Probabilistic Graphical Models 10-708
Homework 4

Daniel Ribeiro Silva
Andrew ID: drsilva

April 16, 2014

Problem 1

Part 1: Stationary Distribution

In this proof, we assume (without loss of generality) that the Gibbs sampling process will follow the order $x_1, x_2, \ldots, x_n$ where $\mathbf{X} = (x_1, \ldots, x_n)$. That means,

$$\begin{aligned}
&x_1' \sim p(x_1'|x_2, \ldots, x_n)\\
&x_2' \sim p(x_2'|x_1', x_3, \ldots, x_n)\\
&\ldots\\
&x_n' \sim p(x_n'|x_1', \ldots, x_{n-1}')
\end{aligned}$$

We also assume the continuous variables, but the proof is the same for discrete variables (all it takes is to replace the integrals by sums over the variables).

The transition kernel is given by:

$$\mathcal{T}(\mathbf{x} \to \mathbf{x}') = p(x_1'|x_2, \ldots, x_n)p(x_2'|x_1', x_3, \ldots, x_n) \ldots p(x_n'|x_2', \ldots, x_n')$$

Now let's compute the transition of a given generic state distribution for our particular transition kernel and prove that it is stationary. we wish to compute

$$\int p(\mathbf{x}|\mathbf{e})\mathcal{T}(\mathbf{x} \to \mathbf{x}')d\mathbf{x}$$

Using the assumptions above, we have

$$\int p(\mathbf{x}|\mathbf{e})\mathcal{T}(\mathbf{x} \to \mathbf{x}'|\mathbf{e})d\mathbf{x} = \int p(x_1, \ldots, x_n|\mathbf{e})p(x_1'|x_2, \ldots, x_n, \mathbf{e}) \ldots p(x_n'|x_2', \ldots, x_n', \mathbf{e})dx_1 \ldots dx_n$$

Now observe how only the first term depends on x_1, and only the first two terms depend on x_2. It is easy to notice that only the first k terms depend on x_k. This observation allows us to rearrange the order of integration, and brings us to a sequential method for reducing this expression. The method for 1 iteration is shown below:

$$\begin{aligned}\int p(\mathbf{x}|\mathbf{e})\mathcal{T}(\mathbf{x}\to\mathbf{x}'|\mathbf{e})d\mathbf{x} &= \int p(x_1,\dots,x_n|\mathbf{e})p(x_1'|x_2,\dots,x_n,\mathbf{e})\dots p(x_n'|x_2',\dots,x_n',\mathbf{e})dx_1\dots dx_n\\ &= \int p(x_1,\dots,x_n|\mathbf{e})dx_1 p(x_1'|x_2,\dots,x_n,\mathbf{e})\dots p(x_n'|x_2',\dots,x_n',\mathbf{e})dx_2\dots dx_n\\ &= \int p(x_2,\dots,x_n|\mathbf{e})p(x_1'|x_2,\dots,x_n,\mathbf{e})\dots p(x_n'|x_2',\dots,x_n',\mathbf{e})dx_2\dots dx_n\\ &= \int p(x_1',x_2,\dots,x_n|\mathbf{e})\dots p(x_n'|x_2',\dots,x_n',\mathbf{e})dx_2\dots dx_n\end{aligned}$$

Observe how first we use marginalization of the target integrating variable

$$\int p(x_1,\dots,x_n|\mathbf{e})dx_1 = p(x_2,\dots,x_n|\mathbf{e})$$

and then we use the definition of conditional probability to merge two terms

$$p(x_2,\dots,x_n|\mathbf{e})p(x_1'|x_2,\dots,x_n,\mathbf{e}) = p(x_1',x_2,\dots,x_n|\mathbf{e})$$

Observe how by performing those two operations we are back to the initial form. We can perform this pair of operations for all n variables, as shown below

$$\begin{aligned}\int p(\mathbf{x}|\mathbf{e})\mathcal{T}(\mathbf{x}\to\mathbf{x}'|\mathbf{e})d\mathbf{x} &= \int p(x_1,\dots,x_n|\mathbf{e})p(x_1'|x_2,\dots,x_n,\mathbf{e})\dots p(x_n'|x_2',\dots,x_n',\mathbf{e})dx_1\dots dx_n\\ &= \int p(x_1',x_2,\dots,x_n|\mathbf{e})\dots p(x_n'|x_2',\dots,x_n',\mathbf{e})dx_2\dots dx_n\\ &= \int p(x_1',x_2',x_3\dots,x_n|\mathbf{e})\dots p(x_n'|x_2',\dots,x_n',\mathbf{e})dx_3\dots dx_n\\ &= \dots\\ &= \int p(x_1',\dots,x_{n-1}',x_n|\mathbf{e})p(x_n'|x_2',\dots,x_n',\mathbf{e})dx_n\\ &= \int p(x_1',\dots,x_{n-1}',x_n|\mathbf{e})dx_n p(x_n'|x_2',\dots,x_n',\mathbf{e})dx_n\\ &= p(x_1',\dots,x_{n-1}'|\mathbf{e})p(x_n'|x_2',\dots,x_n',\mathbf{e})dx_n\\ &= p(x_1',\dots,x_n'|\mathbf{e})\\ &= p(\mathbf{x}'|\mathbf{e})\end{aligned}$$

Thus, the posterior distribution is an invariant distribution for the given transition kernel.

Part 2: Special Case of Metropolis-Hastings

Let's denote $\mathbf{x_{-i}}$ as being the set of all variables except x_i.

In Gibbs sampling, we sample the variable x_i' from $P(x_i'|\mathbf{x_{-i}})$, and the samples are never rejected.

Now consider the following proposal distribution:

$$Q(x_i',\mathbf{x_{-i}}|x_i,\mathbf{x_{-i}}) = P(x_i'|\mathbf{x_{-i}})$$

If we use the Q proposal sample from above, we will always sample from the same distribution of the Gibbs Sampling algorithm. All we need to show now is that for this proposal distribution, the Metropolis-Hasting algorithm will always accept the samples.

