Variational Bayesian Dirichlet-Multinomial Allocation for Exponential Family Mixtures

Shipeng Yu[1,2], Kai Yu[2], Volker Tresp[2], and Hans-Peter Kriegel[1]

[1] Institute for Computer Science, University of Munich, Germany
[2] Siemens Corporate Technology, Munich, Germany

Abstract. This paper studies a Bayesian framework for density modeling with mixture of exponential family distributions. *Variational Bayesian Dirichlet-Multinomial allocation* (VBDMA) is introduced, which performs inference and learning efficiently using variational Bayesian methods and performs automatic model selection. The model is closely related to Dirichlet process mixture models and demonstrates similar automatic model selection in the variational Bayesian context.

1 Introduction

In statistical analysis and artificial intelligence, there has been a strong interest in finite mixture distributions for density estimation. The model offers a natural framework to handle the heterogeneity in clustering analysis, which is often of central importance in many applications. Among all the choices, exponential family mixtures are extremely useful in practice, since they cover a broader scope of characteristics of random variables, and the existence of conjugate priors often makes inference easier [1, 7].

Previously much work has been done with a fixed number of components. The efforts include estimating parameters of each component by EM algorithms or via MCMC in a Bayesian way. Model selection, i.e., choosing the number of components, remains a fundamental challenge for mixture modeling. A frequentist treatment typically tests the hypotheses about this number. On the other side, a Bayesian way computes the *a posteriori* over the model space. Recently, there are increasing interests in *Bayesian nonparametric statistics*, which apply *Dirichlet process* to handle infinite number of components (e.g., [5, 2]).

This paper focuses on a fully Bayesian mixture model with finite K exponential family components. The interesting point is that variational learning in the model tends to end up with a sparsity of mixing weights when K is sufficiently large. This is because the model approaches a Dirichlet process mixture model in the limiting case. A few authors explored this point in Bayesian statistics [8, 9], but it is not sufficiently noticed. In this paper we propose the variational Bayesian Dirichlet-Multinomial allocation (VBDMA) for model selection in finite mixture models. This on one hand offers tractability because of the finite dimensionality and variational methods, and on the other hand provides general solutions to mixture of exponential-family distributions which covers a wide range of real-world problems.

2 Mixture of Exponential Family Distributions

Exponential Family: A probability distribution of $\mathbf{x} \in \mathcal{X}$ given parameters $\boldsymbol{\theta}$ is in the *exponential family* if it takes the form

$$P(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x}) \exp\left\{\boldsymbol{\theta}^\top \phi(\mathbf{x}) - A(\boldsymbol{\theta})\right\}, \tag{1}$$

where $\phi(\mathbf{x})$ is the *sufficient statistics* of $\mathbf{x}$, and $\boldsymbol{\theta}$ is called the *natural parameter*. The quantity $A(\boldsymbol{\theta})$, known as the *log partition function*, is defined as a normalization factor independent of $\mathbf{x}$: $A(\boldsymbol{\theta}) = \log \int_{\mathcal{X}} h(\mathbf{x}) \exp\left\{\boldsymbol{\theta}^\top \phi(\mathbf{x})\right\} d\mathbf{x}$. It is well-known that $A(\boldsymbol{\theta})$ plays an important role for exponential family distributions. In particular, it can be identified as the *moment generating function* of $\phi(\mathbf{x})$. One important example of this is given as:

$$\frac{\partial A(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbb{E}_{\boldsymbol{\theta}}[\phi(\mathbf{x})] := \int_{\mathcal{X}} \phi(\mathbf{x}) P(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x}, \tag{2}$$

which gives the mean of the sufficient statistics.

Conjugate Family: The *conjugate family* defines a prior family for exponential family distributions as

$$P(\boldsymbol{\theta}|\boldsymbol{\gamma}, \eta) = g(\boldsymbol{\theta}) \exp\left\{\boldsymbol{\theta}^\top \boldsymbol{\gamma} - \eta A(\boldsymbol{\theta}) - B(\boldsymbol{\gamma}, \eta)\right\}, \tag{3}$$

where $(\boldsymbol{\gamma}, \eta)$ are the parameters for the prior, i.e., *hyperparameters*, with $\boldsymbol{\gamma}$ having dimensionality $\dim(\boldsymbol{\theta})$, and η a scalar. It is conjugate in that the posterior distribution takes the same form as the prior, calculated by Bayes' rule:

$$P(\boldsymbol{\theta}|\mathbf{x}, \boldsymbol{\gamma}, \eta) \propto P(\mathbf{x}|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\boldsymbol{\gamma}, \eta) \propto g(\boldsymbol{\theta}) \exp\left\{\boldsymbol{\theta}^\top (\boldsymbol{\gamma} + \phi(\mathbf{x})) - (\eta + 1) A(\boldsymbol{\theta}) - B(\boldsymbol{\gamma}, \eta)\right\}.$$

