trees, alternative mixture forms, and "dependent" nonparametric processes. These basic forms are used as components in sophisticated hierarchical models; for example, an extension of the Dirichlet process to the mixture of Dirichlet processes (Antoniak 1974). The methods have proven enormously successful for a wide range of problems. Three main features of inference are often present in examples: (i) use of the methods to control for variation, as when one allows an arbitrary error distribution (e.g., MacEachern and Guha (2011)), (ii) estimation of parameters in a number of component problems (e.g., Escobar (1994)), and (iii) attention to the clustering of observations (e.g., Quintana and Iglesias (2003); Quintana (2006)). Hjort et al. (2010) provide an introduction to the techniques, many applications, and a recent introduction to the literature.

The technical means by which diffuse nonparametric Bayesian prior distributions have been created has been to extend methods developed for the low-dimensional parametric setting to the nonparametric setting, often with some compromise to facilitate computation. Thus, in the parametric case, the thick-tailed prior distributions recommended for the robust inference they bring to the normal means problem (Berger 1993) are replaced by an overdispersed conjugate form, weakening the strength of the prior distribution and allowing the data's likelihood to dominate the prior. The traditional extension to the nonparametric setting for Dirichlet-based models is to replace the base measure (which specifies the prior distribution) with an overdispersed base measure, while leaving the mass of the base measure or the distribution on the mass parameter unchanged. Examples with lucid arguments supporting this practice include Escobar (1994) who suggested a uniform prior with support much larger than the range of the data, Escobar and West (1995) who place a prior on the mass of the base measure, and Ishwaran and James (2002) who suggested use of a base measure with four times the dispersion of the data.

Bush et al. (2010) show that the dispersion of the base measure cannot be increased indefinitely without producing an improper posterior distribution. With this in mind, a specific choice of overdispersion must be made. An open question is what impact a relatively arbitrary choice of overdispersion has on posterior inference. A second question is whether, by replacing the traditional structure of the prior distribution with an alternative structure, we can produce posterior inference that is less sensitive to the choice of overdispersion. In this paper, we investigate these questions, focusing on the robustness, or stability, of posterior inference as the prior distribution is weakened. We compare two forms of prior distribution in the context of the mixture of Dirichlet

processes model–the traditional form that takes the dispersion of the base measure to be independent of the mass of the base measure, and a novel form in which the dispersion is independent of the "local mass". We then extend the proposed modeling strategy to more general two-parameter Poisson-Dirichlet processes. We find that the stability of inference depends both on the form of the prior distribution and on the inference being considered. The new prior distributions show more stability for inference on clustering of observations, while the traditional and new forms are roughly equivalent for certain estimation problems.

The remainder of the paper is organized as follows: Section 2 describes the new class of local-mass preserving prior distributions and presents technical details with an application to the two-parameter Poisson-Dirichlet process model. Section 3 presents analyses of two data sets: a data set from a functional proteomics profiling experiment and an allometric data set, and contrasts inference under similar models with the two forms of prior distribution. The final section contains conclusions.

2 Models

2.1 Dirichlet Process

Let $\boldsymbol{X}_n = \{X_1, \ldots, X_n\}$ be a collection of n objects. We consider a distribution for clustering the objects in $\boldsymbol{X}_n$. We introduce an n-dimensional cluster membership indicator vector $\boldsymbol{s}_n = (s_1, \ldots, s_n)$ to denote a partition of X_i's into K_n ($\leq n$) clusters. The vector $\boldsymbol{s}_n$ is defined by the relation $s_i = j$ if and only if X_i is in cluster j. In other words, any two X_i and $X_{i'}$ ($i \neq i'$) are in a cluster if $s_i = s_{i'}$. For partition $\boldsymbol{s}_n$, we let c_j be the size of cluster j and represent the sizes of the K_n clusters as a K_n-dimensional vector, $\boldsymbol{c} = (c_1, \ldots, c_{K_n})$. The Dirichlet process (DP) provides a distribution on the set of partitions:

$$\begin{aligned} G \,|\, \alpha &\sim \mathrm{DP}(\alpha), \\ X_i \,|\, G &\overset{iid}{\sim} \mathrm{G}, \end{aligned} \tag{1}$$

where α is the base measure of the DP (Ferguson 1973). The parameter of the DP, α, is a measure which may be split into two parts, the total mass of the measure, M and the marginal distribution for X_i, $G_0(\cdot \,|\, \boldsymbol{\nu})$ where $\boldsymbol{\nu}$ is the hyperparameter vector for G_0. Thus α is often written as MG_0.

G in (1) is almost surely a discrete distribution function and it yields positive probability for ties in X_i's. The ties among X_i's can be utilized to cluster the X_i's. We let $X_j^\star$, $1 \leq j \leq K_n$ represent the location of cluster j and $X_i = X_j^\star$ for all X_i's in cluster j

The DP implies much about the distribution of $(K_n, \boldsymbol{X}^\star)$. The prior distribution of partition $\boldsymbol{s}_n$ with K_n clusters and cluster sizes $\boldsymbol{c}$ is

$$\mathrm{p}(\boldsymbol{s}_n \,|\, M) = M^{K_n} \frac{\prod_{j=1}^{K_n} \Gamma(c_j)}{\prod_{i=1}^{n} (M + i - 1)} \tag{2}$$

(Antoniak 1974). The DP also implies that given $\boldsymbol{s}_n$, the $X_j^\star$'s form a random sample of size K_n from G_0, that is, $X_j^\star \,|\, \boldsymbol{s}_n \overset{iid}{\sim} G_0(\cdot \,|\, \boldsymbol{\nu})$. (2) shows that the distribution of the number and sizes of the clusters is determined by M and n. G_0 determines features of G such as shape, location and dispersion. G_0 may be characterized with hyperparameters $\boldsymbol{\nu}$; for example, $\boldsymbol{\nu} = (\mu, \tau^2)$ where μ and τ^2 are location and scale parameters.

The division of X_i's into clusters under the DP is closely tied to the description of X_i's arising from a Pólya urn scheme (Blackwell and MacQueen 1973);

$$\mathrm{P}(s_{n+1} = j \,|\, \boldsymbol{s}_n) = \begin{cases} \dfrac{c_j}{M+n}, & 1 \le j \le K_n \\ \dfrac{M}{M+n}, & j = K_n + 1. \end{cases}$$

2.2 Independence Prior

The mixture of Dirichlet processes (MDP) model uses the DP as a prior on a latent mixing distribution (Antoniak 1974);

$$\begin{aligned} G \,|\, \alpha &\sim \mathrm{DP}(\alpha), \\ X_i \,|\, G &\overset{iid}{\sim} \mathrm{G}, \\ Y_i \,|\, X_i &\overset{ind}{\sim} \mathrm{F}(\cdot \,|\, X_i), \qquad i = 1, \ldots, n, \end{aligned} \tag{3}$$

where $\alpha = MG_0$ and G_0 is a distribution function with hyperparameters $\boldsymbol{\nu}$. For example, Escobar and West (1995) used an MDP model for density estimation assuming the normal likelihood. We now extend the model in (3) by placing prior distributions on M and $\boldsymbol{\nu}$ due to uncertainty about their values. The traditional approach is to declare M and $\boldsymbol{\nu}$ to be independent (Escobar 1994; Escobar and West 1995). We refer to this prior structure as the "independence prior structure" in the subsequent discussion. The independence prior structure leads to conditional (on the partition) posterior independence of M and $\boldsymbol{\nu}$. This independence structure is popular in applications of the DP, and almost all papers that place a distribution on M and $\boldsymbol{\nu}$ use it.

The calibration of M and $\boldsymbol{\nu}$ (or the priors for them) can be obtained using prior information from diverse sources, such as expert knowledge or previous studies. A prior for M can be specified by focusing on the number of clusters. For a given data set of size n, the expectation and variance of the number of clusters under the DP are $\mathrm{E}(K_n \,|\, M) = \sum_{i=1}^{n} M/(i-1+M)$ and $\mathrm{Var}(K_n \,|\, M) = \sum_{i=1}^{n} M(i-1)/(M+i-1)^2$ (Liu 1996). A popular prior for M is a gamma distribution,

$$M \sim \mathrm{Ga}(a, b), \tag{4}$$

where a and b are shape and scale parameters, respectively (hence, M has mean ab) (Escobar and West 1995). a and b are calibrated using the prior expected number of clusters and its variance (Kottas et al. 2005).

