Görür D, Rasmussen CE. Dirichlet process Gaussian mixture models: Choice of the base distribution. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 25(4): 615–626 July 2010/DOI 10.1007/s11390-010-1051-1

Dirichlet Process Gaussian Mixture Models: Choice of the Base Distribution

Dilan Görür[1] and Carl Edward Rasmussen[2,3]

[1] *Gatsby Computational Neuroscience Unit, University College London, London WC1N 3AR, U.K.*

[2] *Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, U.K.*

[3] *Max Planck Institute for Biological Cybernetics, D-72076 Tübingen, Germany*

E-mail: dilan@gatsby.ucl.ac.uk; cer54@cam.ac.uk

Received May 2, 2009; revised February 10, 2010.

Abstract In the Bayesian mixture modeling framework it is possible to infer the necessary number of components to model the data and therefore it is unnecessary to explicitly restrict the number of components. Nonparametric mixture models sidestep the problem of finding the "correct" number of mixture components by assuming infinitely many components. In this paper Dirichlet process mixture (DPM) models are cast as infinite mixture models and inference using Markov chain Monte Carlo is described. The specification of the priors on the model parameters is often guided by mathematical and practical convenience. The primary goal of this paper is to compare the choice of conjugate and non-conjugate base distributions on a particular class of DPM models which is widely used in applications, the Dirichlet process Gaussian mixture model (DPGMM). We compare computational efficiency and modeling performance of DPGMM defined using a conjugate and a conditionally conjugate base distribution. We show that better density models can result from using a wider class of priors with no or only a modest increase in computational effort.

Keywords Bayesian nonparametrics, Dirichlet processes, Gaussian mixtures

1 Introduction

Bayesian inference requires assigning prior distributions to all unknown quantities in a model. The uncertainty about the *parametric form* of the prior distribution can be expressed by using a nonparametric prior. The Dirichlet process (DP) is one of the most prominent random probability measures due to its richness, computational ease, and interpretability. It can be used to model the uncertainty about the functional form of the distribution for parameters in a model.

The DP, first defined using Kolmogorov consistency conditions[1], can be defined in several different perspectives. It can be obtained by normalising a gamma process[1]. Using exchangeability, a generalisation of the Pólya urn scheme leads to the DP[2]. A closely related sequential process is the Chinese restaurant process (CRP)[3-4], a distribution over partitions, which also results in the DP when each partition is assigned an independent parameter with a common distribution. A constructive definition of the DP was given by [5], which leads to the characterisation of the DP as a stick-breaking prior[6].

The hierarchical models in which the DP is used as a prior over the distribution of the parameters are referred to as the Dirichlet process mixture (DPM) models, also called mixture of Dirichlet process models by some authors due to [7]. Construction of the DP using a stick-breaking process or a gamma process represents the DP as a countably infinite sum of atomic measures. These approaches suggest that a DPM model can be seen as a mixture model with infinitely many components. The distribution of parameters imposed by a DP can also be obtained as a limiting case of a parametric mixture model[8-11]. This approach shows that a DPM can easily be defined as an extension of a parametric mixture model without the need to do model selection for determining the number of components to be used. Here, we take this approach to extend simple finite mixture models to Dirichlet process mixture models.

The DP is defined by two parameters, a positive scalar α and a probability measure G_0, referred to as the concentration parameter and the base measure, respectively. The base distribution G_0 is the parameter on which the nonparametric distribution is centered, which can be thought of as the prior guess[7]. The

Regular Paper
This work is supported by Gatsby Charitable Foundation and PASCAL2.

concentration parameter α expresses the strength of belief in G_0. For small values of α, samples from a DP is likely to be composed of a small number of atomic measures with large weights. For large values, most samples are likely to be distinct, thus *concentrated* on G_0.

The form of the base distribution and the value of the concentration parameter are important parts of the model choice that will effect the modeling performance. The concentration parameter can be given a vague prior distribution and its value can be inferred from the data. It is harder to decide on the base distribution since the model performance will heavily depend on its parametric form even if it is defined in a hierarchical manner for robustness. Generally, the choice of the base distribution is guided by mathematical and practical convenience. In particular, conjugate distributions are preferred for computational ease. It is important to be aware of the implications of the particular choice of the base distribution. An important question is whether the modeling performance is weakened by using a conjugate base distribution instead of a less restricted distribution. A related interesting question is whether the inference is actually computationally cheaper for the conjugate DPM models.

The Dirichlet process Gaussian mixture model (DPGMM) with both conjugate and non-conjugate base distributions has been used extensively in applications of the DPM models for density estimation and clustering[11-15]. However, the performance of the models using these different prior specifications have not been compared. For Gaussian mixture models the conjugate (Normal-Inverse-Wishart) priors have some unappealing properties with prior dependencies between the mean and the covariance parameters, see e.g., [16]. The main focus of this paper is empirical evaluation of the differences between the modeling performance of the DPGMM with conjugate and non-conjugate base distributions.

Markov chain Monte Carlo (MCMC) techniques are the most popular tools used for inference on the DPM models. Inference algorithms for the DPM models with a conjugate base distribution are relatively easier to implement than for the non-conjugate case. Nevertheless several MCMC algorithms have been developed also for the general case[17-18]. In our experiments, we also compare the computational cost of inference on the conjugate and the non-conjugate model specifications.

We define the DPGMM model with a non-conjugate base distribution by removing the undesirable dependency of the distribution of the mean and the covariance parameters. This results in what we refer to as a conditionally conjugate base distribution since one of the parameters (mean or covariance) can be integrated out conditioned on the other, but both parameters cannot simultaneously be integrated out. In the following, we give formulations of the DPGMM with both a conjugate and a conditionally conjugate base distribution G_0. For both prior specifications, we define hyperpriors on G_0 for robustness, building upon [11]. We refer to the models with the conjugate and the conditionally conjugate base distributions shortly as the conjugate model and the conditionally conjugate model, respectively. After specifying the model structure, we discuss in detail how to do inference on both models. We show that mixing performance of the non-conjugate sampler can be improved substantially by exploiting the conditional conjugacy. We present experimental results comparing the two prior specifications. Both predictive accuracy and computational demand are compared systematically on several artificial and real multivariate density estimation problems.

2 Dirichlet Process Gaussian Mixture Models

A DPM model can be constructed as a limit of a parametric mixture model[8-11]. We start with setting out the hierarchical Gaussian mixture model formulation and then take the limit as the number of mixture components approaches infinity to obtain the Dirichlet process mixture model. Throughout the paper, vector quantities are written in bold. The index i always runs over observations, $i = 1, \ldots, n$, and index j runs over components, $j = 1, \ldots, K$. Generally, variables that play no role in conditional distributions are dropped from the condition for simplicity.

