Gibbs sampling for fitting finite and infinite Gaussian mixture models

Herman Kamper
h.kamper@sms.ed.ac.uk

14 November 2013

1 Introduction

This document gives a high-level summary of the necessary details for implementing collapsed Gibbs sampling for fitting Gaussian mixture models (GMMs) following a Bayesian approach. The document structure is as follows. After notation and reference sections (Sections 2 and 3), the case for sampling the parameters of a finite Gaussian mixture model is described in Section 4. This is then extended to the infinite case in Section 5.

Much of this document is based on content from [1]. I recommend reading the document in conjunction with Sections 24.2 and 25.2 in [1] while consulting the other references given throughout this text.

2 Notation

We aim to follow generally the same notation as that used in [1]. Below is a (limited) summary of the notation used.

2.1 Data

N	Number of data vectors.
D	Dimension of data vectors.
$\mathbf{x}_i \in \mathbb{R}^D$	The i^{th} data vector.
$\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$	Set of data vectors.
$\mathcal{X}_{\setminus i}$	All data vectors apart from $\mathbf{x}_i$.
$\mathcal{X}_k$	Set of data vectors from mixture component k.
$\mathcal{X}_{k\setminus i}$	Set of data vectors from mixture component k, without taking $\mathbf{x}_i$ into account.
N_k	Number of data vectors from mixture component k.
$N_{k\setminus i}$	Number of data vectors from mixture component k, without taking $\mathbf{x}_i$ into account.

2.2 Model parameters

K	Number of components in a finite mixture model.
$z_i \in \{1, 2, \ldots, K\}$	Discrete latent state indicating which component observation $\mathbf{x}_i$ belongs to.
$\mathbf{z} = (z_1, z_2, \ldots, z_N)$	Latent states for all observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$.
$\mathbf{z}_{\setminus i}$	All latent states excluding z_i.
$\boldsymbol{\mu}$	Mean vector of a multivariate Gaussian density. A subscript is used to for a particular component in a mixture model, e.g. $\boldsymbol{\mu}_k$.
$\boldsymbol{\Sigma}$	Covariance matrix of a multivariate Gaussian density. A subscript is used for a particular component in a mixture model, e.g. $\boldsymbol{\Sigma}_k$.
$\pi_k = P(z_i = k)$	Prior probability that data vector $\mathbf{x}_i$ will be assigned to mixture component k.
$\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)$	Prior assignment probability for all K components.

2.3 Hyper-parameters

$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$	Parameter for Dirichlet prior on the mixing weights $\boldsymbol{\pi}$.
$\boldsymbol{\beta} = (\mathbf{m}_0, \kappa_0, \nu_0, \mathbf{S}_0)$	Parameters for the Gaussian-inverse-Wishart prior on mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ of a multivariate Gaussian distribution. The interpretation for the individual parameters are given below.
$\mathbf{m}_0$	Prior mean for $\boldsymbol{\mu}$.
κ_0	How strongly we believe the above prior.
$\mathbf{S}_0$	Proportional to prior mean for $\boldsymbol{\Sigma}$.
ν_0	How strongly we believe the above prior.

3 The multivariate Gaussian with fully conjugate prior

This section serves as reference for the rest of the document. The content is based on [1, Ch. 4.6], [2] and [3]. In this section, hyper-parameters are implied and not explicitly noted on the right of the conditioning bar of densities. The rest of the document will be more explicit.

3.1 Likelihood

The likelihood of random vectors $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ being generated by a multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ is

$$\begin{aligned} p(\mathcal{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= (2\pi)^{-ND/2}|\boldsymbol{\Sigma}|^{-N/2} \exp\left(-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})\right) \qquad (1) \\ &= (2\pi)^{-ND/2}|\boldsymbol{\Sigma}|^{-N/2} \exp\left(-\frac{N}{2}(\boldsymbol{\mu} - \overline{\mathbf{x}})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \overline{\mathbf{x}})\right) \cdot \\ &\quad \exp\left(-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S}_{\overline{x}})\right) \qquad (2) \end{aligned}$$

where

$$\begin{aligned} \mathbf{S}_{\overline{x}} &\triangleq \sum_{n=1}^{N}(\mathbf{x}_n - \overline{\mathbf{x}})(\mathbf{x}_n - \overline{\mathbf{x}})^{\mathrm{T}} \\ \overline{\mathbf{x}} &\triangleq \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \end{aligned}$$

The equivalence of (1) and (2) follows from the identity [1, p. 132]:

$$\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) = \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S}_{\overline{x}}) + N(\overline{\mathbf{x}} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu})$$

3.2 Prior on parameters

For the mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ of a multivariate Gaussian, the Gaussian-inverse-Wishart (GIW) prior is fully conjugate [1, p. 133]:

$$\begin{aligned} \mathrm{GIW}(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{m}_0, \kappa_0, \nu_0, \mathbf{S}_0) &\triangleq \mathcal{N}(\boldsymbol{\mu}|\mathbf{m}_0, \frac{1}{\kappa_0}\boldsymbol{\Sigma}) \cdot \mathrm{IW}(\boldsymbol{\Sigma}|\mathbf{S}_0, \nu_0) \\ &= \frac{1}{Z_{\mathrm{GIW}}(D, \kappa_0, \nu_0, \mathbf{S}_0)}|\boldsymbol{\Sigma}|^{-1/2} \exp\left(\frac{\kappa_0}{2}(\boldsymbol{\mu} - \mathbf{m}_0)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \mathbf{m}_0)\right) \cdot \\ &\quad |\boldsymbol{\Sigma}|^{-\frac{\nu_0+D+1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S}_0)\right) \\ &= \frac{1}{Z_{\mathrm{GIW}}(D, \kappa_0, \nu_0, \mathbf{S}_0)}|\boldsymbol{\Sigma}|^{-\frac{\nu_0+D+2}{2}} \cdot \\ &\quad \exp\left(-\frac{\kappa_0}{2}(\boldsymbol{\mu} - \mathbf{m}_0)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \mathbf{m}_0) - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S}_0)\right) \qquad (3) \end{aligned}$$

