1

(a)

Introduce a binary random variable $Y_i \sim \text{Bernoulli}(p)$ for each X_i, so that the conditional distributions become

$$\begin{aligned} P(p;\alpha,\beta) &\propto p^{\alpha-1}(1-p)^{\beta-1} \\ P(y_i \mid p) &= p^{y_i}(1-p)^{1-y_i} \\ P(x_i \mid y_i) &= (f_0(x_i))^{y_i}(f_1(x_i))^{1-y_i} \end{aligned}$$

Hence the Gibbs sampling equations are

$$\begin{aligned} P(p \mid y^n, x^n;\alpha,\beta) &= P(p \mid y^n;\alpha,\beta) \\ &\propto P(y^n \mid p;\alpha,\beta)\,P(p;\alpha,\beta) \\ &= P(y^n \mid p)\,P(p;\alpha,\beta) \\ &= \left[\prod_{i=1}^n P(y_i \mid p)\right] P(p;\alpha,\beta) \\ &= p^{\sum_{i=1}^n y_i+\alpha-1}(1-p)^{\beta+n-\sum_{i=1}^n y_i-1} \end{aligned}$$

which is just a $\text{Beta}\left(\alpha+\sum_{i=1}^n y_i, \beta+n-\sum_{i=1}^n y_i\right)$ distribution, and

$$\begin{aligned} P\left(y_i \mid y^{-i}, x^n, p;\alpha,\beta\right) &= P(y_i \mid x_i, p) \\ &\propto P(x_i \mid y_i, p)\,P(y_i \mid p) \\ &= P(x_i \mid y_i)\,P(y_i \mid p) \\ &= (pf_0(x_i))^{y_i}((1-p)f_1(x_i))^{1-y_i} \end{aligned}$$

which is just a $\text{Bernoulli}\left(\frac{pf_0(x_i)}{pf_0(x_i)+(1-p)f_1(x_i)}\right)$ distribution.

To derive a random walk MCMC algorithm, we transform p according to

$$\begin{aligned} \psi &= \log\frac{p}{1-p} \\ (1-p)\exp\psi &= p \\ p &= \frac{\exp\psi}{1+\exp\psi} \end{aligned}$$

Hence the prior becomes

$$\begin{aligned} P(\psi;\alpha,\beta) &= f_\psi(\psi;\alpha,\beta) \\ &= f_p(p(\psi);\alpha,\beta)\left|\frac{dp(\psi)}{d\psi}\right| \\ &\propto \left(\frac{\exp\psi}{1+\exp\psi}\right)^{\alpha-1}\left(\frac{1}{1+\exp\psi}\right)^{\beta-1}\left|\exp(\psi)(-1)(1+\exp\psi)^{-2}\exp(\psi)+\exp(\psi)(1+\exp\psi)^{-1}\right| \\ &= \frac{(\exp\psi)^{\alpha-1}}{(1+\exp\psi)^{\alpha+\beta-2}}\left|\frac{\exp(\psi)}{1+\exp\psi}-\left(\frac{\exp\psi}{1+\exp\psi}\right)^2\right| \end{aligned}$$

while the likelihood becomes

$$P(x^n \mid p(\psi)) = \prod_{i=1}^n\left(\frac{\exp\psi}{1+\exp\psi}f_0(x_i)+\frac{1}{1+\exp\psi}f_1(x_i)\right)$$

The posterior in terms of ψ is therefore

$$\begin{aligned} &\pi(\psi) = P(x^n \mid p(\psi))\,P(\psi;\alpha,\beta) \\ \propto\ &\left[\prod_{i=1}^n\left(\frac{\exp\psi}{1+\exp\psi}f_0(x_i)+\frac{1}{1+\exp\psi}f_1(x_i)\right)\right]\frac{(\exp\psi)^{\alpha-1}}{(1+\exp\psi)^{\alpha+\beta-2}}\left|\frac{\exp\psi}{1+\exp\psi}-\left(\frac{\exp\psi}{1+\exp\psi}\right)^2\right| \end{aligned}$$