The Gaussian mixture model with K components may be written as:

$$p(\boldsymbol{x}|\theta_1, \ldots, \theta_K) = \sum_{j=1}^{K} \boldsymbol{\pi}_j \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, S_j) \qquad (1)$$

where $\theta_j = \{\boldsymbol{\mu}_j, S_j, \pi_j\}$ is the set of parameters for component j, $\boldsymbol{\pi}$ are the *mixing proportions* (which must be positive and sum to one), $\boldsymbol{\mu}_j$ is the mean vector for component j, and S_j is its *precision* (inverse covariance matrix).

Defining a joint prior distribution G_0 on the component parameters and introducing indicator variables c_i, $i = 1, \ldots, n$, the model can be written as:

$$\begin{aligned} \boldsymbol{x}_i \,|\, c_i, \Theta &\sim \mathcal{N}(\boldsymbol{\mu}_{c_i}, S_{c_i}) \\ c_i \,|\, \boldsymbol{\pi} &\sim Discrete(\pi_1, \ldots, \pi_K) \\ (\boldsymbol{\mu}_j, S_j) &\sim G_0 \\ \boldsymbol{\pi} \,|\, \alpha &\sim Dir(\alpha/K, \ldots, \alpha/K). \end{aligned} \qquad (2)$$

Given the mixing proportions $\boldsymbol{\pi}$, the distribution of the

number of observations assigned to each component, called the *occupation number*, is multinomial

$$p(n_1, \ldots, n_K|\boldsymbol{\pi}) = \frac{n!}{n_1!n_2!\cdots n_K!}\prod_{j=1}^{K}\pi_j^{n_j},$$

and the distribution of the indicators is

$$p(c_1, \ldots, c_n|\boldsymbol{\pi}) = \prod_{j=1}^{K}\pi_j^{n_j}. \tag{3}$$

The indicator variables are stochastic variables whose values encode the class (or mixture component) to which observation $\boldsymbol{y}_i$ belongs. The actual values of the indicators are arbitrary, as long as they faithfully represent which observations belong to the same class, but they can conveniently be thought of as taking values from $1, \ldots, K$, where K is the total number of classes, to which observations are associated.

2.1 Dirichlet Process Mixtures

Placing a symmetric Dirichlet distribution with parameter α/K and treating all components as equivalent is the key in defining the DPM as a limiting case of the parametric mixture model. Taking the product of the prior over the mixing proportions $p(\boldsymbol{\pi})$, and the indicator variables $p(\boldsymbol{c}|\boldsymbol{\pi})$ and integrating out the mixing proportions, we can write the prior on $\boldsymbol{c}$ directly in terms of the Dirichlet parameter α:

$$p(\boldsymbol{c}|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)}\prod_{j=1}^{K}\frac{\Gamma(n_j+\alpha/K)}{\Gamma(\alpha/K)}. \tag{4}$$

Fixing all but a single indicator c_i in (4) and using the fact that the datapoints are exchangeable a priori, we may obtain the conditional for the individual indicators:

$$p(c_i = j|\boldsymbol{c}_{-i}, \alpha) = \frac{n_{-i,j}+\alpha/K}{n-1+\alpha}, \tag{5}$$

where the subscript $-i$ indicates all indices except for i, and $n_{-i,j}$ is the number of datapoints, excluding $\boldsymbol{x}_i$, that are associated with class j. Taking the limit $K \to \infty$, the conditional prior for c_i reach the following limits:

components for which $n_{-i,j} > 0$:

$$p(c_i = j|\boldsymbol{c}_{-i}, \alpha) = \frac{n_{-i,j}}{n-1+\alpha}, \tag{6a}$$

all other components combined:

$$p(c_i \neq c_{i'} \text{ for all } i \neq i'|\boldsymbol{c}_{-i}, \alpha) = \frac{\alpha}{n-1+\alpha}. \tag{6b}$$

Note that these probabilities are the same as the probabilities of seating a new customer in a Chinese restaurant process[4]. Thus, the infinite limit of the model in (2) is equivalent to a DPM which we define by starting with Gaussian distributed data $\boldsymbol{x}_i|\theta \sim \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_i, S_i)$, with a random distribution of the model parameters $(\boldsymbol{\mu}_i, S_i) \sim G$ drawn from a DP, $G \sim DP(\alpha, G_0)$.

We need to specify the base distribution G_0 to complete the model. The base distribution of the DP is the prior guess of the distribution of the parameters in a model. It corresponds to the distribution of the component parameters in an infinite mixture model. The mixture components of a DPGMM are specified by their mean and precision (or covariance) parameters, therefore G_0 specifies the prior on the joint distribution of $\boldsymbol{\mu}$ and S. For this model, a conjugate base distribution exists however it has the unappealing property of prior dependency between the mean and the covariance as detailed below. We proceed by giving the detailed model formulation for both conjugate and conditionally conjugate cases.

2.2 Conjugate DPGMM

The natural choice of priors for the mean of the Gaussian is a Gaussian, and a Wishart① distribution for the precision (inverse-Wishart for the covariance). To accomplish conjugacy of the joint prior distribution of the mean and the precision to the likelihood, the distribution of the mean has to depend on the precision. The prior distribution of the mean $\boldsymbol{\mu}_j$ is Gaussian conditioned on S_j:

$$(\boldsymbol{\mu}_j|S_j, \boldsymbol{\xi}, \rho) \sim \mathcal{N}(\boldsymbol{\xi}, (\rho S_j)^{-1}), \tag{7}$$

and the prior distribution of S_j is Wishart:

$$(S_j|\beta, W) \sim \mathcal{W}(\beta, (\beta W)^{-1}). \tag{8}$$

The joint distribution of $\boldsymbol{\mu}_j$ and S_j is the Normal/Wishart distribution denoted as:

$$(\boldsymbol{\mu}_j, S_j) \sim \mathcal{NW}(\boldsymbol{\xi}, \rho, \beta, \beta W), \tag{9}$$

with $\boldsymbol{\xi}, \rho, \beta$ and W being hyperparameters common to all mixture components, expressing the belief where the component parameters should be similar, centered around some particular value. The graphical representation of the hierarchical model is depicted in Fig.1.

Note in (7) that the data precision and the prior on the mean are linked, as the precision for the mean is a multiple of the component precision itself. This dependency is probably not generally desirable since it means

①There are multiple parameterizations used for the density function of the Wishart distribution. We use $\mathcal{W}(\beta, (\beta W)^{-1}) = \frac{(|W|(\beta/2)^D)^{\beta/2}}{\Gamma_D(\beta/2)}|S|^{(\beta-D-1)/2}\exp\Big(-\frac{\beta}{2}tr(SW)\Big)$, where $\Gamma_D(z) = \pi^{D(D-1)/4}\prod_{d=1}^{D}\Gamma\Big(z+(d-D)/2\Big)$.

