ASSIGNMENT 3: MONTE CARLO METHODS

In this assignment you will implement both Markov chain Monte Carlo (MCMC) and sequential Monte Carlo (SMC) methods for a Bayesian Gaussian mixture model.

Background reading: Bishop chapter 11; Murphy sections 23.4–23.6

GETTING STARTED

- Download the skeleton code for the assignment (hw_3.tar.gz)
- Unzip the downloaded material in an appropriate folder, something like
  `~/Documents/AIMS_CDT_ML/HW3/`
- Open MATLAB and navigate to the folder containing the downloaded material

BAYESIAN MIXTURE OF GAUSSIANS

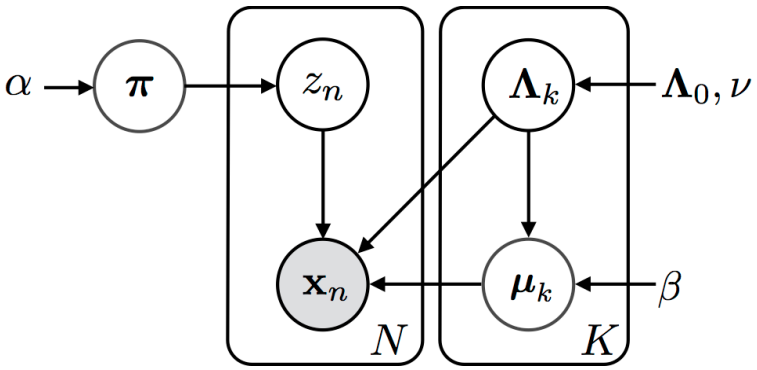
In the last assignment we used EM to estimate parameters π_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ in a classical Gaussian mixture model. Now we consider instead the fully Bayesian case, in which we place *prior distributions* on what were previously our parameters

$$\begin{aligned} \boldsymbol{\pi} &\sim \text{Dirichlet}(\alpha) \\ \boldsymbol{\Lambda}_k &\sim \text{Wishart}(\boldsymbol{\Lambda}_0, \nu) \\ \boldsymbol{\mu}_k \,\big|\, \boldsymbol{\Lambda}_k &\sim \text{Normal}\big(\mathbf{0}, (\beta\boldsymbol{\Lambda}_k)^{-1}\big) \end{aligned}$$

and treat them as latent variables. The likelihood of the data is the same as in the previous homework, with

$$\begin{aligned} z_n \,\big|\, \boldsymbol{\pi} &\sim \text{Categorical}(\boldsymbol{\pi}) \\ \mathbf{x}_n \,\big|\, z_n = k,\, \boldsymbol{\mu}_k,\, \boldsymbol{\Lambda}_k &\sim \text{Normal}\big(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\big)\,. \end{aligned}$$

Here is the graphical model for the Bayesian mixture of Gaussians:



Here we have parameratized the multivariate normal distribution with a *precision matrix*, or inverse covariance matrix. The priors in this model are specifically chosen so as to be conjugate to the likelihood (http://en.wikipedia.org/wiki/Conjugate_prior).

We are going to perform inference using both MCMC and SMC methods. In both cases, we can take advantage of conjugacy to reduce the dimensionaliy of the space on which we need to sample.

PART 1: GIBBS SAMPLING

Gibbs sampling is a variety of MCMC sampling in which we cycle through all our latent random variables, resampling each conditioned on the currently sampled values of all other random variables. See Bishop chapter 11.3 for details, and Figure 11.11 for a useful illustration.

One could write a Gibbs sampler for the Gaussian mixture model which alternates between sampling each of the cluster weight prior $\boldsymbol{\pi}$, the cluster means $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$, the cluster precisions $\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_K$, and the per-data-point cluster assignments $z_1, \ldots, z_N$. Here we will instead take advantage of the conjugacy in the model to integrate out most of the random variables analytically, and construct a collapsed Gibbs sampler (http://en.wikipedia.org/wiki/Gibbs_sampling#Collapsed_Gibbs_sampler) which only needs sample the assignments $z_1, \ldots, z_N$.

1. Write down and understand the full conditional distribution necessary for implementing a collapsed Gibbs sampler on this model. That is, integrate out (analytically) the random variables $\boldsymbol{\pi}, \boldsymbol{\mu}_k$, and $\boldsymbol{\Lambda}_k$ to find the functional form of $p(z_n | \mathbf{z}_{-n}, \mathbf{x}, \alpha, \beta, \nu)$ where $\mathbf{z}_{-n}$ is a vector of length $N-1$ containing all assignments *except* z_n.