We know that for the MH algorithm, the portability of accepting a sample is given by:

$$A(x_i', \mathbf{x_{-i}}|x_i, \mathbf{x_{-i}}) = \min\left(1, \frac{P(x_i', \mathbf{x_{-i}})Q(x_i, \mathbf{x_{-i}}|x_i, \mathbf{x_{-i}})}{P(x_i, \mathbf{x_{-i}})Q(x_i', \mathbf{x_{-i}}|x_i', \mathbf{x_{-i}})}\right)$$

In our case, we have

$$\begin{aligned}
A(x_i', \mathbf{x_{-i}}|x_i, \mathbf{x_{-i}}) &= \min\left(1, \frac{P(x_i', \mathbf{x_{-i}})Q(x_i, \mathbf{x_{-i}}|x_i, \mathbf{x_{-i}})}{P(x_i, \mathbf{x_{-i}})Q(x_i', \mathbf{x_{-i}}|x_i', \mathbf{x_{-i}})}\right) \\
&= \min\left(1, \frac{P(x_i', \mathbf{x_{-i}})P(x_i|\mathbf{x_{-i}})}{P(x_i, \mathbf{x_{-i}})P(x_i'|\mathbf{x_{-i}})}\right) \\
&= \min\left(1, \frac{P(x_i'|\mathbf{x_{-i}})P(\mathbf{x_{-i}})P(x_i|\mathbf{x_{-i}})}{P(x_i|\mathbf{x_{-i}})P(\mathbf{x_{-i}})P(x_i'|\mathbf{x_{-i}})}\right) \\
&= \min(1, 1) \\
&= 1
\end{aligned}$$

Thus, the sample is always accepted, as required by the Gibbs Sampling algorithm.

We conclude that the Gibbs Sampling algorithm is a special case of the Metropolis-Hasting algorithm.

Part 3: Variants of Gibbs Sampling

Let $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ be the set of random variables (nodes) in a PGM. The regular Gibbs Sampling algorithm will sample $\mathbf{X^{(t)}}$ based on $\mathbf{X^{(t-1)}}$ by sampling one node at a time:

$$\begin{aligned}
&x_1^{(t)} \sim p(x_1^{(t)}|x_2^{(t-1)}, \ldots, x_n^{(t-1)}) \\
&x_2^{(t)} \sim p(x_2^{(t)}|x_1^{(t)}, x_3^{(t-1)}, \ldots, x_n^{(t-1)}) \\
&\cdots \\
&x_n^{(t)} \sim p(x_n^{(t)}|x_1^{(t)}, \ldots, x_{n-1}^{(t)})
\end{aligned}$$

The Block Gibbs Sampling variant of the Gibbs Sampling algorithm will divide the set of nodes into blocks (subsets of nodes) and will sample all variables from the same block/subset at the same time (i.e. from the joint probability). Let B_i be a block of random variables (a subset of all the nodes $\mathbf{X}$ in the PGM) and k be the total number of blocks, such that $\bigcup_{i=1}^k B_i = \mathbf{X}$ and $B_i \cap B_j = \emptyset, \forall i \neq j$. The Block Gibbs Sampling is such that:

$$\begin{aligned}
&B_1^{(t)} \sim p(B_1^{(t)}|B_2^{(t-1)}, \ldots, B_k^{(t-1)}) \\
&B_2^{(t)} \sim p(B_2^{(t)}|B_1^{(t)}, B_3^{(t-1)}, \ldots, B_k^{(t-1)}) \\
&\cdots \\
&B_k^{(t)} \sim p(B_k^{(t)}|B_1^{(t)}, \ldots, B_{k-1}^{(t)})
\end{aligned}$$

Finally, the Collapsed Gibbs Sampling will eliminate one or more variables of the PGM (by marginalizing over them and thus yielding a collapsed PGM), before sampling for a given variable. That will improve accuracy of the samples since only a sub-space is sampled[1].

For example, if we choose to integrate out the variable x_2, then $x_1^{(t)}$ will be sampled as

$$x_1^{(t)} \sim p(x_1^{(t)}|x_3^{(t-1)}, \dots, x_n^{(t-1)})$$

Problem 2

1.

$$P(\mathbf{z}) = \int P(\mathbf{z}|\theta)P(\theta)d\theta$$

But we know that $\mathbf{z_{ji}}|\theta_\mathbf{i} \sim \mathit{Discrete}(\theta_\mathbf{i})$ and $\theta_\mathbf{i} \sim \mathit{Dirichlet}(\alpha)$. Thus, the product $P(\mathbf{z}|\theta)P(\theta)$ also follows a Dirichlet distribution, since the Dirichlet distribution is the conjugate prior of the categorical distribution.

$$P(\mathbf{z}) = \int \theta \frac{1}{B(\alpha)} \prod_{k=1}^{K} x_k^{\alpha_k - 1} d\theta$$

We can prove [3] that the value of this integral is:

$$P(\mathbf{z}) = \left(\frac{\Gamma(K\alpha)}{\Gamma(\beta)^W}\right)^D \prod_{d=1}^{D} \frac{\prod_j \Gamma(N_{ki} + \alpha)}{\Gamma(N_i + K\alpha)}$$

where D is the total number of documents, K is the number of topics, and W the number of unique words.

In a similar way we have

$$P(\mathbf{d}|\mathbf{z}) = \int P(\mathbf{d}|\mathbf{z}, \phi)P(\phi)d\phi$$

Once again, we know that $\mathbf{d_{ji}}|\mathbf{z_{ji}}, \phi_{\mathbf{z_{ji}}} \sim \mathit{Discrete}(\phi_\mathbf{i})$ and $\phi_\mathbf{k} \sim \mathit{Dirichlet}(\beta)$. Thus, the product $P(\mathbf{d}|\mathbf{z}, \phi)P(\phi)$ also follows a Dirichlet distribution, since the Dirichlet distribution is the conjugate prior of the categorical distribution.

$$P(\mathbf{z}) = \int \phi \frac{1}{B(\beta)} \prod_{k=1}^{K} x_k^{\beta_k - 1} d\phi$$

We can prove [3] that the value of this integral is:

$$P(\mathbf{d}|\mathbf{z}) = \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^K \prod_{k=1}^{K} \frac{\prod_w \Gamma(N_{kw} + \beta)}{\Gamma(N_k + W\beta)}$$

2. The posterior is given by

$$P(\mathbf{z}|\mathbf{d}) = \frac{P(\mathbf{d},\mathbf{z})}{\sum_{\mathbf{z}} P(\mathbf{d},\mathbf{z})} = \frac{P(\mathbf{d}|\mathbf{z})P(\mathbf{z})}{\sum_{\mathbf{z}} P(\mathbf{d}|\mathbf{z})P(\mathbf{z})}$$

The main problem here is that the denominator does not factorize and consists of the sum of $K^{\sum_i N_i}$, where N_i is the number of word instances in the document i. This space is too large to be computed. Without the computation of this partition function, we cannot convert the numerator value into a probability. One of the solutions for this problem is to perform sampling using MCMC.

3. We'll start by computing the general expression for $p(z_{ji}|z^{(-ji)}, \mathbf{d}, \boldsymbol{\alpha}, \boldsymbol{\beta})$, instead of the case for a particular value $p(z_{ji} = k|z^{(-ji)}, \mathbf{d}, \boldsymbol{\alpha}, \boldsymbol{\beta})$.

