Stat Comput (2006) 16: 5–14
DOI 10.1007/s11222-006-5196-2

Nonparametric empirical Bayes for the Dirichlet process mixture model

Jon D. McAuliffe · David M. Blei · Michael I. Jordan

Received: January 2005 / Accepted: September 2005

Abstract The Dirichlet process prior allows flexible nonparametric mixture modeling. The number of mixture components is not specified in advance and can grow as new data arrive. However, analyses based on the Dirichlet process prior are sensitive to the choice of the parameters, including an infinite-dimensional distributional parameter G_0. Most previous applications have either fixed G_0 as a member of a parametric family or treated G_0 in a Bayesian fashion, using parametric prior specifications. In contrast, we have developed an adaptive nonparametric method for constructing smooth estimates of G_0. We combine this method with a technique for estimating α, the other Dirichlet process parameter, that is inspired by an existing characterization of its maximum-likelihood estimator. Together, these estimation procedures yield a flexible empirical Bayes treatment of Dirichlet process mixtures. Such a treatment is useful in situations where smooth point estimates of G_0 are of intrinsic interest, or where the structure of G_0 cannot be conveniently modeled with the usual parametric prior families. Analysis of simulated and real-world datasets illustrates the robustness of this approach.

Keywords Bayesian nonparametrics · Density estimation · Clusetring · Kernel density estimates · Gibbs sampling · Pólya urn schemes

J. D. McAuliffe
Statistics Department, University of California, Berkeley, CA 94720

D. M. Blei
Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213

M. I. Jordan
Statistics Department and Computer Science Division, University of California, Berkeley, CA 94720

1. Introduction

Mixture modeling is an effective and widely practiced density estimation method, capable of representing the phenomena that underlie many real-world datasets. However, a long-standing difficulty in mixture analysis is choosing the number of mixture components. Model selection methods, such as cross-validation or methods based on Bayes factors, treat the number of components as an unknown constant and set its value based on the observed data. Such an approach can be too rigid, particularly if we wish to model the possibility that new observations come from as yet unseen components.

The Dirichlet process (DP) mixture model, studied in Bayesian nonparametrics (Ferguson 1973, Blackwell and MacQueen 1973, Berry and Christensen 1979, Escobar and West 1995, Liu 1996, MacEachern and Müller 1998, Neal 2000, Ishwaran and James 2002), allows for exactly this possibility. When a new data point arrives, it either shares the component of some previously drawn value, or it uses a newly generated component realized from a distribution G_0. The frequency with which new components are generated is controlled by a parameter $\alpha > 0$. The DP mixture model has found widespread application in recent statistical research (Carota and Parmigiani 2002, Gelfand and Kottas 2002, Ishwaran and James 2002, Brown and Ibrahim 2003, Iorio *et al*. 2004, Müller, Quintana and Rosner 2004, Teh *et al*. 2004).

Analyses based on DP mixtures are sensitive to the choice of the Dirichlet process parameters G_0 and α; the need to treat these parameters carefully is discussed by Escobar and West (1995). In applications, some previous authors have chosen a fixed G_0, such as a normal with large variance relative to the data (Ishwaran and James 2002) or a Beta(a, b), where a and b are stipulated without further discussion (Liu 1996). Other authors have used a parametric family of priors for G_0, such

as uniform-inverse Wishart (MacEachern and Müller 1998) or normal-inverse gamma (MacEachern and Müller 1998), with fixed hyperprior distributions on the parameters of the prior family. The hyperprior distributions are usually chosen by sensitivity analysis, or else they are chosen to be diffuse with respect to the observed data.

Such approaches can succeed in circumstances where the true G_0 has a form that is well approximated by a simple fixed choice or by some member of a typical parametric prior model. On the other hand, the parametric approach can have adverse consequences when the true G_0 exhibits complicated structure. For example, G_0 could be multimodal, skewed, heavy-tailed, or all of these in combination. Furthermore, there is sometimes scientific interest in the structure of G_0 itself, not just in inferences using a DP mixture based on G_0. In such cases, parametric families can create obstacles. For example, a Bayes point estimate inside a normal family for G_0 will necessarily be symmetric and unimodal.

In this work we present a procedure for smooth, adaptive estimation of an arbitrary continuous G_0. This poses some difficulties, because G_0 is an infinite-dimensional distributional parameter whose realizations are never directly observed. We address this problem using approximate posterior inference over flexible, model-free nonparametric density estimators. We also employ an approximate maximum-likelihood estimator (MLE) for α, based closely on a characterization of the MLE due to Liu (1996). When combined in parallel, these two estimation procedures yield a nonparametric empirical Bayes approach to handling the parameters (G_0, α) of the DP mixture model. As we will see in Section 4, this approach is robust to a variety of choices for the true G_0, and works especially well when the true α and the number of observations are not too small.

There have been studies of related ideas. Berry and Christensen (1979) used an empirical Bayes approach in the context of binomial DP mixtures. But they assume that G_0 and α are known, in order to focus on posterior point estimates of G, the realization of the Dirichlet process in the mixture model. By contrast, our method treats G as a fully random quantity to be marginalized for purposes of inference, and so incorporates more of the structured uncertainty from the DP mixture. The same comment applies to Yu, Tresp and Yu (2004) and Ishwaran and James (2002), which both describe methods for constructing point estimates of G. Yu, Tresp and Yu (2004) also assume a fixed G_0. Teh *et al*. (2004) put a further Dirichlet process prior on G_0, and so still face the issue of a hyperprior distributional parameter G_{00}. This is also true of Tomlinson and Escobar (1999), though these authors use a Dirichlet process mixture, rather than a Dirichlet process, in order to obtain smooth realizations of G_0. In both cases, the focus is on specifying hierarchical nonparametric models for grouped data. To our knowledge, smooth nonparametric point estimates of G_0 in the DP mixture model have not been studied.