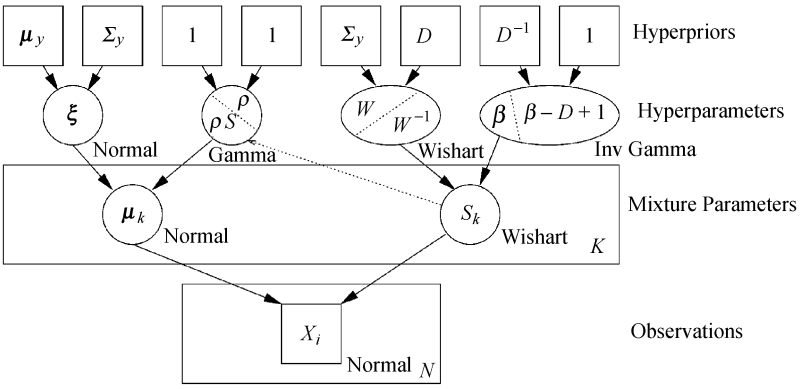

Fig.1. Graphical representation of hierarchical GMM model with conjugate priors. Note the dependency of the distribution of the component mean on the component precision. Variables are labelled below by the name of their distribution, and the parameters of these distributions are given above. The circles/ovals are segmented for the parameters whose prior is defined not directly on that parameter but on a related quantity.

that the prior distribution of the mean of a component depends on the covariance of that component but it is an unavoidable consequence of requiring conjugacy.

2.3 Conditionally Conjugate DPGMM

If we remove the undesired dependency of the mean on the precision, we no longer have conjugacy. For a more realistic model, we define the prior on $\boldsymbol{\mu}_j$ to be

$$p(\boldsymbol{\mu}_j|\boldsymbol{\xi}, R) \sim \mathcal{N}(\boldsymbol{\xi}, R^{-1}) \tag{10}$$

whose mean vector $\boldsymbol{\xi}$ and precision matrix R are hyperparameters common to all mixture components. Keeping the Wishart prior over the precisions as in (8), we obtain the *conditionally* conjugate model. That is, the prior of the mean is conjugate to the likelihood conditional on S and the prior of the precision is conjugate conditional on $\boldsymbol{\mu}$. See Fig.2 for the graphical representation of the hierarchical model.

A model is required to be flexible to be able to deal with a large range of datasets. Furthermore, robustness is required so that the performance of the model does not change drastically with small changes in its parameter values. Generally it is hard to specify good parameter values that give rise to a successful model fit for a given problem and misspecifying parameter values may lead to poor modeling performance. Using hierarchical priors guards against this possibility by allowing to express the uncertainty about the initial parameter values, leading to flexible and robust models. We put hyperpriors on the hyperparameters in both prior specifications to have a robust model. We use the hierarchical model specification of [11] for the conditionally conjugate model, and a similar specification for the conjugate case. Vague priors are put on the hyperparameters, some of which depend on the observations which technically they ought not to. However, only the empirical mean $\boldsymbol{\mu}_y$ and the covariance Σ_y of the data are used in such a way that the full procedure becomes invariant to translations, rotations and rescaling of the

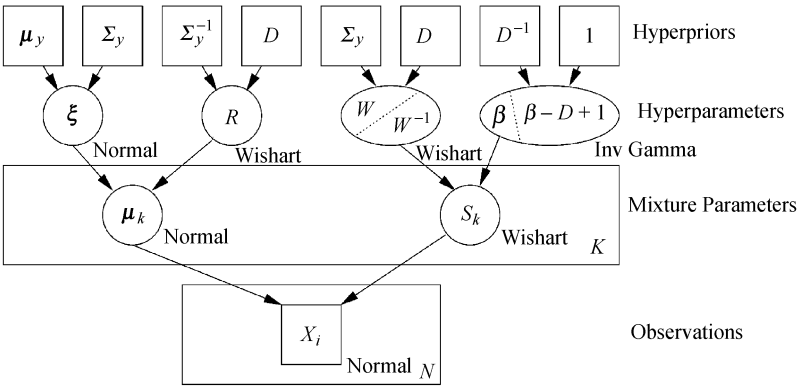

Fig.2. Graphical representation of hierarchical GMM model with conditionally conjugate priors. Note that the mean $\boldsymbol{\mu}_k$ and the precision S_k are independent conditioned on the data.

data. One could equivalently use unit priors and scale the data before the analysis since finding the overall mean and covariance of the data is not the primary concern of the analysis, rather, we wish to find structure within the data.

In detail, for both models the hyperparameters W and β associated with the component precisions S_j are given the following priors, keeping in mind that β should be greater than $D-1$:

$$W \sim \mathcal{W}\Big(D, \frac{1}{D}\Sigma_y\Big), \quad \Big(\frac{1}{\beta - D + 1}\Big) \sim \mathcal{G}\Big(1, \frac{1}{D}\Big). \tag{11}$$

For the conjugate model, the priors for the hyperparameters $\boldsymbol{\xi}$ and ρ associated with the mixture means $\boldsymbol{\mu}_j$ are Gaussian and gamma②:

$$\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{\mu}_y, \Sigma_y), \quad \rho \sim \mathcal{G}(1/2,\, 1/2). \tag{12}$$

For the non-conjugate case, the component means have a mean vector $\boldsymbol{\xi}$ and a precision matrix R as hyperparameters. We put a Gaussian prior on $\boldsymbol{\xi}$ and a Wishart prior on the precision matrix:

$$\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{\mu}_y, \Sigma_y), \quad R \sim \mathcal{W}(D, (D\Sigma_y)^{-1}). \tag{13}$$

The prior number of components is governed by the concentration parameter α. For both models we chose the prior for α^{-1} to have Gamma shape with unit mean and 1 degree of freedom:

$$p(\alpha^{-1}) \sim \mathcal{G}(1, 1). \tag{14}$$

This prior is asymmetric, having a short tail for small values of α, expressing our prior belief that we do not expect that a very small number of active classes (say $K^\dagger = 1$) is overwhelmingly likely a priori.

The probability of the occupation numbers given α and the number of active components, as a function of α, is the likelihood function for α. It can be derived from (6) by reinterpreting this equation to draw a sequence of indicators, each conditioned only on the previously drawn ones. This gives us the following likelihood function:

$$\alpha^{K^\dagger} \prod_{i=1}^{n} \frac{1}{i - 1 + \alpha} = \frac{\alpha^{K^\dagger}\Gamma(\alpha)}{\Gamma(n+\alpha)}, \tag{15}$$

where $K^\dagger$ is the number of active components, that is the components that have nonzero occupation numbers. We notice, that this expression depends only on the total number of active components, $K^\dagger$, and not on how the observations are distributed among them.

3 Inference Using Gibbs Sampling

We utilise MCMC algorithms for inference on the models described in the previous section. The Markov chain relies on Gibbs updates, where each variable is updated in turn by sampling from its posterior distribution conditional on all other variables. We repeatedly sample the parameters, hyperparameters and the indicator variables from their posterior distributions conditional on all other variables. As a general summary, we iterate:

- update mixture parameters $(\boldsymbol{\mu}_j, S_j)$;
- update hyperparameters;
- update the indicators, conditional on the other indicators and the (hyper) parameters;
- update DP concentration parameter α.