We can sample from this distribution using Metropolis-Hastings with a Gaussian proposal distribution

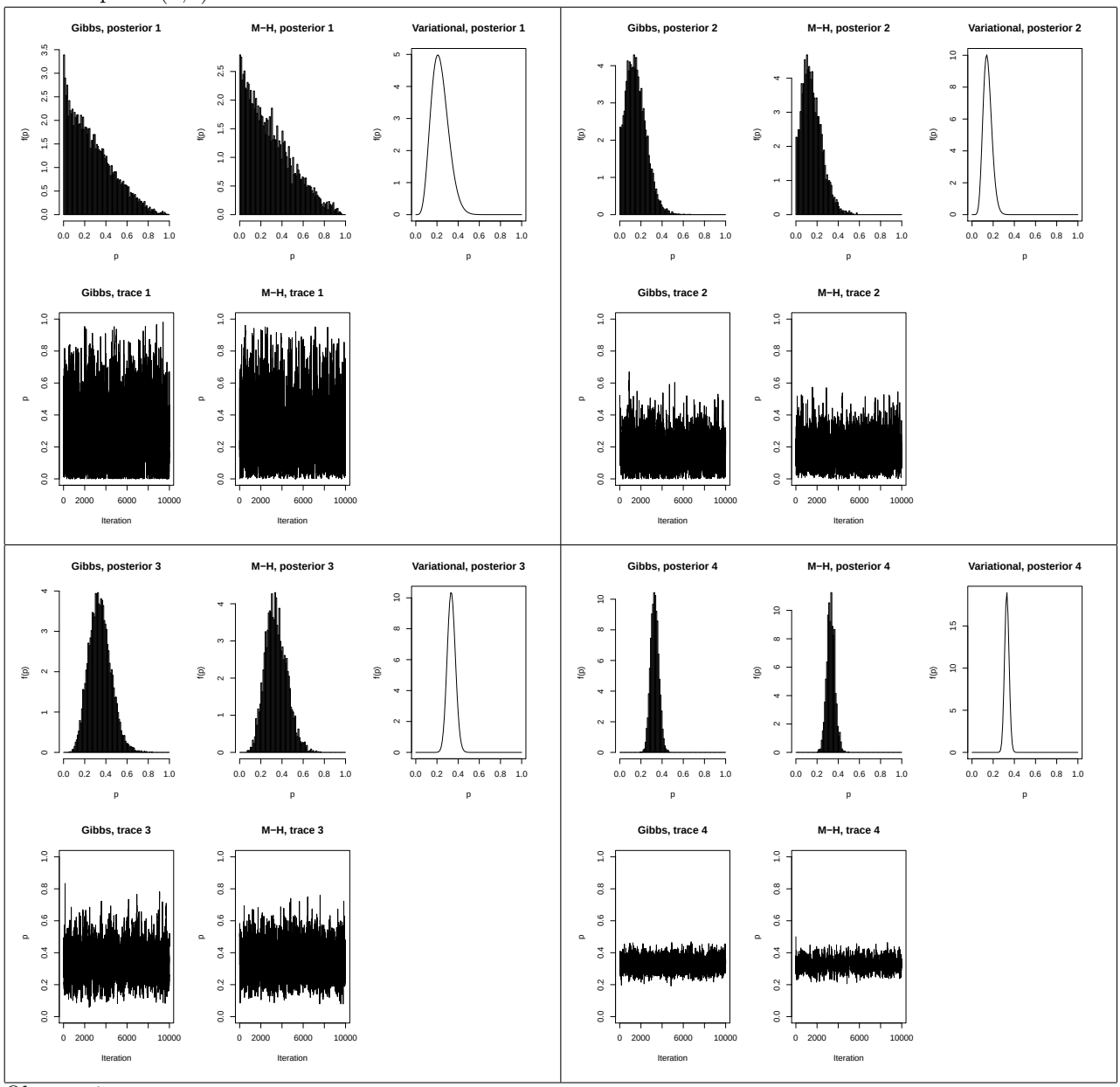
$$Q(y \mid \psi) \quad \sim \quad \text{Normal}(\psi, 1)$$

where the acceptance probability is

$$r(\psi, y) \quad = \quad \min\left\{\frac{\pi(y)}{\pi(\psi)}\frac{Q(\psi \mid y)}{Q(y \mid \psi)}, 1\right\} = \min\left\{\frac{\pi(y)}{\pi(\psi)}, 1\right\}$$