with

$$Z_{\mathrm{GIW}}(D, \kappa_0, \nu_0, \mathbf{S}_0) = 2^{\frac{(\nu_0+1)D}{2}}\pi^{D(D+1)/4}\kappa_0^{-D/2}|\mathbf{S}_0|^{-\nu_0/2}\prod_{i=1}^{D}\Gamma\left(\frac{\nu_0 + 1 - i}{2}\right) \qquad (4)$$

Thus the fully conjugate prior density is

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathrm{GIW}(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{m}_0, \kappa_0, \nu_0, \mathbf{S}_0)$$

An intuitive interpretation of the hyper-parameters are as follows [1, p. 133]: $\mathbf{m}_0$ is our prior mean for $\boldsymbol{\mu}$, κ_0 is how strongly we believe this prior for $\boldsymbol{\mu}$, $\mathbf{S}_0$ is proportional to our prior mean for $\boldsymbol{\Sigma}$, and ν_0 is how strongly we believe this prior for $\boldsymbol{\Sigma}$. Because the Gamma function is not defined for negative integers and zero, from (4) we require $\nu_0 > D - 1$.

3.3 Full joint

An expression for the full joint of the data $\mathcal{X}$ and the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be obtained as follows [1, p. 143]:

$$\begin{aligned}
p(\mathcal{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= p(\mathcal{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\, p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= \frac{(2\pi)^{-ND/2}}{Z_{\text{GIW}}(D, \kappa_0, \nu_0, \mathbf{S}_0)} |\boldsymbol{\Sigma}|^{-\frac{\nu_0+N+D+2}{2}} \cdot \\
&\quad \exp\left(-\frac{N}{2}(\boldsymbol{\mu} - \overline{\mathbf{x}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \overline{\mathbf{x}}) - \frac{\kappa_0}{2}(\boldsymbol{\mu} - \mathbf{m}_0)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \mathbf{m}_0) \right. \\
&\qquad \left. - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S}_{\overline{x}}) - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S}_0) \right) \\
&= \frac{(2\pi)^{-ND/2}}{Z_{\text{GIW}}(D, \kappa_0, \nu_0, \mathbf{S}_0)} |\boldsymbol{\Sigma}|^{-\frac{\nu_0+N+D+2}{2}} \cdot \\
&\quad \exp\left\{ -\frac{\kappa_0 + N}{2} \left(\boldsymbol{\mu} - \frac{\kappa_0 \mathbf{m}_0 + N\overline{\mathbf{x}}}{\kappa_N} \right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\mu} - \frac{\kappa_0 \mathbf{m}_0 + N\overline{\mathbf{x}}}{\kappa_N} \right) \right. \\
&\qquad \left. - \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1} \left(\mathbf{S}_0 + \mathbf{S}_{\overline{x}} + \frac{\kappa_0 N}{\kappa_0 + N} (\overline{\mathbf{x}} - \mathbf{m}_0)(\overline{\mathbf{x}} - \mathbf{m}_0)^{\mathrm{T}} \right) \right] \right\} \qquad (5)
\end{aligned}$$

where we used the form in (2) for the likelihood.

3.4 Posterior of parameters

The posterior of the $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ parameters is

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{X}) \propto p(\mathcal{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\, p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad (6)$$

The right hand side of (6) is the full joint given in (5). By comparing (3) and (5), it follows that the posterior is a GIW density with updated parameters [1, p. 134]:

$$\begin{aligned}
p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{X}) &= \text{GIW}(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{m}_N, \kappa_N, \nu_N, \mathbf{S}_N) \qquad (7) \\
\mathbf{m}_N &= \frac{\kappa_0 \mathbf{m}_0 + N\overline{\mathbf{x}}}{\kappa_N} \\
\kappa_N &= \kappa_0 + N \\
\nu_N &= \nu_0 + N \\
\mathbf{S}_N &= \mathbf{S}_0 + \mathbf{S}_{\overline{x}} + \frac{\kappa_0 N}{\kappa_0 + N} (\overline{\mathbf{x}} - \mathbf{m}_0)(\overline{\mathbf{x}} - \mathbf{m}_0)^{\mathrm{T}} \\
&= \mathbf{S}_0 + \mathbf{S} + \kappa_0 \mathbf{m}_0 \mathbf{m}_0^{\mathrm{T}} - \kappa_N \mathbf{m}_N \mathbf{m}_N^{\mathrm{T}} \qquad (8)
\end{aligned}$$

where we define $\mathbf{S} \triangleq \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}}$.