For the models we consider, the conditional posteriors for all parameters and hyperparameters except for α, β and the indicator variables c_i are of standard form, thus can be sampled easily. The conditional posteriors of $\log(\alpha)$ and $\log(\beta)$ are both log-concave, so they can be updated using Adaptive Rejection Sampling (ARS)[19] as suggested in [11].

The likelihood for components that have observations associated with them is given by the parameters of that component, and the likelihood pertaining to currently inactive classes (which have no mixture parameters associated with them) is obtained through integration over the prior distribution. The conditional posterior class probabilities are calculated by multiplying the likelihood term by the prior.

The conditional posterior class probabilities for the DPGMM are:

$$\begin{aligned}
&\text{components for which } n_{-i,j} > 0: \\
&\quad p(c_i = j | \boldsymbol{c}_{-i}, \boldsymbol{\mu}_j, S_j, \alpha) \\
&\quad \propto p(c_i = j | \boldsymbol{c}_{-i}, \alpha) p(\boldsymbol{x}_i | \boldsymbol{\mu}_j, S_j) \\
&\quad \propto \frac{n_{-i,j}}{n - 1 + \alpha} \mathcal{N}(\boldsymbol{x} | \boldsymbol{\mu}_j, S_j), \qquad (16a) \\
&\text{all others combined:} \\
&\quad p(c_i \neq c_{i'} \text{ for all } i \neq i' | \boldsymbol{c}_{-i}, \boldsymbol{\xi}, \rho, \beta, W, \alpha) \\
&\quad \propto \frac{\alpha}{n - 1 + \alpha} \\
&\qquad \times \int p(\boldsymbol{x}_i | \boldsymbol{\mu}, S) p(\boldsymbol{\mu}, S | \boldsymbol{\xi}, \rho, \beta, W) \mathrm{d}\boldsymbol{\mu} \mathrm{d}S. \qquad (16b)
\end{aligned}$$

We can evaluate this integral to obtain the conditional posterior of the inactive classes in the conjugate case, but it is analytically intractable in the non-conjugate case.

②Gamma distribution is equivalent to a one-dimensional Wishart distribution. Its density function is given by $\mathcal{G}(\alpha, \beta) = \frac{(\frac{\alpha}{2\beta})^{\alpha/2}}{\Gamma(\alpha/2)} s^{\alpha/2-1} \exp(-\frac{s\alpha}{2\beta})$.

For updating the indicator variables of the conjugate model, we use Algorithm 2 of [9] which makes full use of conjugacy to integrate out the component parameters, leaving only the indicator variables as the state of the Markov chain. For the conditionally conjugate case, we use Algorithm 8 of [9] utilizing auxiliary variables.

3.1 Conjugate Model

When the priors are conjugate, the integral in (16b) is analytically tractable. In fact, even for the active classes, we can marginalise out the component parameters using an integral over their posterior, by analogy with the inactive classes. Thus, in all cases the log likelihood term is:

$$\log\ p(\boldsymbol{x}_i \,|\, c_{-i}, \rho, \boldsymbol{\xi}, \beta, W) = \frac{D}{2}\log\frac{\rho + n_j}{\rho + n_j + 1} -$$

$$\frac{D}{2}\log\pi + \log\Gamma\Big(\frac{\beta + n_j + 1}{2}\Big) -$$

$$\log\Gamma\Big(\frac{\beta + n_j + 1 - D}{2}\Big) + \frac{\beta + n_j}{2}\log|W^*| -$$

$$\frac{\beta + n_j + 1}{2}\log|W^* + \frac{\rho + n_j}{\rho + n_j + 1}(\boldsymbol{x}_i - \boldsymbol{\xi}^*)(\boldsymbol{x}_i - \boldsymbol{\xi}^*)^{\mathrm{T}}|, \tag{17}$$

where

$$\boldsymbol{\xi}^* = \Big(\rho\boldsymbol{\xi} + \sum_{l:c_l=j}\boldsymbol{x}_l\Big)\Big/(\rho + n_j)$$

and

$$W^* = \beta W + \rho\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}} + \sum_{l:c_l=j}\boldsymbol{x}_l\boldsymbol{x}_l^{\mathrm{T}} - (\rho + n_j)\boldsymbol{\xi}^*\boldsymbol{\xi}^{*\mathrm{T}},$$

which simplifies considerably for the inactive classes.

The sampling iterations become:

• update $\boldsymbol{\mu}_j$ and S_j conditional on the data, the indicators and the hyperparameters;

• update hyperparameters conditional on $\boldsymbol{\mu}_j$ and S_j;

• remove the parameters, $\boldsymbol{\mu}_j$ and S_j from representation;

• update each indicator variable, conditional on the data, the other indicators and the hyperparameters;

• update the DP concentration parameter α, using ARS.

3.2 Conditionally Conjugate Model

As a consequence of not using fully conjugate priors, the posterior conditional class probabilities for inactive classes cannot be computed analytically. Here, we give details for using the auxiliary variable sampling scheme of [9] and also show how to improve this algorithm by making use of the conditional conjugacy.

`SampleBoth`

The auxiliary variable algorithm of [9] (Algorithm 8) suggests the following sampling steps. For each observation $\boldsymbol{x}_i$ in turn, the updates are performed by: "invent" ζ auxiliary classes by picking means $\boldsymbol{\mu}_j$ and precisions S_j from their priors. Update c_i using Gibbs sampling, i.e., sample from the discrete conditional posterior class distribution), and finally remove the components that are no longer associated with any observations. Here, to emphasize the difference between the other sampling schemes that we will describe, we call it `SampleBoth` scheme since both means and precisions are sampled from their priors to *represent* the inactive components.

The auxiliary classes represent the effect of the inactive classes, therefore using (6b), the prior for each auxiliary class is

$$\frac{\alpha/\zeta}{n - 1 + \alpha}. \tag{18}$$

In other words, we define $n_{-i,j} = \alpha/\zeta$ for the auxiliary components.

The sampling iterations are as follows:

• Update $\boldsymbol{\mu}_j$ and S_j conditional on the indicators and hyperparameters.

• Update hyperparameters conditional on $\boldsymbol{\mu}_j$ and S_j

• For each indicator variable:

  – If c_i is a singleton, assign its parameters $\boldsymbol{\mu}_{c_i}$ and S_{c_i} to one of the auxiliary parameter pairs;

  – Invent other auxiliary components by sampling values for $\boldsymbol{\mu}_j$ and S_j from their priors, (10) and (8) respectively;

  – Update the indicator variable, conditional on the data, the other indicators and hyperparameters using (16a) and defining $n_{-i,j} = \alpha/\zeta$ for the auxiliary components;

  – Discard the empty components;

• Update DP concentration parameter α.