It is easy to check that conjugate family (3) also belongs to exponential family, with sufficient statistics $\binom{\boldsymbol{\theta}}{-A(\boldsymbol{\theta})}$ and natural parameter $\binom{\boldsymbol{\gamma}}{\eta}$. Then we have

$$\frac{\partial B(\boldsymbol{\gamma}, \eta)}{\partial \boldsymbol{\gamma}} = \mathbb{E}_{\boldsymbol{\gamma}, \eta}[\boldsymbol{\theta}], \quad \frac{\partial B(\boldsymbol{\gamma}, \eta)}{\partial \eta} = \mathbb{E}_{\boldsymbol{\gamma}, \eta}[-A(\boldsymbol{\theta})] \tag{4}$$

by applying (2). These results turn out to be useful for subsequent sections.

Exponential Family Mixtures: In mixture modeling, each data point is sampled from a fixed but unknown *component distribution*, which belongs to exponential family here. At the moment we fix the number of components in the mixture to be K, a finite positive integer. We will focus on the case that all the component distributions take the same form, e.g., Gaussian. Then the likelihood given N *i.i.d.* sampled data points $\mathcal{D} := \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ is formally written as

$$P(\mathcal{D}|\boldsymbol{\pi}, \boldsymbol{\Theta}) = \textstyle\prod_{i=1}^{N} \sum_{k=1}^{K} P(c_i = k|\boldsymbol{\pi}) P(\mathbf{x}_i|\boldsymbol{\theta}_k),$$

where $P(c_i = k|\boldsymbol{\pi}) = \pi_k$ is a Multinomial with parameters $\boldsymbol{\pi}$, and $P(\mathbf{x}_i|\boldsymbol{\theta}_k)$ takes the general form (1). The K-dimensional vector $\boldsymbol{\pi} := \{\pi_k\}_{k=1}^K$ gives the weights for the component distributions and sums to 1. $\boldsymbol{\Theta} := \{\boldsymbol{\theta}_k\}_{k=1}^K$ contain the natural parameters of all component distributions. c_i is seen as a random variable of *indicator* for data $\mathbf{x}_i$, saying which component $\mathbf{x}_i$ is sampled from.

We need to assign priors to all the parameters. For $\boldsymbol{\Theta}$ we assign conjugate prior (3) to each $\boldsymbol{\theta}_k$ independently, with the same hyperparameters $(\boldsymbol{\gamma}_0, \eta_0)$: $P(\boldsymbol{\theta}_k|\boldsymbol{\gamma}_0, \eta_0) = g(\boldsymbol{\theta}_k) \exp\left\{\boldsymbol{\theta}_k^\top \boldsymbol{\gamma}_0 - \eta_0 A(\boldsymbol{\theta}_k) - B(\boldsymbol{\gamma}_0, \eta_0)\right\}$. For the Multinomial parameters $\boldsymbol{\pi}$ we assign a Dirichlet distribution $\boldsymbol{\pi} \sim \mathrm{Dir}(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K})$. Here we make the constraint that all the parameters in this Dirichlet are the same and sum to a scalar that is independent of K, the number of components.

With these priors, the final data likelihood can be obtained by integrating out *latent variables* $\boldsymbol{\pi}$ and $\boldsymbol{\Theta}$ (see plate model in Fig. 1 left):

$$P(\mathcal{D}) = \int_{\boldsymbol{\pi}} P(\boldsymbol{\pi}|\alpha) \int_{\boldsymbol{\Theta}} \prod_{k=1}^K P(\boldsymbol{\theta}_k|\boldsymbol{\gamma}_0, \eta_0) \left\{\prod_{i=1}^N \sum_{k=1}^K \pi_k P(\mathbf{x}_i|\boldsymbol{\theta}_k)\right\} d\boldsymbol{\Theta} d\boldsymbol{\pi}.$$

The model has two parameters: α is a positive scalar; $(\boldsymbol{\gamma}_0, \eta_0)$ has dimensionality $\dim(\phi(\mathbf{x})) + 1$. [1, 4] studied the special case of Gaussian mixtures.