3.5 Marginal likelihood of data

Using the full joint in (5), the marginal likelihood of the data can be obtained as follows [2]:

$$p(\mathcal{X}) = \int_{\boldsymbol{\mu}} \int_{\boldsymbol{\Sigma}} p(\mathcal{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \, \mathrm{d}\boldsymbol{\mu} \, \mathrm{d}\boldsymbol{\Sigma} = \int_{\boldsymbol{\mu}} \int_{\boldsymbol{\Sigma}} p(\mathcal{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \, p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \, \mathrm{d}\boldsymbol{\mu} \, \mathrm{d}\boldsymbol{\Sigma} \tag{9}$$

$$= \frac{(2\pi)^{-ND/2}}{Z_{\mathrm{GIW}}(D, \kappa_0, \nu_0, \mathbf{S}_0)} \int_{\boldsymbol{\mu}} \int_{\boldsymbol{\Sigma}} |\boldsymbol{\Sigma}|^{-\frac{\nu_0+N+D+2}{2}} \cdot$$

$$\exp\left(-\frac{\kappa_N}{2} (\boldsymbol{\mu} - \mathbf{m}_N) \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \mathbf{m}_N) - \frac{1}{2} \mathrm{tr}(\boldsymbol{\Sigma}^{-1} \mathbf{S}_N) \right) \mathrm{d}\boldsymbol{\mu} \, \mathrm{d}\boldsymbol{\Sigma} \tag{10}$$

$$= (2\pi)^{-ND/2} \frac{Z_{\mathrm{GIW}}(D, \kappa_N, \nu_N, \mathbf{S}_N)}{Z_{\mathrm{GIW}}(D, \kappa_0, \nu_0, \mathbf{S}_0)} \tag{11}$$

$$= \pi^{-ND/2} \, \frac{\kappa_0^{D/2} |\mathbf{S}_0|^{\nu_0/2}}{\kappa_N^{D/2} |\mathbf{S}_N|^{\nu_N/2}} \prod_{i=1}^{D} \frac{\Gamma\left(\frac{\nu_N+1-i}{2}\right)}{\Gamma\left(\frac{\nu_0+1-i}{2}\right)} \tag{12}$$

Equation (11) follows from (10) since the integral reduces to the normalizing constant of the GIW density of the posterior given in (7). The final result in (12) is obtained by substituting in the GIW normalizing constants as defined in (4).

3.6 Posterior predictive

Suppose we observe a new data vector $\mathbf{x}^*$. Then the posterior predictive for this vector is

$$p(\mathbf{x}^*|\mathcal{X}) = \frac{p(\mathbf{x}^*, \mathcal{X})}{p(\mathcal{X})} \tag{13}$$

An expression for the denominator in (13) is given in (12). The numerator can be obtained in a similar way from (12) by considering the marginal likelihood of the new set $\{\mathcal{X}, \mathbf{x}^*\}$. Thus, using the notation $\mathbf{S}_{N*}$ to denote the calculation of (8) on this new set, we can calculate (13) as

$$p(\mathbf{x}^*|\mathcal{X}) = \frac{(2\pi)^{-(N+1)D/2}}{(2\pi)^{-ND/2}} \frac{Z_{\mathrm{GIW}}(D, \kappa_N + 1, \nu_N + 1, \mathbf{S}_{N*})}{\cancel{Z_{\mathrm{GIW}}(D, \kappa_0, \nu_0, \mathbf{S}_0)}} \frac{\cancel{Z_{\mathrm{GIW}}(D, \kappa_0, \nu_0, \mathbf{S}_0)}}{Z_{\mathrm{GIW}}(D, \kappa_N, \nu_N, \mathbf{S}_N)}$$

$$= (2\pi)^{-D/2} \, \frac{Z_{\mathrm{GIW}}(D, \kappa_N + 1, \nu_N + 1, \mathbf{S}_{N*})}{Z_{\mathrm{GIW}}(D, \kappa_N, \nu_N, \mathbf{S}_N)}$$

$$= \pi^{-D/2} \frac{(\kappa_N + 1)^{-D/2} |\mathbf{S}_{N*}|^{-(\nu_N+1)/2} \prod_{i=1}^{D} \Gamma\left(\frac{\nu_N+2-i}{2}\right)}{\kappa_N^{-D/2} |\mathbf{S}_N|^{-\nu_N/2} \prod_{i=1}^{D} \Gamma\left(\frac{\nu_N+1-i}{2}\right)} \tag{14}$$

where we used the form in (11) for the marginals and then substituted (4) in.

As an alternative form, it can be shown that this posterior predictive has a multivariate Student's t distribution [1, p. 135]:

$$p(\mathbf{x}^*|\mathcal{X}) = \mathcal{T}(\mathbf{x}^*|\mathbf{m}_N, \frac{\kappa_N + 1}{\kappa_N(\nu_N - D + 1)} \mathbf{S}_N, \nu_N - D + 1) \tag{15}$$

In practical implementations, both (14) and (15) are used to calculate the posterior predictive.

For example, in Yee Whye Tey's Matlab code[1], a variation of (14) is used.[2] On the other hand, in Frank Wood's Matlab implementation,[3] equation (15) is used directly.

Another predictive value of interest is the predictive of a new data vector $\mathbf{x}^*$ before any data has been observed, i.e. the posterior predictive of $\mathbf{x}^*$ under the prior alone. This can be obtained as a special case of (12) or (14), yielding

$$\begin{aligned} p(\mathbf{x}^*) &= \int_{\boldsymbol{\mu}} \int_{\boldsymbol{\Sigma}} p(\mathbf{x}^*|\boldsymbol{\mu}, \boldsymbol{\Sigma})\, p(\boldsymbol{\mu}, \boldsymbol{\Sigma})\, \mathrm{d}\boldsymbol{\mu}\, \mathrm{d}\boldsymbol{\Sigma} \\ &= \pi^{-D/2} \frac{\kappa_0^{D/2}}{(\kappa_0+1)^{D/2}} |\mathbf{S}_0|^{-1/2} \prod_{i=1}^{D} \frac{\Gamma\left(\frac{\nu_0+2-i}{2}\right)}{\Gamma\left(\frac{\nu_0+1-i}{2}\right)} \end{aligned}$$

or as a special case of (15) yielding the equivalent result

$$p(\mathbf{x}^*) = \mathcal{T}(\mathbf{x}^*|\mathbf{m}_0, \frac{\kappa_0+1}{\kappa_0(\nu_0-D+1)}\mathbf{S}_0, \nu_0 - D + 1)$$

4 Bayesian finite Gaussian mixture model

This section is primarily based on [1, Section 24.2.4] and [4].