The integral we want to evaluate is over two parameters, $\boldsymbol{\mu}_j$ and S_j. Exploiting the *conditional* conjugacy, it is possible to integrate over one of these parameters given the other. Thus, we can pick only one of the parameters randomly from its prior, and integrate over the other, which might lead to faster mixing. The log likelihood for the `SampleMu` and the `SampleS` schemes are as follows:

`SampleMu`

Sampling $\boldsymbol{\mu}_j$ from its prior and integrating over S_j give the conditional log likelihood:

$$\log p(\boldsymbol{x}_i|\boldsymbol{c}_{-i}, \boldsymbol{\mu}_j, \beta, W) = -\frac{D}{2}\log\pi + \frac{\beta+n_j}{2}\log|W^*| - \frac{\beta+n_j+1}{2}\log|W^* + \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}| + \log\Gamma\Big(\frac{\beta+n_j+1}{2}\Big) - \log\Gamma\Big(\frac{\beta+n_j+1-D}{2}\Big), \tag{19}$$

where $W^* = \beta W + \sum_{l:c_l=j}(\boldsymbol{x}_l - \boldsymbol{\mu}_j)(\boldsymbol{x}_l - \boldsymbol{\mu}_j)^{\mathrm{T}}$.

The sampling steps for the indicator variables are:

• Remove all precision parameters S_j from the representation;

• If c_i is a singleton, assign its mean parameter $\boldsymbol{\mu}_{c_i}$ to one of the auxiliary parameters;

• Invent other auxiliary components by sampling values for the component mean $\boldsymbol{\mu}_j$ from its prior, (10);

• Update the indicator variable, conditional on the data, the other indicators, the component means and hyperparameters using the likelihood given in (19) and the prior given in (6a) and (18).

• Discard the empty components.

`SampleS`

Sampling S_j from its prior and integrating over $\boldsymbol{\mu}_j$, the conditional log likelihood becomes:

$$\begin{aligned}&\log p(\boldsymbol{x}_i|\boldsymbol{c}_{-i}, S_j, R, \boldsymbol{\xi})\\ =& -\frac{D}{2}\log(2\pi) - \frac{1}{2}\boldsymbol{x}_i^{\mathrm{T}}S_j\boldsymbol{x}_i + \\ &\frac{1}{2}(\boldsymbol{\xi}^* + S_j\boldsymbol{x}_i)^{\mathrm{T}}\big((n_j+1)S_j + R\big)^{-1}(\boldsymbol{\xi}^* + S_j\boldsymbol{x}_i) - \\ &\frac{1}{2}\boldsymbol{\xi}^{*\mathrm{T}}(n_jS_j + R)^{-1}\boldsymbol{\xi}^* + \frac{1}{2}\log\frac{|S_j||n_jS_j+R|}{|(n_j+1)S_j+R|},\end{aligned}$$

where

$$\boldsymbol{\xi}^* = S_j\sum_{l:c_l=j}\boldsymbol{x}_j + R\boldsymbol{\xi}. \tag{20}$$

The sampling steps for the indicator variables are:

• Remove the component means $\boldsymbol{\mu}_j$ from the representation;

• If c_i is a singleton, assign its precision S_{c_i} to one of the auxiliary parameters;

• Invent other auxiliary components by sampling values for the component precision S_j from its prior, (8);

• Update the indicator variable, conditional on the data, the other indicators, the component precisions and hyperparameters using the likelihood given in (20) and the prior given in (6a) and (18);

• Discard the empty components.

The number of auxiliary components ζ can be thought of as a free parameter of these sampling algorithms. It is important to note that the algorithms will converge to the same true posterior distribution regardless of how many auxiliary components are used. The value of ζ may effect the convergence and mixing time of the algorithm. If more auxiliary components are used, the Markov chain may mix faster, however using more components will increase the computation time of each iteration. In our experiments, we tried different values of ζ and found that using a single auxiliary component was enough to get a good mixing behaviour. In the experiments, we report results using $\zeta = 1$.

Note that `SampleBoth`, `SampleMu` and `SampleS` are three different ways of doing inference for the same model since there are no approximations involved. One should expect only the computational cost of the inference algorithm (e.g., the convergence and the mixing time) to differ among these schemes, whereas the conjugate model discussed in the previous section is a different *model* since the prior distribution is different.

4 Predictive Distribution

The predictive distribution is obtained by averaging over the samples generated by the Markov Chain. For a particular sample in the chain, the predictive distribution is a mixture of the finite number of Gaussian components which have observations associated with them, and the infinitely many components that have no data associated with them given by the integral over the base distribution:

$$\int p(\boldsymbol{x}_i|\boldsymbol{\mu}, S)p(\boldsymbol{\mu}, S|\boldsymbol{\xi}, \rho, \beta, W)\,\mathrm{d}\boldsymbol{\mu}\,\mathrm{d}S. \tag{21}$$

The combined mass of the represented components is $n/(n+\alpha)$ and the combined mass of the unrepresented components is the remaining $\alpha/(n+\alpha)$.

For the conjugate model, the predictive distribution can be given analytically. For the non-conjugate models, since the integral for the unrepresented classes (21) is not tractable, it is approximated by a mixture of a finite number $\zeta^\dagger$ of components, with parameters ($\boldsymbol{\mu}$ and S, or $\boldsymbol{\mu}$ or S, depending on which of the three sampling schemes is being used) drawn from the base distribution $p(\boldsymbol{\mu}, S)$. Note that the larger the $\zeta^\dagger$, the better the approximation will get. The predictive performance of the model will be underestimated if we do not use enough components to approximate the integral. Therefore the predictive performance calculated using a finite $\zeta^\dagger$ can be thought of as a lower bound on the actual performance of the model. In our experiments, we used $\zeta^\dagger = 10$.

5 Experiments

We present results on simulated and real datasets with different dimensions to compare the predictive accuracy and computational cost of the different model specifications and sampling schemes described above.

We use the duration of consecutive eruptions of the Old Faithful geyser[20] as a two dimensional example. The three dimensional Spiral dataset used in [11], the four dimensional "Iris" dataset used in [21] and the 13 dimensional "Wine" dataset[22] are modeled for assessing the model performance and the computational cost in higher dimensions.

The density of each dataset has been estimated using the conjugate model (CDP), the three different sampling schemes for the conditionally conjugate model (CCDP), `SampleBoth, SampleMu, and SampleS`, and by kernel density estimation③ (KDE) using Gaussian kernels.

5.1 Density Estimation Performance

As a measure for modeling performance, we use the average leave one out predictive densities. That is, for all datasets considered, we leave out one observation, model the density using all others, and calculate the log predictive density on the left-out datapoint. We repeat this for all datapoints in the training set and report the average log predictive density. We choose this as a performance measure because the log predictive density gives a quantity proportional to the KL divergence which is a measure of the discrepancy between the actual generating density and the modeled density. The three different sampling schemes for the conditionally conjugate model all have identical equilibrium distributions, therefore the result of the conditionally conjugate model is presented only once, instead of discriminating between different schemes when the predictive densities are considered.