3 Model Inference and Learning

Inference in the proposed model is intractable and needs Markov chain Monte Carlo (MCMC) sampling. In this paper we instead focus on *variational Bayesian* methods, which are motivated by approximating the *a posteriori* distribution of latent variables with a tractable family, and then maximizing a lower-bound of data likelihood with respect to some *variational parameters* [10, 7]. One common way of achieving this is to assume a *factorized* distribution for the latent variables, which indicates that for exponential family mixtures we use distribution

$$Q(\boldsymbol{\pi}, \boldsymbol{\theta}, \mathbf{c}|\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\varphi}) := Q(\boldsymbol{\pi}|\boldsymbol{\lambda}) \prod_{k=1}^K Q(\boldsymbol{\theta}_k|\boldsymbol{\gamma}_k, \eta_k) \prod_{i=1}^N Q(c_i|\boldsymbol{\varphi}_i)$$

to approximate the true posterior $P(\boldsymbol{\pi}, \boldsymbol{\theta}, \mathbf{c}|\mathcal{D}, \alpha, \boldsymbol{\gamma}_0, \eta_0)$. Here $Q(\boldsymbol{\pi}|\boldsymbol{\lambda})$ is K-dim. Dirichlet, $Q(\boldsymbol{\theta}_k|\boldsymbol{\gamma}_k, \eta_k)$ the conjugate family (3), and $Q(c_i|\boldsymbol{\varphi}_i)$ K-dim. Multinomial. Applying Jensen's inequality yields a lower bound of the log-likelihood: $\mathcal{L}(\mathcal{D}) = \mathbb{E}_Q[\log P(\boldsymbol{\pi}|\alpha)] + \sum_{k=1}^K \mathbb{E}_Q[\log P(\boldsymbol{\theta}_k|\boldsymbol{\gamma}_0, \eta_0)] + \sum_{i=1}^N \mathbb{E}_Q[\log P(c_i|\boldsymbol{\pi})] + \sum_{i=1}^N \mathbb{E}_Q[\log P(\mathbf{x}_i|\boldsymbol{\theta}, c_i)] - \mathbb{E}_Q[\log Q(\boldsymbol{\pi}, \boldsymbol{\theta}, \mathbf{c})]$. Variational Bayesian methods in the literature maximize this lower bound *only* with respect to variational parameters $\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\varphi}$, and thus fix the model parameters $\alpha, \boldsymbol{\gamma}_0, \eta_0$ (see [1, 7]). This paper will however treat it as the E-step of the algorithm, and estimate the model parameters in the M-step.

In the E-step, it is straightforward to obtain the following updates by setting the partial derivatives with respect to each variational parameter to be zero:

$$\varphi_{i,k} \propto \exp\left\{\mathbb{E}_{\boldsymbol{\gamma}_k, \eta_k}[\boldsymbol{\theta}_k^\top \phi(\mathbf{x}_i) - A(\boldsymbol{\theta}_k)] + \mathbb{E}_{\boldsymbol{\lambda}}[\log \pi_k]\right\}, \tag{5}$$

$$\boldsymbol{\gamma}_k = \sum_{i=1}^N \varphi_{i,k} \phi(\mathbf{x}_i) + \boldsymbol{\gamma}_0, \quad \eta_k = \sum_{i=1}^N \varphi_{i,k} + \eta_0, \quad \lambda_k = \sum_{i=1}^N \varphi_{i,k} + \frac{\alpha}{K}, \tag{6}$$

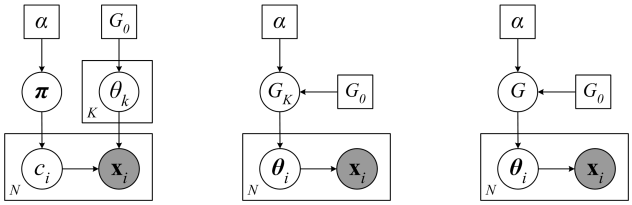


Fig. 1. Plate models for exponential family finite mixtures (left and middle), and the DP mixture model (right). G_K denotes the finite discrete prior for $\boldsymbol{\theta}_k$'s.

where the first expectation in (5) can be calculated using (4), and $\mathbb{E}_{\boldsymbol{\lambda}}[\log \pi_k] = \left\{\Psi(\lambda_k) - \Psi\left(\sum_{j=1}^{K} \lambda_j\right)\right\}$, with $\Psi(\cdot)$ the digamma function. This expectation is obtained by applying (2) to Dirichlet distribution $Q(\boldsymbol{\pi}|\boldsymbol{\lambda})$. Since these equations are coupled, they should be updated iteratively until convergence. In variational Bayes, (5) is called *variational E-step*, and (6) is called *variational M-step*. This yields the algorithm given in [1] for mixture of Gaussians.