4.1 The model

The Bayesian finite Gaussian mixture model is illustrated in Figure 1. See Section 2 for details on notation. For each observed data vector $\mathbf{x}_i$, we have a latent variable $z_i \in \{1, 2, \ldots, K\}$ indicating which of the K components $\mathbf{x}_i$ belongs to. $\pi_k = P(z_i = k)$ is the prior probability that $\mathbf{x}_i$ belongs to component k. Given $z_i = k$, $\mathbf{x}_i$ is generated by the k^{th} Gaussian mixture component with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$.

We use a Dirichlet distribution for the mixture weights $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)$:

$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \tag{16}$$

where

$$\mathrm{Dir}(\mathbf{x}|\boldsymbol{\alpha}) \triangleq \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1} \tag{17}$$

such that $\sum_{k=1}^{K} x_k = 1$, $x_k \in [0, 1]$ and

$$B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\alpha_0)} \tag{18}$$

[1] http://www.stats.ox.ac.uk/~teh/software.html, partially described in [2].

[2] In Yee Whye Teh's Matlab code, a mysterious function `ZZ()` is defined, which is actually a helper function for computing the log of (13). The helper function is defined as

$$z(D, N, \kappa, \nu, \mathbf{S}) = -\frac{ND}{2}\log\pi - \frac{D}{2}\log\kappa - \frac{\nu}{2}\log|\mathbf{S}| + \sum_{i=1}^{D}\log\Gamma\left(\frac{\nu+i-1}{2}\right)$$

The log posterior predictive can then be calculated as

$$\log p(\mathbf{x}^*|\mathcal{X}) = z(D, N+1, \kappa_N+1, \nu_N+1, \mathbf{S}_{N*}) - z(D, N, \kappa_N, \nu_N, \mathbf{S}_N)$$

[3] http://www.robots.ox.ac.uk/~fwood/Code/index.html, partially described in [4].

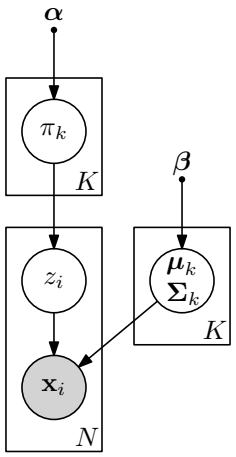

Figure 1: A Bayesian finite GMM. The hyper-parameter $\boldsymbol{\beta} = (\mathbf{m}_0, \kappa_0, \nu_0, \mathbf{S}_0)$.

where $\alpha_0 = \sum_{k=1}^{K} \alpha_k$. We use a Dirichlet distribution for $p(\boldsymbol{\pi}|\boldsymbol{\alpha})$ since the Dirichlet distribution is a conjugate prior for the multinomial distribution and $P(\mathbf{z}|\boldsymbol{\pi})$ in (20) will have the same form as a multinomial, up to an irrelevant constant factor [1, Section 3.4].

For the mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$ of each of the K Gaussian mixture components, we use a GIW distribution with hyper-parameters $\boldsymbol{\beta} = (\mathbf{m}_0, \kappa_0, \nu_0, \mathbf{S}_0)$:

$$p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k|\boldsymbol{\beta}) = \text{GIW}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k|\mathbf{m}_0, \kappa_0, \nu_0, \mathbf{S}_0)$$

We do this since the GIW is fully conjugate to the multivariate Gaussian likelihood (see Sections 3.2 and 3.4) and the $p(\mathcal{X}_k|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ term in (28) will be a Gaussian likelihood.

4.2 Collapsed Gibbs sampling

Since we chose $p(\boldsymbol{\pi}|\boldsymbol{\alpha})$ and $p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k|\boldsymbol{\beta})$ to be conjugate, we are able to analytically integrate out the model parameters $\boldsymbol{\pi}$, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ and only sample the component assignments $\mathbf{z}$. This is done as follows:

$$\begin{aligned} P(z_i = k|\mathbf{z}_{\setminus i}, \mathcal{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &\propto P(z_i = k|\mathbf{z}_{\setminus i}, \boldsymbol{\alpha}, \cancel{\boldsymbol{\beta}})\, p(\mathcal{X}|z_i = k, \mathbf{z}_{\setminus i}, \cancel{\boldsymbol{\alpha}}, \boldsymbol{\beta}) \\ &= P(z_i = k|\mathbf{z}_{\setminus i}, \boldsymbol{\alpha})\, p(\mathbf{x}_i|\mathcal{X}_{\setminus i}, z_i = k, \mathbf{z}_{\setminus i}, \boldsymbol{\beta})\, p(\mathcal{X}_{\setminus i}|\cancel{z_i = k}, \mathbf{z}_{\setminus i}, \boldsymbol{\beta}) \\ &\propto P(z_i = k|\mathbf{z}_{\setminus i}, \boldsymbol{\alpha})\, p(\mathbf{x}_i|\mathcal{X}_{\setminus i}, z_i = k, \mathbf{z}_{\setminus i}, \boldsymbol{\beta}) \end{aligned} \tag{19}$$

Typically the $\boldsymbol{\alpha}$ hyper-parameter is set to $\alpha_k = \alpha/K$. For this setting $\alpha_0 = \sum_{k=1}^{K} \alpha_k = \alpha$. In the following, we will give both the general solution and the solution when using this standard setting for $\boldsymbol{\alpha}$.

In the following two sections we respectively find expressions for the first and second terms on the right hand side of (19).