Note moreover that several authors (Escobar and West 1995, Ishwaran and James 2002) fix a G_0 from a parametric family, or choose fixed hyperprior distributions for a parametric family of G_0's, based on the observed data. When this strategy is not accompanied by sensitivity analysis, it is in effect a form of parametric empirical Bayes in the DP mixture model, to which our method is a nonparametric counterpart.

2. Dirichlet process mixture models

Recall that in the *finite* mixture model, each data point is drawn from one of k fixed, unknown distributions. For example, the simplest Gaussian mixture assumes that each observation has been drawn from one of k Gaussians, parameterized by k different means.

To allow the number of mixture components to grow with the data, we move to the *general* mixture model setting. This is best understood as the hierarchical graphical model of Fig. 1(A); the conditional hierarchical relationships are summarized as follows:

$$X_n \mid \theta_n \sim p(\cdot \mid \theta_n), \quad n = 1, \ldots, N, \tag{1}$$

$$\theta_n \mid G \sim G(\cdot), \quad n = 1, \ldots, N. \tag{2}$$

If the unknown G is a discrete distribution on a finite set of values, this setup reduces to the finite mixture model.

Of course, estimation and inference in the general mixture model must somehow account for G. Bayesian nonparametric methods view G as an (infinite-dimensional) parameter, requiring a prior distribution in the usual way. Since G itself is a probability distribution on the θ parameter space, this prior is a distribution on probability distributions, making G (viewed as a random variable) a *random distribution*. In other

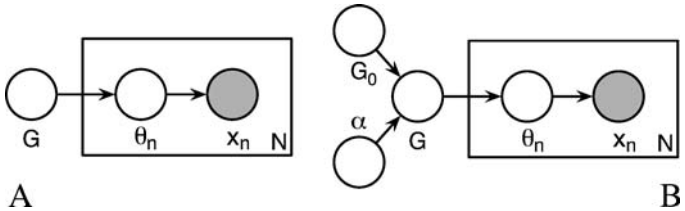

Fig. 1 (A) The general mixture model and (B) the Dirichlet process mixture model, as graphical models. Panel (B) adds to Panel (A) a nonparametric Dirichlet process prior on G, with parameters G_0 and α

words, the fully Bayesian framework treats the probability distribution G as a random quantity whose realizations are members of $\mathcal{G}$, the family of all probability distributions on the θ parameter space.

One class of Bayesian nonparametric techniques uses a distribution on distributions called the *Dirichlet process* (Ferguson 1973). A Dirichlet process (DP) is parameterized by a scaling parameter $\alpha > 0$ and a base measure G_0, which, like realizations of G, is a probability distribution on the θ parameter space. Using this prior on G leads to the *Dirichlet process mixture model (DP mixture)* of Fig. 1(B), corresponding to the following extension of the hierarchical specification in Eqs. (1) and (2):

$$X_n \mid \theta_n \sim p(\cdot \mid \theta_n), \quad n = 1, \ldots, N, \tag{3}$$

$$\theta_n \mid G \sim G(\cdot), \quad n = 1, \ldots, N, \tag{4}$$

$$G \mid G_0, \alpha \sim DP(\cdot \mid G_0, \alpha). \tag{5}$$

For example, a real-valued dataset may be modeled by a Gaussian DP mixture: θ_n is the mean parameter to the Gaussian from which X_n was drawn, the Gaussian variance is fixed, and G_0 is an arbitrary distribution on $\Re$.

Blackwell and MacQueen (1973) show that the joint distribution induced on $\theta_{1:N}$ by the DP mixture model corresponds to a generalized Pólya urn scheme. Thus, the θ_n's are exchangeable and exhibit a so-called clustering effect: given $\theta_{1:(n-1)}$, θ_n is either equal to one of $\theta_{1:(n-1)}$, or it is an independent draw from G_0:

$$p(\theta_n \mid \theta_{1:n-1}) \propto \alpha G_0(\theta_n) + \sum_{i=1}^{n-1} \delta(\theta_i, \theta_n). \tag{6}$$

This means the $\theta_{1:N}$ are randomly partitioned into subsets, or *classes*, within which all parameters take on the same value. The distribution over partitions is obtained from the Pólya urn scheme. Letting $\theta^*_{1:k}$ denote the k *distinct* values in $\theta_{1:(n-1)}$, the next draw from the DP follows the urn distribution

$$\theta_n \mid \theta_{1:n-1} = \begin{cases} \theta_i^* & \text{with prob.} \quad \dfrac{n_i}{n-1+\alpha} \\ \theta, \theta \sim G_0 & \text{with prob.} \quad \dfrac{\alpha}{n-1+\alpha}, \end{cases} \tag{7}$$

where n_i is the number of times the value θ_i^* occurs in $\theta_{1:(n-1)}$. We can uniquely specify $\theta_{1:N}$ using the $\theta^*_{1:k}$ and class assignments $z_{1:N}$, such that $\theta_n = \theta^*_{z_n}$ for all n.

From the clustering properties of the DP described above, the data $x_{1:N}$ can also be partitioned into classes, each of whose members share the same distinct parameter value θ_i^*. Thus, as with a finite mixture model, the data are drawn from k unique distributions $p(\cdot \mid \theta_i^*)$. However, in a DP mixture, the number k of such distributions is randomly determined by the urn model. Furthermore, a new observation x_{new} might be drawn using a parameter value which did not appear during the generation of $x_{1:N}$, i.e., x_{new} might come from a new, previously unseen class.