These equations recover the theorem in [7] for exponential family mixture models, and turn out to be very intuitive and explainable. For instance, $\varphi_{i,k}$ measures the *a posteriori* probability that data $\mathbf{x}_i$ comes from component k, and can be written from (5) as $\varphi_{i,k} \propto \exp\left\{\mathbb{E}_{\boldsymbol{\gamma}_k,\eta_k}[\log P(\mathbf{x}_i|\boldsymbol{\theta}_k)]\right\} \exp\left\{\mathbb{E}_{\boldsymbol{\lambda}}[\log \pi_k]\right\}$ which can be seen as a *likelihood* term (left term) multiplied by a *prior* (right term), with other parameters fixed. This is analogous to a direct application of Bayes' rule. Other updates (6) also combine the *empirical observations* (the sum terms) with the prior (the model parameters).

In the M-step, we maximize the lower-bound with respect to the model parameters. For α we obtain $\sum_{k=1}^{K} \mathbb{E}_{\alpha}[\log \pi_k] = \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\lambda}}[\log \pi_k]$, which turns out to match the *sufficient statistics* of Dirichlet distributions. Similar results hold for $\boldsymbol{\gamma}_0$ and η_0: $\sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\gamma}_0,\eta_0}[\boldsymbol{\theta}_k] = \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\gamma}_k,\eta_k}[\boldsymbol{\theta}_k]$, $\sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\gamma}_0,\eta_0}[-A(\boldsymbol{\theta}_k)] = \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\gamma}_k,\eta_k}[-A(\boldsymbol{\theta}_k)]$. These expectations can be calculated using (4). Analytical solutions for these equations are in general not obtainable, so we need computational methods such as Newton-Raphson method to solve the problem.

4 Variational Bayesian Dirichlet-Multinomial Allocation

Model selection for mixture modeling, i.e., choosing the number K, is an important problem. This can be done via cross-validation; a Bayesian way selects the model with the largest *a posteriori* likelihood. However in both cases we have to retrain the model with different K's, which is normally very expensive.

In this section we investigate the functionality of α and show that the learning algorithm in Sec. 3 can lead to sparse mixtures. The algorithm has strong connections to *Dirichlet process* (DP) [6], and can be viewed as a variational algorithm for inference in DP mixture models. Therefore we call it *variational Bayesian Dirichlet-Multinomial allocation* (VBDMA), and it turns out that K can be automatically obtained after training, with the sparsity controlled by α.

Connections to Dirichlet Process: Denote $\boldsymbol{\theta}$ the natural parameter that generates data $\mathbf{x}$. In the mixture model we see that $\boldsymbol{\theta}$ is sampled from distribution $G_K(\boldsymbol{\theta}) := P(\boldsymbol{\theta}|\boldsymbol{\pi}, \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k \delta_{\boldsymbol{\theta}_k}(\boldsymbol{\theta})$, where $\delta_{\boldsymbol{\theta}_k}(\boldsymbol{\theta})$ is the point mass distribution and takes value 1 for $\boldsymbol{\theta} = \boldsymbol{\theta}_k$ and 0 otherwise. $G_K(\cdot)$ defines a *discrete prior* for $\boldsymbol{\theta}$, and model parameters α and $(\boldsymbol{\gamma}_0, \eta_0)$ now take the role of tuning the discrete but *unknown* distribution $G_K(\cdot)$.