4.2.1 First term

We find an expression for $P(z_i = k|\mathbf{z}_{\setminus i}, \boldsymbol{\alpha})$ in (19) using

$$P(z_i = k|\mathbf{z}_{\setminus i}, \boldsymbol{\alpha}) = \frac{P(z_i = k, \mathbf{z}_{\setminus i}|\boldsymbol{\alpha})}{P(\mathbf{z}_{\setminus i}|\boldsymbol{\alpha})} = \frac{P(\mathbf{z}|\boldsymbol{\alpha})}{P(\mathbf{z}_{\setminus i}|\boldsymbol{\alpha})}$$

with $z_i = k$ in the numerator. We can calculate both the numerator and denominator above if we can find an expression for the marginal $P(\mathbf{z}|\boldsymbol{\alpha})$.

We do this by marginalizing out $\boldsymbol{\pi}$:

$$P(\mathbf{z}|\boldsymbol{\alpha}) = \int_{\boldsymbol{\pi}} P(\mathbf{z}|\boldsymbol{\pi})\, p(\boldsymbol{\pi}|\boldsymbol{\alpha})\, \mathrm{d}\boldsymbol{\pi}$$

The first term in the integrand is

$$P(\mathbf{z}|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{N_k} \tag{20}$$

where N_k is the count of component k in $\mathbf{z}$. The second term in the integrand is given in (16). We can thus marginalize [1, p. 842]:

$$\begin{aligned}
P(\mathbf{z}|\boldsymbol{\alpha}) &= \int_{\boldsymbol{\pi}} P(\mathbf{z}|\boldsymbol{\pi})\, p(\boldsymbol{\pi}|\boldsymbol{\alpha})\, \mathrm{d}\boldsymbol{\pi} \\
&= \int_{\boldsymbol{\pi}} \prod_{k=1}^{K} \pi_k^{N_k} \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1}\, \mathrm{d}\boldsymbol{\pi} \\
&= \frac{1}{B(\boldsymbol{\alpha})} \int_{\boldsymbol{\pi}} \prod_{k=1}^{K} \pi_k^{N_k + \alpha_k - 1}\, \mathrm{d}\boldsymbol{\pi} && (21) \\
&= \frac{\Gamma(\alpha_0)}{\Gamma(N + \alpha_0)} \prod_{k=1}^{K} \frac{\Gamma(N_k + \alpha_k)}{\Gamma(\alpha_k)} && (22) \\
&= \frac{\Gamma(\alpha)}{\Gamma(N + \alpha)} \prod_{k=1}^{K} \frac{\Gamma(N_k + \alpha/K)}{\Gamma(\alpha/K)} && (23)
\end{aligned}$$

Equation (22) follows from (21) since the integral reduces to the normalizing constant of the Dirichlet distribution proportional to $\prod_{k=1}^{K} \pi_k^{N_k + \alpha_k - 1}$. In (23) we used the standard $\boldsymbol{\alpha}$ setting.

From (22) we can find an expression for the desired term [1, p. 843]:

$$\begin{aligned}
P(z_i = k|\mathbf{z}_{\setminus i}, \boldsymbol{\alpha}) &= \frac{P(z_i = k, \mathbf{z}_{\setminus i}|\boldsymbol{\alpha})}{P(\mathbf{z}_{\setminus i}|\boldsymbol{\alpha})} = \frac{P(\mathbf{z}|\boldsymbol{\alpha})}{P(\mathbf{z}_{\setminus i}|\boldsymbol{\alpha})} && (24) \\
&= \frac{\frac{\Gamma(\alpha_0)}{\Gamma(N+\alpha_0)} \frac{\Gamma(N_k+\alpha_k)}{\Gamma(\alpha_k)} \prod_{j=1, j\neq k}^{K} \frac{\Gamma(N_j+\alpha_j)}{\Gamma(\alpha_j)}}{\frac{\Gamma(\alpha_0)}{\Gamma(N+\alpha_0-1)} \frac{\Gamma(N_{k\setminus i}+\alpha_k)}{\Gamma(\alpha_k)} \prod_{j=1, j\neq k}^{K} \frac{\Gamma(N_j+\alpha_j)}{\Gamma(\alpha_j)}} \\
&= \frac{\Gamma(N + \alpha_0 - 1)}{\Gamma(N + \alpha_0)} \frac{\Gamma(N_k + \alpha_k)}{\Gamma(N_{k\setminus i} + \alpha_k)} \\
&= \frac{N_{k\setminus i} + \alpha_k}{N + \alpha_0 - 1} && (25) \\
&= \frac{N_{k\setminus i} + \alpha/K}{N + \alpha - 1} && (26)
\end{aligned}$$

where we used $\Gamma(x+1) = x\Gamma(x)$ and $N_{k\setminus i} = N_k - 1$. Note that the latter statement is not true in general, but comes from the fact that for the numerator in (24) we have $z_i = k$. In (26) we used the standard $\boldsymbol{\alpha}$ setting.