Given data $x_{1:N}$ and parameters (G_0, α), the DP mixture yields a posterior distribution on $\theta_{1:N}$, the parameters associated with each data point. This posterior is analytically intractable, but Markov chain Monte Carlo (MCMC) methods such as Gibbs sampling are used in practice to construct approximations (Neal 2000).

The parameters G_0 and α play important roles in the distribution of $\theta_{1:N}$. The probability that θ_n differs from all previously drawn parameter values is proportional to α; thus, larger values of this parameter lead to a higher probability of many unique values of θ_n (i.e., a larger k). More importantly, the θ_i^* are iid draws from G_0, which is also the distribution of the next unique parameter value. While the estimation of α has been previously addressed in several ways, a smooth nonparametric point estimate of G_0 has not been studied. Such a treatment is the primary focus of this work.

3. Empirical Bayes and Dirichlet process mixtures

The empirical Bayes (EB) method synthesizes inferential calculations common in Bayesian analysis with point estimation of the frequentist type. The terminology and ideas originate with Robbins (1955) and have been developed extensively in the statistical literature; a readable reference is Maritz and Lwin (1989). Typical applications treat various specializations of the general mixture model given in Eqs. (1) and (2) and Fig. 1(A).

In the classical EB approach to the general mixture model, we first construct a point estimate $\hat{G}$ of G (construed for the moment as a fixed, unknown distribution rather than a random measure), using realizations $x_{1:N}$ from the model. For example, if G is assumed to lie in a parametric class $\{G(\cdot \mid \lambda) : \lambda \in \Re^d\}$, we may obtain a point estimate $\hat{G}(\cdot) = G(\cdot \mid \hat{\lambda})$ by computing the maximum-likelihood estimate (MLE) $\hat{\lambda}$ under the marginal likelihood $\ell(\lambda) = p(x_{1:N} \mid \lambda) = \int p(x_{1:N} | \theta_{1:N}) G(\theta_1 \mid \lambda) \ldots G(\theta_N \mid \lambda) d\theta_{1:N}$. Then, probabilistic quantities of interest, such as the predictive density $p(x_{\text{new}} \mid x_{1:N})$ or the marginal posteriors $p(\theta_n \mid x_{1:N})$, are computed after replacing G with $\hat{G}$.

Subjecting the DP mixture model of Fig. 1(B) to the EB method is conceptually straightforward. Though the DP mixture hierarchy has one more level than the general mixture model of Fig. 1(A), the empirical Bayes rationale nonetheless continues to apply at the top. Point estimates $\hat{G}_0$ and $\hat{\alpha}$ result from an appropriate procedure; then some form of

approximate inference proceeds with $\hat{G}_0$ and $\hat{\alpha}$ supplied in place of G_0 and α.

The empirical Bayes paradigm can make probabilistic inferences in the DP mixture more robust by avoiding the need to pre-specify top-level parameters when there is little prior knowledge available. An alternative means of achieving robustness in this situation is a fully Bayesian specification of vague priors for G_0 and α. Since vague priors on infinite-dimensional parameters like G_0 can be difficult to formulate, a choice is usually made to put prior probability one on a parametric subfamily for G_0. A hierarchy of hyperpriors can then be employed to increase robustness, but the corresponding marginal prior for G_0 will still put no mass on a large class of smooth distributions. This is undesirable when we don't have strong prior beliefs about G_0 beyond smoothness. Furthermore, while a misspecified prior for G_0 may not greatly affect inferential quantities like the predictive density, its impact on a posterior point estimate of G_0 can be considerable.

We now describe our nonparametric empirical Bayes procedure for the DP mixture model. The approach alternates between an estimation phase and an inference phase. In the inference phase, fixed point estimates $(\hat{G}_0, \hat{\alpha})$ from the previous estimation phase are used to sample from the posterior distribution $p(\theta_{1:N} \mid x_{1:N})$, with G marginalized out. In general, $\hat{G}_0$ will not be conjugate to the mixture kernel $p(x \mid \theta)$, which creates additional difficulties in sampling. To handle this, we use the nonconjugate Gibbs sampler called Algorithm 8 in Neal (2000). The estimation phase uses the newly sampled values of $\theta_{1:N}$ to construct updated point estimates $(\hat{G}_0, \hat{\alpha})$, as we now describe. Initialization of our procedure is discussed in Section 4.1.

To motivate our estimate of G_0, consider how we might proceed if the true $\theta^*_{1:k}$, which generated the observed data, were revealed. Since these variables are independent and identically distributed according to G_0, ideas from frequentist nonparametric density estimation can be applied (Silverman 1986). Specifically, one could construct a kernel density estimate (KDE) of the form $\hat{G}^*_0(\cdot) = (1/k)\sum_{i=1}^{k} \kappa_h(\theta^*_i, \cdot)$. Here, $\kappa_h(\theta^*_i, \cdot)$ is a smooth, positive, integrable function, one copy centered at each θ^*_i, with dispersion controlled by the *bandwidth* parameter h. The KDE spreads the concentrated mass of the empirical distribution on $\theta^*_{1:k}$ smoothly over the θ domain.