When we let $K \to \infty$, it is known in statistics that the unknown distribution G_K tends to be a sample from a Dirichlet process, constrained by the *concentration parameter* (a positive scalar) and a *base distribution* [11]. In our model, the concentration parameter is just α, and the base distribution G_0 is given by (3). This model is illustrated in Fig. 1 middle. Following the convention for Dirichlet process, all the parameters $\boldsymbol{\theta}_i$ for data $\mathbf{x}_i$ are sampled as:

$$\boldsymbol{\theta}_i \overset{\text{iid}}{\sim} G, \quad \text{for } i = 1, \ldots, N; \qquad G \sim \text{DP}(\alpha, G_0).$$

Dirichlet process is well-known for the property of obtaining a nonparametric and discrete prior, and thus is widely applied for mixture modeling (see, e.g., [9]). When K is finite, however, the model is not equivalent to defining a Dirichlet process prior for $\boldsymbol{\theta}_i$'s, but is shown to be a good approximation if K is sufficiently large. This finite approximation is sometimes referred to as *Dirichlet-Multinomial allocation* (DMA), and is used for approximated sampling for Dirichlet processes [8]. In both DP and DMA, model selection can be done automatically via sampling methods, and the concentration parameter α is known to control the flexibility of generating new mixture components.

Sparsity of Infinite Mixture: Let us first fix α and focus on the E-step (5)$\sim$(6). With an uninformative initialization of variational parameters (e.g., we choose $\boldsymbol{\gamma}_k = \boldsymbol{\gamma}_0$, $\eta_k = \eta_0$ and $\lambda_k = \alpha/K$, for all k), we first fit the mixture membership $\varphi_{i,k}$ from (5), and then update the Dirichlet parameters using (6). Since all the components have the same prior terms $\mathbb{E}_{\boldsymbol{\lambda}}[\log \pi_k]$ initially, in (5) the assignment probabilities $\varphi_{i,k}$ will solely depend on the *empirical explanation* of $\mathbf{x}_i$ given component parameter $\boldsymbol{\theta}_k$. This will make the updated $\varphi_{i,k}$ unevenly distributed, and the constraints $\sum_k \varphi_{i,k} = 1, \forall i$ will lead to some "unlikely" components with very small assignment probabilities, i.e., $\sum_{i=1}^{N} \varphi_{i,k}$. When these values are fed into (6), these components will get smaller values for λ_k, and thus the prior term $\mathbb{E}_{\boldsymbol{\lambda}}[\log \pi_k]$ in (5) will also get smaller, which makes $\varphi_{i,k}$ more sharply distributed. Eventually, these components will get $\varphi_{i,k} = 0$, for all data points $\mathbf{x}_i$. This in turn leads to $\boldsymbol{\gamma}_k = \boldsymbol{\gamma}_0$, $\eta_k = \eta_0$ and $\lambda_k = \alpha/K$, all equal to the hyperparameters. When K is very large, α/K is very small, and these components almost have no chance to get bigger $\varphi_{i,k}$ in the future for some data $\mathbf{x}_i$, as seen from (5). Finally when the algorithm converges, we obtain only a small number of *effective* components.

This phenomenon is illustrated in Fig. 2 (upper row) for Gaussian mixtures, where we sampled 250 data points from 5 Gaussians. When we fix $\alpha = 1$, sparsity is obtained for all K's, even if K is only 10. When K becomes larger, the fitted

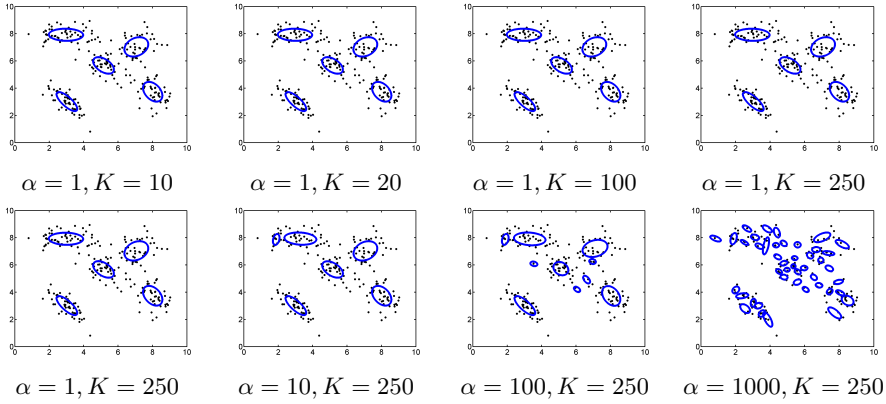


Fig. 2. Fitting VBDMA on a toy Gaussian mixture data with different α and K values.

number does not vary, but tends to be stable. As will be seen next, the strength of sparsity is not random, but depends strongly on parameter α.

