

# Implementing the Infinite GMM

Michael Mandel

May 6, 2005

## Abstract

Rasmussen [2000] describes a hierarchical Bayesian model for a mixture of Gaussians with a possibly infinite number of components. I have implemented his model for univariate data, along with the Adaptive Rejection Sampling method of Gilks and Wild [1992]. In this paper I explain some of the difficulties in implementing Rasmussen's model and clarify some of the points he leaves vague in his paper. I also explain my own difficulties in implementing the multivariate infinite Gaussian mixture model and propose future work towards modelling audio signals including music.

## 1 Introduction

Dirichlet processes, also known as Chinese Restaurant processes, have recently garnered much attention for their flexibility in modeling mixture processes with an indeterminate, possible infinite, number of components. Rasmussen [2000] proposes their application to the modeling of data generated from a mixture of an infinite number of Gaussians. Of course, a finite amount of data cannot come from an infinite number of sources, but the number of Gaussian components in this model is neither bounded nor set *a priori*. It is instead inferred from the data by a hierarchical Bayesian model.

For this project, I implemented the infinite GMM (IGMM) as described by Rasmussen [2000] for both univariate and multivariate data. In addition, these two systems require the Adaptive Rejection Sampling method of Gilks and Wild [1992], which I also implemented. Due to time constraints and Rasmussen's

incomplete description, I have not completely debugged the multivariate IGMM. As a warm up, I implemented a Gibbs sampler for a predetermined number of Gaussian components and fixed priors on the model parameters.

## 2 Adaptive Rejection Sampling

Gilks and Wild [1992] describe a system for generating samples from an arbitrary log-concave probability distribution function (pdf). They further flesh out their algorithm's description in [Wild and Gilks, 1993], from which I implemented my routine.

The most straightforward way to draw a sample,  $x$ , from a pdf,  $p(x)$ , is to draw a sample from the uniform distribution  $u \sim [0, 1]$  and then to transform it according to the cumulative distribution function of  $x$ ,  $x = F^{-1}(u)$ , where  $F(x) = p(X < x) = \int_{-\infty}^x p(x) dx$ . For pdfs that do not allow analytical integration,  $F(x)$  could be computed numerically at significant computational expense. Such an approach is feasible if the parameters of  $p(x)$  do not change, but the integral must be re-calculated for every setting of the distributions parameters.

A computationally more efficient means of sampling from complex distributions is known as *rejection sampling*. A sample,  $x_0$ , is drawn from a proposal distribution,  $q(x)$ , which shares the support of  $p(x)$  and upper bounds  $p(x)$  for all  $x$ . Another sample is then drawn from the uniform distribution  $u \sim [0, q(x_0)]$  and is kept as a sample from  $p(x)$  if  $u < p(x_0)$ , otherwise it is rejected and the process repeated.

Adaptive rejection sampling (ARS) uses a piecewise exponential approximation to  $p(x)$  as  $q(x)$ .

Since  $p(x)$  is log-concave, the tangents to  $\log p(x)$  upper bound it. By calculating the tangents only at proposed points (after two initial evaluations), ARS reduces the number of times  $p(x)$  must be evaluated. Furthermore,  $q(x)$  approximates  $p(x)$  most accurately around the  $x$ s that are most likely, leading to fewer rejected samples. Even if  $p(x)$  is analytically difficult to integrate,  $\log p(x)$  is generally easy to differentiate, because ARS requires both  $p(x_0)$  and  $\frac{\partial p}{\partial x}|_{x=x_0}$  to describe the tangents.

The original algorithm only included upper-bounding the pdf with the tangents to the log-pdf, but [Wild and Gilks, 1993] includes a lower bound of the secants inside the log-pdf to further reduce the number of evaluations of the original pdf. I did not implement this addition because the IGMM only draws one sample for a particular parameter setting, and the bookkeeping associated with the extra set piecewise functions would surely outweigh any reductions in pdf evaluation.

One challenge in implementing ARS was avoiding problems with numerical precision. Since ARS only requires its evaluation results to be proportional to  $p(x)$ , there is a degree of freedom in the offset of the log-pdf. At one point in the algorithm, points on the log-pdf must be exponentiated and subtracted from one another. If their log values are off by more than 36 ( $\epsilon \approx 10^{-16} \approx e^{-36}$ ) then the addition has no effect and bad things happen. In order to try to avoid this, I calculate an offset for all of the log-pdf values that tries to center them around 0. This generally works, but occasionally will still fail.