Notation:
$z^{(-ji)} = \mathbf{z} \backslash z_{ji}$
N_i: total words in document i
$B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)}$

$$\begin{aligned}
p(z_{ji}|z^{(-ji)}, \mathbf{d}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \frac{p(z_{ji}, z^{(-ji)}, \mathbf{d}|\boldsymbol{\alpha}, \boldsymbol{\beta})}{p(z^{(-ji)}, \mathbf{d}|\boldsymbol{\alpha}, \boldsymbol{\beta})} && (1)\\
&\propto p(z_{ji}, z^{(-ji)}, \mathbf{d}|\boldsymbol{\alpha}, \boldsymbol{\beta}) && (2)\\
&= p(\mathbf{z}, \mathbf{d}|\boldsymbol{\alpha}, \boldsymbol{\beta}) && (3)\\
&= \int\int p(\mathbf{z}, \mathbf{d}, \theta, \phi|\boldsymbol{\alpha}, \boldsymbol{\beta}) d\phi d\theta && (4)\\
&= \int\int p(\phi|\boldsymbol{\beta}) p(\theta|\boldsymbol{\alpha}) p(\mathbf{z}|\theta) p(\mathbf{d}|\mathbf{z}, \phi) d\phi d\theta && (5)\\
&= \int p(\mathbf{z}|\theta) p(\theta|\boldsymbol{\alpha}) d\theta \int p(\phi|\boldsymbol{\beta}) p(\mathbf{d}|\mathbf{z}, \phi) d\phi && (6)\\
&= \int \prod_{m=1}^{M} p(z_m|\theta_m) p(\theta_m|\boldsymbol{\alpha}) d\theta \int \prod_{k=1}^{K} p(\phi_k|\boldsymbol{\beta}) \prod_{m=1}^{M} \prod_{n=1}^{N_i} p(d_{mn}|\phi_{z_{mn}}, z_{mn}) d\phi && (7)
\end{aligned}$$

Step-by-step explanation:
(1): Conditional probability definition
(2): Denominator independent of z_{ji}
(3): $(z_{ji}, z^{(-ji)}) = \mathbf{z}$
(4): add and marginalize over θ, ϕ
(5): factor according to PGM
(6): separate by integrating vriable
(7): independence of each documents and each word

$$= \prod_{m=1}^{M} \int p(z_m|\theta_m) p(\theta_m|\boldsymbol{\alpha}) d\theta_m \prod_{k=1}^{K} \int p(\phi_k|\boldsymbol{\beta}) \prod_{m=1}^{M} \prod_{n=1}^{N_i} p(d_{mn}|\phi_{z_{mn}}, z_{mn}) d\phi_k \tag{8}$$

$$= \prod_{m=1}^{M} \int \prod_{n=1}^{N_m} \theta_{m,z_{mn}} \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_{mk}^{\alpha_k - 1} d\theta_m \prod_{k=1}^{K} \int \frac{1}{B(\beta)} \prod_{j=1}^{J} \phi_{kj}^{\beta_j - 1} \prod_{m=1}^{M} \prod_{n=1}^{N_i} \phi_{z_{mn},d_{mn}} d\phi_k \tag{9}$$

$$= \prod_{m=1}^{M} \int \prod_{n=1}^{K} \theta_{m,z_{mn}}^{N_{ki}} \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_{mk}^{\alpha_k - 1} d\theta_m \prod_{k=1}^{K} \int \frac{1}{B(\beta)} \prod_{j=1}^{J} \phi_{kj}^{\beta_j - 1} \prod_{j=1}^{J} \phi_{kj}^{N_{wk}} d\phi_k \tag{10}$$

$$= \prod_{m=1}^{M} \int \frac{1}{B(\alpha)} \prod_{n=1}^{K} \theta_{m,z_{mn}}^{N_{ki} + \alpha_k - 1} d\theta_m \prod_{k=1}^{K} \int \frac{1}{B(\beta)} \prod_{j=1}^{J} \phi_{kj}^{N_{wk} + \beta_j - 1} d\phi_k \tag{11}$$

$$= \prod_{m=1}^{M} \frac{B(\alpha + N_{.i})}{B(\alpha)} \int \frac{1}{B(\alpha + N_{.i})} \prod_{n=1}^{K} \theta_{m,z_{mn}}^{N_{ki} + \alpha_k - 1} d\theta_m \prod_{k=1}^{K} \frac{B(\beta + N_{.k})}{B(\beta)} \int \frac{1}{B(\beta + N_{.k})} \prod_{j=1}^{J} \phi_{kj}^{N_{wk} + \beta_j - 1} d\phi_k \tag{12}$$

$$= \prod_{m=1}^{M} \frac{B(\alpha + N_{.i})}{B(\alpha)} \prod_{k=1}^{K} \frac{B(\beta + N_{.k})}{B(\beta)} \tag{13}$$

$$= \prod_{m=1}^{M} \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\prod_{k=1}^{K} \Gamma(N_{ki} + \alpha_k)}{\Gamma\left(\sum_{k=1}^{K} N_{ki} + \alpha_k\right)} \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{j=1}^{J} \beta_j\right)}{\prod_{j=1}^{J} \Gamma(\beta_j)} \frac{\prod_{j=1}^{J} \Gamma(N_{wk} + \beta_j)}{\Gamma\left(\sum_{j=1}^{J} N_{wk} + \beta_j\right)} \tag{14}$$

$$\propto \prod_{m=1}^{M} \frac{\prod_{k=1}^{K} \Gamma(N_{ki} + \alpha_k)}{\Gamma\left(\sum_{k=1}^{K} N_{ki} + \alpha_k\right)} \prod_{k=1}^{K} \frac{\prod_{j=1}^{J} \Gamma(N_{wk} + \beta_j)}{\Gamma\left(\sum_{j=1}^{J} N_{wk} + \beta_j\right)} \tag{15}$$

Step-by-step explanation:
(8): distribute multivariate integral over dimensions
(9): distribution probabilities according to PGM
(10): change index of product and use counts
(11): merge products with same indexes
(12): multiply by $\frac{B(N_{.i}+\alpha)}{B(N_{.i}+\alpha)} \frac{B(N_{.k}+\beta)}{B(N_{.k}+\beta)}$ where $B(N_{.x} + \alpha) = \frac{\prod_{i=1}^{K} \Gamma(N_{ix}+\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i + N_{ix}\right)}$
(13): integral over dirichlet distribution equals 1
(14): expand terms
(15): drop terms that depend only on α or β

Now let's analyse that expression for a specific case z_{ab} (notice that we change the indexes of the original expression z_{ji} so we avoid confusions).