Functionality of $\boldsymbol{\alpha}$: Now we investigate the situation that K is fixed, and α is allowed to change. When α is small, updates for λ_k will mostly depend on the empirical assignments $\sum_{i=1}^{N} \varphi_{i,k}$ in (6), and thus quickly get unbalanced. Then similar to the previous discussion, $\varphi_{i,k}$ will get an even sharp distribution in the next update, and the algorithm quickly converges to a small number of components that fit the data best. In the limiting case that $\alpha = 0$, λ_k's are purely determined by empirical updates, and we are making a *maximum likelihood* estimate for the mixing weights $\boldsymbol{\pi}$.

On the other hand when α is relatively large, the prior term α/K will dominate the update equation (6), and thus λ_k will not be very unbalanced in one step. This will in turn make the update equation (5) smooth for $\varphi_{i,k}$, and more components will survive than that with small α. As the iteration continues, certainly some components will be "dead" because of their poor fit to the data, but the death rate is much slower and we could expect more components left after convergence. A limiting case for this is to let $\alpha \to \infty$, which corresponds to fix the $\boldsymbol{\pi}$ *a priori* to be $\{\frac{1}{K}, \ldots, \frac{1}{K}\}$, and does not change it in the whole learning process. This normally leads to non-sparsity of the learned model.

Fig. 2 (bottom row) shows how α controls the sparsity of mixture modeling. With K fixed as 250, smaller α (e.g., 1) leads to higher sparsity, and larger α (e.g., 1000) results in more components. Therefore choosing a suitable α means choosing a desired number of mixture components.

Discussions: Previous discussions suggest that the algorithm in Sec. 3 can be viewed as a variational algorithm for DP mixture model, which we call the VBDMA. A nice property of VBDMA is that decrease of K is a natural consequence of model fitting with the data, and can be controlled by α. This is in contrast to

Table 1. The number of learned mixture components (means and standard deviations) in VBDMA (top) and VBTDP (bottom) for the toy Gaussian data with different initial K and α values. The experiments are repeated 20 times independently.

	$K = 5$	$K = 10$	$K = 20$	$K = 50$	$K = 100$	$K = 250$
$\alpha = 1$	4.45 ± 0.60	6.00 ± 1.03	6.70 ± 0.86	7.15 ± 1.27	6.85 ± 1.42	6.25 ± 1.16
$\alpha = 10$	4.95 ± 0.22	7.80 ± 1.01	8.65 ± 1.14	7.35 ± 1.04	7.10 ± 1.37	6.45 ± 1.10
$\alpha = 100$	5.00 ± 0.00	10.00 ± 0.00	19.90 ± 0.31	21.20 ± 1.58	11.40 ± 1.76	7.80 ± 1.40
$\alpha = 1000$	5.00 ± 0.00	10.00 ± 0.00	20.00 ± 0.00	49.65 ± 0.49	69.05 ± 2.19	45.05 ± 2.06
$\alpha = 10000$	5.00 ± 0.00	10.00 ± 0.00	20.00 ± 0.00	49.90 ± 0.31	85.10 ± 2.47	87.75 ± 2.07

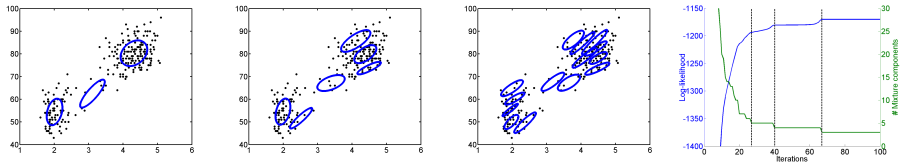
	$K = 5$	$K = 10$	$K = 20$	$K = 50$	$K = 100$	$K = 250$
$\alpha = 1$	4.50 ± 0.61	6.30 ± 1.03	7.35 ± 1.46	8.15 ± 1.39	8.55 ± 1.23	9.00 ± 1.62
$\alpha = 10$	4.65 ± 0.49	6.75 ± 0.91	7.85 ± 1.14	8.50 ± 1.24	8.80 ± 1.32	9.15 ± 1.09
$\alpha = 100$	4.60 ± 0.60	7.55 ± 1.15	8.95 ± 1.79	9.60 ± 1.70	9.90 ± 1.21	10.10 ± 1.33
$\alpha = 1000$	4.65 ± 0.49	7.80 ± 1.01	10.45 ± 1.47	10.80 ± 2.07	11.15 ± 2.06	11.10 ± 2.31
$\alpha = 10000$	4.60 ± 0.50	7.75 ± 1.02	10.20 ± 1.32	11.05 ± 2.01	11.50 ± 1.82	11.40 ± 2.19

post-processing methods (e.g., [12]) where heuristics must be used. VBDMA also provides explanations to [4], and can be extended to more complicated mixture models like mixture of factor analyzers [7].