### 3 The Infinite GMM

Rasmussen [2000] proposes a hierarchical Bayesian model for sampling from the posterior of a Gaussian mixture model with a possibly infinite number of components given a collection of data. Figure 1 shows a graphical model representation of the infinite Gaussian mixture model. The graphical representation has difficulty showing the dependence of one  $c_i$  on all of the other  $c$  variables and the integration of  $k$  into the number of means and precisions.

The data,  $\{y_i\}$ , are assumed to have come from the

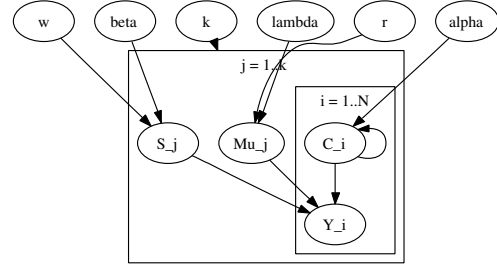


Figure 1: A graphical model representing the infinite gmm.

following generative model.

$$y_i | c_i \sim \mathcal{N}(\mu_{c_i}, s_{c_i}^{-1}), \quad (1)$$

where  $c_i$  is an integer from 1 to  $k$ , the number of mixtures, and  $s_j$  are the *precisions* of the data, or inverse variances. The following priors are put on the parameters to  $y_i$ ,

$$\mu_j \sim \mathcal{N}(\lambda, r) \quad s_j \sim \mathcal{G}(\beta, w^{-1}), \quad (2)$$

where  $\mathcal{G}(\cdot)$  is the gamma distribution having shape parameter  $\beta$  and mean  $w^{-1}$ . These hyper-parameters are controls by a second level of hyper-parameters,

$$\lambda \sim \mathcal{N}(\mu_y, \sigma_y^2) \quad r \sim \mathcal{G}(1, \sigma_y^{-2}) \quad (3)$$

$$\beta^{-1} \sim \mathcal{G}(1, 1) \quad w \sim \mathcal{G}(1, \sigma_y^2) \quad \alpha^{-1} \sim \mathcal{G}(1, 1), \quad (4)$$

where  $\mu_y$  and  $\sigma_y^2$  are the mean and variance of the data itself.

Rasmussen apparently uses a differently parameterized version of the gamma pdf than Matlab does, causing me a number of difficulties. Using his definition of the mean of the gamma pdf, and the pdf of  $\beta$  when  $\beta^{-1} \sim \mathcal{G}(1, 1)$ , I was able to infer the proper transformation between them two. It appears that his definition of the gamma pdf is

$$p(x | \alpha, \theta) = \mathcal{G}_R(\alpha, \theta) = \frac{x^{\alpha/2-1} e^{-\alpha x/2\theta}}{\Gamma(\alpha/2)(2\theta/\alpha)^{\alpha/2}}, \quad (5)$$

having  $E[X] = \theta$ . Matlab's `gamrnd`, on the other hand, uses the pdf,

$$p(x | \alpha, \theta) = \mathcal{G}_M(\alpha, \theta) = \frac{x^{\alpha-1} e^{-x/\theta}}{\Gamma(\alpha)\theta^\alpha}, \quad (6)$$

having  $E[X] = \alpha\theta$ . The parameterizations of these two pdfs differ by the factor of  $\frac{1}{2}$  on  $\alpha$  and in the factor of  $\frac{1}{\alpha}$  on  $\theta$ .

All of these priors are conjugate, except for those on  $\alpha$  and  $\beta$ , which need to be sampled using ARS, as described in Section 2. The log-posterior of  $\beta$  is

$$\begin{aligned} \log p(\beta | s_1, \dots, s_k) &= C - k \log \Gamma\left(\frac{\beta}{2}\right) - \frac{1}{2\beta} \\ &+ \frac{k\beta - 3}{2} \log \frac{\beta}{2} + \sum_{j=1}^k \frac{\beta}{2} (\log s_j + \log w) - \frac{\beta s_j w}{2} \end{aligned} \quad (7)$$

and the log-posterior of  $\alpha$  is

$$\begin{aligned} \log p(\alpha | k, n) &= C + (k - 3/2) \log \alpha \\ &- \frac{1}{2\alpha} \log \Gamma(\alpha) - \log \Gamma(n + \alpha). \end{aligned} \quad (8)$$

The derivatives of both of these with respect to  $\beta$  and  $\alpha$  are easily computable, noting the presence in Matlab of the function `psi` for computing the digamma function,  $\psi(x) \equiv \frac{\partial}{\partial x} \log \Gamma(x)$ .