We use the Old Faithful geyser dataset to visualise the estimated densities, Fig.3. Visualisation is important for giving an intuition about the behaviour of the different algorithms. Convergence and mixing of all samplers is fast for the two dimensional Geyser dataset. There is also not a significant difference in the predictive performance, see Tables 1 and 2. However, we can see form the plots in Fig.3 and the average entropy values in Table 3 that the resulting density estimates are different for all models. CDP uses fewer components to fit the data, therefore the density estimate has fewer modes compared to the estimates obtained by CCDP or KDE. In Fig.3 we see three almost Gaussian modes for CDP. Since KDE places a kernel on each datapoint, its resulting density estimate is less smooth. Note that both CDP and CCDP have the potential to use one component per datapoint and return the same estimate as the KDE, however their prior as well as the likelihood of the data does not favor this. The density fit by CCDP can be seen as an interpolation between the CDP and the KDE results as it utilizes more mixture components than CDP but has a smoother density estimate than KDE. For all datasets, the KDE model has the lowest average leave one out predictive density, and the conditionally conjugate model has the best, with the difference between the models getting larger in higher dimensions, see Table 1. For instance, on the Wine data CCDP is five fold better than KDE.

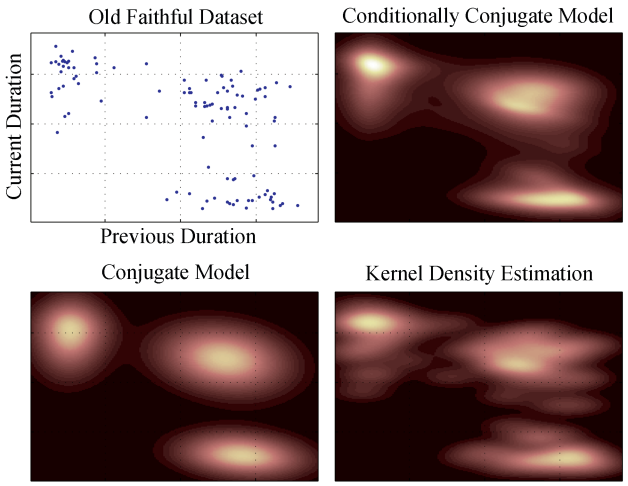

Fig.3. Old Faithful geyser dataset and its density modelled by CDP, CCDP and KDE. The two dimensional data consists of the durations of the consecutive eruptions of the Old Faithful geyser.

Table 1. Average Leave One Out Log-Predictive Densities for Kernel Density Estimation (KDE), Conjugate DP Mixture Model (CDP), Conditionally Conjugate DP Mixture Model (CCDP) on Different Datasets

Dataset	KDE	CDP	CCDP
Geyser	−1.906	−1.902	−1.879
Spiral	−7.205	−7.123	−7.117
Iris	−1.860	−1.577	−1.546
Wine	−18.979	−17.595	−17.341

Table 2. Paired t-Test Scores of Leave One Out Predictive Densities (The test does not give enough evidence in case of the Geyser data, however it shows that KDE is statistically significantly different from both DP models for the higher dimensional datasets. Also, CDP is significantly different from CCDP for the Wine data.)

Dataset	KDE/CDP	KDE/CCDP	CDP/CCDP
Geyser	0.95	0.59	0.41
Spiral	< 0.01	< 0.01	0.036
Iris	< 0.01	< 0.01	0.099
Wine	< 0.01	< 0.01	< 0.01

③Kernel density estimation is a classical non parametric density estimation technique which places kernels on each training datapoint. The kernel bandwidth is adjusted separately on each dimension to obtain a smooth density estimate, by maximising the sum of leave-one-out log densities.

p-values for a paired t-test are given in Table 2 to compare the distribution of the leave one out densities. There is no statistically significant difference between any of the models on the Geyser dataset. For the Spiral, Iris and Wine datasets, the difference between the predictions of KDE and both DP models are statistically significant. CCDP is significantly different from CDP in terms of its predictive density only on the Wine dataset.

Furthermore, for all datasets, the CCDP consistently utilizes more components than CDP. The average number of datapoints assigned to different components and the distribution over the number of active components are given in Fig.4 and Fig.5, respectively. Fig.5 shows that the distribution of the number of components used by the CCDP is much broader and centered on higher values. The difference in the density estimates is also reflected in the average entropy values reported in Table 3.

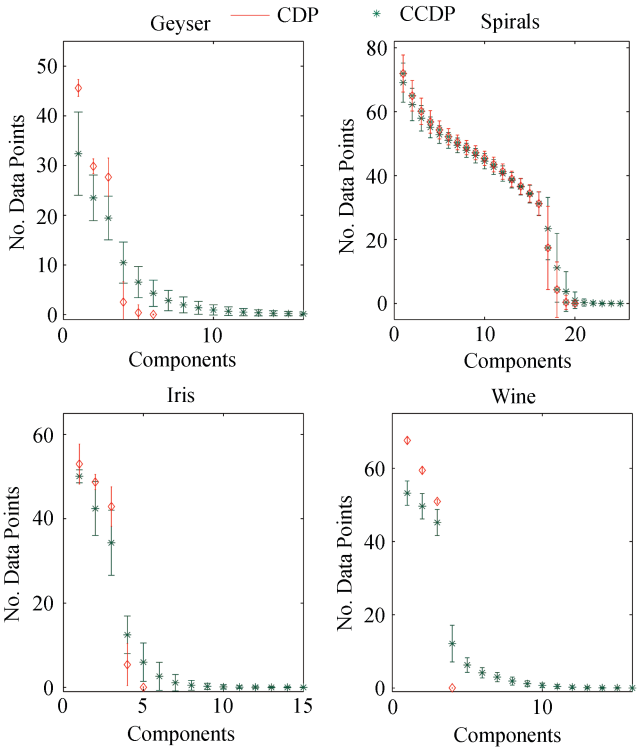

Fig.4. Number of datapoints assigned to the components averaged over different positions in the chain. The standard deviations are indicated by the error bars. Note the existence of many small components for the CCDP model.