Another variational algorithm for DP is proposed in [2] which is based on truncated DP (we denote it VBTDP). The idea is similar to VBDMA, but they put variational distributions directly on the stick-breaking parameters (see the definition in [9]). It is known that the variational form in VBTDP induces a *generalized Dirichlet distribution* to weights $\boldsymbol{\pi}$, and uses twice as many parameters as a Dirichlet distribution [3]. Some properties of generalized Dirichlet distribution include that each dimension of $\boldsymbol{\pi}$ is not always *negatively correlated* to other dimensions (i.e,. observing a sample from one dimension will surely increase the expected value of the parameter for this dimension, but decrease those for the other dimensions) as in Dirichlet distribution, and that the order of these dimensions is important for sampling and learning [13]. Both properties are however unnecessary for mixture modeling, and the latter is even contradictory to Bayesian exchangeability in this context.

In Tab. 1 we show the numbers of learned components for VBDMA and VBTDP on the toy data with different α and K values. For both methods, increasing α leads to more components, and sparsity is achieved for all K's when α is small. However, while varying α yields quite different sparsity for VBDMA, in VBTDP α seems to be insensitive to the results. Please refer to [14] for more detailed discussion about these two methods.

Empirical Study: Due to space limit we only consider the VBDMA with Gaussian mixtures on the "Old Faithful" data set. For more results on real data sets please refer to [14]. "Old Faithful" contains 272 2D observations from the Old Faithful Geyser in the Yellowstone National Park. Each observation consists of the duration of the eruption and the waiting time to the next eruption. We set K to 272 initially, and setting α to 100, 500 and 1000 results in 3, 6 and 15 Gaussians, respectively (see Fig. 3). All of them fit the data well, but in different granularities. The final log likelihoods of the three model fitting are -1174.55,



$\alpha = 100, K = 272$ $\alpha = 500, K = 272$ $\alpha = 1000, K = 272$ $\alpha = 100$

Fig. 3. Fitting a mixture of Gaussians on the "Old Faithful" data set.

-1187.75 and -1253.17, respectively. One can do a model selection using this likelihood and prefer the first one, but now there is no need to choose K *a priori* because this number is automatically determined by the VBDMA algorithm with a learned α which is approximately 100. We also see that each time the effective mixture number decreases, the likelihood has a noticeable increase (we mark three of them using dashed lines).

References

1. H. Attias. A variational Bayesian framework for graphical models. In *Advances in Neural Information Processing Systems 12*. MIT Press, 2000.
2. D. M. Blei and M. I. Jordan. Variational methods for the Dirichlet process. Proceedings of the 21st International Conference on Machine Learning, 2004.
3. R. J. Connor and J. E. Mosimann. Concepts of independence for proportions with a generalization of the Dirichlet distribution. *J. Amer. Stat. Ass.*, 64, 1969.
4. A. Corduneanu and C. M. Bishop. Variational Bayesian model selection for mixture distributions. In *Workshop AI and Statistics*, pages 27–34, 2001.
5. M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430), June 1995.
6. T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–230, 1973.
7. Z. Ghahramani and M. J. Beal. Graphical models and variational methods. In *Advanced mean Field Methods — Theory and Practice*. MIT Press, 2000.
8. P. J. Green and S. Richardson. Modelling heterogeneity with and without the Dirichlet process. unpublished paper, 2000.
9. H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.
10. M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.
11. R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265, 2000.
12. N. Ueda, R. Nakano, Z. Ghahramani, and G. E. Hinton. SMEM algorithm for mixture models. In *Neural Computation*, 1999.
13. T.-T. Wong. Generalized Dirichlet distribution in Bayesian analysis. *Appl. Math. Comput.*, 97(2-3):165–181, 1998.
14. S. Yu. *Advanced Probabilistic Models for Clustering and Projection*. PhD thesis, University of Munich, 2006.