   To do this, we take advantage of Dirichlet-Multinomial conjugacy (http://en.wikipedia.org/wiki/Dirichlet-multinomial_distribution) and Multivariate Normal – Wishart conjugacy (http://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf) (see section 8).

2. Implement the collapsed Gibbs sampler in MATLAB. The full conditional distribution and the collapsed joint density over $z_{1:N}, \mathbf{x}_{1:N}$ are implemented for you already in `log_predictive_collapsed.m` and `log_joint_collapsed.m`. You only need to implement the Gibbs sampling inner loop in `run_gibbs_sweep.m`. Test your sampler on the Iris dataset from Assignment 2 using `main_gibbs_sampler.m`.

   Hint: for the collapsed sampler, you will need to keep track of three sets of statistics related to the distributions we have marginalized over analytically. These include cluster assignment counts N_k, per-cluster means $\bar{\mathbf{x}}_k$, and per-cluster scatter matrix $\mathbf{S}_k$, where

$$N_k = \sum_{n=1}^{N} \mathbf{I}[z_n = k]$$

$$\bar{\mathbf{x}}_k = \frac{1}{N_k} \sum_{\{\, n \,:\, z_n = k \,\}} \mathbf{x}_n$$

$$\mathbf{S}_k = \sum_{\{\, n \,:\, z_n = k \,\}} (\mathbf{x}_n - \bar{\mathbf{x}}_k)(\mathbf{x}_n - \bar{\mathbf{x}}_k)^T.$$

3. One parameter to the model is the dimensionality K of the Dirichlet $\boldsymbol{\pi}$; the values of z_n range over $1, \ldots, K$. What is K set to for your sampler? How many different cluster labels do you see once the sampler has converged? What was K set to in Assignment 2, when we ran EM?

   What you see here is known as a *Bayesian Occam's Razor*; for more information on this (and on model comparison more broadly) see MacKay chapter 28.

4. The Dirichlet prior $\boldsymbol{\pi}$, the Gaussian prior $\boldsymbol{\mu}_k$, and the Wishart prior $\boldsymbol{\Lambda}_k$ all have parameters of their own. The parameters of our prior distributions (that is, of our "parameters") are known as *hyperparameters*. What happens if you change the value of α from 1 to 1,000,000? How would you suggest choosing a value for this hyperparameter?

PART 2: SEQUENTIAL MONTE CARLO

Sequential Monte Carlo (SMC) methods are an alternate approach to approximating posterior distributions and computing posterior expectations. For a brief but thorough review, see Murphy sections 23.4–23.6.

1. Implement a Rao-Blackwellized particle filter for this the Gaussian mixture model by completing the function `run_smc_sweep.m`. As with the Gibbs sampler, we take advantage of conjugacy to reduce the dimensionality of the space we need importance sample at every data point, and only sample $z_1, \ldots, z_N$.

   You have complete freedom in selecting a proposal distribution $q(z_n | z_{1:n-1}, \mathbf{x}_{1:n})$ but be prepared to defend your choice, and be sure to calculate your importance weights correctly. Again, you will probably find the supplied functions `log_predictive_collapsed.m` and `log_joint_collapsed.m` helpful.

   Test your sampler on the Iris dataset by running `main_smc.m`.

2. Particle MCMC methods (http://www.stats.ox.ac.uk/~doucet/andrieu_doucet_holenstein_PMCMC.pdf) use sequential Monte Carlo sweeps as proposal distributions within a MCMC sampler. Extend your SMC sampler to a *Particle independent Metropolis-Hastings* (PIMH) sampler by adding an outer loop to `main_smc.m` (i.e. by implementing equation 11 in the PMCMC paper). This should take fewer than five additional lines of code.

3. In all our examples so far, we have chosen some fixed K to be the dimension of $\boldsymbol{\pi}$, which effectively enforces a maximum possible number of clusters. What happens if we want to allow unboundedly many clusters, i.e. take $K \to \infty$? This yields a *Bayesian nonparametric model* known as the Dirichlet Process mixture of Gaussians. Extending our samplers to sample from a mixture of potentially infinitely many Gaussians requires only minor changes to the code; see e.g. this nice tutorial (http://snippyhollow.github.io/blog/2013/03/10/collapsed-gibbs-sampling-for-dirichlet-process-gaussian-mixture-models/).