(b,d)

Plots for parts (b,d):



Observations:

- All trace plots are well-mixed, even during the early interations. This implies that the Gibbs and Metropolis-Hastings samplers reach their stationary distributions quickly.

- For the samplers, the variance in the posterior distribution and trace plots decreases noticably as n increases from 25 (experiment 1) to 500 (experiment 4). In contrast, the variational posterior has a small variance even with $n = 25$, and this variance does not decrease as dramatically with increasing n — in fact, all 3 posteriors have the same shape in experiment 4. The variational approximation is known to underestimate the posterior variance, so these observations are expected.
- In experiment 1 (small value of $n = 25$), the posterior estimated by either sampler does not reveal the true value of $p = 0.18$. This is to be expected: for small n, the prior on p dominates the posterior distribution, so the true value of p is not apparent. In contrast, the variational approximation accurately captures the true value, but at the same time it poorly approximates the Bayesian posterior.
- In experiments 2-4, the samplers and variational inference algorithm produce roughly the same posterior distributions, and are equally accurate at finding the true value of p.
- In conclusion, the variational approximation is accurate in the frequentist sense; it produces a good approximation to the true value of p in all experiments. However, it is a poor approximation in the Bayesian sense, especially for small n. In particular, because the variational distribution is unimodal, it cannot accurately represent the bimodal Beta $(0.8, 0.8)$ prior.

(c)

Using Jensen's inequality, the joint distribution can be lower bounded as follows:

$$\begin{aligned} p(x^n, p; \alpha, \beta) &= \text{Beta}(p; \alpha, \beta) \prod_{i=1}^{n} \left(p f_0(x_i) + (1-p) f_1(x_i) \right) \\ &\geq \text{Beta}(p; \alpha, \beta) \prod_{i=1}^{n} \left(\frac{p f_0(x_i)}{q_i} \right)^{q_i} \left(\frac{(1-p) f_1(x_i)}{1-q_i} \right)^{1-q_i} \\ &= Q(x^n, p; q^n) \end{aligned}$$

where $0 \leq q_i \leq 1$ for all i. Observe that the variational distribution Q can be maximized via EM, with p as the latent variable and q^n as the parameters. The EM step is

$$q_{k+1} \leftarrow \arg\max_q \mathbb{E}_{q_k} \left[\log Q(X^n, P; q) \mid X^n = x^n \right]$$