The specification of G_0 is trickier. The distribution G_0 generates locations of clusters induced by the DP and, jointly with M, controls the smoothness of estimates of G. Specification of G_0 follows patterns set out for parametric models. One common way to express weak prior information while maintaining conjugacy in a parametric model is to use a conjugate prior and inflate its dispersion. An advantage of using a conjugate prior with large dispersion is to circumvent mathematical or computational problems involved with non-conjugate priors and improper priors. For example, assuming a normal likelihood we let G_0 be a normal distribution with parameter vector $\boldsymbol{\nu} = (\mu, \tau^2)$ and consider priors for the mean μ and the variance τ^2; $\mu \sim \mathrm{N}(\mu_0, \tau_0^2)$ and $\tau^2 \sim \mathrm{IG}(a_0, b_0)$ where a_0 and b_0 are the shape and scale parameters of an inverse-gamma distribution (hence, τ^2 has mean $b_0/(a_0-1)$ provided $a_0 > 1$). The dispersion of G_0 is then inflated through its scale parameter τ^2 to reflect vague prior information, although explicit description of this inflation is rarely stated. In addition to papers already mentioned, Hirano (2002) and Ji et al. (2009) contain examples of overdispersed priors in nonparametric models.

In this standard example which we have described, M is *independent* of the dispersion parameter of G_0. Under this independence prior structure negligent inflation of the base distribution's dispersion can result in unreasonable inference on clustering due to the interplay between shrinkage of X_i toward μ (driven by τ^2) and clustering of X_i (driven by M): For any fixed M, α spreads its mass more widely with a diffuse G_0 than with a less-diffuse G_0. Consequently the diffuse G_0 pushes more mass into the tails, and so extreme observations tend to fall in smaller clusters. Furthermore, under the diffuse G_0, α assigns smaller mass to the central portion of the parameter space. The impact of less mass is that the model puts any two X_i's in the central region into the same cluster with larger probability. Taken together, these effects yield poor (and unintended) inference on the clustering structure of the X_i's. The impact appears greatest in the central region. To alleviate this problem, we introduce dependence between M and the dispersion of G_0. This leads to the concept of local mass and a different structure for the prior distribution.

2.3 Local-Mass Preserving Prior

Bush et al. (2010) defined local mass as the mass assigned by a measure to a small measurable set, confined to some region of the parameter space. They showed that preserving local mass allows one to develop a limiting improper version of the DP which leads to effective inference. We apply the concept of local mass and construct *proper* local-mass preserving prior distributions. We first define the middle region as a region of interest in the parameter space prior to analysis. Let $\mathcal{L} = (\ell_1, \ell_2) \subset \mathbb{R}$, $\ell_1 < \ell_2$ denote the middle region.

Remark 1. *Let $\mathcal{L} = (\ell_1, \ell_2) \subset \mathbb{R}$, $\ell_1 < \ell_2$ be the middle region. We assume the model in (1) for clustering $\boldsymbol{X}_n$. Conditional on the event that $X_i \in \mathcal{L}$ for all i ($i = 1, \ldots, n$), the distribution over partitions of $\boldsymbol{X}_n$ is*

$$\mathrm{p}(\boldsymbol{s}_n \mid M, X_1, \ldots, X_n \in \mathcal{L}) \propto M^{\star K_n} \prod_{j=1}^{K_n} \Gamma(c_j), \tag{5}$$

where $M^\star = \alpha(\mathcal{L}) = MG_0(\mathcal{L} \mid \boldsymbol{\nu})$ *represents the mass assigned to* $\mathcal{L}$ *under* α.

Proof. Let $\boldsymbol{s}_i$ and K_i represent the cluster configuration of $X_{i'} \in \mathcal{L}$ for $i' = 1, \ldots, i (< n)$ and its number of clusters, respectively. Following the Pólya urn scheme, $\mathrm{P}(s_1 = 1 \mid X_1 \in \mathcal{L}) = 1$. Given $\boldsymbol{s}_i$ $(i > 1)$, the distribution of the cluster membership of X_{i+1} is

$$\begin{aligned}
\mathrm{P}(s_{i+1} = j \mid M, \boldsymbol{s}_i, X_1, \ldots, X_{i+1} \in \mathcal{L}) &\propto \mathrm{P}(s_{i+1} = j, X_{i+1} \in \mathcal{L} \mid M, \boldsymbol{s}_i, X_1, \ldots, X_i \in \mathcal{L}) \\
&= \mathrm{P}(s_{i+1} = j \mid M, \boldsymbol{s}_i, X_1, \ldots, X_i \in \mathcal{L}) \\
&\quad \times \mathrm{P}(X_{i+1} \in \mathcal{L} \mid M, \boldsymbol{s}_i, s_{i+1} = j, X_1, \ldots, X_i \in \mathcal{L}) \\
&\propto \begin{cases} c_j, & 1 \le j \le K_i \\ M\mathrm{P}(X_{n+1} \in \mathcal{L}), & j = K_i + 1. \end{cases}
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathrm{p}(\boldsymbol{s}_n \mid M, X_1, \ldots, X_n \in \mathcal{L}) &\propto (M\mathrm{P}(X \in \mathcal{L}))^{K_n} \prod_{j=1}^{K_n} \Gamma(c_j) \\
&= (\alpha(\mathcal{L}))^{K_n} \prod_{j=1}^{K_n} \Gamma(c_j).
\end{aligned}$$

□

Remark 1 states that clustering of $\boldsymbol{X}_n$ conditional on all X_i's lying in $\mathcal{L}$ depends on the local mass assigned to $\mathcal{L}$, not on the total mass, M. In Definition 1, we describe a prior for $M^\star$ instead of M to preserve the local mass in $\mathcal{L}$ regardless of the specification of G_0.

Definition 1. *Let* $\mathcal{L} = (\ell_1, \ell_2) \subset \mathbb{R}$, $\ell_1 < \ell_2$ *be the middle region. We let* $M^\star = \alpha(\mathcal{L})$. *A prior of* $M^\star$,

$$M^\star \sim \mathrm{Ga}(a^\star, b^\star), \tag{6}$$

is termed a local-mass preserving prior for the MDP model.

From the definition, $M^\star$ under the local-mass preserving prior does not depend on $\boldsymbol{\nu}$. Combining this with the result in Remark 1, the distribution on clustering components of $\boldsymbol{X}_n$ given that they all lie in $\mathcal{L}$ does not depend on $\boldsymbol{\nu}$ under the local-mass preserving prior.

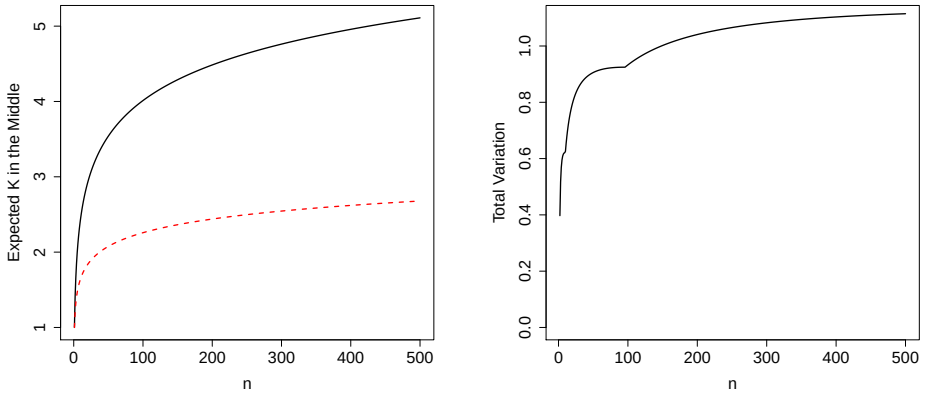
Remark 2. *Assume the local-mass preserving prior structure in Definition 1.* (6) *implies* $M \sim \mathrm{Ga}(a^\star, 1/p_0(\boldsymbol{\nu})b^\star)$ *where* $p_0(\boldsymbol{\nu}) = G_0(\mathcal{L} \mid \boldsymbol{\nu})$. M *and the dispersion of* G_0 *are positively associated under the local-mass preserving prior structure. In particular, as the dispersion of* G_0 *increases,* M *increases at a rate which keeps the mass in* $\mathcal{L}$ *constant.*

Remark 3. *Assume the independence prior structure of* M *and* $\boldsymbol{\nu}$ *in* (4). *The mass assigned by the independence prior to* $\mathcal{L}$ *depends on* $\boldsymbol{\nu}$. *Furthermore, conditional on the event that* $X_i \in \mathcal{L}$ *for all* i $(i = 1, \ldots, n)$, *the distribution over partitions of* $\boldsymbol{X}_n$ *depends on* $\boldsymbol{\nu}$.

Proof. Assume (4) as a prior for M, $\mathrm{p}(M)$. The mass assigned to $\mathcal{L}$ is:

$$\begin{aligned} M^\star &= \int_{\ell_1}^{\ell_2} \int_0^\infty M g_0(x \,|\, \boldsymbol{\nu}) \mathrm{p}(M) \mathrm{d}M \mathrm{d}x \\ &= ab \int_{\ell_1}^{\ell_2} g_0(x \,|\, \boldsymbol{\nu}) \mathrm{d}x \\ &= abG_0(\mathcal{L} \,|\, \boldsymbol{\nu}), \end{aligned}$$

where g_0 is the density function of G_0. Thus, $M^\star$ under the independence prior depends on $\boldsymbol{\nu}$. Combining this with the result in Remark 1, the distribution on clustering of X_i's in $\mathcal{L}$ depends on $\boldsymbol{\nu}$ under the independence prior. □



(a) Expected number of clusters in $\mathcal{L}$ (b) TV of distributions on K_n in $\mathcal{L}$

Figure 1: Dirichlet process: (a) the expected number of clusters in the middle region where the black solid line and the red dashed line represent the expected number of clusters in $\mathcal{L}$ under a baseline prior distribution and an overdispersed prior distribution, respectively; (b) the total variation distance (TV) of the distributions of the number of clusters in the middle region under the independence prior.