4.2.2 Second term

To find an expression for $p(\mathbf{x}_i|\mathcal{X}_{\setminus i}, z_i = k, \mathbf{z}_{\setminus i}, \boldsymbol{\beta})$ in (19) we use [1, p. 843]:

$$p(\mathbf{x}_i|\mathcal{X}_{\setminus i}, z_i = k, \mathbf{z}_{\setminus i}, \boldsymbol{\beta}) = p(\mathbf{x}_i|\mathcal{X}_{k\setminus i}, \boldsymbol{\beta})$$

where $\mathcal{X}_{k\backslash i}$ is the set of vectors assigned to component k without taking $\mathbf{x}_i$ into account. Thus the second term in (19) can be written as

$$p(\mathbf{x}_i|\mathcal{X}_{k\backslash i}, \boldsymbol{\beta}) = \frac{p(\mathbf{x}_i, \mathcal{X}_{k\backslash i}|\boldsymbol{\beta})}{p(\mathcal{X}_{k\backslash i}|\boldsymbol{\beta})} = \frac{p(\mathcal{X}_k|\boldsymbol{\beta})}{p(\mathcal{X}_{k\backslash i}|\boldsymbol{\beta})} \tag{27}$$

where $\mathbf{x}_i$ is assumed to be assigned to component k in the numerator. As in the previous section, we can calculate both the numerator and denominator above if we can find an expression for the marginal $p(\mathcal{X}_k|\boldsymbol{\beta})$.

We do this by marginalizing out $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$:

$$\begin{aligned} p(\mathcal{X}_k|\boldsymbol{\beta}) &= \int_{\boldsymbol{\mu}_k} \int_{\boldsymbol{\Sigma}_k} p(\mathcal{X}_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k|\boldsymbol{\beta}) \, \mathrm{d}\boldsymbol{\mu}_k \, \mathrm{d}\boldsymbol{\Sigma}_k \\ &= \int_{\boldsymbol{\mu}_k} \int_{\boldsymbol{\Sigma}_k} p(\mathcal{X}_k|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \, p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k|\boldsymbol{\beta}) \, \mathrm{d}\boldsymbol{\mu}_k \, \mathrm{d}\boldsymbol{\Sigma}_k \end{aligned} \tag{28}$$

Now note that the marginalization in (28) is exactly equivalent to the marginalization performed in (9). In fact, (27) is equivalent to the posterior predictive in (13). The desired expression for (27) is thus given in (14), with appropriate changes to the numerator and denominator sets. Alternatively, (15) can be used directly.

4.2.3 Pseudo code

Pseudo code for the collapsed Gibbs sampler for a finite GMM is given in Algorithm 1.

Algorithm 1 Collapsed Gibbs sampler for a finite Gaussian mixture model.

```
Choose an initial z.
for T iterations do                                   ▷ Gibbs sampling iterations
    for i = 1 to N do
        Remove x_i's statistics from component z_i.          ▷ Old assignment for x_i
        for k = 1 to K do                                    ▷ Every possible component
            Calculate P(z_i = k|z_{\i}, α) using (25).
            Calculate p(x_i|X_{k\i}, β) in (27) using (14) or (15).
            Calculate P(z_i = k|z_{\i}, X, α, β) ∝ P(z_i = k|z_{\i}, α) p(x_i|X_{k\i}, β).
        end for
        Sample k_new from P(z_i|z_{\i}, X, α, β) after normalizing.
        Add x_i's statistics to the component z_i = k_new.   ▷ New assignment for x_i
    end for
end for
```

4.3 Marginal of data and component assignment

In order to evaluate the Gibbs sampling procedure and to ensure that mixing is taking place it is useful to have some metric to calculate over the sampling iterations. The value of the marginal of the data and component assignments $p(\mathcal{X}, \mathbf{z}|\boldsymbol{\alpha})$ are useful in this regard since it captures both changes in the likelihood of the data under the current assignment through $p(\mathcal{X}|\mathbf{z}, \boldsymbol{\beta})$, as well as the probability of the current component assignment $P(\mathbf{z}|\boldsymbol{\alpha})$. The marginal can be calculated

as follows:

$$\begin{aligned} p(\mathcal{X}, \mathbf{z}|\boldsymbol{\alpha}, \boldsymbol{\beta}) &= p(\mathcal{X}|\mathbf{z}, \boldsymbol{\beta})\, P(\mathbf{z}|\boldsymbol{\alpha}) \\ &= \left(\prod_{k=1}^{K} p(\mathcal{X}_k|\boldsymbol{\beta}) \right) P(\mathbf{z}|\boldsymbol{\alpha}) \end{aligned} \tag{29}$$

where $\mathcal{X}_k$ is the set of data vectors assigned to component k. The terms in the product in (29) can each be calculated using (12), while the second term in (29) can be calculated using (22).

5 Infinite Gaussian mixture model

This section is primarily based on [1, Section 25.2] and [4]. The infinite Gaussian mixture model is also sometimes referred to as a Dirichlet process Gaussian mixture model (DP GMM).

5.1 The Chinese restaurant process

Before describing the model, we first give an overview of the Chinese restaurant process (CRP).

The CRP is a simple stochastic process that is exchangeable.In the analogy from which this process takes its name, customers seat themselves at a restaurant with an infinite number of tables. The first customer enters and sits at the first table. The second customer enters and sits at the first table with probability $\frac{1}{1+\alpha}$ and at a new table with probability $\frac{\alpha}{1+\alpha}$. The i^{th} customer sits at an occupied table with probability proportional to the number of customers already seated at that table, or sits at a new table with a probability proportional to α. Formally, if z_i is the table chosen by the i^{th} customer, then

$$P(z_i = k|\mathbf{z}_{\setminus i}, \alpha) = \begin{cases} \frac{N_k}{N+\alpha-1} & \text{if } k \text{ is occupied, i.e. } N_k > 0 \\ \frac{\alpha}{N+\alpha-1} & \text{if } k \text{ is a new table, i.e. } k = k^* = K+1 \end{cases} \tag{30}$$

where $\mathbf{z}_{\setminus i} = (z_1, z_2, \ldots, z_{i-1})$ and N_k is the number of customers already seated at table k.