The difficulty with this idea, of course, is that we do not observe $\theta^*_{1:k}$. However, the inference phase supplies a set of B approximate realizations from the posterior distribution on $\theta^*_{1:k}$. Thus it is natural first to replace the KDE $\hat{G}^*_0(\cdot)$ with its posterior mean, then substitute a Monte Carlo approximation thereto. Let the random variable $K(\theta_{1:N})$ be the true, unknown number of θ^*_i's, and let K_b equal the number of constituents in the bth draw $\theta^{b*}_{1:K_b}$ from the Gibbs sampler. Then we have

$$
\begin{aligned}
E(\hat{G}^*_0(\cdot) \mid X_{1:N}) &= E\left(\frac{1}{K(\theta_{1:N})}\sum_{i=1}^{K(\theta_{1:N})} \kappa_h(\theta^*_i, \cdot) \mid X_{1:N}\right) \\
&\approx \frac{1}{B}\sum_{b=1}^{B}\left(\frac{1}{K_b}\sum_{i=1}^{K_b} \kappa_{h_b}\left(\theta^{b*}_i, \cdot\right)\right) \\
&=: \hat{G}_0(\cdot).
\end{aligned}
$$

In words, our estimate $\hat{G}_0$ of G_0 is obtained by constructing B KDEs, one for each of the Gibbs sampler draws $\theta^{b*}_{1:K_b}$, then combining them in an average. In this definition, we need to select a kernel κ_h. There is optimality theory based on asymptotics for this issue (Silverman 1986), but a Gaussian density function with variance h is a convenient choice, and that same theory suggests it comes with little loss in efficiency. We also need to choose an appropriate bandwidth h_b for each KDE in the average. Numerous methods exist to accomplish this automatically based on each draw's θ^*_i values. For the work reported in this paper, we used the "direct plug-in" method $\hat{h}_{\mathrm{2P}}$ of Sheather and Jones (1991), which is known to have superior theoretical and empirical properties. However, when we repeated our analyses with other bandwidth-selection protocols, including the simple formula given by (3.31) in Silverman (1986), the results remained qualitatively the same.

One may ask why the kernel density estimation method was not applied directly to the observed data $x_{1:N}$ at the outset. One reason is that the DP mixture model imposes concrete probabilistic structure which may correspond to our understanding of the stochastic mechanism underlying the data: in particular, there could be a rationale for repeatedly using the same conditional data density to generate a subset of the observations, in which case the partition of the dataset given by the class assignments has intrinsic content.

We turn now to our estimate $\hat{\alpha}$. Point estimates of α have been studied previously. In particular, Liu (1996) proves that the marginal maximum likelihood estimate of α in the DP mixture model must satisfy the stationarity condition

$$
\sum_{n=1}^{N} \frac{\alpha}{\alpha + n - 1} = E(K(\theta_{1:N}) \mid X_{1:N}, \alpha), \tag{8}
$$

with $K(\theta_{1:N})$ as defined above. Here the dependence of the posterior mean on α has been made explicit in the notation. This condition does not have a closed-form solution in α; indeed, exact evaluation of the posterior mean in (8) is generally not feasible. Liu notes that the posterior mean can be approximated using a Gibbs sampling estimate, but instead he goes on to employ a sequential imputation method. Here

we pursue the Gibbs sampling approximation, given by

$$E(K(\theta_{1:N}) \mid X_{1:N}, \alpha) \approx \frac{1}{B} \sum_{b=1}^{B} K_b, \tag{9}$$

with K_b as previously defined. In the estimation phase, given the Gibbs draws from the previous inference phase, we solve numerically the approximation to (8) based on (9). This yields our estimate $\hat{\alpha}$.

4. Results

4.1. Simulations

We applied our nonparametric empirical Bayes procedure to simulated data from DP mixture models under several settings of (G_0, α). In all simulations, we drew 500 observations; the conditional data density is Gaussian with mean θ_n and standard deviation fixed to $0.1 \cdot \mathrm{SD}(G_0)$. Because our interest in these simulations lies foremost with estimating G_0 and α, we treat the class-conditional SD as known to the procedure. In order to include examples where G_0 has properties not easily modelled by the parametric priors of previous authors, we consider the following four G_0 distributions: (i) a standard normal; (ii) a t_3 distribution, which exhibits heavy tails; (iii) the distribution of $-\log Z$, $Z \sim \chi^2_1$, which has strong positive skew; and (iv) an equal mixture of two well-separated t_5 distributions, having both heavy tails and pronounced bimodality. We choose $\alpha \in \{1, 5, 10\}$, corresponding to a small, moderate, and large typical number of classes in the data.

To obtain an initial estimate $\hat{G}_0$ for our procedure, we fit a KDE to the observed data, resulting in a peaky, multimodal distribution. In order to understand the initialization of $\hat{\alpha}$, note that for a moderate observation count n, the prior expected number of classes in the data (i.e., the number of θ_i^*) is approximately $\alpha \log n$. Thus, a plausible range of candidate initial values for $\hat{\alpha}$ is from $1/\log n$, corresponding to a single class, to $n/\log n$, corresponding to one class per observation. We tried initializing $\hat{\alpha}$ at the bottom, middle, and top of this range; the final converged value was nearly the same in all cases.

Each column of Fig. 2 plots the estimate $\hat{G}_0$ as a solid line after 0, 5, and 200 rounds of alternation between the estimation phase and inference phase, with a true α of 5. The true G_0 for a given simulation is named above the column and plotted with dashes in the column's top three panels. In addition, the KDE $\hat{G}_0^*$ constructed from the true θ_i^*'s in the simulation is given as a dashed line in those panels. We see that, as the rounds progress, $\hat{G}_0$ changes from a rapidly fluctuating function to a smooth curve which closely approximates $\hat{G}_0^*$. Given that $\hat{G}_0^*$ is the ideal kernel density estimate we would construct if the θ_i^*'s were revealed to us, this behavior is striking. The bottom row of Fig. 2 demonstrates this property more clearly, using the total variation distance (TVD) metric. The TVD between two probability distributions P and Q is defined as $\sup_A |P(A) - Q(A)|$, with the supremum taken over measurable sets A. Each panel plots with a solid line the TVD between $\hat{G}_0$ and G_0 as the rounds progress, and gives the TVD between $\hat{G}_0^*$ and G_0 as a horizontal dotted line for reference. As can be seen, after no more than 15 rounds $\hat{G}_0$ begins to fluctuate around a density which is almost as good an estimate as $\hat{G}_0^*$. The fluctuation arises due to sampling variability in the Gibbs draws from the inference phase. This stabilization occurs despite the kurtosis of G_0 in column 2, its skewness in column 3, and its multimodality in column 4, all of which are captured by $\hat{G}_0$. The behavior of $\hat{\alpha}$ is very similar: after a small number of rounds, $\hat{\alpha}$ settles down to fluctuate around a value near the true α.