$$= \prod_{m \neq a} \frac{\prod_{k=1}^{K} \Gamma(N_{ki} + \alpha_k)}{\Gamma\left(\sum_{k=1}^{K} N_{ki} + \alpha_k\right)} \frac{\prod_{k=1}^{K} \Gamma(N_{ka} + \alpha_k)}{\Gamma\left(\sum_{k=1}^{K} N_{ka} + \alpha_k\right)} \prod_{k=1}^{K} \frac{\prod_{j \neq y_{ab}}^{J} \Gamma(N_{wk} + \beta_j) \Gamma(N_{y_{ab}k} + \beta_{y_{ab}})}{\Gamma\left(\sum_{j=1}^{J} N_{jk} + \beta_j\right)} \tag{16}$$

$$\propto \frac{\prod_{k=1}^{K} \Gamma(N_{ka} + \alpha_k)}{\Gamma\left(\sum_{k=1}^{K} N_{ka} + \alpha_k\right)} \prod_{k=1}^{K} \frac{\Gamma(N_{y_{ab}k} + \beta_{y_{ab}})}{\Gamma\left(\sum_{j=1}^{J} N_{jk} + \beta_j\right)} \tag{17}$$

$$\propto \frac{\prod_{k \neq z_{ab}} \Gamma(N_{ka}^{-(ab)} + \alpha_k) \Gamma(N_{z_{ab}a}^{-(ab)} + \alpha_{z_{ab}} + 1)}{\Gamma\left(1 + \sum_{k=1}^{K} N_{ka}^{-(ab)} + \alpha_k\right)} \prod_{k \neq z_{ab}} \frac{\Gamma(N_{y_{ab}k}^{-(ab)} + \beta_{y_{ab}})}{\Gamma\left(\sum_{j=1}^{J} N_{jk} + \beta_j\right)} \frac{\Gamma(N_{y_{ab}z_{ab}}^{-(ab)} + \beta_{y_{ab}} + 1)}{\Gamma\left(1 + \sum_{j=1}^{J} N_{jz_{ab}}^{-(ab)} + \beta_j\right)} \tag{18}$$

Step-by-step explanation:
(16): separate terms containing y_{ab} or z_{ab}
(17): get rid of what does not depend on a or b
(18): counts for (a, b)

If we use the property $\Gamma(x+1) = x\Gamma(x)$, we get:

$$\frac{\prod_{k \neq z_{ab}} \Gamma(N_{ka}^{-(ab)} + \alpha_k)\Gamma(N_{z_{ab}a}^{-(ab)} + \alpha_{z_{ab}})(N_{z_{ab}a}^{-(ab)} + \alpha_{z_{ab}})}{\Gamma\left(1 + \sum_{k=1}^{K} N_{ka}^{-(ab)} + \alpha_k\right)} \times$$
$$\prod_{k \neq z_{ab}} \frac{\Gamma(N_{y_{ab}k}^{-(ab)} + \beta_{y_{ab}})}{\Gamma\left(\sum_{j=1}^{J} N_{wk} + \beta_j\right)} \frac{\Gamma(N_{y_{ab}z_{ab}}^{-(ab)} + \beta_{y_{ab}})(N_{y_{ab}z_{ab}}^{-(ab)} + \beta_{y_{ab}})}{\Gamma\left(\sum_{j=1}^{J} N_{jz_{ab}}^{-(ab)} + \beta_j\right)\left(\sum_{j=1}^{J} N_{jz_{ab}}^{-(ab)} + \beta_j\right)}$$

Now that we have the (a, b) terms, we regroup the other terms:

$$\frac{\prod_{k=1}^{K} \Gamma(N_{ka}^{-(ab)} + \alpha_k)(N_{z_{ab}a}^{-(ab)} + \alpha_{z_{ab}})}{\Gamma\left(1 + \sum_{k=1}^{K} N_{ka}^{-(ab)} + \alpha_k\right)} \times$$
$$\prod_{k=1}^{K} \frac{\Gamma(N_{y_{ab}k}^{-(ab)} + \beta_{y_{ab}})}{\Gamma\left(\sum_{j=1}^{J} N_{wk} + \beta_j\right)} \frac{(N_{y_{ab}z_{ab}}^{-(ab)} + \beta_{y_{ab}})}{\sum_{j=1}^{J}\left(N_{jz_{ab}}^{-(ab)} + \beta_j\right)}$$

Since those regrouped terms and the topic denominator are constant, they can be dropped, giving us

$$\propto \frac{(N_{y_{ab}z_{ab}}^{-(ab)} + \beta_{y_{ab}})(N_{z_{ab}a}^{-(ab)} + \alpha_{z_{ab}})}{\sum_{j=1}^{J}\left(N_{jz_{ab}}^{-(ab)} + \beta_j\right)} = \frac{(N_{y_{ab}z_{ab}}^{-(ab)} + \beta_{y_{ab}})(N_{z_{ab}a}^{-(ab)} + \alpha_{z_{ab}})}{N_{z_{ab}}^{-(ab)} + \sum_{j=1}^{J}(\beta_j)}$$

Now, if we consider our original example $z_{ji} = k$, we get:

$$p(z_{ji} = k | z^{(-ji)}, \mathbf{d}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \frac{(N_{ki}^{-(ij)} + \alpha_k)(N_{wk}^{-(ij)} + \beta_w)}{N_{z_{ab}}^{-(ij)} + \sum_w (\beta_w)}$$

4.
$$\theta_{ik} = \frac{N_{ki} + \alpha_k}{N_i + \sum_k \alpha_k}$$

$$\phi_{kw} = \frac{N_{wk} + \beta_w}{N_k + \sum_w \beta_w}$$

5. Note that we can normalise the probability distribution found previously. We obtain a closed form for the probability distribution of $p(z_{ji} = k | z^{(-ji)}, \mathbf{d}, \boldsymbol{\alpha}, \boldsymbol{\beta})$:

$$p(z_{ji} = k | z^{(-ji)}, \mathbf{d}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{Z} \frac{(N_{ki}^{-(ij)} + \alpha_k)(N_{wk}^{-(ij)} + \beta_w)}{N_k^{-(ij)} + \sum_w (\beta_w)}$$

with

$$Z = \sum_{k=1}^{K} \frac{(N_{ki}^{-(ij)} + \alpha_k)(N_{wk}^{-(ij)} + \beta_w)}{N_k^{-(ij)} + \sum_w (\beta_w)}$$

It is a simple discrete probability, and is thus easy to sample from. The pseudo-code for LDA collapsed Gibbs Sampling is:

Algorithm 1 LDA collapsed Gibbs Sampling

```
Given parameters: $\alpha, \beta$
$K$ = total number of topics
$I$ = vocabulary number of documents
$W$ = vocabulary size
initialize array $N_k[]$ of size $K$
initialize array $N_{wk}[][]$ of size $W \times K$
initialize array $N_{ik}[]$ of size $I \times K$
while not enough samples do
    for each document $i$ do
        for each word position $j$ in document do
            $w_{current} = d_{ij}$
            $k_{past}$ = past $k$ sample for $(i, j)$
            $N_k[k_{past}] - -$
            $N_{wk}[w_{current}][k_{past}] - -$
            $N_{ik}[i][k_{past}] - -$
            $Z = \sum_{k=1}^{K} \frac{(N_{ik}[k]+\alpha_k)(N_{wk}[w_{current}][k]+\beta_w)}{N_k[k]+\sum_w(\beta_w)}$
            $p(z_{ji} = k|z^{(-ji)}, \mathbf{d}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{(N_{ik}[k]+\alpha_k)(N_{wk}[w_{current}][k]+\beta_w)}{N_k[k]+\sum_w(\beta_w)}$
            sample $k_{sample}$ from $p(z_{ji}|z^{(-ji)}, \mathbf{d}, \boldsymbol{\alpha}, \boldsymbol{\beta})$
            $N_k[k_{sample}] + +$
            $N_{wk}[w_{current}][k_{sample}] + +$
            $N_{ik}[i][k_{sample}] + +$
        end for
    end for
end while
```

Problem 3

Part 1: An Improper Prior

Let's first observe that the integral of the prior is not finite. The reason is that this prior is constant (and strictly positive) over a domain of non-finite measure.

$$\int p(\mu_1, \mu_2) d\mu_1 d\mu_2 = \int p(\mu_1, \mu_2) d\mu_1 = \int p(\mu_1, \mu_2) d\mu_2 = +\infty$$

Now let's observe the integral of the posterior probability. Since the prior is constant across the entire domain, we have:

$$\begin{aligned}\int p(\mu_1, \mu_2|\mathbf{x}) d\mu_1 d\mu_2 &\propto \int\int p(\mathbf{x}|\mu_1, \mu_2) d\mu_1 d\mu_2 \\ &= \int\int \prod_{i=1}^{N} \left(\frac{1}{2\sqrt{2\pi}} e^{\frac{-(\mu_1 - x_i)}{2}} + \frac{1}{2\sqrt{2\pi}} e^{\frac{-(\mu_2 - x_i)}{2}} \right) d\mu_1 d\mu_2\end{aligned}$$

Observe that we can decompose the expression above into 2^N terms. Moreover, one of the terms will only contain functions of μ_1 (and no μ_2). More precisely, that term is

$$A(\mu_1) = K e^{\sum_{i=1}^{N} \frac{-(\mu_1 - x_i)}{2}}$$

with $K = \left(\frac{1}{2\sqrt{2\pi}}\right)^N$.

In a similar way, we'll also have a term that is a function of μ_2 only:

$$B(\mu_2) = K e^{\sum_{i=1}^{N} \frac{-(\mu_2 - x_i)}{2}}$$

Observe that, since all terms correspond to probabilities (and are thus non-negative for any parameter (μ_1, μ_2) or any point sample $\mathbf{x}$) the integral of any term is non-negative. Thus,

$$\int p(\mu_1, \mu_2|\mathbf{x}) d\mu_1 d\mu_2 \leq \int\int \left(A(\mu_1) + B(\mu_2)\right) d\mu_1 d\mu_2$$

But it is well known (and easy to demonstrate) that

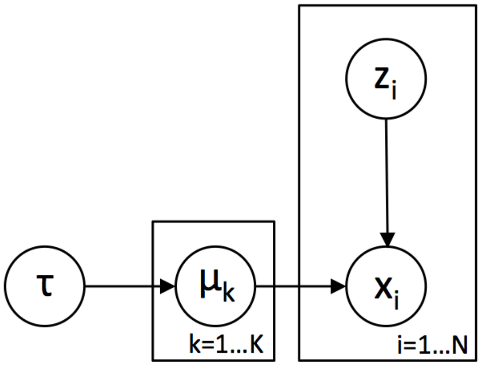
$$\int A(\mu_1) d\mu_1 = \int B(\mu_2) d\mu_2 = K\sqrt{2\pi}$$

Thus,

$$\begin{aligned}\int p(\mu_1, \mu_2|\mathbf{x}) d\mu_1 d\mu_2 &\geq \int\int \left(A(\mu_1) + B(\mu_2)\right) d\mu_1 d\mu_2 \\ &= \int\int A(\mu_1) d\mu_1 d\mu_2 + \int\int B(\mu_2) d\mu_2 d\mu_1 \\ &= \int K\sqrt{2\pi} d\mu_2 + \int K\sqrt{2\pi} d\mu_1 \\ &= K\sqrt{2\pi} \left(\int d\mu_2 + \int d\mu_1 \right) \\ &= +\infty\end{aligned}$$

Part 2: Deriving Gibbs Sampling

In order to perform the Gibbs Sampling for (μ_1, μ_2) we will introduce a new variable z_i into the model. The new variable z_i represents the Gaussian component from which the sample point x_i was originally sampled. If we have K components, then $z_i \in \{1, 2, \ldots, K\}$. In our particular case, $K = 2$ so $z_i \in \{1, 2\}$. Our new PGM can be represented by



We'll start by sampling for z_i. Let's find an expression for this sampling:

$$\begin{aligned} p(z_i|z_{-i},\mathbf{x},\mu) &= p(z_i|\mathbf{x},\mu) \\ &= p(z_i|x_i,\mu) \\ &\propto p(x_i|z_i,\mu)p(z_i|\mu) \end{aligned}$$

Thus,

$$p(z_i=k|z_{-i},\mathbf{x},\mu) \propto p(x_i|\mu_k) \times p(z_i=k)$$

In our case, since $K=2$, we have

$$\begin{aligned} p(z_i=1|z_{-i},\mathbf{x},\mu) \propto p(x_i|\mu_1)p(z_i=1) &= \frac{1}{\sqrt{2\pi}}e^{\frac{-(x_i-\mu_1)^2}{2}} \times \frac{1}{2} \propto e^{-(x_i-\mu_1)^2} \\ p(z_i=2|z_{-i},\mathbf{x},\mu) \propto p(x_i|\mu_2)p(z_i=2) &= \frac{1}{\sqrt{2\pi}}e^{\frac{-(x_i-\mu_2)^2}{2}} \times \frac{1}{2} \propto e^{-(x_i-\mu_2)^2} \end{aligned}$$

Thus,

$$\begin{aligned} p(z_i=1|z_{-i},\mathbf{x},\mu) &= \frac{1}{K_i}e^{-(x_i-\mu_1)^2} \\ p(z_i=2|z_{-i},\mathbf{x},\mu) &= \frac{1}{K_i}e^{-(x_i-\mu_2)^2} \end{aligned}$$

where $K_i = e^{-(x_i-\mu_1)^2} + e^{-(x_i-\mu_1)^2}$

It's a simple Bernoulli distribution with extremely low computation requirements. Thus, we can easily sample points for z_i.