Table 3. Average Entropies of Mixing Proportions for the Conjugate DP Mixture Model (CDP), Conditionally Conjugate DP Mixture Model (CCDP) on Various Datasets

Dataset	CDP	CCDP
Geyser	1.64 (0.14)	2.65 (0.51)
Spiral	4.03 (0.06)	4.10 (0.08)
Iris	1.71 (0.13)	2.13 (0.28)
Wine	1.58 (0.004)	2.35 (0.17)

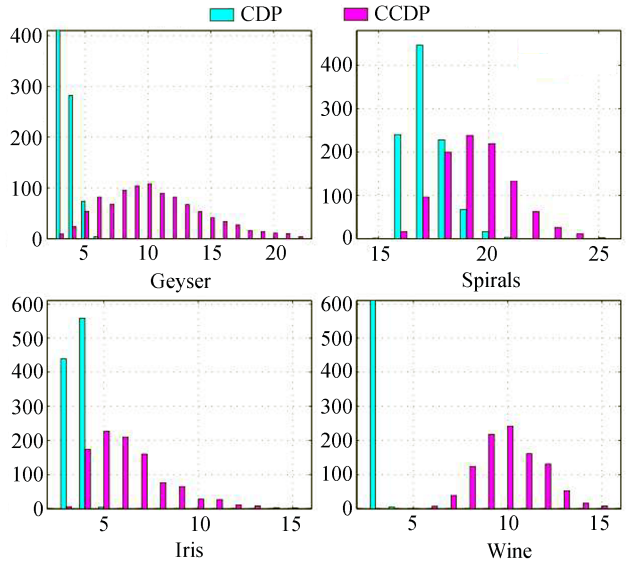

Fig.5. Distribution of number of active components for DPGMM from 1000 iterations. The CDP model favors a lower number of components for all datasets. The average number of components for the CCDP model is larger, with a more diffuse distribution. Note that histogram for the CDP model for the Geyser dataset and the Wine dataset has been cut off on the y-axis.

5.2 Clustering Performance

The main objective of the models presented in this paper is density estimation, but the models can be used for clustering (or classification where labels are available) as well by observing the assignment of datapoints to model components. Since the number of components change over the chain, one would need to form a confusion matrix showing the frequency of each data pair being assigned to the same component for the entire Markov chain, see Fig.6.

Class labels are available for the Iris and Wine datasets, both datasets consisting of 3 classes. The CDP model has 3∼4 active components for the Iris data and 3 active components for the Wine dataset. The assignment of datapoints to the components shows successful clustering. The CCDP model has more components on average for both datasets, but datapoints with different labels are generally not assigned to the same component, resulting in successful clustering which can be seen by the block diagonal structure of the confusion matrices given in Fig.6, and comparing to the true labels given on the right hand side figures. The confusion matrices were constructed by counting the number of times each datapoint was assigned to the same component with another datapoint. Note that the CCDP utilizes more clusters, resulting in less extreme values

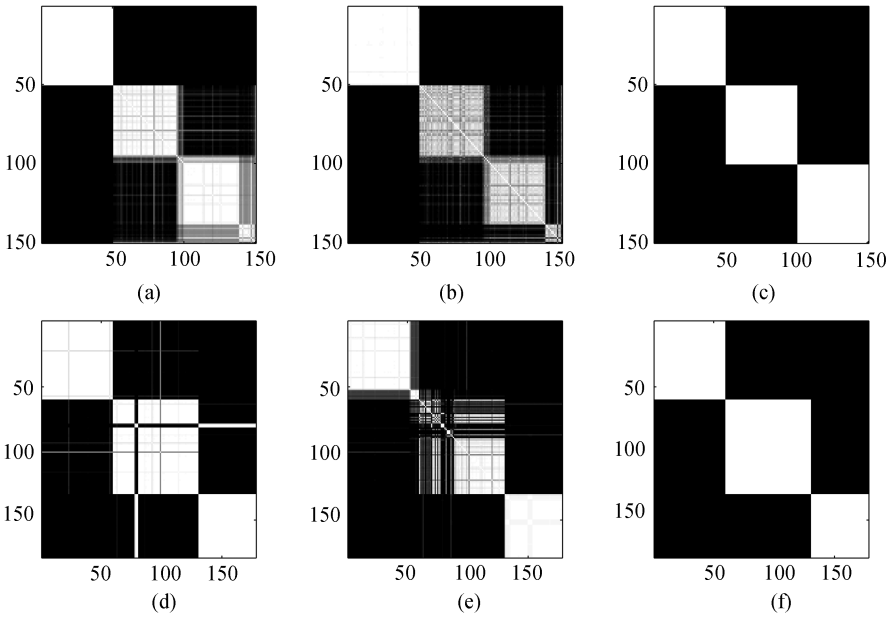

Fig.6. Confusion matrices for the Iris dataset: (a) CDP; (b) CCDP; (c) Correct labels. Wine dataset: (d) CDP; (e) CCDP; (f) Correct labels. Brighter means higher, hence the darker gray values for the datapoints that were not always assigned to the same component.

for the confusion matrix entries (darker gray values of the confusion matrix) which expresses the uncertainty in cluster assignments for some datapoints. Furthermore, we can see in Fig.6(d) that there is one datapoint in the Wine dataset that CDP assigns to the wrong cluster for almost all MCMC iterations, whereas this datapoint is allowed to have its own cluster by CCDP.

The Spiral dataset is generated by sampling 5 points form each of the 160 Gaussians whose means lie on a spiral. For this data, the number of active components of CDP and CCDP do not go beyond 21 and 28, respectively. This is due to the assumption of independence of component means for both models, which does not hold for this dataset, therefore it is not surprising that the models do not find the correct clustering structure although they can closely estimate the density.

5.3 Computational Cost

The differences in density estimates and predictive performances show that different specification of the base distribution leads to different behaviour of the model on the same data. It is also interesting to find out if there is a significant gain (if at all) in computational efficiency when conjugate base distribution is used rather than the non-conjugate one. The inference algorithms considered only differ in the way they update the indicator variables, therefore the computation time per iteration is similar for all algorithms.

We use the convergence and burn-in time as measures of computational cost. Convergence was determined by examining various properties of the state of the Markov chain, and mixing time was calculated as the sum of the auto-covariance coefficients of the slowest mixing quantities from lag -1000 to 1000.

The slowest mixing quantity was the number of active components in all experiments. Example auto-covariance coefficients are shown in Fig.7. The convergence time for the CDP model is usually shorter than the `SampleBoth` scheme of CCDP but longer than the two other schemes. For the CCDP model, the `SampleBoth` scheme is the slowest in terms of both converging and mixing. `SampleS` has comparable convergence time to the `SampleMu` scheme.

6 Conclusions

The Dirichlet process mixtures of Gaussians model is one of the most widely used DPM models. We have presented hierarchical formulations of DPGMM with conjugate and conditionally conjugate base distributions. The only difference between the two models is the prior on $\boldsymbol{\mu}$ and the related hyperparameters. We kept the specifications for all other parameters the same so as to make sure only the presence or absence of conjugacy would effect the results. We compared the two model specifications in terms of their modeling properties for density estimation and clustering and the computational cost of the inference algorithms on several datasets with differing dimensions.