Figure 3 shows how the Gibbs sampler in the inference phase also stabilizes as the rounds progress. Each panel depicts multiple draws (after the diagnosed convergence of the sampler) from the posterior distribution on $\theta_{1:N}$, at rounds 1 (left), 5 (middle), and 175 (right). Within a panel, each draw is represented as a row of black circles, with rows arranged from earlier (bottom row) to later (top row) in the Gibbs run. Within a row, each circle corresponds to a distinct θ_i^*. The value of θ_i^* is indicated by the circle's location on the x-axis, and the number of observations assigned to θ_i^* is proportional to the radius of the circle. The true θ's in this simulation are shown as the bottommost row of grey circles, and the observed data appear as a rug plot underneath. Notice that the Gibbs samples from round 1 reflect much uncertainty about the number and location of the $\theta_{1:k}^*$ underlying the data. As $\hat{G}_0$ and $\hat{\alpha}$ improve in later rounds, the posterior distribution becomes sharply peaked around the true values of θ.

We also carried out a comparison between our nonparametric EB procedure and a simpler parametric approach to adaptive point estimation of G_0, called the data-dependent Gaussian method. The latter takes $\hat{G}_0$ to be a fixed Gaussian distribution, with mean set to the sample mean of the data and SD set to four times the sample SD. In addition, rather than utilizing a point estimate of α, we place a Gamma(2,2) prior on it. Both approaches then estimate the predictive distribution using $\hat{G}_0$ whenever G_0 is required. The data-dependent Gaussian method is similar to a procedure described by Ishwaran and James (2002), in which the SD of G_0 is set to four times the sample SD and a highly diffuse zero-mean normal prior is put on G_0's mean. This prior has the effect of centering the posterior mean of G_0 with respect to the θ_i^*'s, much like our method centers G_0 on the data. These authors use a Gamma(2,2) prior for α.