Note that for any two clustering configurations with the same K_n, $\boldsymbol{s}_n$ and $\boldsymbol{s}'_n$, $\mathrm{P}(\boldsymbol{s}_n \,|\, M, X_1, \ldots, X_n \in \mathcal{L})/\mathrm{P}(\boldsymbol{s}'_n \,|\, M, X_1, \ldots, X_n \in \mathcal{L})$ does not depend on $M^\star$ from (5). Thus, if any two distributions on K_n are identical, the distributions on $\boldsymbol{s}_n$ implied by them under the DP are identical. We examine distributions on K_n to compare the two prior structures instead of the more complicated distributions on $\boldsymbol{s}_n$.

Figure 1(a) and (b) illustrate the expected number of clusters and the total variation distance between distributions on K_n. The distributions on K_n come from a baseline prior distribution and an overdispersed prior distribution and are conditional on the event that $X_i \in \mathcal{L}$ $(1 \leq i \leq n)$. For the baseline prior, let $G_0 = \mathrm{N}(0, 1)$ and $M = 1$. For the overdispersed prior, let the variance of G_0 be inflated by a factor of 9 to express

weak prior information, i.e. $G_0' = \mathrm{N}(0, 3^2)$. Define $\mathcal{L} = (-1, 1)$ as the middle. As in Remark 3, conditional on the event that all X_i's lie in $\mathcal{L}$, the probability of starting a new singleton cluster under the independence prior structure is lessened when the base measure is overdispersed. Therefore, the independence prior structure induces larger clusters in the middle (equivalently, fewer clusters in the middle). As a result, in Figure 1(a) $\mathrm{E}(K_n|M, X_1, \ldots, X_n \in \mathcal{L})$ with G_0 (the solid line) is greater than that with G_0' (the dashed line). The black solid line in Figure 1(b) shows the total variation distance in $\mathrm{P}(K_n|M, X_1, \ldots, X_n \in \mathcal{L})$ under the independence prior structure, one with M and G_0 and the other with M and G_0'. The total variation distance is $TV = \sum_{K_n=1}^{n} |\mathrm{P}(K_n) - \mathrm{P}'(K_n)|$ and P and P′ are distributions of K_n.

On the other hand, following Definition 1, we focus on local mass $M^\star$ under the local-mass preserving prior structure. G_0 and G_0' are $\mathrm{N}(0, 1)$ and $\mathrm{N}(0, 3^2)$ but now the mass in the middle, $M^\star$, is held fixed. Note that $M^\star = M(\Phi(1) - \Phi(-1))$ where Φ is the standard normal distribution function, so $M' = M^\star/(\Phi(1/3) - \Phi(-1/3))$. Then $\alpha' = M'G_0'$ assigns the same amount of mass to $\mathcal{L}$ as $\alpha = MG_0$ does. The conditional distribution of K_n with α' is identical with that with α and so the total variation distance between the two is 0 for all n in contrast to the substantial difference under the independence prior (Figure 1(b)). Additionally, the baseline and overdispersed expected number of clusters are identical as well (both follow the solid line in Figure 1(a)). Thus that preserves the clustering pattern in X_i's lying in the middle, irrespective of any arbitrary overdispersion of the base measure.

We also note that the local-mass preserving prior structure requires only minor modifications in the Markov chain Monte Carlo (MCMC) posterior simulation. The full conditional for τ^2 is not in a closed form and so we use other sampling techniques such as the Metropolis-Hastings algorithm. Escobar and West (1995)'s sampling technique can be used to sample M by replacing b with $1/p_0(\boldsymbol{\nu})b^\star$ as defined in Definition 1.

2.4 Extension to the Two-Parameter Poisson-Dirichlet Process

The two-parameter Poisson-Dirichlet process, denoted by $\mathrm{PD}(\sigma, \theta)$, for $0 < \sigma < 1$ and $\theta > 0$ (often called the Pitman-Yor process) (Pitman 1996; Pitman and Yor 1997) generalizes the DP and provides more modeling flexibility. The DP in (3) can be replaced by the two-parameter Poisson Dirichlet process, and a gamma distribution $\mathrm{Ga}(a, b)$ is often considered as a prior of θ. The two-parameter Poisson-Dirichlet process can be characterized by means of the predictive probability function and the baseline distribution:

$$\mathrm{P}(s_{n+1} = j \,|\, \sigma, \theta, \boldsymbol{s}_n) = \begin{cases} \dfrac{c_j - \sigma}{n + \theta}, & 1 \le j \le K_n \\ \dfrac{\theta + K_n\sigma}{n + \theta}, & j = K_n + 1, \end{cases}$$

and given $\boldsymbol{s}_n$, $X_j^\star$'s are i.i.d. from $G_0(\cdot \,|\, \boldsymbol{\nu})$. This model includes the DP as a special case when $\sigma = 0$. Similar to the DP, it can be shown that $\mathrm{p}(\boldsymbol{s}_n \,|\, \theta, \sigma, X_1, \ldots, X_n \in \mathcal{L}) \propto$

$(p_0(\boldsymbol{\nu}))^{K_n} \mathrm{p}(\boldsymbol{s}_n \,|\, \theta, \sigma)$ where

$$\mathrm{p}(\boldsymbol{s}_n \,|\, \sigma, \theta) = \frac{\prod_{j=1}^{K_n}(\theta + j\sigma)}{(\theta+1)_{n-1\uparrow}} \prod_{j=1}^{K_n} (1-\sigma)_{c_j - 1\uparrow},$$

and $(x)_{n\uparrow} = x(x+1)\cdots(x+n-1)$ denotes the Pochhammer symbol, with $(x)_{0\uparrow} = 1$. The distribution of $\boldsymbol{s}_n$ conditional on $X_i \in \mathcal{L}$ for all i changes with $p_0(\boldsymbol{\nu})$. This will be illustrated through a small simulation later in this section.

For $\sigma \neq 0$, θ and G_0 cannot be collected into a meaningful measure as in the DP. We cannot preserve the entire conditional distribution on $\boldsymbol{s}_n$ as the dispersion of G_0 is increased. Instead, we settle for preserving $\mathrm{E}(K_n \,|\, X_1, \ldots, X_n \in \mathcal{L})$. To do so, we keep $p_0(\boldsymbol{\nu})$ independent of $\mathrm{E}(K_n \,|\, X_1, \ldots, X_n \in \mathcal{L})$. We propose the following distribution as a prior for θ in the two-parameter Poisson-Dirichlet process.

Definition 2. *Let $\mathcal{L} = (\ell_1, \ell_2) \subset \mathbb{R}$, $\ell_1 < \ell_2$ be the middle region. A prior for θ,*

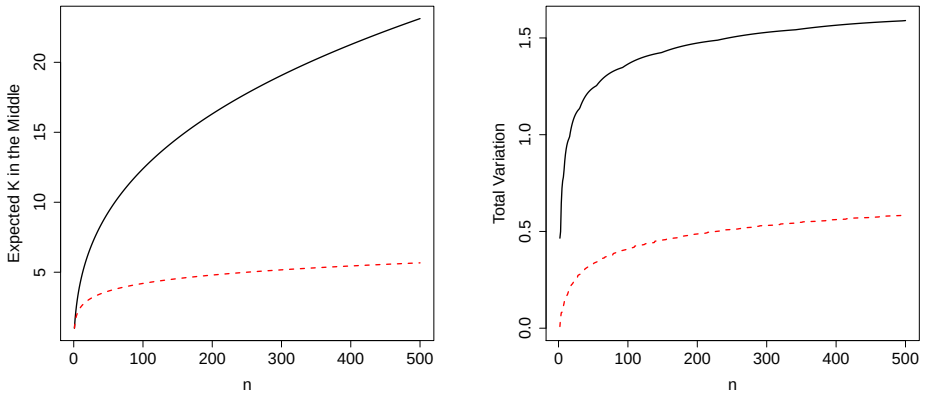
$$\theta \quad \sim \quad \mathrm{Ga}(a^\star, b^\star), \tag{7}$$

where $a^\star$ is fixed and $b^\star$ is a function of $\boldsymbol{\nu}$ by matching $\mathrm{E}(K_n \,|\, b^\star, X_1, \ldots, X_n \in \mathcal{L})$ to the assumed prior expected number of clusters in $\mathcal{L}$ is termed a local-mass preserving prior for the two-parameter Poisson-Dirichlet process.