The exact order in which the posteriors should be sampled during Gibbs sampling is unclear from Rasmussen [2000]. Sampling from most of the posteriors can happen in any order, but gets tricky with the addition and subtraction of new Gaussian components. In order to speed sampling and convergence to the true posterior, one would like to sample all of the  $c_i$  parameters at once. The sampling of the  $c_i$ s could involve adding or removing a Gaussian, which should in turn affect all subsequent samples. Any time a component is not added or removed, however, it does not affect the other samples, and no changes need to be made.

One possible solution to this problem is to sample only one  $c_i$  each iteration, but this wastes computation and time as the values for other parameters will be highly correlated in adjacent samples. The opposite approach would be to loop over the  $c_i$ s, adding and removing Gaussians as necessary, before sampling  $c_{i+1}$ . This method is not particularly well-suited to a Matlab implementation, as that loop would be very slow.

I chose instead a middle path, in which I drew as many  $c_i$ s as possible up to the first addition of

a Gaussian. The uniformity of these calculations makes them easily vectorizable. Once a Gaussian was added, I resampled all of the other parameters and then started from the  $c_i$  where I had left off. This procedure strikes a balance between efficiency of implementation in Matlab and speed of convergence.

## 4 Multivariate Case

On the whole, adapting the models of Section 3 to multivariate data is relatively straightforward. The normal variables  $y_i$ ,  $\mu_j$ , and  $\lambda$  become multinormal random vectors. The gamma variables  $s_j$ ,  $r$ , and  $w$  become Wishart random matrices. And the variables dealing with mixtures,  $\alpha$ ,  $c_i$ , and  $k$  remain the same. Certain care must be taken in the order of matrix multiplications, for example in the posterior mean of a multinormal, and the posterior mean of a Wishart variable involves a sum of outer products instead of just a sum of squares, but there aren't many concerns beyond these except for the exact parameterization of the Wishart distribution used.

As unclear as Rasmussen [2000] was on the gamma pdf, when generalizing to multivariate data, he gives even less information about his parameterization of the Wishart distribution. The Wishart distribution is the conjugate prior to the precision matrix of the multinormal distribution. Rasmussen claims that the gamma priors on  $s_j$ ,  $r$ , and  $w$  may be replaced by Wishart priors without any further changes.

The pdf of the Wishart written out explicitly was quite hard to find. Mardia et al. [1979] mention it almost in passing and do not describe its mean or any properties that would be useful for matching Matlab's implementation to Rasmussen's description. Matlab's documentation was also sorely lacking, as it didn't even describe its own parameterization, although one can experiment with the function (`wishrnd`) to estimate its mean. I did manage to find a slightly more complete description of the Wishart pdf in Box and Tiao [1973], which includes its mean and describes its conjugacy to the multinormal and its posterior density given observations.

For conjugate priors, an incorrect parameterization meant drawing from a slightly inaccurate pdf. The

biggest problem, however, was in defining the posterior on  $\beta$ , which remains a scalar. In  $d$  dimensions,  $(\beta + d - 1)^{-1} \sim \mathcal{G}(1, 1)$ , so if we define a new variable  $y \equiv \beta - d + 1$ , then  $p(y) \propto y^{-3/2} \exp\left(\frac{1}{2y}\right)$ . The posterior density over  $y$  is then,

$$p(y \mid S_1, \dots, S_k, W) \propto (y + d - 1)^{-3/2} |W|^{(y+d-1)k/2} \times \exp\left(\frac{d}{2(y + d - 1)}\right) \left(\frac{y + d - 1}{2}\right)^{(y+d-1)kd/2} \times \prod_{j=1}^k \frac{|S_j|^{y/2-1} \exp\left(-\frac{(y+d-1)Tr(W S_j)}{2}\right)}{\prod_{i=0}^{d-1} \Gamma\left(\frac{y+i}{2}\right)}. \quad (9)$$

$\beta$  can be recovered by sampling  $y$  from the above posterior and then taking  $\beta = y + d - 1$ .

As mentioned above, I wasn't able to fully debug the multivariate IGMM, particularly the Wishart distribution. The current bugs include  $\beta$  being too small, `wishrnd` drawing matrices that aren't positive definite, and numerical precision problems in ARS.

The function  $\mathcal{W}(\beta, W)$  requires that  $\beta \geq d$ .  $\beta$ 's posterior, however, keeps it in the range  $\beta \geq d - 1$ , which often generates errors from `wishrnd`. The inverse Wishart distribution, however, does not place such a restriction on its shape parameter, so perhaps either I could switch to using that as the prior on  $\Sigma_j = S_j^{-1}$  and the other Wishart variables, or I could further manipulate  $\beta$  to remain in the valid region.