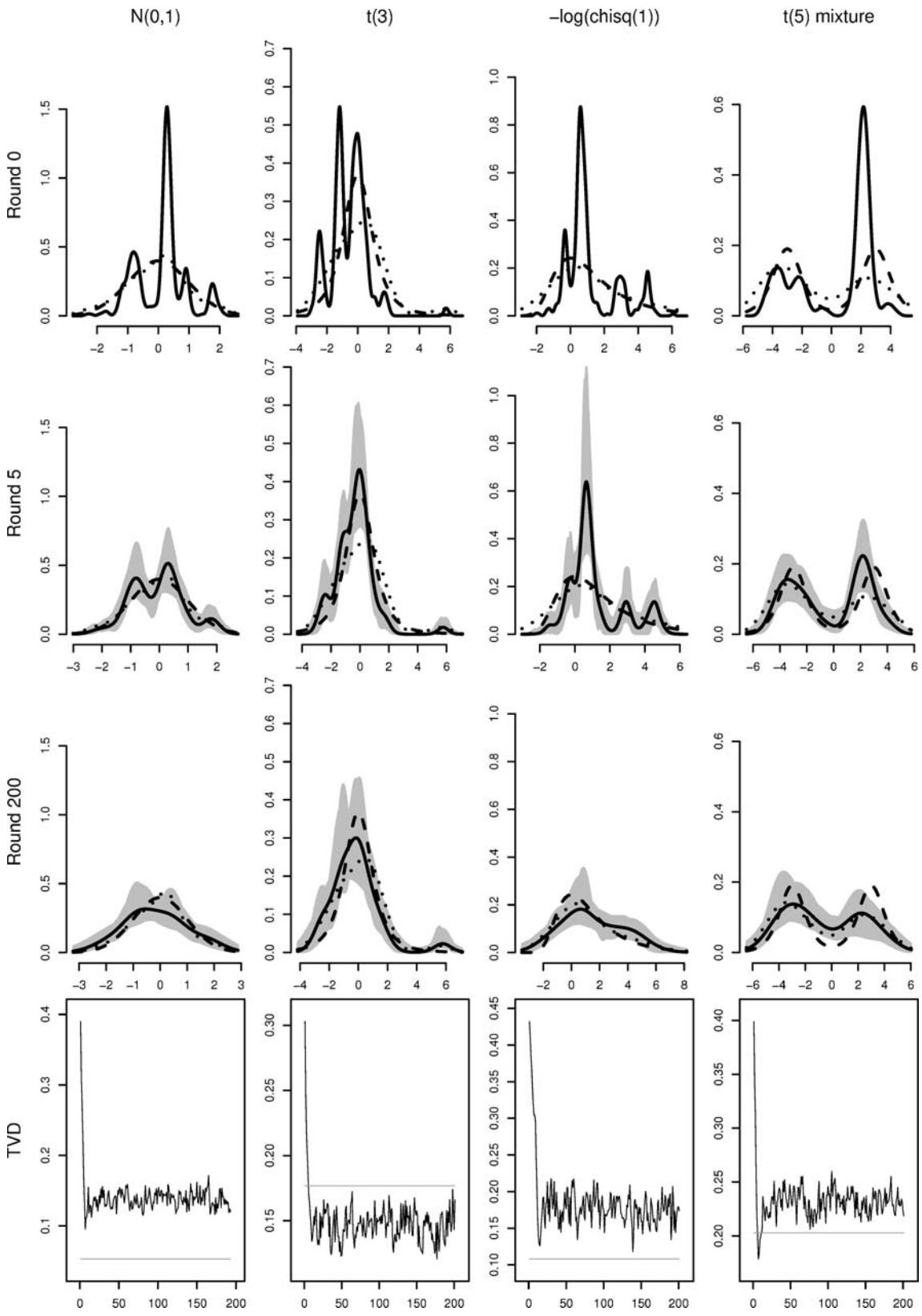

Fig. 2 Each column depicts the evolution of $\hat{G}_0$ in a simulation using some true G_0 (top label) and $\alpha = 5$. The density plots in rows 1–3 show the true G_0 (dashed), $\hat{G}_0$ (solid), and $\hat{G}_0^*$ (dotted), a KDE built from the true θ_i^*'s. Grey lines give the individual component KDE's of $\hat{G}_0$, one constructed from each Gibbs sample of $\theta_{1:k}^*$. Each plot in the bottom row gives the progression of TVD(G_0, $\hat{G}_0$), with TVD(G_0, $\hat{G}_0^*$) as a grey reference line

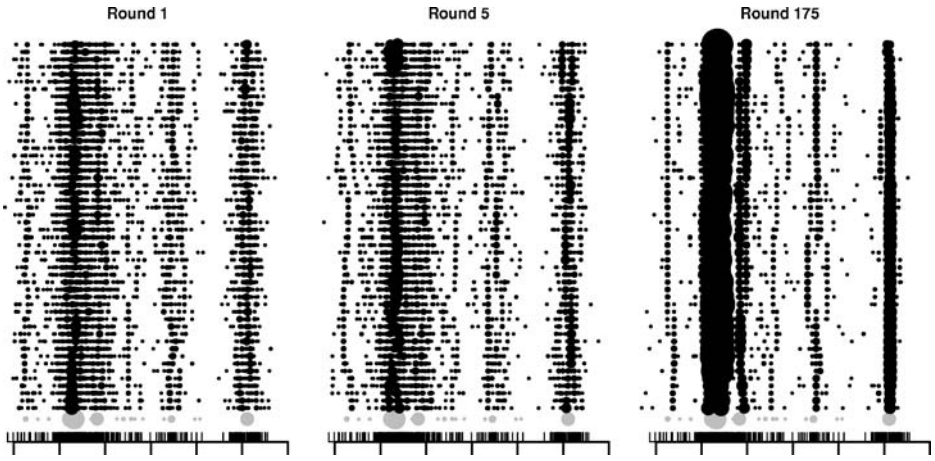

Fig. 3 Flow plots of Gibbs sampler output at early rounds (left, middle) and after convergence (right) of the DP EB procedure. At bottom, a rug plot of the observed data (black vertical ticks). Each grey circle is centered at a true θ_i^*, with radius proportional to the number of observations it generated. Each row of black circles gives the same depiction for one draw of $\theta_{1:N}$ from the Gibbs sampler. From a simulation with $\alpha = 5$ and $G_0 = N(0, 1)$

Table 1 Total variation distances (%) in simulation

G_0	α	G_0:DG	G_0:NP	PD:DG	PD:NP
$N(0, 1)$	1	41.63(5.54)	50.15(12.61)	2.93(0.47)	4.42(0.84)
	5	46.94(1.72)	13.56(1.29)	2.84(0.39)	3.15(0.42)
	10	52.77(4.26)	13.24(1.40)	3.63(0.26)	2.81(0.43)
t_3	1	33.59(5.20)	41.09(8.71)	2.45(0.45)	3.35(0.48)
	5	58.17(6.68)	19.84(3.41)	3.92(0.49)	2.85(0.30)
	10	60.57(2.24)	17.91(1.56)	4.43(0.14)	2.79(0.22)
$-\log Z$, $Z \sim \chi_1^2$	1	47.67(2.73)	33.92(10.08)	2.37(0.38)	3.76(0.97)
	5	44.14(5.22)	18.75(2.17)	3.39(0.29)	3.12(0.46)
	10	59.02(3.08)	17.11(1.54)	4.02(0.30)	2.55(0.12)
t_5 mixture	1	56.73(5.91)	53.23(10.87)	2.46(0.38)	2.91(0.34)
	5	64.21(2.10)	28.64(1.88)	3.08(0.51)	3.30(0.58)
	10	65.66(0.52)	20.20(3.69)	4.07(0.11)	2.18(0.25)

Each entry contains a mean %TVD and its standard error (parenthesized) between a pair of densities (column label) in one of 12 simulation scenarios (row label). The density pairs are (1) G_0 vs. the $\hat{G}_0$ of the data-dependent Gaussian method (G_0:DG); (2) G_0 vs. our nonparametric $\hat{G}_0$ (G_0:NP); (3) the predictive density using G_0 vs. the one inferred using the data-dependent Gaussian method (PD:DG); (4) the predictive density using G_0 vs. the one based on our nonparametric EB procedure (PD:NP). Mean and SE are over 5 repetitions of each simulation.

Table 1 shows the average percent TVD between the true G_0 and the $\hat{G}_0$ from the data-dependent Gaussian method in column 3; column 4 gives the TVD between G_0 and the nonparametric $\hat{G}_0$. Each row corresponds to one particular combination of true G_0 and α from our simulations; the TVD average and its standard error (SE) are over five independent replications of the relevant simulation. We see that the quality of the $\hat{G}_0$ estimate can depend strongly on the choice of estimation procedure. When $\alpha = 1$, the two approaches to constructing $\hat{G}_0$ perform comparably (after accounting for simulation variability). However, at higher values of α the nonparametric $\hat{G}_0$ improves substantially on the parametric $\hat{G}_0$. The data-dependent Gaussian method also has difficulties because its prior puts more mass on small values of α, as is conventional in other treatments of the Dirichlet process mixture model.