Note that the distribution of $\boldsymbol{s}_n$ conditional on $X_i \in \mathcal{L}$ for all i remains identical regardless of the dispersion of G_0 under the DP. Here, we match the first moment under the baseline and overdispersed two-parameter Poisson-Dirichlet process priors.

Figure 2(a) and (b) show the expected number of clusters and the total variation distance of distributions of K_n conditional on the event that $X_i \in \mathcal{L}$ for all i $(i \leq n)$ under the two-parameter Poisson-Dirichlet process. This lets us examine the stability of the distribution of $\boldsymbol{s}_n$ conditional on $X_i \in \mathcal{L}$ for all i. For Figure 2, let $G_0 = \mathrm{N}(0,1)$ and $G_0' = \mathrm{N}(0, 3^2)$ be a well-calibrated base measure and an overdispersed base measure. Define $\mathcal{L} = (-1, 1)$ as the middle. With overdispersion, the expectation of K_n in the middle drops dramatically as shown in Figure 2(a). The solid and dashed lines are the expected K_n in the middle with G_0 and G_0', respectively. The change in expected K_n in the middle results in a large total variation distance between the distributions of K_n indicated by the black solid line in Figure 2(b).

Following the strategy proposed in Definition 2, we search for θ' such that $E(K_n \,|\, \sigma, \theta', X_1, \ldots, X_n \in \mathcal{L})$ with G_0' is the same as $E(K_n \,|\, \sigma, \theta, X_1, \ldots, X_n \in \mathcal{L})$ with G_0. The search for such a θ' can be done numerically. By matching the expected K_n in the middle, the total variation distance of the distributions of K_n in the middle decreases (shown with the dashed line in Figure 2(b)), leading to a smaller change in the distribution of $\boldsymbol{s}_n$. This shows that θ (or a prior for θ) needs to be recalibrated according to G_0 to yield stable inference on clustering in the middle.



(a) Expected number of clusters in $\mathcal{L}$ (b) TV of distributions of K_n in $\mathcal{L}$

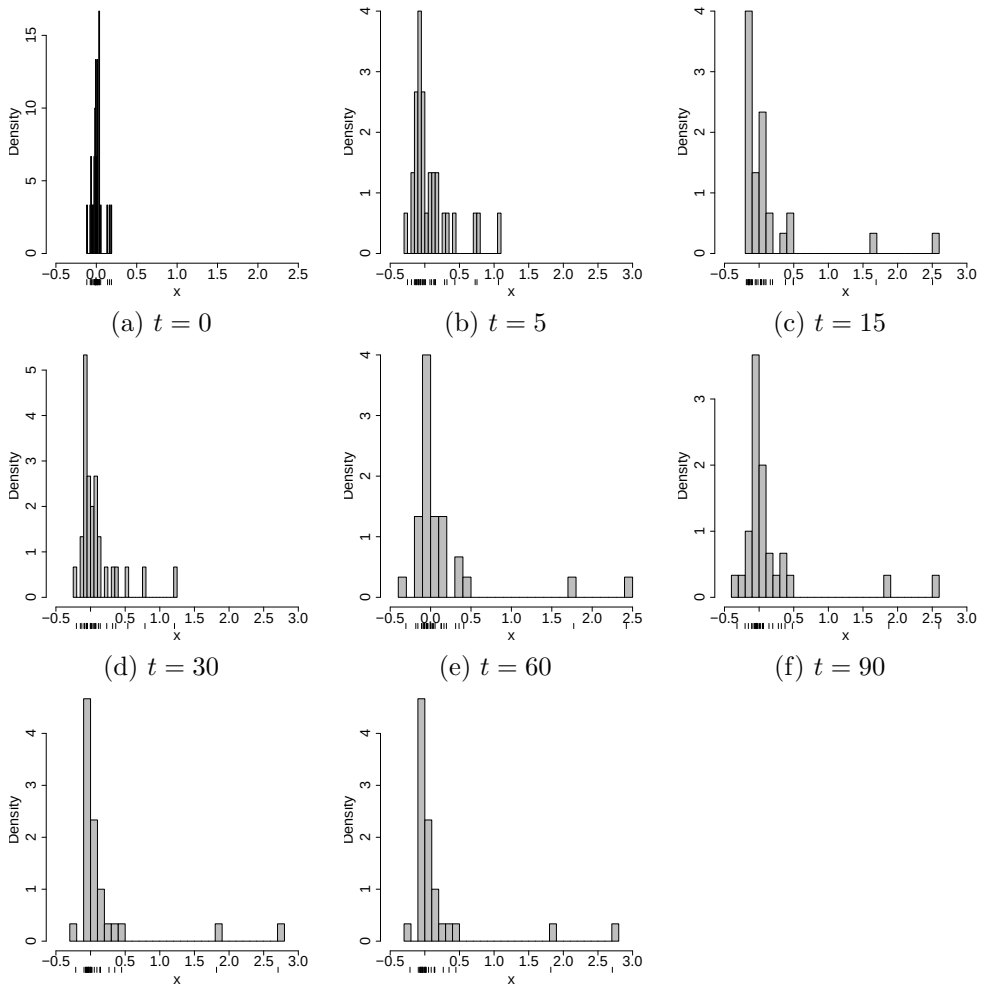
Figure 2: Two-parameter Poisson-Dirichlet process: (a) the expected number of clusters in the middle region where the black solid line and the red dashed line represent the expected number of clusters in $\mathcal{L}$ under a baseline prior distribution and an overdispersed prior distribution, respectively; (b) the total variation distance (TV) of the distributions of the number of clusters in the middle region under the independence prior.

3 Examples

3.1 RPPA Data

We compared the performance of the two prior structures for the MDP model with data from an experiment using reverse phase protein arrays (RPPA) (Tibes et al. 2006). The data set is introduced in Nieto-Barajas et al. (2012). The investigators treated an ovarian cancer cell line with the epidermal growth factor receptor (EGFR) inhibitor, Lapatinib, at the commencement of the experiment. The cell line was stimulated with EGFR over time. Expression intensities of 30 proteins in the ovarian cancer cell line were recorded with three replicates at eight different time points, $t = 0$, 5, 15, 30, 60, 90, 120 and 240 minutes after the initial intervention. For an analysis, the intensities were first normalized so that each protein had a median expression intensity of 1,000, and were then log-transformed. To make a comparison across proteins, difference scores developed in Tusher et al. (2001) were computed. See Nieto-Barajas et al. (2012) for details on the data. Figure 3 shows the histograms of the difference scores from the 30 proteins at the eight time points. The figure shows that the distributions are right skewed except at $t = 0$ and the skewness is different across the time points.

We applied the hierarchical MDP model with a reasonably calibrated base measure at each time point. We then reanalyzed the data with an intentionally inflated base measure and then compared the results with those from the well-calibrated base measure for each of the two prior structures. We compared the two prior structures under the



(a) $t = 0$ (b) $t = 5$ (c) $t = 15$

(d) $t = 30$ (e) $t = 60$ (f) $t = 90$

(g) $t = 120$ (h) $t = 240$

Figure 3: Histograms of difference scores of protein intensities at each time point. Note that the $t = 0$ histogram is qualitatively different from those at other time points.

two-parameter Poisson-Dirichlet model in a similar fashion.

We assumed the normal likelihood, $Y_i \overset{indep}{\sim} \mathrm{N}(X_i, v^2)$, $i = 1, \ldots, 30$. We calibrated the priors as follows: We chose the normal distribution with mean μ and variance τ^2 for a base measure of a MDP model. We fixed μ at $\bar{y}$ for simplicity. We placed inverse gamma priors, $\mathrm{IG}(a_0, b_0)$ and $\mathrm{IG}(a_v, b_v)$, on τ^2 and v^2, respectively. We set $a_0 = a_v = 10$ and chose a value of b_0 and b_v by matching their prior expectations with the sample variances.

For an arbitrarily overdispersed base measure, we increased b_0 by a factor of 25, where 25 is an arbitrary choice. We considered a gamma prior for M, $\text{Ga}(a, b)$. The gamma prior was elicited by setting the expected number of clusters and its variance. For this data set, we set the expected number of clusters and its variance to be 3 and 2 for $t = 0$, and 4 and 3 for the other time points. This results in $\text{Ga}(9.385, 0.065)$ for $t = 0$, and $\text{Ga}(9.275, 0.108)$ for the other time points for the independence prior structure. For the local-mass preserving prior structure, we defined the middle, (ℓ_1, ℓ_2) by finding ℓ_1 and ℓ_2 such that $\Phi(\tau_0^{-1}(\ell_2-\mu))-\Phi(\tau_0^{-1}(\ell_1-\mu)) = 0.68$ with $|\ell_2-\mu|= |\ell_1-\mu|$ where $\tau_0^2 = E(\tau^2)$ and Φ is the cdf of the standard normal distribution. For $M^\star \sim \text{Ga}(a^\star, b^\star)$ let $a^\star = a$ and $b^\star = \{\Phi(\tau_0^{-1}(\ell_2 - \mu)) - \Phi(\tau_0^{-1}(\ell_1 - \mu))\}b$, implying that $\text{E}(K_n \,|\, X_1, \ldots, X_n \in \mathcal{L}) = 2.43$ and $\text{Var}(K_n \,|\, X_1, \ldots, X_n \in \mathcal{L}) = 1.44$ for $t = 0$ and $\text{E}(K_n \,|\, X_1, \ldots, X_n \in \mathcal{L}) = 3.20$ and $\text{Var}(K_n \,|\, X_1, \ldots, X_n \in \mathcal{L}) = 2.20$ for the other time points.

