STATS370: Final Project

Erich Trieschman

2022 Fall Quarter

1 Introduction

1.1 Description

- Compute quantiles, plot histograms, obtain means, variances and correlation matrix.
- Check convergence of the chain: does the variance of the distribution stabilize? Run two markov chains, how do they compare?
- Hamiltonian: Parameter search. Implement over M, eps, L. See section 3.5 for tuning
- Gibbs: basically show the math
- Metropolis Hasting: choice of candidate distribution. Make sure domain is right for each, could optimize parameters of the candidate distributions
- Importance sampling: TBD

Priors

$$\begin{aligned}
\theta &= (\sigma^2, \tau, \mu_1, \mu_2, \gamma_1, \gamma_2) \\
p(\sigma^2) &\propto \frac{1}{\sigma^2} \\
p(\tau) &\sim Unif[0,1] \\
p(\mu) &= p(\mu_1, \mu_2) \propto 1 \text{ (improper uniform)} \\
p(\gamma) &= p(\gamma_1, \gamma_2) \propto 1 \text{ (improper uniform)}
\end{aligned}$$

Gene expression distributions

$$\begin{aligned}
(y_i|g_i = 1) &\sim N(\mu, \sigma^2 I) \\
(y_i|g_i = 2) &\sim N(\gamma, \sigma^2 I) \\
(y_i|g_i = 3) &\sim N(\frac{1}{2}(\mu + \gamma), \sigma^2 I) \\
(y_i|g_i = 4) &\sim N(\tau\mu + (1-\tau)\gamma, \sigma^2 I)
\end{aligned}$$

Likelihood

$$\begin{aligned}
L(Y|\theta) &= \prod_{i=1}^{n} p(y_i|\theta), \text{ for } Y = (y_i, \dots, y_n) \\
&= \prod_{i \in g_1} \frac{1}{\sqrt{2\pi}}((\sigma^2)^2)^{-\frac{1}{2}} \exp[-\frac{1}{2\sigma^2}(y_i - \mu)^T(y_i - \mu)] \times \prod_{i \in g_2} \frac{1}{\sqrt{2\pi}}((\sigma^2)^2)^{-\frac{1}{2}} \exp[-\frac{1}{2\sigma^2}(y_i - \gamma)^T(y_i - \gamma)] \\
&\quad \times \prod_{i \in g_3} \frac{1}{\sqrt{2\pi}}((\sigma^2)^2)^{-\frac{1}{2}} \exp[-\frac{1}{2\sigma^2}(y_i - \frac{1}{2}(\mu + \gamma))^T(y_i - \frac{1}{2}(\mu + \gamma))] \\
&\quad \times \prod_{i \in g_4} \frac{1}{\sqrt{2\pi}}((\sigma^2)^2)^{-\frac{1}{2}} \exp[-\frac{1}{2\sigma^2}(y_i - (\tau\mu + (1-\tau)\gamma))^T(y_i - (\tau\mu + (1-\tau)\gamma))] \\
&= \left(\frac{1}{\sigma^2\sqrt{2\pi}}\right)^n \exp[-\frac{1}{2\sigma^2}(\sum_{i \in g_1}(y_i - \mu)^T(y_i - \mu) + \sum_{i \in g_2}(y_i - \gamma)^T(y_i - \gamma) \\
&\quad + \sum_{i \in g_3}(y_i - \frac{1}{2}(\mu + \gamma))^T(y_i - \frac{1}{2}(\mu + \gamma)) + \sum_{i \in g_4}(y_i - (\tau\mu + (1-\tau)\gamma))^T(y_i - (\tau\mu + (1-\tau)\gamma)))]
\end{aligned}$$

Posterior

$$p(\theta|Y) \propto p(\theta)L(Y|\theta)$$

2 Metropolis Hasting

- Candidate distribution selection, including log-normal for ss with change of varibles transformation
- Grid search for potential parameters within chosen proposal distribution

3 Gibbs sampling

Posterior conditioned on each parameter

$$p(\theta[i]|Y,\theta[-i]) = \frac{p(\theta[i],\theta[-i]|Y)}{p(\theta[-i]|Y)} = f(\theta[i]) \propto p(\theta|Y) \text{ with fixed } \theta[-i], Y$$