Table 1 also reports the average percent TVD between the predictive density based on G_0 and the one based on either the data-dependent Gaussian method (column 5) or the nonparametric EB procedure (column 6). Notice that, in all simulations, both procedures produce rather close approximations to the ideal predictive density estimate based on G_0—the predictive density is not too sensitive to the choice of $\hat{G}_0$. At $\alpha = 1$, the smallest value, the predictive density resulting from the nonparametric EB procedure is worse than its counterpart based on the Gaussian $\hat{G}_0$. This can be understood by recalling that the expected number of distinct θ_i's is about $\alpha \log n \approx 6$ for $\alpha = 1$ and $n = 500$. Thus, even the KDE $\hat{G}_0^*$ built from the true θ_i^*'s, which our nonparametric procedure approximates, is based on only about six values. This is an example of a well-known phenomenon: nonparametric estimates on small samples are often outperformed

by simpler parametric estimates, even when the parametric model is misspecified. On the other hand, at $\alpha = 5$ the two procedures for $\hat{G}_0$ perform about the same (up to simulation variability), and at $\alpha = 10$ the nonparametric method is significantly better in every simulation, even when the Gaussian family for the parametric $\hat{G}_0$ is correctly specified.

4.2. Real-world data analysis

To study the behavior of our nonparametric empirical Bayes procedure on real data, we re-analyzed two well-known datasets.

4.2.1. Galaxies data

The galaxies data (Postman, Huchra and Geller 1986) consist of velocity measurements for 82 galaxies from diverse sections of the Corona Borealis region. There are empirical reasons to expect subpopulation structure in the observations (Roeder 1990). These data have been analyzed using mixture models by Roeder (1990), Chib (1995), Carlin and Chib (1995), Escobar and West (1995), Richardson and Green (1997), Stephens (2000), and Ishwaran and James (2002).

Escobar and West (1995) and Ishwaran and James (2002) analyze the galaxy data using Dirichlet process mixtures of Gaussians to account for the uncertainty about the number of components in the underlying finite mixture. Escobar and West (1995) use a Dirichlet process on both the mean and the variance, with a conjugate G_0 (normal/inverse gamma). This simplifies their computations, though there is no reason to expect in advance that it is otherwise appropriate.

In Ishwaran and James (2002), the variance of the conditional data density is assumed equal for all data, and the base measure G_0 for the DP on the mean parameter is Gaussian. The authors fix a mean and variance for G_0 in a data-dependent way before running DP inference. We adopt their equal-variance model, but estimate G_0 with the empirical Bayes approach of Section 3.

We ran 200 rounds of the DP EB procedure, though stabilization of $\hat{G}_0$ and $\hat{\alpha}$ occurred much sooner. We used the same Gibbs sampler configuration as in Section 4.1. To address the unknown variance of the conditional data density, we introduced a shared parameter σ^2. Draws from the Gibbs sampler were used to effect a straightforward Monte Carlo EM procedure (Wei and Tanner 1990) for its estimation. The MC-EM algorithm operates in parallel with the nonparametric empirical Bayes algorithm for (G_0, α).

Each sampling stage (given a fixed $\hat{G}_0$, $\hat{\alpha}$, and $\hat{\sigma}^2$ from the previous round) took approximately 50 seconds on an Apple 1.3 GHz PowerPC G4 architecture. The time to construct $\hat{G}_0$, $\hat{\alpha}$, and $\hat{\sigma}^2$ in the estimation stage was negligible. Thus, the complete run lasted about three hours. Convergence was apparent after 20 to 30 rounds, or about half an hour. We wrote the sampler entirely in the high-level language R, so we expect that a substantial speedup would result from a more efficient implementation. All of our DP EB analysis source code is freely available upon request, under the terms of the GNU General Public License.

Figure 4 (top) presents the converged estimates $\hat{G}_0$ as well as the predictive densities $p(x_{\text{new}} \mid x_{1:N}, \hat{\alpha}, \hat{G}_0)$ we obtained. It is interesting that $\hat{G}_0$ differs noticeably from the Gaussian family used in Ishwaran and James (2002). In particular, the galaxy data's $\hat{G}_0$ exhibits a small mode in each tail in addition to a main central mode, raising the possibility of additional

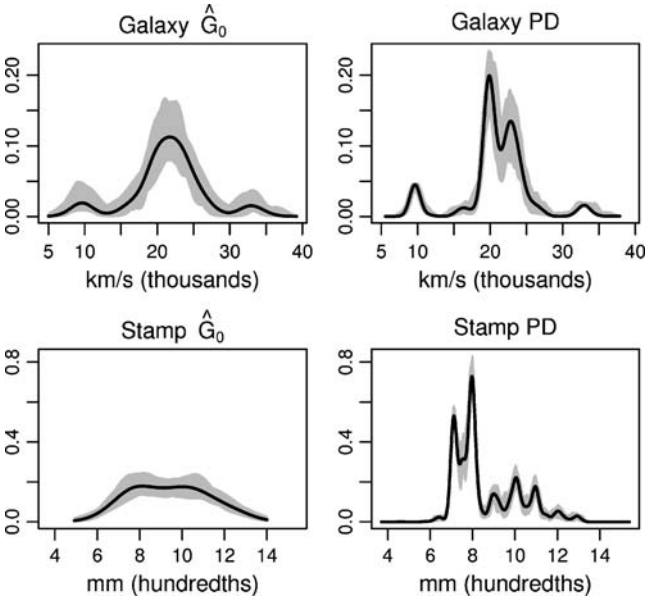

Fig. 4 Graphs of $\hat{G}_0$ (left-hand) and the estimated predictive density (right-hand) for the galaxy and stamp datasets. Grey lines give the individual component curves, one from each Gibbs sample, which are averaged to produce the relevant estimate

higher-level clustering structure beyond that reflected in the sharing of θ values by the observations.