For the two-parameter Poisson-Dirichlet process, we assumed the same base measure. We fixed $\sigma = 0.1$ and let $\theta \sim \text{Ga}(a, b)$ where a is fixed at 15 and b is fixed at 0.278 and 0.05 for the independence prior structure assuming the expected number of clusters to be 3 and 4 for $t = 0$ and the other time points, respectively. For the local-mass preserving prior, we let $a^\star = a$ and $b^\star$ calibrated conditional on τ^2 by matching the expected number of clusters in $\mathcal{L}$ with 2.43 and 3.18 for $t = 0$ and the other time points, respectively.

Tables 1 and 2 show the posterior means of parameters under the local-mass preserving prior structure and the independence prior structure for the MDP model and the two-parameter Poisson-Dirichlet process model, respectively. With the overdispersed base measure, the posterior mean of τ^2 increases, yet the posterior mean of M for the MDP model (the posterior mean of θ for the two-parameter Poisson-Dirichlet process model) does not change much under the independence prior structure (see Table 1(b) and Table 2(b)). On the other hand, under the local-mass preserving prior structure the posterior mean of τ^2 increases and the posterior mean of M also increases to preserve the mass in the middle for the MDP model as in Table 1(a). Similarly, in Table 2(a) θ increases to preserve the prior expected number of clusters in the middle for the two-parameter Poisson-Dirichlet process model.

We examined the stability of inference on the clustering of X_i's as the dispersion of the base measure increases. The difference of clustering was measured by $0.5 \sum_{i \neq j} |\text{P}_1(X_i = X_j) - \text{P}_2(X_i = X_j)|$, where P_1 and P_2 are the posterior distributions with the well-calibrated and overdispersed base measures, respectively. Table 3 shows differences in the posterior probabilities that any two X_i's are in the same cluster. From the table, we observe larger changes in the clustering pattern of X_i's under the independence prior structure at all t for both models. Under the local-mass preserving prior structure, clustering of X_i's in the middle is better preserved and this leads to smaller changes in the posterior pairwise co-clustering probabilities.

Next, we investigated how changes in the clustering of X_i's affect estimation and prediction of X_i's. To study the impact on the estimation of X_i's, we compared the posterior means of X_i with the well-calibrated base measure to those with the overdispersed base measure at each time point. Figure 4 illustrates the comparison for the MDP model at

	Well-Calibrated			Overdispersed		
	M	τ^2	v^2	M	τ^2	v^2
$t = 0$	0.615	0.004	0.003	1.814	0.084	0.003
$t = 5$	1.013	0.093	0.065	2.599	1.779	0.058
$t = 15$	1.016	0.385	0.170	2.496	6.402	0.159
$t = 30$	1.019	0.091	0.059	2.590	1.711	0.053
$t = 60$	1.017	0.379	0.152	2.436	6.052	0.143
$t = 90$	1.016	0.427	0.173	2.450	6.845	0.163
$t = 120$	1.007	0.428	0.170	2.483	6.924	0.159
$t = 240$	1.025	0.088	0.050	2.576	1.591	0.046

(a) Local-Mass Preserving Prior

	Well-Calibrated			Overdispersed		
	M	τ^2	v^2	M	τ^2	v^2
$t = 0$	0.626	0.004	0.003	0.573	0.095	0.003
$t = 5$	1.016	0.092	0.065	0.900	2.082	0.058
$t = 15$	0.975	0.393	0.169	0.891	7.576	0.158
$t = 30$	1.010	0.091	0.059	0.901	2.010	0.054
$t = 60$	0.959	0.390	0.151	0.881	7.158	0.143
$t = 90$	0.961	0.435	0.173	0.882	8.103	0.162
$t = 120$	0.959	0.439	0.169	0.890	8.180	0.158
$t = 240$	1.006	0.087	0.050	0.892	1.864	0.046

(b) Independence Prior

Table 1: Posterior means of M, τ^2 and v^2 for the MDP model.

a selected time point, $t = 240$, under either the local-mass preserving prior structure (panel (a)) or the independence prior structure (panel (b)). The smaller change in the clustering pattern under the local-mass preserving prior structure leads to more robust estimates of the X_i's, especially for those lying in the middle as evidenced by those $\hat{X}_i$'s near 0.1. Panel (a) shows that the $\hat{X}_i$'s in the middle move down almost equally due to the separation of X_i's in the far right tail and fall below the 45 degree line by about the same distance (see insert in the panel). In contrast, panel (b) indicates that in addition to the separation, larger clusters in the middle under the independence prior structure lead to additional shrinkage of X_i's in the middle with the overdispersed base measure. Results from the other time points as well as the results under the two-parameter Poisson-Dirichlet model show the same pattern.

For prediction stability in a central region, we focused on the central 90% of the posterior predictive distribution under the well-calibrated prior and compared the posterior predictive distributions conditional on the region implicitly defined by the central portion. We measured the differences in the posterior predictive densities by the total variation distance, $2\sup\{|\int_A f_1(x)\mathrm{d}x - \int_A f_2(x)\mathrm{d}x|, A \subseteq \mathcal{B}(\mathbb{R})\}$ where $\mathcal{B}(\mathbb{R})$ denotes the Borel sets on $\mathbb{R}$, and by mean integrated squared error, $\mathrm{E}\{(f_1(x) - f_2(x))^2\}$ where f_1 and f_2 are the posterior predictive density estimates under the well-calibrated base measure and

	Well-Calibrated			Overdispersed		
	θ	τ^2	v^2	θ	τ^2	v^2
$t = 0$	0.383	0.004	0.004	1.583	0.083	0.003
$t = 5$	0.619	0.093	0.065	2.156	1.802	0.058
$t = 15$	0.659	0.385	0.170	2.201	6.470	0.158
$t = 30$	0.617	0.092	0.059	2.100	1.741	0.053
$t = 60$	0.658	0.383	0.152	2.141	6.100	0.144
$t = 90$	0.649	0.434	0.172	2.108	6.918	0.163
$t = 120$	0.650	0.438	0.169	2.124	7.009	0.159
$t = 240$	0.613	0.089	0.049	2.077	1.612	0.046

(a) Local-Mass Preserving Prior

	Well-Calibrated			Overdispersed		
	θ	τ^2	v^2	θ	τ^2	v^2
$t = 0$	0.427	0.004	0.003	0.411	0.094	0.003
$t = 5$	0.757	0.092	0.065	0.710	2.080	0.059
$t = 15$	0.738	0.390	0.170	0.711	7.544	0.158
$t = 30$	0.758	0.090	0.059	0.713	2.031	0.054
$t = 60$	0.739	0.387	0.151	0.707	7.146	0.142
$t = 90$	0.737	0.436	0.173	0.704	8.078	0.162
$t = 120$	0.732	0.438	0.170	0.705	8.181	0.158
$t = 240$	0.755	0.088	0.050	0.708	1.860	0.046

(b) Independence Prior

Table 2: Posterior means of θ, τ^2 and v^2 for the two-parameter Poisson-Dirichlet process.