Posterior conditional probability of τ

$$\begin{aligned}
p(\tau|Y,\theta[-\tau]) &\propto p(\theta)L(Y|\theta) \\
&\propto p(\tau)\prod_{i\in g_4}\frac{1}{\sigma^2\sqrt{2\pi}}\exp[-\frac{1}{2\sigma^2}(y_i-(\tau\mu+(1-\tau)\gamma))^T(y_i-(\tau\mu+(1-\tau)\gamma))] \\
&\propto \exp[-\frac{1}{2\sigma^2}\sum_{i\in g_4}(y_i-(\tau\mu+(1-\tau)\gamma))^T(y_i-(\tau\mu+(1-\tau)\gamma))], \text{ for } \tau\in[0,1] \\
&\propto \exp[-\frac{1}{2\sigma^2}\sum_{i\in g_4}(y_i^Ty_i-2y_i^T(\tau\mu+(1-\tau)\gamma)+(\tau\mu+(1-\tau)\gamma)^T(\tau\mu+(1-\tau)\gamma))], \text{ for } \tau\in[0,1] \\
&\propto \exp[-\frac{1}{2\sigma^2}\sum_{i\in g_4}\tau^2(\mu^T\mu-2\mu^T\gamma+\gamma^T\gamma)-2\tau(y_i^T\mu-y_i^T\gamma-\mu^T\gamma+\gamma^T\gamma)], \text{ for } \tau\in[0,1] \\
&\propto \exp[-\frac{1}{2\sigma^2}(n_4\tau^2(\mu-\gamma)^T(\mu-\gamma)-2\tau(\mu-\gamma)^T[\sum_{i\in g_4}(y_i-\gamma), \text{ for } \tau\in[0,1] \\
&\propto \exp\left[-\frac{1}{2}*\frac{n_4(\mu-\gamma)^T(\mu-\gamma)}{\sigma^2}\left(\tau-\frac{(\mu-\gamma)^T(\sum_{i\in g_4}(y_i-\gamma))}{n_4(\mu-\gamma)^T(\mu-\gamma)}\right)^2\right], \text{ for } \tau\in[0,1] \\
&\sim Norm\left(\mu=\frac{(\mu-\gamma)^T(\sum_{i\in g_4}(y_i-\gamma))}{n_4(\mu-\gamma)^T(\mu-\gamma)}, \sigma^2=\frac{\sigma^2}{n_4(\mu-\gamma)^T(\mu-\gamma)}\right), \text{ truncated to } [0,1]
\end{aligned}$$

Posterior conditional probability of σ^2

$$\begin{aligned}
p(\sigma^2|Y,\theta[-\sigma^2]) \propto& p(\theta)L(Y|\theta) \propto p(\sigma^2)\prod_{i=1}^n p(y_i|\theta) \\
\propto& \frac{1}{\sigma^2}*\left(\frac{1}{\sigma^2}\right)^n\exp\left[-\frac{1}{2\sigma^2}M\right], \text{ where } M = \\
&\sum_{i\in g_1}(y_i-\mu)^T(y_i-\mu)+\sum_{i\in g_2}(y_i-\gamma)^T(y_i-\gamma)+\sum_{i\in g_3}(y_i-\frac{1}{2}(\mu+\gamma))^T(y_i-\frac{1}{2}(\mu+\gamma)) \\
&+\sum_{i\in g_4}(y_i-(\tau\mu+(1-\tau)\gamma))^T(y_i-(\tau\mu+(1-\tau)\gamma)) \\
\propto& (\sigma^2)^{-n-1}\exp\left[-\frac{M}{2}\frac{1}{\sigma^2}\right] \\
\sim& InvGamma\left(\alpha=n,\beta=\frac{M}{2}\right)
\end{aligned}$$

Posterior conditional probability of μ

$$
\begin{aligned}
p(\mu|Y,\theta[-\mu]) \propto & p(\theta)L(Y|\theta) \\
& \propto \prod_{i\in g_1} \frac{1}{\sigma^2\sqrt{2\pi}} \exp[-\frac{1}{2\sigma^2}(y_i-\mu)^T(y_i-\mu)] \times \prod_{i\in g_3} \frac{1}{\sigma^2\sqrt{2\pi}} \exp[-\frac{1}{2\sigma^2}(y_i-\frac{1}{2}(\mu+\gamma))^T(y_i-\frac{1}{2}(\mu+\gamma))] \\
& \quad \times \prod_{i\in g_4} \frac{1}{\sigma^2\sqrt{2\pi}} \exp[-\frac{1}{2\sigma^2}(y_i-(\tau\mu+(1-\tau)\gamma))^T(y_i-(\tau\mu+(1-\tau)\gamma))] \\
& \propto \exp[-\frac{1}{2\sigma^2}(n_1\mu^T\mu - 2\mu^T\left(\sum_{i\in g_1} y_i\right) - \mu^T\left(\sum_{i\in g_3} y_i\right) + \frac{n_3}{4}\mu^T\mu + \frac{n_3}{2}\mu^T\gamma \\
& \quad - 2\tau\mu^T\left(\sum_{i\in g_4} y_i\right) + \tau^2 n_4\mu^T\mu + 2\tau(1-\tau)n_4\mu^T\gamma)] \\
& \propto \exp\left[-\frac{1}{2\sigma^2}(n_1+\frac{n_3}{4}+n_4\tau^2)\left(\mu^T\mu - 2\mu^T\frac{\sum_{i\in g_1} y_i + \frac{1}{2}\sum_{i\in g_3} y_i + \tau\sum_{i\in g_4} y_i - (\frac{n_3}{4}+n_4\tau(1-\tau))\gamma}{n_1+\frac{n_3}{4}+n_4\tau^2}\right)\right] \\
& \propto \exp\left[-\frac{1}{2\phi^2}(\mu-\psi)^T(\mu-\psi)\right], \text{ where} \\
& \quad \psi = \frac{\sum_{i\in g_1} y_i + \frac{1}{2}\sum_{i\in g_3} y_i + \tau\sum_{i\in g_4} y_i - (\frac{n_3}{4}+n_4\tau(1-\tau))\gamma}{n_1+\frac{n_3}{4}+n_4\tau^2}, \ \phi^2 = \frac{\sigma^2}{n_1+\frac{n_3}{4}+n_4\tau^2} \\
& \sim N(\mu=\psi, \Sigma=\phi^2 I)
\end{aligned}
$$