Now let's find the expression for sampling (μ_1,μ_2), given the samples for z_i.

Let's start with the analysis of $P(\mu_1|\mu_2,\mathbf{x},\mathbf{z})$.

$$\begin{aligned} P(\mu_1|\mu_2,\mathbf{x},\mathbf{z}) &= P(\mu_1|x_{i:z_i=1},\mathbf{z}) \\ &\propto P(x_{i:z_i=1}|\mu_1,\mathbf{z})P(\mu_1|\mathbf{z}) \\ &= \prod_{i:z_i=1} P(x_i|\mu_1,\mathbf{z})P(\mu_1|\mathbf{z}) \\ &= \prod_{i:z_i=1} \mathcal{N}(x_i|\mu_1,1)\mathcal{N}(\mu_1|0,\tau) \end{aligned}$$

The actual probability takes into account the partition function as the normalisation term:

$$P(\mu_1|\mu_2, \mathbf{x}, \mathbf{z}) = \frac{\prod_{i:z_i=1} \mathcal{N}(x_i|\mu_1, 1)\mathcal{N}(\mu_1|0, \tau)}{\int \prod_{i:z_i=1} \mathcal{N}(x_i|\mu_1, 1)\mathcal{N}(\mu_1|0, \tau) d\mu_1}$$

Let's call $C = \int \prod_{i:z_i=1} \mathcal{N}(x_i|\mu_1, 1)\mathcal{N}(\mu_1|0, \tau) d\mu_1$, then

$$P(\mu_1|\mu_2, \mathbf{x}, \mathbf{z}) = \frac{1}{C} \prod_{i:z_i=1} \mathcal{N}(x_i|\mu_1, 1)\mathcal{N}(\mu_1|0, \tau)$$

is distributed as a Gaussian distribution, since Gaussians are the conjugate priors of Gaussian distributions. Thus, this Gaussian distribution is given by [4]:

$$P(\mu_1|\mu_2, \mathbf{x}, \mathbf{z}) \sim \mathcal{N}\left(\frac{\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i:z_i=1} x_i}{\sigma^2}}{\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)}, \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1} \right) = \mathcal{N}\left(\frac{\tau \sum_{i:z_i=1} x_i}{N_1\tau + 1}, \frac{\tau}{N_1\tau + 1} \right)$$

Where N_1 is the total number of z_i such that $z_i = 1$.

In a similar way,

$$P(\mu_2|\mu_1, \mathbf{x}, \mathbf{z}) \sim \mathcal{N}\left(\frac{\tau \sum_{i:z_i=2} x_i}{N_2\tau + 1}, \frac{\tau}{N_2\tau + 1} \right)$$

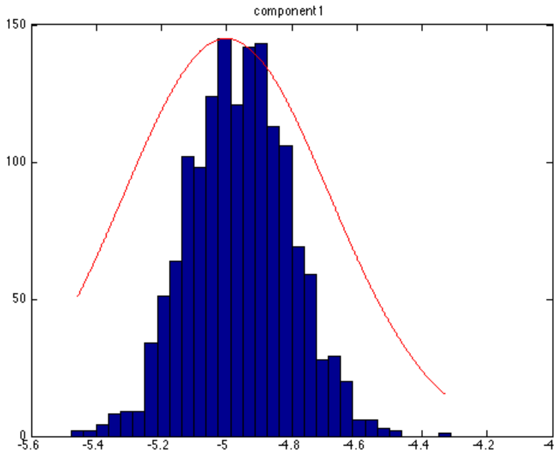
Thus, our Gibbs sampling algorithm is given by

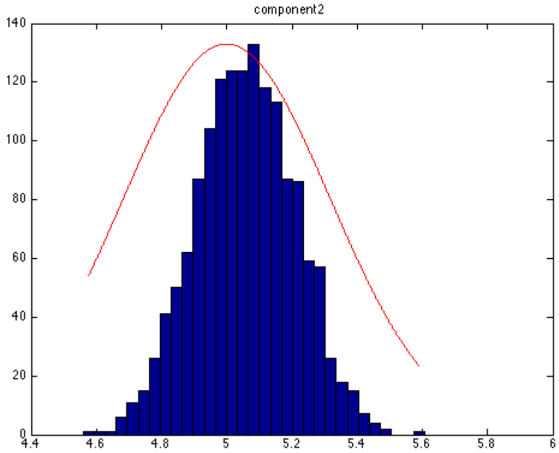
Algorithm 2 Gibbs Sampling for Two Component Mixture of Gaussians

```
Given parameters: $\tau$
Given data points: $\mathbf{x^N} = \{x_1, \ldots, x_N\}$
Initialize $\mu_1^{(0)}, \mu_2^{(0)}$ with some values
$t = 1$
while not enough samples do
    for each datapoint $x_i$ do
        Sample $z_i^{(t)} \sim p(z_i^{(t)} = k|z_{-i}^{(t-1)}, \mathbf{x}^{(t-1)}, \mu^{(t-1)}) = \frac{1}{K_i} e^{-(x_i - \mu_k^{(t-1)})^2}$
    end for
    for each component $k$ do
        $N_k^{(t)} = \sum_{i:z_i^{(t)}=k} 1$
    end for
    Sample $\mu_1^{(t)} \sim \mathcal{N}\left(\frac{\tau \sum_{i:z_i^{(t)}=1} x_i}{N_1^{(t)}\tau+1}, \frac{\tau}{N_1^{(t)}\tau+1}\right)$
    Sample $\mu_2^{(t)} \sim \mathcal{N}\left(\frac{\tau \sum_{i:z_i^{(t)}=2} x_i}{N_2^{(t)}\tau+1}, \frac{\tau}{N_2^{(t)}\tau+1}\right)$
    $t++$
end while
```

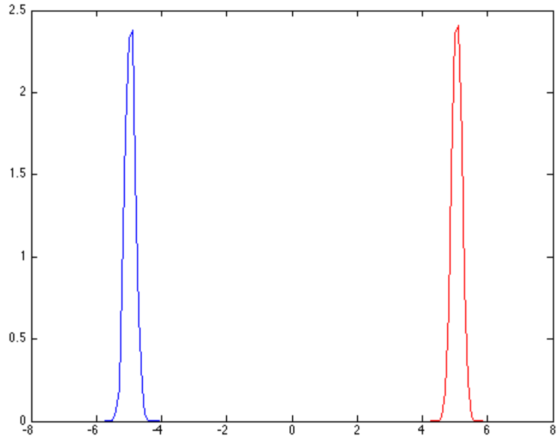
Part 3: Running Gibbs Sampling

The code for this part is attached. The distribution estimate for μ_1 and μ_2 are shown below. The red line represents the true distribution (with adjusted scale to be visible).