Prior	$t = 0$	$t = 5$	$t = 15$	$t = 30$	$t = 60$	$t = 90$	$t = 120$	$t = 240$
Local-mass	43.0	41.4	30.9	41.6	28.2	32.1	30.6	35.3
Indep.	83.5	92.7	77.8	95.1	69.0	69.7	68.4	88.5

(a) MDP Model

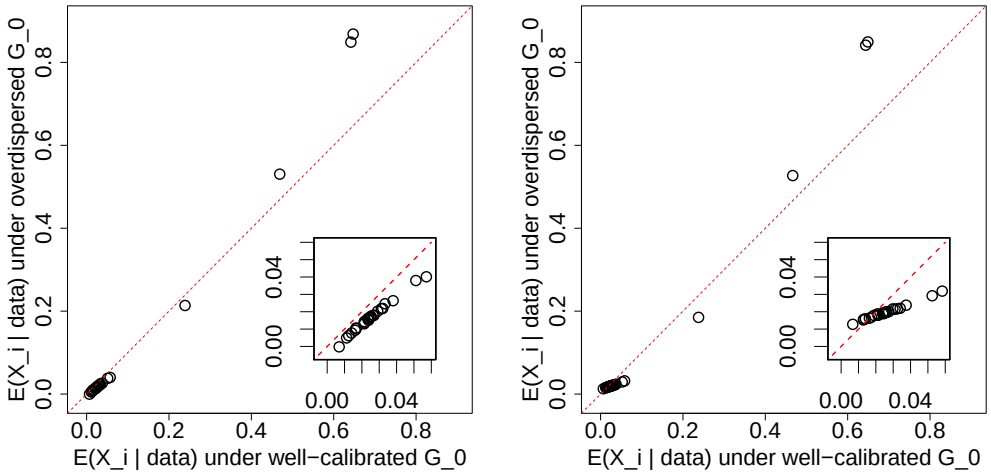
Local-mass	43.0	42.5	23.9	40.0	27.3	27.6	22.6	32.3
Indep.	84.3	96.3	71.4	99.9	65.1	69.9	66.7	88.8

(b) Two-parameter Poisson-Dirichlet Model

Table 3: Sum of the absolute differences in the posterior probabilities that each pair of X_i's is in the same cluster. Large numbers indicate instability of inference on clustering as the prior dispersion is varied.

the overdispersed base measure, respectively. We evaluated both quantities numerically. From Tables 4 and 5 the predictive density estimates in the middle change less under the local-mass preserving prior structure for the both models, except for $t = 0$.

Note that a qualitative difference between $t = 0$ and the other time points explains why



(a) Local-mass preserving prior at $t = 240$ (b) Independence prior at $t = 240$

Figure 4: Plots of the posterior means of X_i for a selected time point, $t = 240$. Panel (a) plots the posterior mean of X_i with the well-calibrated base measure against the overdispersed base measure under the local-mass preserving prior structure. Panel (b) plots the posterior means of X_i under the independence prior structure.

Prior	$t = 0$	$t = 5$	$t = 15$	$t = 30$	$t = 60$	$t = 90$	$t = 120$	$t = 240$
Local-mass	0.164	0.14	0.072	0.114	0.064	0.066	0.068	0.116
Indep.	0.134	0.17	0.114	0.136	0.106	0.112	0.108	0.150

(a) MDP Model

Local-mass	0.177	0.141	0.068	0.115	0.053	0.060	0.060	0.103
Indep.	0.135	0.164	0.116	0.139	0.105	0.111	0.111	0.153

(b) Two-parameter Poisson-Dirichlet Model

Table 4: Total variation distance under the well calibrated base measure and the overdispersed base measure. The differences are computed for middle areas having posterior probability of 0.9 under the predictive density estimates with the well-calibrated base measure.

the differences in prediction at $t = 0$ are not smaller under the local-mass preserving prior structure (see Figure 3). In the case of $t = 0$, there are no aberrant outliers so that τ_0^2 in the prior calibration is small, the area of the pre-defined middle at $t = 0$ is small, and so relatively many observations fall out of the pre-defined middle area. Therefore, the pre-defined middle receives much less posterior probability than 0.9 even with the well-calibrated base measure. Consequently, the difference in the predictive density estimates for those areas remains relatively large under the local-mass preserving prior

Prior	$t = 0$	$t = 5$	$t = 15$	$t = 30$	$t = 60$	$t = 90$	$t = 120$	$t = 240$
L-M	8.1E-05	2.7E-05	7.7E-06	2.0E-05	9.4E-07	6.4E-06	7.0E-06	1.8E-05
Indep.	5.6E-05	4.3E-05	1.9E-05	2.9E-05	2.5E-06	1.8E-05	1.7E-05	3.2E-05

(a) MDP Model

L-M	9.7E-05	2.9E-05	7.0E-06	2.1E-05	6.4E-07	5.4E-06	5.6E-06	1.4E-05
Indep.	5.7E-05	4.0E-05	2.0E-05	3.0E-05	2.5E-06	1.8E-05	1.9E-05	3.4E-05

(b) Two-parameter Poisson-Dirichlet Model

Table 5: Mean integrated squared error of the predictive density estimates under the well calibrated base measure and the overdispersed base measure. The differences are computed for middle areas having posterior probability of 0.9 under the predictive density estimates with the well-calibrated base measure.

structure at $t = 0$.

3.2 Allometric Data

We analyzed an allometric data set from Weisberg (1985). The data include the body mass and brain mass of 62 species of mammals. A simple linear regression model has been used to predict log brain mass from log body mass in previous analyses (for example, see Weisberg (1985)). A logarithmic transformation of the variables suggested by theoretical considerations yields approximate linearity and constant variance. For our analysis we recentered the transformed covariate and placed independent priors on the regression coefficients. See Figure 5 for a plot of the transformed and recentered data with a simple linear regression line. MacEachern and Guha (2011) applied an MDP model with the independence prior structure to this data set to show that allowing an arbitrary distribution for errors through an MDP model achieves a better fit than does a parametric model. Their model is shown below:

$$
\begin{aligned}
\beta \sim \mathrm{N}(\mu_\beta, \sigma_\beta^2); \qquad \gamma \sim \mathrm{N}(\mu_\gamma, \sigma_\gamma^2); \qquad & \tau^2 \sim \mathrm{IG}(a_0, b_0); \qquad v^2 \sim \mathrm{IG}(a_v, b_v) \\
\mathrm{G}|\tau^2 \sim \mathrm{DP}(\alpha) \qquad & \text{where } M = 1 \text{ and } \mathrm{G}_0 = \mathrm{N}(0, \tau^2) \\
\psi_i|\mathrm{G} \overset{iid}{\sim} \mathrm{G}; \qquad & \eta_i|\sigma^2 \overset{iid}{\sim} \mathrm{N}(0, v^2) \\
\epsilon_i &= \gamma + \psi_i + \eta_i \\
Y_i &= X_i\beta + \epsilon_i.
\end{aligned}
$$

To illustrate the performance of the local-mass preserving prior, we extended the above model by placing a prior on M.

We first calibrated the priors as follows: Peters (1983) presents empirical power laws of the form $Y_i' = e^\alpha X_i'^\beta$ to explain the relation of the body mass (X') to some other physiological characteristic of interest such as brain mass (Y'). In an empirical allometric

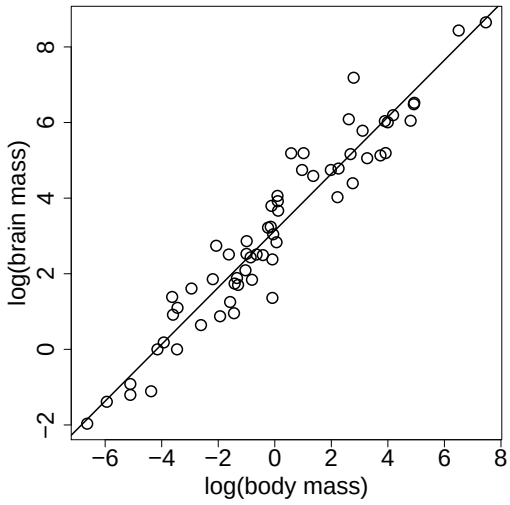
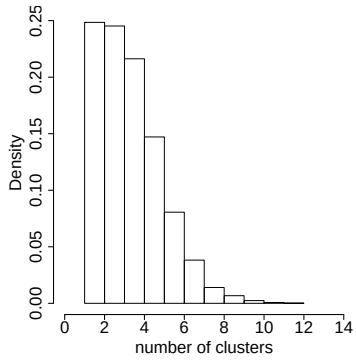


Figure 5: Plot of the allometric data. The line represents a simple linear regression line, $\hat{y}_i = 3.14 + 0.75x_i$ where y_i= log of brain mass and x_i= centered log of body mass.

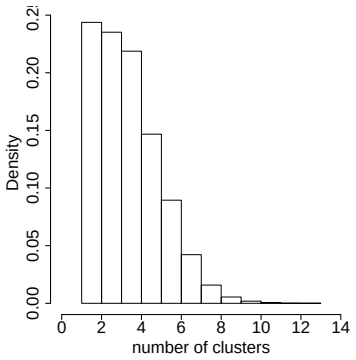
theory, β is stated to be approximately equal to 3/4 for our example. Following this, we set $\mu_\beta = 0.75$ and $\sigma^2_\beta = 1/16$. We let $\mu_\gamma = 3.14$, $\sigma^2_\gamma = 100$ and $a_0 = b_0 = a_\sigma = b_\sigma = 2$, similar to MacEachern and Guha (2011). To make the base measure overdispersed, we increased b_0 by a factor of 25 where 25 was chosen arbitrarily. We chose a gamma distribution as a prior for M. The hyperparameters for the gamma prior were found by setting the expected number of clusters and its variance to be 4 and 3 as we did in Section 3.1. The resulting prior for M is Ga(14.80, 0.05). The prior elicitation under the local-mass preserving prior structure is the same except for the prior on M. To define a prior on $M^\star$ under the local-mass preserving prior structure, the middle, $(-\ell, \ell)$ is defined by finding ℓ such that $\Phi(\ell/\tau_0) - \Phi(-\ell/\tau_0) = 0.68$ with $\tau_0^2 = \mathrm{Var}(y - x\mu_\beta - \mu_\gamma)$. Assuming $M^\star \sim \mathrm{Ga}(a^\star, b^\star)$ let $a^\star = a$ and $b^\star = \{\Phi(\ell/\tau_0) - \Phi(-\ell/\tau_0)\}b$. This prior implies that $\mathrm{E}(K_n \,|\, X_1, \ldots, X_n \in \mathcal{L}) = 3.15$ and $\mathrm{Var}(K_n \,|\, X_1, \ldots, X_n \in \mathcal{L}) = 2.15$.