Posterior conditional probability of γ

By symmetry with posterior conditional probability of μ,

$$
\begin{aligned}
p(\gamma|Y,\theta[-\gamma]) \sim & N(\mu=\psi', \Sigma=\phi'^2 I), \text{ where} \\
& \psi' = \frac{\sum_{i\in g_2} y_i + \frac{1}{2}\sum_{i\in g_3} y_i + (1-\tau)\sum_{i\in g_4} y_i - (\frac{n_3}{4}+n_4\tau(1-\tau))\mu}{n_2+\frac{n_3}{4}+n_4(1-\tau)^2}, \ \phi'^2 = \frac{\sigma^2}{n_2+\frac{n_3}{4}+n_4(1-\tau)^2}
\end{aligned}
$$

4 Hamiltonian Monte Carlo

- Boundary constraints using Rollback HMC: https://arxiv.org/pdf/1709.02855.pdf and Reflection HMC: https://people.cs.umass.edu/papers/2015nips1.pdf

$$
\begin{aligned}
H(q,p) =& U(q) + K(p), \text{ Hamiltonian function} \\
& \text{with } U(\cdot), \text{ potential energy, } K(\cdot), \text{ kinetic energy, } q\in\mathbb{R}^k, \text{ position, } p\in\mathbb{R}^k \text{ momentum} \\
& \text{let } U(\theta) := -\log f(\theta), \text{ where } p(\theta|Y) = \frac{1}{C}f(\theta), \text{ and } p\sim N(0,\mathbf{M})
\end{aligned}
$$

5 Importance sampling

Observe that I can draw samples from the following form of our posterior distribution

$$
\begin{aligned}
\theta = (\sigma^2,\tau,\mu,\gamma) \sim q(\sigma^2,\tau,\mu,\gamma \mid Y_{1,2,3,4}) &= q(\tau \mid \sigma^2,\mu,\gamma,Y_{1,2,3,4}) \times q(\sigma^2,\mu,\gamma \mid Y_{1,2,3,4}) \\
&\propto \ q(\tau \mid \sigma^2,\mu,\gamma,Y_4) \times q(\sigma^2,\mu,\gamma \mid Y_{1,2}) \times \int_0^1 \prod_{i\in g_3,g_4} L(y_i \mid \sigma^2,\mu,\gamma,Y_{1,2})d\tau
\end{aligned}
$$

Since

$$
\begin{aligned}
&q(\sigma^2, \mu, \gamma \mid Y_{1,2}) = q(\sigma^2 \mid \mu, \gamma, Y_{1,2}) \times q(\mu, \gamma \mid Y_{1,2}), \text{ where} \\
&\quad q(\mu, \gamma \mid Y_{1,2}) \sim ... \\
&q(\sigma^2 \mid \mu, \gamma, Y_{1,2}) \sim InvGamma\left(\alpha = n_1 + n_2, \beta = \frac{1}{2}\sum_{i \in g_1} \|y_i - \mu\|^2 + \sum_{i \in g_2} \|y_i - \gamma\|^2\right) \\
&q(\tau \mid \sigma^2, \mu, \gamma, Y_4) \sim Norm\left(\mu = \frac{(\mu-\gamma)^T(\sum_{i \in g_4}(y_i - \gamma))}{n_4(\mu-\gamma)^T(\mu-\gamma)}, \sigma^2 = \frac{\sigma^2}{n_4(\mu-\gamma)^T(\mu-\gamma)}\right), \text{ truncated to } [0,1]
\end{aligned}
$$

$$
\int_0^1 \prod_{i \in g_3, g_4} L(y_i \mid \sigma^2, \mu, \gamma, Y_{1,2}) d\tau \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i \in g_3} \left\|y_i - \frac{1}{2}(\mu - \gamma)\right\|^2\right]
$$