The probability of a particular sequence of table assignments can be obtained from (30) as follows (see [5] and [6] for details):

$$\begin{aligned} P(\mathbf{z}) &= \prod_{i=1}^{N} P(z_i|\mathbf{z}_{\setminus i}) \\ &= \alpha^K \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \prod_{k=1}^{K} (N_k - 1)! \\ &= \frac{\alpha^K \prod_{k=1}^{K} (N_k - 1)!}{\prod_{i=1}^{N} (i - 1 + \alpha)} \end{aligned} \tag{31}$$

Note that since (31) only depends on the number of tables K and the number of customers seated at each table N_k, the probability of a particular seating arrangement does not depend on the order in which the customers arrived. The random variables z_i in $\mathbf{z}$ is therefore exchangeable.

5.2 The Model

The Bayesian infinite Gaussian mixture model is illustrated in Figure 2. The model is very similar to the finite GMM described in Section 4.1. However, for the infinite model the possible number of mixture components could be infinite where for the finite model the number of mixture components K were known beforehand.

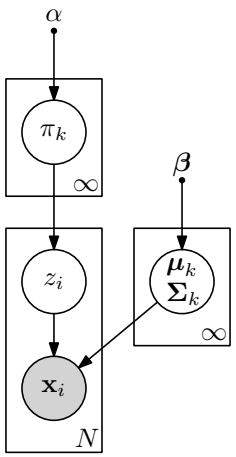

Figure 2: A Bayesian infinite GMM. The hyper-parameter $\boldsymbol{\beta} = (\mathbf{m}_0, \kappa_0, \nu_0, \mathbf{S}_0)$.

We used a Dirichlet distribution as the prior on $\boldsymbol{\pi}$ for the finite model. Here we use a Dirichlet process (DP) prior with concentration parameter α and a GIW base distribution with hyper-parameters $\boldsymbol{\beta}$ for the model parameters. It can be shown that the by choosing the prior in this way, the model is equivalent to a CRP mixture model [5]. For formal discussions of the DP mixture model and how it relates to the CRP, see [1, Section 25.2.2], [5], [6, Section 3.6], [7] and [8].

In the following we thus use the CRP formulation of the DP. We show that the resulting equations also follows from the finite case as the number of components $K \to \infty$. We do this for the collapsed case (after integrating out $\boldsymbol{\pi}$, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$) as is also done (more completely) in [1, Section 25.2.4] and [4]. Consult the relevant references for more complete derivations of the equations given in this section.

5.3 Collapsed Gibbs sampling

As in Section 4.2, we are able to analytically integrate out the parameters $\boldsymbol{\pi}$, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ and sample the component assignment $\mathbf{z}$ directly:

$$P(z_i = k|\mathbf{z}_{\setminus i}, \mathcal{X}, \alpha, \boldsymbol{\beta}) \propto P(z_i = k|\mathbf{z}_{\setminus i}, \alpha)\, p(\mathbf{x}_i|\mathcal{X}_{\setminus i}, z_i = k, \mathbf{z}_{\setminus i}, \boldsymbol{\beta}) \tag{32}$$

In the following two sections we respectively give expressions for the two terms on the right hand side of (32).

5.3.1 First term

The probability $P(z_i = k|\mathbf{z}_{\setminus i}, \alpha)$ in (32) is governed by the CRP. From (30) we can thus write

$$P(z_i = k|\mathbf{z}_{\setminus i}, \alpha) = \begin{cases} \frac{N_{k\setminus i}}{N+\alpha-1} & \text{if } k \text{ is an existing component, i.e. } N_{k\setminus i} > 0 \\ \frac{\alpha}{N+\alpha-1} & \text{if } k \text{ is a new component, i.e. } k = k^* = K+1 \end{cases} \tag{33}$$

where we have assumed by exchangeability that z_i is the last "customer" to arrive at the "restaurant".

The first condition in (33) also follows directly from (26) as $K \to \infty$. The second condition also follows from (26) (although maybe not as clearly); although the probability of a single new component is only $\frac{\alpha/K}{N+\alpha-1}$ (which appears to go to zero), if we lump the probability of all the possible empty components together, we have $\frac{\alpha}{N+\alpha-1}$ as $K \to \infty$.[4] It thus follows that (33) and (26) are equivalent as $K \to \infty$ [9].

From (31) the component assignments of all the data vectors are given by

$$P(\mathbf{z}|\alpha) = \alpha^K \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \prod_{k=1}^{K} (N_k - 1)! = \frac{\alpha^K \prod_{k=1}^{K} (N_k - 1)!}{\prod_{i=1}^{N} (i - 1 + \alpha)} \tag{34}$$

Similar to the discussion above about (33), it can be shown that (34) results in the limit from (23) as $K \to \infty$ [10].

5.3.2 Second term

Exactly as in Section 4.2.2, we can find an expression for $p(\mathbf{x}_i|\mathcal{X}_{\setminus i}, z_i = k, \mathbf{z}_{\setminus i}, \boldsymbol{\beta})$ in (32) by writing it as

$$p(\mathbf{x}_i|\mathcal{X}_{\setminus i}, z_i = k, \mathbf{z}_{\setminus i}, \boldsymbol{\beta}) = p(\mathbf{x}_i|\mathcal{X}_{k\setminus i}, \boldsymbol{\beta}) = \frac{p(\mathcal{X}_k|\boldsymbol{\beta})}{p(\mathcal{X}_{k\setminus i}|\boldsymbol{\beta})} \tag{35}$$

The above can be calculated using (14) or (15) if $z_i = k$ is an existing component. If $z_i = k^*$ is a new component then we have [1, p. 886]:

$$p(\mathbf{x}_i|\mathcal{X}_{\setminus i}, z_i = k^*, \mathbf{z}_{\setminus i}, \boldsymbol{\beta}) = p(\mathbf{x}_i|\boldsymbol{\beta}) = \int_{\boldsymbol{\mu}} \int_{\boldsymbol{\Sigma}} p(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma})\, p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{\beta})\, \mathrm{d}\boldsymbol{\mu}\, \mathrm{d}\boldsymbol{\Sigma} \tag{36}$$

which is just the prior predictive distribution and can be calculated using (14) or (15) with $\mathcal{X} = \emptyset$.