We investigated changes in the clustering of ψ_i's under the two prior structures as the base measure becomes overdispersed. Figure 6 shows histograms of the number of clusters of ψ_i's, K. Note that the inferences with the well-calibrated base measure are identical under the two prior structures. Histograms of K with the well-calibrated base measure are shown in panels (a) and (c) under the local-mass preserving prior structure and the independence prior structure, respectively. Panels (b) and (d) show histograms of K with the overdispersed base measure. The number of clusters slightly decreases under the local-mass preserving prior structure and yet it greatly decreases under the independence prior structure. In the latter, the ψ_i's are partitioned into one or two clusters most of time. This shows that the clustering pattern of the ψ_i's changes much more under the independence prior structure than under the local mass prior structure.

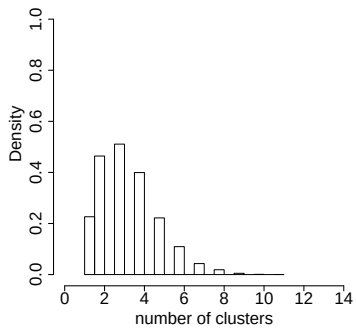
The changes in the clustering of the ψ_i's affects the estimation of $\gamma + \psi_i$. Figure 7 plots



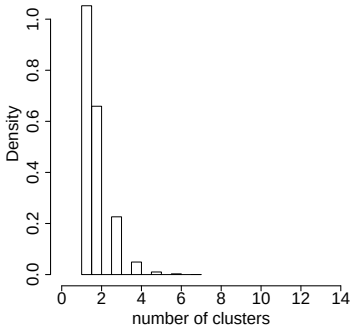
(a) Local-mass with well calibrated G_0



(b) Local-mass with overdispersed G_0



(c) Indep. with well calibrated G_0

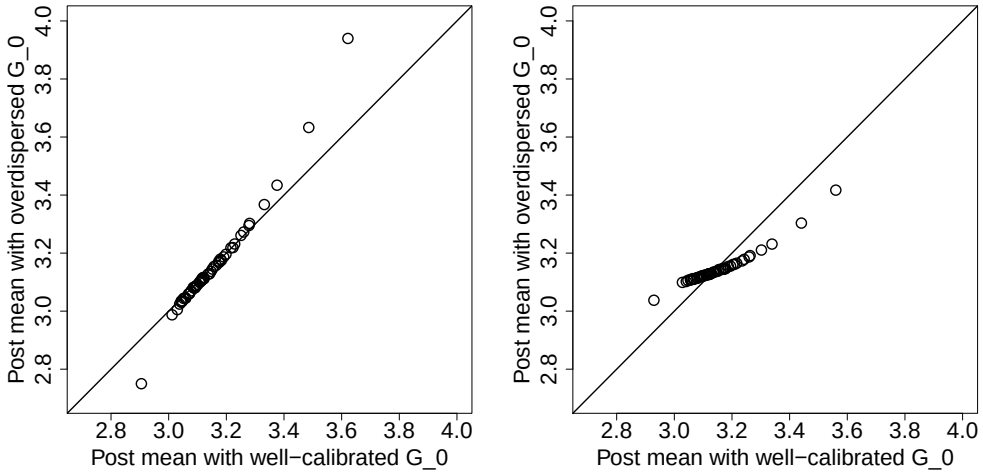


(d) Indep. with overdispersed G_0

Figure 6: Histograms of the number of clusters (K). Panels (a) and (b) are histograms of K under the local-mass preserving prior structure with the well-calibrated base measure and the overdispersed base measure, respectively. Panels (c) and (d) are histograms of Kunder the independence prior structure.

the posterior mean of $\gamma+\psi_i$ with the well-calibrated base measure against a similar value under the overdispersed base measure. In Figure 7 (a), the posterior mean of $\gamma + \psi_i$ in the middle changes a little under the local mass preserving prior structure as the base measure is overdispersed. In addition, since ψ_i's in the tails are less frequently clustered with the others, the estimate of $\gamma + \psi_i$ increases (decreases) for those ψ_i's in the right (left) tail. However, Figure 7 (b) demonstrates a large change in the posterior mean of $\gamma + \psi_i$ under the independence prior structure. Assigning ψ_i's to one or two clusters with the overdispersed base measure under the independence prior structure results in more shrinkage of $\gamma + \psi_i$ toward their center, resulting in the lack of robustness in the estimation of $\gamma + \psi_i$.

Figure 8 illustrates the posterior density estimates of β and v^2 where the top panels are from the local-mass preserving prior structure and the bottom panels from the

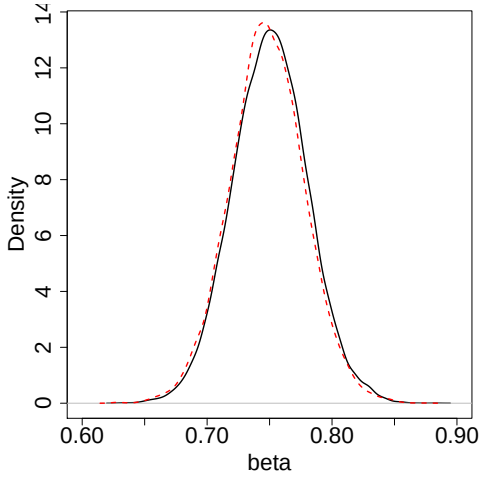


(a) Local-mass preserving prior (b) Independence prior

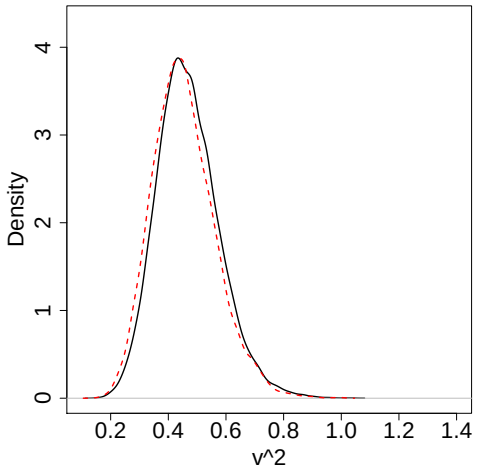
Figure 7: Plots of the posterior mean of $\gamma + \psi_i$. The left panel plots the posterior mean of $\gamma + \psi_i$ with the well-calibrated base measure against the overdispersed base measure under the local-mass preserving prior structure. The right panel plots the posterior mean of $\gamma + \psi_i$ under the independence prior structure.

independence prior structure. The figure shows how the change in the clustering of the ψ_i's affects estimation of the other parameters such as β and v^2. In particular, the posterior distribution of β changes very little under the two prior structures as the base measure is overdispersed as shown in panels (a) and (c). Interestingly, the posterior distributions of v^2 in panels (b) and (d) behave differently under the two prior structures as the base measure becomes overdispersed. Under the independence prior structure, all of the $\hat{\psi}_i$ shrink more toward their center, many of the ϵ_i become larger and so the posterior distribution of v^2 moves to the right as shown in (d). However, under the local mass preserving prior structure, the posterior distribution of v^2 stays approximately the same as shown in (b).

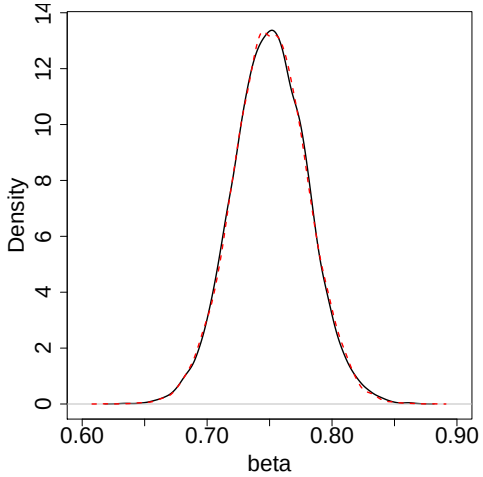
We also applied the two-parameter Poisson-Dirichlet model with the two prior structures to the allometric data. Similar conclusions are obtained. The clustering of the ψ_i is approximately preserved under the local-mass preserving prior structure as the dispersion of the base measure increases and yet the clustering greatly changes under the conventional model.