4.2.2. *Stamp data*

The stamp data (Wilson 1983) are the thicknesses of 485 Hidalgo postage stamps issued in Mexico from 1872–1874. During this time, the same stamp issue was printed on papers of different thickness. Izenman and Sommer (1988), Basford *et al*. (1997) and Ishwaran and James (2002) use mixtures with an unknown number of components to determine the number of paper types used for this stamp. Ishwaran and James (2002) again use DP mixtures of Gaussians, with Gaussian base measure and equal variance across components.

In Fig. 4 (bottom), we illustrate the density estimate under the DP mixture with G_0 fit by our empirical Bayes procedure. The configuration was the same as for the galaxy data analysis, and we used the same hardware. With the larger number of observations in the stamp data, the sampling procedure slowed down to about five minutes per stage, with convergence again evident by 30 iterations (two and a half hours). Again, our choice to implement in *R* slowed the computation down considerably.

The high-probability, slightly bimodal region in the $\hat{G}_0$ of the stamp data suggests a more disperse relationship among the θ_i^*'s than a normality assumption implies. As we noted in Section 4.1, the differences between our predictive densities and those of Ishwaran and James (2002) are small, though some of the modes in our predictive densities are more pronounced.

5. Discussion

We have treated the problem of Dirichlet process mixture modeling in the context of nonparametric estimation for the distributional parameter G_0. After motivating a nonparametric point estimate $\hat{G}_0$ based on a posterior mean kernel density estimator, we studied the effectiveness of the resulting empirical Bayes procedure in a suite of simulations. Analyses of two real-world datasets suggested previously undiscovered structure in the DP mixture setting. We emphasize that application of this nonparametric empirical Bayes method to real data required no specification of the Dirichlet process parameters α and G_0.

Acknowledgments

We would like to acknowledge support from the Intel Corporation, Microsoft Research, and the National Science Foundation.

References

Basford K., McLachlan G. and York M. 1997. Modelling the distribution of stamp paper thickness via finite normal mixtures: The 1872 Hidalgo stamp issue of Mexico revisited. Journal of Applied Statistics 24(2): 169–180.

Berry D.A. and Christensen R. 1979. Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes. Annals of Statistics 7: 558–568.

Blackwell D. and MacQueen J.B. 1973. Ferguson distributions via Pólya urn schemes. Annals of Statistics 1: 353–355.

Brown E.R. and Ibrahim J. G. 2003. A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. Biometrics 59: 221–228.

Carlin B. and Chib S. 1995. Bayesian model choice via Markov chain Monte Carlo methods. Journal of the Royal Statistical Society. Series B (Methodological) 57(3): 473–484.

Carota C. and Parmigiani G. 2002. Semiparametric regression for count data. Biometrika 89: 265–281.

Chib S. 1995. Marginal likelihood from the Gibbs output. Journal of the American Statistical Association 90(432): 1313–1321.

Escobar M.D. and West M. 1995. Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association 90: 577–588.

Ferguson T.S. 1973. A Bayesian analysis of some nonparametric problems. Annals of Statistics 1: 209–230.

Gelfand A.E. and Kottas A. 2002. A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. Journal of Computational and Graphical Statistics 11: 289–305.

Iorio M.D., 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.

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: 508–532.

Izenman A. and Sommer C. 1988. Philatelic mixtures and multimodal densities. Journal of the American Statistical Association 83(404): 941–953.

Liu J.S. 1996. Nonparametric hierarchical Bayes via sequential imputations. Annals of Statistics 24: 911–930.

MacEachern S.N. and Müller P. 1998. Estimating mixture of Dirichlet process models. Journal of Computational and Graphical Statistics 7: 223–238.

Maritz J.S. and Lwin T. 1989. Empirical Bayes Methods. Monographs on Statistics and Applied Probability. Chapman & Hall, London.

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.

Neal R.M. 2000. Markov chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical Statistics 9: 249–265.

Postman M., Huchra J. and Geller M. 1986. Probes of large-scale structure in the Corona Borealis region. The Astronomical Journal 92: 1238–1247.

Richardson S. and Green P. 1997. On Bayesian analysis of mixtures with unknown number of components. Journal of the Royal Statistical Society. Series B (Methodological) 59(4): 731–792.

Robbins H. 1955. An empirical Bayes approach to statistics. In Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, pp. 131–148.

Roeder K. 1990. Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. Journal of the American Statistical Association 85: 617–624.

Sheather S.J. and Jones M.C. 1991. A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society Series B 53: 683–690.

Silverman B.W. 1986. Density estimation for statistics and data analysis. Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Stephens M. 2000. Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. The Annals of Statistics 28(1): 40–74.

Teh Y.W., Jordan M.I., Beal M.J. and Blei D.M. 2004. Hierarchical Dirichlet processes. Technical Report 653, U.C. Berkeley, Dept. of Statistics.

Tomlinson G. and Escobar M. 1999. Analysis of Densities. Technical report, University of Toronto.

Wei G.C.G. and Tanner M.A. 1990. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. Journal of the American Statistical Association 85: 699–704.

Wilson I.G. 1983. Add a new dimension to your philately. The American Philatelist 97: 342–349.

Yu K., Tresp V. and Yu S. 2004. A nonparametric hierarchical Bayesian framework for information filtering. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). ACM Press.