If we use a kernel smoothing estimate for the distribution, we get



Part 4: Running Metropolis-Hastings

For the Metropolis-Hastings method, we'll use the original PGM (without the insertion of the nodes z_i). The posterior probability is given by:

$$P(\mu_1, \mu_2|\mathbf{x}) = \frac{1}{Z}P(\mathbf{x}|\mu_1, \mu_2)P(\mu_1, \mu_2) = \frac{1}{Z}\prod_{i=1}^{N}\left[\frac{1}{2}\mathcal{N}(x_i|\mu_1, 1) + \frac{1}{2}\mathcal{N}(x_i|\mu_2, 1)\right]\mathcal{N}(\mu_1|0, \tau)\mathcal{N}(\mu_2|0, \tau)$$

with

$$Z = \int\int\prod_{i=1}^{N}\left[\frac{1}{2}\mathcal{N}(x_i|\mu_1, 1) + \frac{1}{2}\mathcal{N}(x_i|\mu_2, 1)\right]\mathcal{N}(\mu_1|0, \tau)\mathcal{N}(\mu_2|0, \tau)d\mu_1 d\mu_2$$

But this normalization term doesn't really matter for the Metropolis-Hastings algorithm, since we only consider the ratio of probabilities for the acceptance probability of a sample:

$$A(x, x') = min\left(1, \frac{p(x')q(x', x)}{p(x)q(x, x')}\right)$$

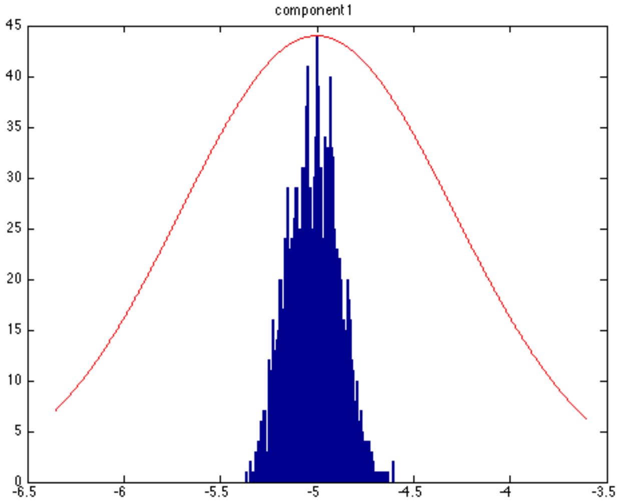
We'll choose as the proposal distribution

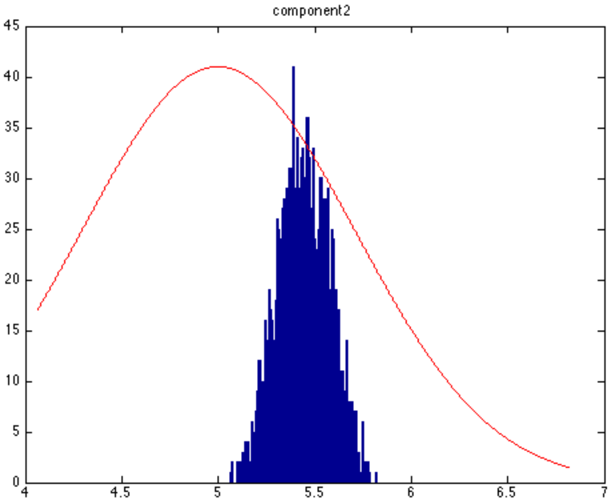
$$q(\mu^{(\mathbf{t})}) \sim \mathcal{N}(\mu^{(\mathbf{t-1})}, 10I_2)$$

Observe that this distribution is symmetrical and thus the acceptance probability can be reduced to

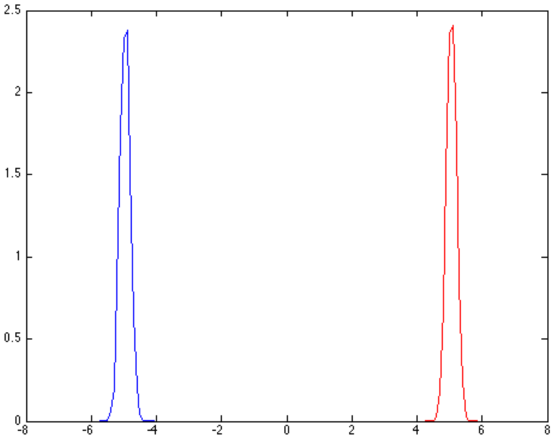
$$A(x, x') = min\left(1, \frac{p(x')}{p(x)}\right)$$

The code for this part is attached. The distribution estimate for μ_1 and μ_2 are shown below. The red line represents the true distribution (with adjusted scale to be visible).





If we use a kernel smoothing estimate for the distribution, we get



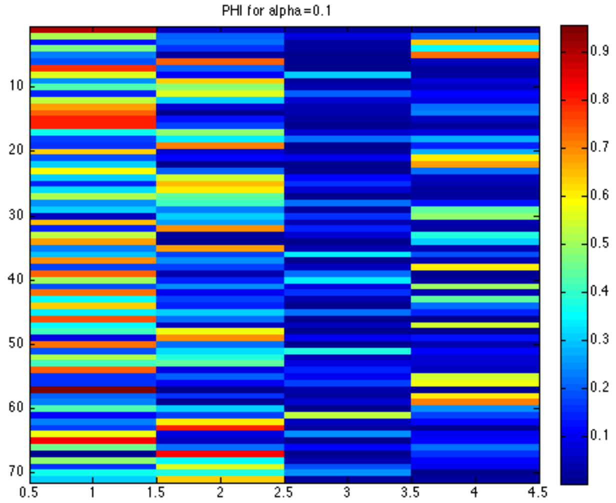
Problem 4

1. The update equations [5, 6] are given by:

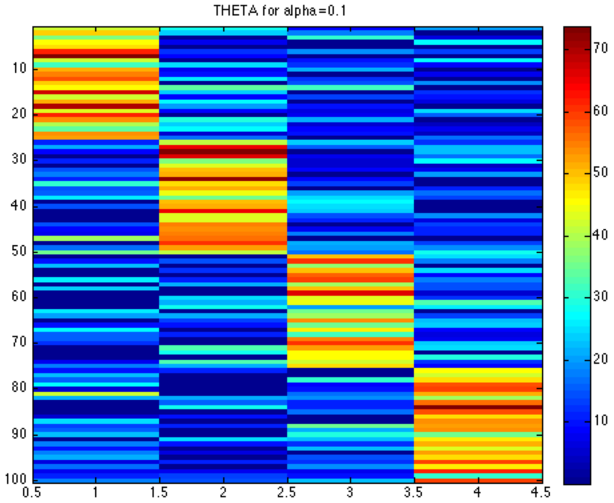
$$\begin{cases} \phi_{ink}^{(t)} = \beta_{in} e^{\Psi(\gamma_{ik}^{(t-1)})} \\ \gamma_{ik}^{(t)} = \alpha_{ik} + \sum_{n=1}^{N_i} D_{in} \phi_{ink}^{(t)} \end{cases}$$

   where i is the index of the documents (individuals), n is the index over the existing word types (genotype loci) for that document, and k is the index over the topics (genotype ancestry).