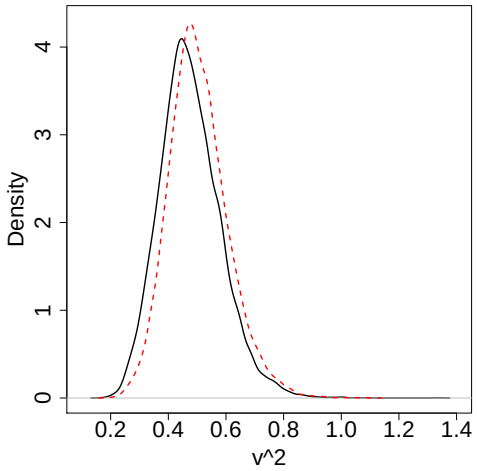
(a) Posterior density estimates of β under the local-mass



(b) Posterior density estimates of v^2 under the local-mass



(c) Posterior density estimates of β under the indep.



(d) Posterior density estimates of v^2 under the indep.

Figure 8: Plots of kernel density estimates of the posterior densities of β and v^2. The black solid lines and the red dashed lines represent density estimates with the well-calibrated base measure and the overdispersed base measure, respectively.

4 Conclusions

The local-mass preserving prior robustifies the conventional hierarchical MDP model. The concept of preserving local mass is extended to the two-parameter Dirichlet-Poisson

process model. Through this modeling strategy, we can preserve local mass in regions of interest under an overdispersed base measure. This leads to more stable inference for clustering, and it leads to much more stable predictive inference in the central region, arguably the region of greatest interest. If interest focuses on these issues, we recommend use of this prior structure.

The essence in the development can be easily extended to other models and other settings. Although illustrated most easily for real-valued X_i, the notion of a "middle" and clustering can be generalized in straightfoward fashion. Remark 1 can be recast with $\mathcal{L}$ representing a set into which X_i falls or does not fall. Definition 1 need not be tied to a gamma distribution, and so on. As an example of another model, countable mixture models whose distribution of mixing weights depends on a different rule than the two-parameter Poisson-Dirichlet process (Navarrete et al. 2008) can be modified appropriately. The key is to decrease the weights at an appropriate rate as the dispersion of the base measure increases.

Finally, high dimensional problems are quite complex. The problem of specifying a reasonable joint prior distribution over many parameters is often difficult, and the requisite computations complicated and time consuming. In these cases, conjugacy is helpful. The strategy developed here allows us to retain a conjugate structure, leading to simplification of the calculation of the posterior distribution. Moreover, by focusing on the local behavior of parameters of interest, the resulting posterior inference is less sensitive to prior misspecification. This is especially valuable in settings where prior elicitation is difficult. We would argue that in the infinite dimensional setting of nonparametric Bayesian models the prior distribution is always misspecified. Here again, we recommend use of a prior structure that lends stability to the most important inferences.

Acknowledgments

Steven N. MacEachern's work was supported by the NSF under grant numbers DMS-1007682, DMS-1209194 and H98230-10-1-0202. The views in this paper are not necessarily those of the NSF.

References

Antoniak, C. E. (1974). "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems." *The Annals of Statistics*, 2: 1152–1174. 308, 310

Berger, J. O. (1993). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag Inc. 308

Berger, J. O. and Bernardo, J. M. (1992). "On the development of the reference prior method." In Bernardo, J. M. e., Berger, J. O. e., Dawid, A. P. e., and Smith, A. F. M. e. (eds.), *Bayesian Statistics 4. Proceedings of the Fourth Valencia International Meeting*, 859. Clarendon Press [Oxford University Press]. 307

Berger, J. O., Bernardo, J. M., and Sun, D. (2009). "The Formal Definition of Reference Priors." *The Annals of Statistics*, 37(2): 905–938. 307

Bernardo, J. M. (1979). "Reference Posterior Distributions for Bayesian Inference (C/R P128-147)." *Journal of the Royal Statistical Society, Series B: Methodological*, 41: 113–128. 307

Blackwell, D. and MacQueen, J. B. (1973). "Ferguson Distributions Via Pólya Urn Schemes." *The Annals of Statistics*, 1: 353–355. 310

Bush, C. A., Lee, J., and MacEachern, S. N. (2010). "Minimally informative prior distributions for non-parametric Bayesian analysis." *Journal of the Royal Statistical Society, Series B: Methodological*, 72: 253–268. 308, 311

De Finetti, B. (1975). *Theory of Probability: A Critical Introductory Treatment, Vol. 2*. John Wiley & Sons. 307

Escobar, M. D. (1994). "Estimating Normal Means with a Dirichlet Process Prior." *Journal of the American Statistical Association*, 89: 268–277. 308, 310

Escobar, M. D. and West, M. (1995). "Bayesian Density Estimation and Inference Using Mixtures." *Journal of the American Statistical Association*, 90: 577–588. 308, 310, 314

Ferguson, T. S. (1973). "A Bayesian Analysis of Some Nonparametric Problems." *The Annals of Statistics*, 1: 209–230. 308, 309

Ghosal, S. and van der Vaart, A. W. (2001). "Entropies and Rates of Convergence for Maximum Likelihood and Bayes Estimation for Mixtures of Normal Densities." *The Annals of Statistics*, 29(5): 1233–1263. 308

Hirano, K. (2002). "Semiparametric Bayesian Inference in Autoregressive Panel Data Models." *Econometrica*, 70(2): 781–799. 311

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (eds.) (2010). *Bayesian Nonparametrics*. Cambridge University Press. 308

Ishwaran, H. and James, L. F. (2002). "Approximate Dirichlet Process Computing in Finite Normal Mixtures: Smoothing and Prior Information." *Journal of Computational and Graphical Statistics*, 11(3): 508–532. 308

Jeffreys, H. (1998). *Theory of Probability*. Oxford University Press. 307

Ji, C., Merl, D., Kepler, T. B., and West, M. (2009). "Spatial Mixture Modelling for Unobserved Point Processes: Examples in Immunofluorescence Histology." *Bayesian Analysis*, 4(2): 297–316. 311

Kleijn, B. J. K. and van der Vaart, A. W. (2006). "Misspecification in Infinite-dimensional Bayesian Statistics." *The Annals of Statistics*, 34(2): 837–877. 307

Kottas, A., Müller, P., and Quintana, F. (2005). "Nonparametric Bayesian modeling for multivariate ordinal data." *Journal of Computational and Graphical Statistics*, 14(3): 610–625. 310

Lindley, D. (1965). *Introduction to Probability and Statistics*. Cambridge University. 307

Liu, J. S. (1996). "Nonparametric hierarchical Bayes via sequential imputations." *The Annals of Statistics*, 24(3): 911–930. 310

MacEachern, S. N. and Guha, S. (2011). "Parametric and Semiparametric Hypotheses in the Linear Model." *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 39(1): 165–180. 308, 322, 323

Navarrete, C., Quintana, F., and Müller, P. (2008). "Some Issues on Nonparametric Bayesian Modeling Using Species Sampling Models." *Statistical Modelling International Journal*, 8(1): 3–21. 327

Nieto-Barajas, L., Müller, P., Ji, Y., Lu, Y., and Mills, G. (2012). "A Time-Series DDP for Functional Proteomics Profiles." *Biometrics*, 68(3): 859–868. 316

Peters, R. H. (1983). *The Ecological Implications of Body Size*. Cambridge: Cambridge University Press. 322

Pitman, J. (1996). "Some developments of the Blackwell-MacQueen urn scheme." *Lecture Notes-Monograph Series*, 245–267. 308, 314

Pitman, J. and Yor, M. (1997). "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator." *The Annals of Probability*, 25(2): 855–900. 308, 314

Quintana, F. A. (2006). "A Predictive View of Bayesian Clustering." *Journal of Statistical Planning and Inference*, 136(8): 2407–2429. 308

Quintana, F. A. and Iglesias, P. L. (2003). "Bayesian Clustering and Product Partition Models." *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 65(2): 557–574. 308

Rousseau, J. (2010). "Rates of Convergence for the Posterior Distributions of Mixtures of Betas and Adaptive Nonparametric Estimation of the Density." *The Annals of Statistics*, 38(1): 146–180. 308

Salinetti, G. (2003). "New Tools for Consistency in Bayesian Nonparametrics." In *Bayesian Statistics 7*, 369–384. Oxford University Press. 307

Savage, L. J. (1972). *The Foundations of Statistics*. Dover Publications, Inc. 307

Tibes, R., Qiu, Y., Lu, Y., Hennessy, B., Andreeff, M., Mills, G. B., and Kornblau, S. M. (2006). "Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells." *Molecular cancer therapeutics*, 5(10): 2512–2521. 316

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). "Significance analysis of microarrays applied to the ionizing radiation response." In *Proceedings of the National Academy of Sciences of the United States of America, National Academy of Sciences*, volume 98, 5116–5121. Washington, D.C. 316

Weisberg, S. (1985). *Applied Linear Regression*. John Wiley & Sons. 322