where k denotes the iteration number, and

$$\begin{aligned} & \mathbb{E}_{q_k} \left[\log Q(X^n, P; q) \mid X^n = x^n \right] \\ = \; & \mathbb{E}_{q_k} \left[\log \left[\text{Beta}(P; \alpha, \beta) \prod_{i=1}^{n} \left(\frac{P f_0(x_i)}{q_i} \right)^{q_i} \left(\frac{(1-P) f_1(x_i)}{1-q_i} \right)^{1-q_i} \right] \right] \\ = \; & \mathbb{E}_{q_k} \left[\log \left[\text{Beta}(P; \alpha, \beta) P^{\sum_i q_i} (1-P)^{\sum_i 1-q_i} \prod_{i=1}^{n} \left(\frac{f_0(x_i)}{q_i} \right)^{q_i} \left(\frac{f_1(x_i)}{1-q_i} \right)^{1-q_i} \right] \right] \\ = \; & \mathbb{E}_{q_k} \left[\log \left[\text{Beta}(P; \alpha, \beta) P^{\sum_i q_i} (1-P)^{n-\sum_i q_i} \right] + \log \left[\prod_{i=1}^{n} \left(\frac{f_0(x_i)}{q_i} \right)^{q_i} \left(\frac{f_1(x_i)}{1-q_i} \right)^{1-q_i} \right] \right] \\ = \; & \mathbb{E}_{q_k} \left[\log \left[\text{Beta}(P; \alpha, \beta) P^{\sum_i q_i} (1-P)^{n-\sum_i q_i} \right] \right] + \sum_{i=1}^{n} q_i \log \frac{f_0(x_i)}{q_i} + (1-q_i) \log \frac{f_1(x_i)}{1-q_i} \\ = \; & \mathbb{E}_{q_k} \left[\log \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} + \log \left[P^{\alpha+\sum_i q_i} (1-P)^{\beta+n-\sum_i q_i} \right] \right] + \sum_{i=1}^{n} q_i \log \frac{f_0(x_i)}{q_i} + (1-q_i) \log \frac{f_1(x_i)}{1-q_i} \\ = \; & \log \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} + \mathbb{E}_{q_k} \left[\left(\alpha + \sum_i q_i \right) \log P + \left(\beta + n - \sum_i q_i \right) \log (1-P) \right] + \sum_{i=1}^{n} q_i \log \frac{f_0(x_i)}{q_i} + (1-q_i) \log \frac{f_1(x_i)}{1-q_i} \\ = \; & \log \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} + \left(\alpha + \sum_i q_i \right) \mathbb{E}_{q_k} \left[\log P \right] + \left(\beta + n - \sum_i q_i \right) \mathbb{E}_{q_k} \left[\log (1-P) \right] + \sum_{i=1}^{n} q_i \log \frac{f_0(x_i)}{q_i} + (1-q_i) \log \frac{f_1(x_i)}{1-q_i} \end{aligned}$$

We note the following:

1. The posterior $P \mid X^n = x^n$ has a Beta $\left(\alpha + \sum_i q_i, \beta + n - \sum_i q_i\right)$ distribution, because

$$\begin{aligned} Q\left(P \mid x^n; q\right) &= \frac{Q\left(x^n, P; q\right)}{Q\left(x^m; q\right)} \\ &\propto Q\left(x^n, P; q\right) \\ &\propto \text{Beta}\left(P; \alpha, \beta\right) P^{\sum_i q_i} \left(1-P\right)^{\sum_i 1-q_i} \quad \text{(ignoring the constant of proportionality in } q, x^n\text{)} \\ &\propto \text{Beta}\left(P; \alpha + \sum_i q_i, \beta + n - \sum_i q_i\right) \end{aligned}$$

2. $\int \left(\log P\right) \text{Beta}\left(\gamma_1, \gamma_2\right) dP = \Psi\left(\gamma_1\right) - \Psi\left(\gamma_1 + \gamma_2\right)$ where Ψ is the digamma function. I took this fact from Wikipedia; the provided variational inference notes appear to have a mistake.

Hence

$$\begin{aligned} \mathbb{E}_{q_k}\left[\log P\right] &= \int \left(\log P\right) \text{Beta}\left(P; \alpha + \sum_i q_{(k)i}, \beta + n - \sum_i q_{(k)i}\right) \\ &= \Psi\left(\alpha + \sum_i q_{(k)i}\right) - \Psi\left(\alpha + \beta + n\right) \end{aligned}$$

and

$$\mathbb{E}_{q_k}\left[\log\left(1-P\right)\right] = \Psi\left(\beta + n - \sum_i q_{(k)i}\right) - \Psi\left(\alpha + \beta + n\right)$$

where we note that $1 - P$ has a Beta $\left(\beta + n - \sum_i q_i, \alpha + \sum_i q_i\right)$ distribution conditioned on x^n. Therefore