2. The code is attached. Parameters: $\epsilon = 0.001, \alpha = 0.1$



3. The code is attached. Parameters: $\epsilon = 0.001, \alpha = 0.1$

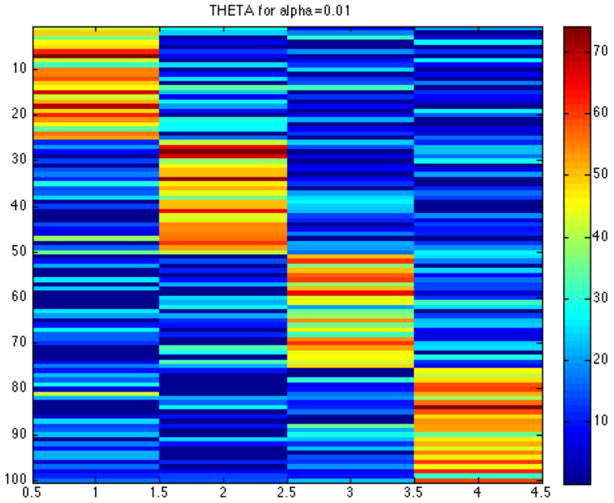


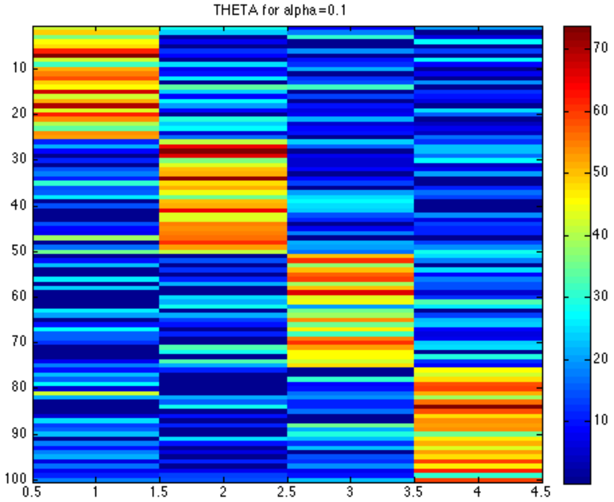
4. The results for each value of α are shown below.

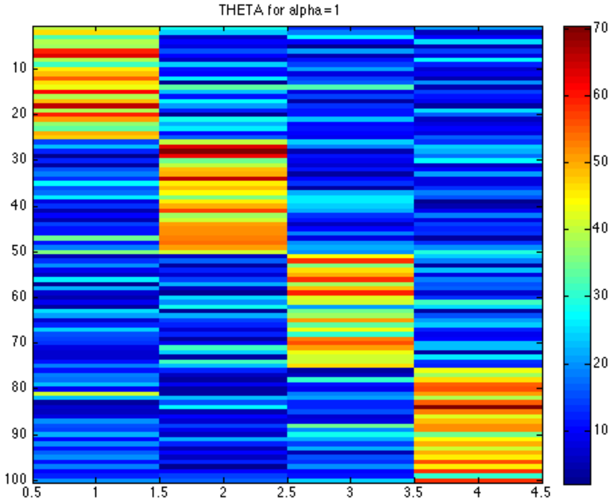
α	0.01	0.1	1	10
iterations	6044	5935	3434	1516
run time (s)	9.89	9.44	5.77	2.56

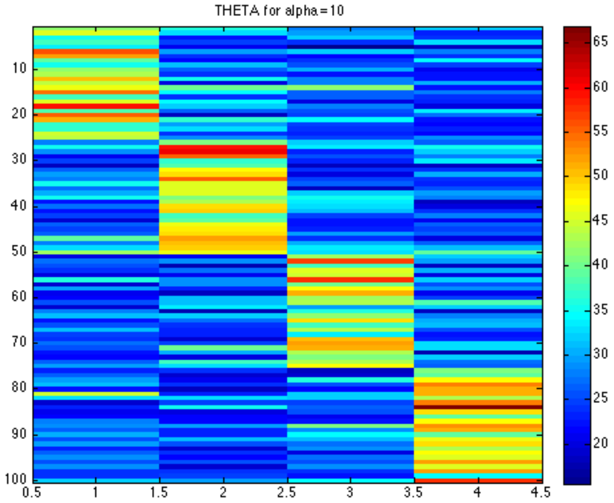
5. Results are shown on the table above.

6. The results for each value of α are shown below. By analyzing the contrast (or difference of the color level) we observe that for smaller values of α the likelihood of each individual belonging to a specific ancestry is more accentuated. Basically, it means that we have a higher confidence in predicting the ancestry of a given individual. That comes with the price of many more iterations to go through (and, as a consequence, a larger run time).


THETA for alpha=0.01
10
20
30
40
50
60
70
80
90
100
0.5
1
1.5
2
2.5
3
3.5
4
4.5
70
60
50
40
30
20
10


THETA for alpha=0.1
10
20
30
40
50
60
70
80
90
100
0.5
1
1.5
2
2.5
3
3.5
4
4.5
70
60
50
40
30
20
10


THETA for alpha=1
10
20
30
40
50
60
70
80
90
100
0.5
1
1.5
2
2.5
3
3.5
4
4.5
70
60
50
40
30
20
10



References

[1] Venugopal, Deepak, and Vibhav Gogate. ”Dynamic blocking and collapsing for gibbs sampling.” arXiv preprint arXiv:1309.6870 (2013).

[2] Carpenter, Bob. Integrating out multinomial parameters in latent Dirichlet allocation and naive bayes for collapsed Gibbs sampling. Technical report, LingPipe, 2010.

[3] Griffiths, Thomas L., and Mark Steyvers. ”Finding scientific topics.” Proceedings of the National academy of Sciences of the United States of America 101.Suppl 1 (2004): 5228-5235.

[4] Murphy, Kevin P. ”Conjugate Bayesian analysis of the Gaussian distribution.” def 1.2?2 (2007): 16.

[5] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. ”Latent dirichlet allocation.” the Journal of machine Learning research 3 (2003): 993-1022.

[6] Reed, Colorado. ”Latent Dirichlet Allocation: Towards a Deeper Understanding”
$http://obphio.us/pdfs/lda\_tutorial.pdf$