- Step 1: draw $(\sigma^2, \mu, \gamma) \sim q(\sigma^2, \mu, \gamma \mid Y_{1,2})$
- Step 2: draw $\tau \sim q(\tau \mid \sigma^2, \mu, \gamma, Y_4)$
- Step

Derivation of group 3 and 4 likelihood for candidate distribution

$$
\begin{aligned}
\int_0^1 \prod_{i \in g_3, g_4} L(y_i \mid \sigma^2, \mu, \gamma, Y_{1,2}) d\tau \propto\ & \int_0^1 \exp\left[-\frac{1}{2\sigma^2}\sum_{i \in g_3} \left\|y_i - \frac{1}{2}(\mu-\gamma)\right\|^2 - \frac{1}{2\sigma^2}\sum_{i \in g_4} \|y_i - \tau\mu - (1-\tau)\gamma\|^2\right] d\tau \\
\propto\ & \exp\left[-\frac{1}{2\sigma^2}\sum_{i \in g_3} \left\|y_i - \frac{1}{2}(\mu-\gamma)\right\|^2\right] \\
& \times \int_0^1 \exp\left[-\frac{1}{2} * \frac{n_4(\mu-\gamma)^T(\mu-\gamma)}{\sigma^2}\left(\tau - \frac{(\mu-\gamma)^T(\sum_{i \in g_4}(y_i-\gamma))}{n_4(\mu-\gamma)^T(\mu-\gamma)}\right)^2\right] d\tau \\
\propto\ & \exp\left[-\frac{1}{2\sigma^2}\sum_{i \in g_3} \left\|y_i - \frac{1}{2}(\mu-\gamma)\right\|^2\right]
\end{aligned}
$$

Derivation of sigma squared conditional on mu, gamma, and groups 1/2

$$
\begin{aligned}
q(\sigma^2 \mid \mu, \gamma, Y_{1,2}) \propto\ & \frac{1}{\sigma^2} * \left(\frac{1}{\sigma^2}\right)^{n_1+n_2} \exp\left[-\frac{1}{2\sigma^2}M\right], \text{ where } M = \sum_{i \in g_1}(y_i-\mu)^T(y_i-\mu) + \sum_{i \in g_2}(y_i-\gamma)^T(y_i-\gamma) \\
& \propto\ (\sigma^2)^{-n_1-n_2-1} \exp\left[-\frac{M}{2}\frac{1}{\sigma^2}\right] \\
& \sim InvGamma\left(\alpha = n_1 + n_2, \beta = \frac{M}{2}\right)
\end{aligned}
$$

Derivation of marginal distribution of mu and gamma conditional on groups 1/2

$$
\begin{aligned}
q(\mu, \gamma \mid Y_{1,2}) &= \int_0^\infty q(\mu, \gamma, \sigma^2 \mid Y_{1,2}) d\sigma^2 \\
&\propto \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{n_1+n_2+n_3+1} \exp\left[-\frac{1}{2\sigma^2}(\sum_{i \in g_1}\|y_i-\mu\|^2 + \sum_{i \in g_2}\|y_i - \gamma\|^2\right] d\sigma^2 \\
&\propto \Gamma(n_1+n_2)\left(-\frac{1}{2}\sum_{i \in g_1}\|y_i-\mu\|^2 + \sum_{i \in g_2}\|y_i-\gamma\|^2\right)^{-(n_1+n_2)}, \text{ where } \frac{\Gamma(\alpha)}{\beta^\alpha} = \int_0^\infty (\sigma^2)^{-\alpha-1}\exp\left(\frac{\beta}{\sigma^2}\right) d\sigma^2 \\
&\propto (M_1 + n_1\|\overline{Y_1} - \mu\|^2 + n_2\|\overline{Y_2} - \gamma\|^2)^{-(n_1+n_2)}, \text{ where } M_1 = \sum_{i \in g_1}\|y_i - \overline{Y_1}\|^2 + \sum_{i \in g_2}\|y_i - \overline{Y_2}\|^2 \\
&\propto \left(M_1 + M_2 + n_1\mu^T\mu - 2n_1\mu^T\overline{Y_1} + n_1\overline{Y_1}^T\overline{Y_1} + n_2\gamma^T\gamma - 2n_2\gamma^T\overline{Y_2} + n_2\overline{Y_2}^T\overline{Y_2}\right)^{-(n_1+n_2)} \\
&\propto \ldots \\
&\propto \left[1 + \frac{1}{\nu}(x-\eta)^T\Sigma^{-1}(x-\eta)\right]^{-\frac{\nu+4}{2}}, \text{where } \nu = XX, \Sigma = XXX, \eta = XX
\end{aligned}
$$