I'm not at all sure why `wishrnd` would generate matrices that are not positive definite. The functions from Matlab's statistics toolbox and from David Shera's MCMC Matlab toolbox both have the same problem. It could be that the covariance argument to the function is itself not positive definite, but the functions would complain if that were the case. If  $W$  is not positive definite, the  $|W|$  term in Equation (9) will explode when taking the log of the posterior, leading to failed ARS runs.

## 5 Results

In order to test the system, I gave it some simple data from a GMM with 2 components. 500 data points

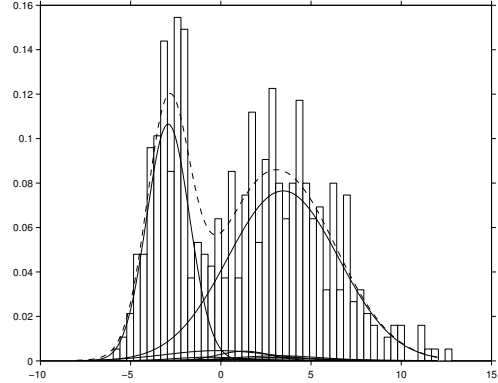


Figure 2: One sample drawn from the posterior of the IGMM

were drawn from the following distribution

$$y_i \sim \frac{1}{3}\mathcal{N}(-3, 1) + \frac{2}{3}\mathcal{N}(3, 10), \quad (10)$$

where  $\mathcal{N}(\mu, \sigma^2)$  describes a single Gaussian.

Since Gibbs sampling only draws one parameter at a time, samples close to one another in time will tend to be correlated. In order to measure this correlation, which depends on the dataset at hand among other things, I measured the autocovariance of each of the model parameters as a function of lag between samples. See Figure 3 for plots of autocovariance versus lag for all of the model hyper-parameters. It seems from this plot that samples separated by 700-800 other samples are independent of one another.

Also see Figure 2 for an example of a sample drawn from the IGMM given this dataset plotted over the histogram of the data points. For this particular sample, there are 10 represented components, although the two correct components dominate the others. Also for this sample  $\alpha = 1.79$ , implying that the unrepresented components make up only  $\frac{\alpha}{n+\alpha} = .4\%$  of the probability mass.

## 6 Discussion and Future Work

This project shows a working version of the Infinite Gaussian Mixture Model for univariate data and

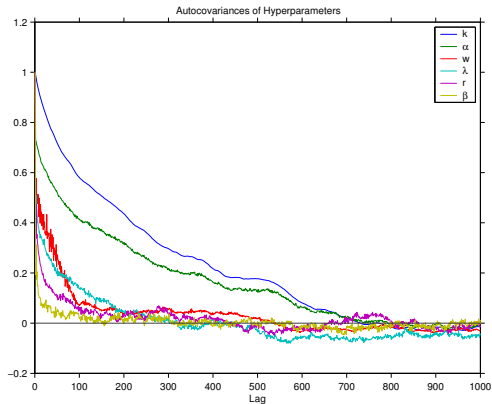


Figure 3: The autocovariance of model hyperparameters as a function of lag between samples.

an implementation of Adaptive Rejection Sampling. The IGMM performs well in estimating the original parameters of a two-Gaussian mixture without any externally supplied prior information.

Once I get the bugs worked out of the multivariate IGMM, I plan to apply it to the case of clustering audio data. Using mel-frequency cepstral coefficients (MFCCs) to represent audio, I would like to address the questions of how many Gaussians best describe a song, an artist, a style of music, all music, assuming that the MFCC frames are IID draws from a GMM. Since the IGMM has no prior assumptions and does not suffer from over-fitting, it would not require cross-validation to determine such numbers.

Another question I would like to answer is whether or not there are real clusters of MFCC frames, or whether they are spread out on a continuum. This question could be asked about any dataset. The IGMM should help answer it by either consistently picking the same parameters for certain Gaussians, if there are clusters, or by spreading the distribution of parameters around, in the case of a continuum.

## References

George E. P. Box and George C. Tiao. *Bayesian Inference in Statistical Analysis*. Addison–Wesley, Reading, Massachusetts, 1st edition, 1973.

W. R. Gilks and P. Wild. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41(2):337–348, 1992. ISSN 0035-9254.

K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, London, 1979.

Carl E. Rasmussen. The infinite gaussian mixture model. In S.A. et al. Solla, editor, *Advances in information processing systems 12*, pages 554–560. MIT Press, 2000.

P. Wild and W. R. Gilks. Algorithm AS 287: Adaptive rejection sampling from log-concave density functions. *Applied Statistics*, 42(4):701–709, 1993.