Experiments show that the behaviour of the two base distributions differ. For density estimation, the predictive accuracy of the CCDP model is found to be better than the CDP model for all datasets considered, the

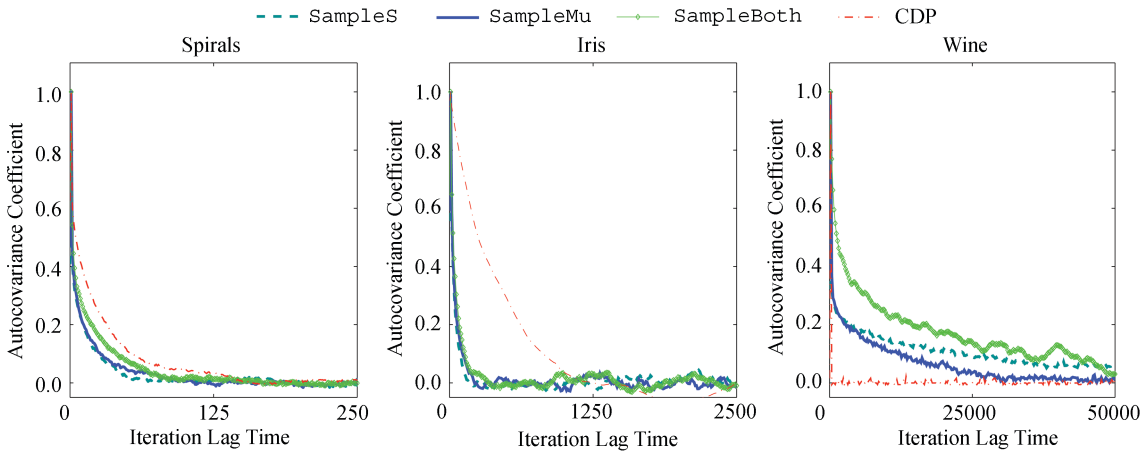

Fig.7. Autocorrelation coefficients of the number of active components for CDP and different sampling schemes for CCDP, for the Spiral data based on 5×10^5 iterations, the Iris data based on 10^6 iterations and Wine data based on 1.5×10^6 iterations.

difference being larger in high dimensions. The CDP model tends to use less components than the CCDP model, having smaller entropy. The clustering performance of both cases are comparable, with the CCDP expressing more uncertainty on some datapoints and allowing some datapoints to have their own cluster when they do not match the datapoints in other clusters. From the experimental results we can conclude that the more restrictive form of the base distribution forces the conjugate model to be more parsimonious in the number of components utilized. This may be a desirable feature if only a rough clustering is adequate for the task in hand, however it has the risk of overlooking the outliers in the dataset and assigning them to a cluster together with other datapoints. Since it is more flexible in the number of components, the CCDP model may in general result in more reliable clusterings.

We adjusted MCMC algorithms from [9] for inference on both specifications of the model and proposed two sampling schemes for the conditionally conjugate base model with improved convergence and mixing properties. Although MCMC inference on the conjugate case is relatively easier to implement, experimental results show that it is not necessarily computationally cheaper than inference on the conditionally conjugate model when conditional conjugacy is exploited.

In the light of the empirical results, we conclude that marginalising over one of the parameters by exploiting conditional conjugacy leads to considerably faster mixing in the conditionally conjugate model. When using this trick, the fully conjugate model is not necessarily computationally cheaper in the case of DPGMM. The DPGMM with the more flexible prior specification (conditionally conjugate prior) can be used on higher dimensional density estimation problems, resulting in better density estimates than the model with conjugate prior specification.

References

[1] Ferguson T S. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1973, 1(2): 209-230.

[2] Blackwell D, MacQueen J B. Ferguson distributions via Pólya urn schemes. *Annals of Statistics*, 1973, 1(2): 353-355.

[3] Aldous D. Exchangeability and Related Topics. Ecole d'Été de Probabilités de Saint-Flour XIII–1983, Berlin: Springer, 1985, pp.1-198.

[4] Pitman J. Combinatorial Stochastic Processes Ecole d'Eté de Probabilités de Saint-Flour XXXII – 2002, Lecture Notes in Mathematics, Vol. 1875, Springer, 2006.

[5] Sethuraman J, Tiwari R C. Convergence of Dirichlet Measures and the Interpretation of Their Parameter. Statistical Decision Theory and Related Topics, III, Gupta S S, Berger J O (eds.), London: Academic Press, Vol.2, 1982, pp.305-315.

[6] Ishwaran H, James L F. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, March 2001, 96(453): 161-173.

[7] Antoniak C E. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, 1974, 2(6): 1152–1174.

[8] Neal R M. Bayesian mixture modeling. In *Proc. the 11th International Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis*, Seattle, USA, June, 1991, pp.197-211.

[9] Neal R M. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 2000, 9(2): 249-265.

[10] Green P, Richardson S. Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics*, 2001, 28: 355-375.

[11] Rasmussen C E. The infinite Gaussian mixture model. *Advances in Neural Information Processing Systems*, 2000, 12: 554-560.

[12] Escobar M D. Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 1994, 89(425): 268-277.

[13] MacEachern S N. Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics: Simulation and Computation*, 1994, 23(3): 727-741.

[14] Escobar M D, West M. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical*

Association, 1995, 90(430): 577-588.

[15] Müller P, Erkanli A, West M. Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 1996, 83(1): 67-79.

[16] West M, Müller P, Escobar M D. Hierarchical Priors and Mixture Models with Applications in Regression and Density Estimation. Aspects of Uncertainty, Freeman P R, Smith A F M (eds.), John Wiley, 1994, pp.363-386.

[17] MacEachern S N, Müller P. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 1998, 7(2): 223-238.

[18] Neal R M. Markov chain sampling methods for Dirichlet process mixture models. Technical Report 4915, Department of Statistics, University of Toronto, 1998.

[19] Gilks W R, Wild P. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 1992, 41(2): 337-348.

[20] Scott D W. Multivariate Density Estimation: Theory, Practice and Visualization, Wiley, 1992.

[21] Fisher R A. The use of multiple measurements in axonomic problems. *Annals of Eugenics*, 1936, 7: 179-188.

[22] Forina M, Armanino C, Castino M, Ubigli M. Multivariate data analysis as a discriminating method of the origin of wines. *Vitis*, 1986, 25(3): 189-201.

Dilan Görür is a post-doctoral research fellow in the Gatsby Computational Neuroscience Unit, University College London. She completed her Ph.D. on machine learning in 2007 at the Max Planck Institute for Biological Cybernetics, Germany under the supervision of Carl Edward Rasmussen. She received the B.S. and M.Sc. degrees from Electrical and Electronics Engineering Department of Middle East Technical University, Turkey in 2000 and 2003, respectively. Her research interests lie in the theoretical and practical aspects of machine learning and Bayesian inference.

Carl Edward Rasmussen is a lecturer in the Computational and Biological Learning Lab at the Department of Engineering, University of Cambridge and an adjunct research scientist at the Max Planck Institute for Biological Cybernetics, Tübingen, Germany. His main research interests are Bayesian inference and machine learning. He received his Masters' degree in engineering from the Technical University of Denmark and his Ph.D. degree in computer science from the University of Toronto in 1996. Since then he has been a post doc at the Technical University of Denmark, a senior research fellow at the Gatsby Computational Neuroscience Unit at University College London from 2000 to 2002 and a junior research group leader at the Max Planck Institute for Biological Cybernetics in Tübingen, Germany, form 2002 to 2007.