5.4 Pseudo code

Pseudo code for the collapsed Gibbs sampler for an infinite GMM is given in Algorithm 2.

5.5 Marginal of data and component assignment

Again the marginal of the data and component assignments $p(\mathcal{X}, \mathbf{z}|\alpha, \boldsymbol{\beta})$ can be used as evaluation of the Gibbs sampling process:

$$\begin{aligned} p(\mathcal{X}, \mathbf{z}|\alpha, \boldsymbol{\beta}) &= p(\mathcal{X}|\mathbf{z}, \boldsymbol{\beta})\, P(\mathbf{z}|\alpha) \\ &= \left(\prod_{k=1}^{K} p(\mathcal{X}_k|\boldsymbol{\beta}) \right) P(\mathbf{z}|\alpha) \end{aligned} \tag{37}$$

The terms in the product can each be calculated using (12), while the second term can be calculated as in (34).

[4] Yee Whye Teh gives a nice overview of this in his tutorial talk available online at http://videolectures.net/mlss09uk_teh_nbm/ at around 41:27.

Algorithm 2 Collapsed Gibbs sampler for an infinite Gaussian mixture model.

Choose an initial $\mathbf{z}$.
for T iterations **do** ▷ Gibbs sampling iterations
  for $i = 1$ to N **do**
    Remove $\mathbf{x}_i$'s statistics from component z_i. ▷ Old assignment for $\mathbf{x}_i$
    for $k = 1$ to K **do** ▷ Every possible existing component
      Calculate $P(z_i = k|\mathbf{z}_{\setminus i}, \alpha) = \frac{N_{k\setminus i}}{N+\alpha-1}$ as in (34).
      Calculate $p(\mathbf{x}_i|\mathcal{X}_{k\setminus i}, \boldsymbol{\beta})$ in (35) using (14) or (15).
      Calculate $P(z_i = k|\mathbf{z}_{\setminus i}, \mathcal{X}, \alpha, \boldsymbol{\beta}) \propto P(z_i = k|\mathbf{z}_{\setminus i}, \alpha)\, p(\mathbf{x}_i|\mathcal{X}_{k\setminus i}, \boldsymbol{\beta})$.
    end for
    Calculate $P(z_i = k^*|\mathbf{z}_{\setminus i}, \alpha) = \frac{\alpha}{N+\alpha-1}$ as in (34). ▷ Consider a new component
    Calculate $p(\mathbf{x}_i|\boldsymbol{\beta})$ in (36) using (14) or (15).
    Calculate $P(z_i = k^*|\mathbf{z}_{\setminus i}, \mathcal{X}, \alpha, \boldsymbol{\beta}) \propto P(z_i = k^*|\mathbf{z}_{\setminus i}, \alpha)\, p(\mathbf{x}_i|\boldsymbol{\beta})$.
    Sample k_{new} from $P(z_i|\mathbf{z}_{\setminus i}, \mathcal{X}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ after normalizing.
    Add $\mathbf{x}_i$'s statistics to the component $z_i = k_{\text{new}}$. ▷ New assignment for $\mathbf{x}_i$
    If any component is empty, remove it and decrease K.
  end for
end for

6 Notes on initialization

For both the finite and infinite GMMs, an initial clustering $\mathbf{z}$ is required for Gibbs sampling. Several options exist:

- Start with all data vectors assigned to one component. This is the same as the option below with $K = 1$.
- Randomly assign all data vectors to K components. Yee Whye Teh does this in his Matlab implementation.
- For the infinite case, the data vectors can be assigned one component at a time, with clustering assignment then determined by the CRP. Frank Wood does this in his Matlab implementation.
- Use some initial clustering, e.g. that obtained from a run of K-means. One issue in this case might be the choice of the number of components for this preclustering run.

References

[1] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press, 2012.

[2] Y. W. Teh, "Exponential families: Gaussian, Gaussian-gamma, Gaussian-Wishart, multinomial," 2007. [Online]. Available: http://www.stats.ox.ac.uk/~teh/notes.html

[3] T. S. F. Haines, "Gaussian conjugate prior cheat sheet," 2011. [Online]. Available: http://thaines.com/content/blogcategory/18/24/

[4] F. Wood and M. J. Black, "A nonparametric Bayesian alternative to spike sorting," *J. Neurosci. Methods*, vol. 173, no. 1, pp. 1–12, 2012.

[5] S. J. Gershman and D. M. Blei, "A tutorial on Bayesian nonparametric models," *J. Math. Psychol.*, vol. 56, no. 1, pp. 1–12, 2012.

[6] S. J. Goldwater, "Nonparameteric Bayesian models for lexical acquisition," Ph.D. dissertation, Brown University, Providence, RI, 2007.

[7] E. B. Sudderth, "Graphical models for visual object recognition and tracking," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 2006.

[8] B. A. Frigyik, A. Kapila, and M. R. Gupta, "Introduction to the Dirichlet distribution and related processes," University of Washington, Seattle, WA, Tech. Rep., 2010.

[9] C. E. Rasmussen, "The infinite Gaussian mixture model," in *Proc. NIPS*, Denver, CO, 1999.

[10] T. Griffiths and Z. Ghahramani, "Infinite latent feature models and the Indian buffet process," University College London, London, UK, Tech. Rep., 2005.