$$\begin{aligned} \mathbb{E}_{q_k}\left[\log Q\left(X^n, P; q\right) \mid X^n = x^n\right] &= \log \frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)} + \left[\sum_{i=1}^n q_i \log \frac{f_0\left(x_i\right)}{q_i} + \left(1-q_i\right) \log \frac{f_1\left(x_i\right)}{1-q_i}\right] \\ &\quad + \left(\alpha + \sum_i q_i\right)\left(\Psi\left(\alpha + \sum_i q_{(k)i}\right) - \Psi\left(\alpha+\beta+n\right)\right) \\ &\quad + \left(\beta + n - \sum_i q_i\right)\left(\Psi\left(\beta + n - \sum_i q_{(k)i}\right) - \Psi\left(\alpha+\beta+n\right)\right) \\ &= \left[\sum_{i=1}^n q_i\left(\log f_0\left(x_i\right) - \log q_i\right) + \left(1-q_i\right)\left(\log f_1\left(x_i\right) - \log\left(1-q_i\right)\right)\right] + C \\ &\quad + \left(\sum_i q_i\right)\Psi\left(\alpha + \sum_i q_{(k)i}\right) - \left(\sum_i q_i\right)\Psi\left(\beta + n - \sum_i q_{(k)i}\right) \end{aligned}$$

where C does not depend on q. Maximizing with respect to q gives

$$\begin{aligned} \frac{\partial}{\partial q_i}\mathbb{E}_{q_k}\left[\log Q\left(X^n, P; q\right) \mid X^n = x^n\right] &= 0 \\ \log f_0\left(x_i\right) - \log q_i - 1 + \frac{1}{1-q_i} - \log f_1\left(x_i\right) + \log\left(1-q_i\right) - \frac{q_i}{1-q_i} & \\ + \Psi\left(\alpha + \sum_i q_{(k)i}\right) - \Psi\left(\beta + n - \sum_i q_{(k)i}\right) &= 0 \\ \log f_1\left(x_i\right) - \log f_0\left(x_i\right) + \Psi\left(\beta + n - \sum_i q_{(k)i}\right) - \Psi\left(\alpha + \sum_i q_{(k)i}\right) &= \log \frac{1-q_i}{q_i} \end{aligned}$$

$$\begin{aligned} \frac{f_1(x_i)\exp\Psi\left(\beta+n-\sum_i q_{(k)i}\right)}{f_0(x_i)\exp\Psi\left(\alpha+\sum_i q_{(k)i}\right)} &= \frac{1-q_i}{q_i} \\ \frac{f_0(x_i)\exp\Psi\left(\alpha+\sum_i q_{(k)i}\right)}{f_0(x_i)\exp\Psi\left(\alpha+\sum_i q_{(k)i}\right)+f_1(x_i)\exp\Psi\left(\beta+n-\sum_i q_{(k)i}\right)} &= q_i \end{aligned}$$

so the EM algorithm is just

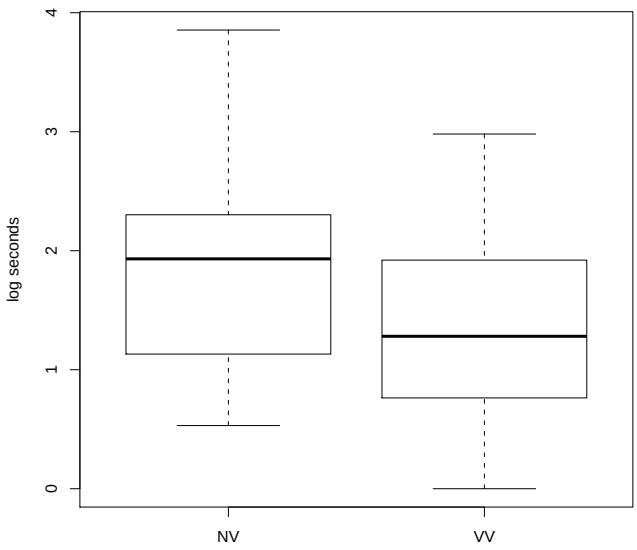
$$q_{(k+1)i} \leftarrow \frac{f_0(x_i)\exp\Psi\left(\alpha+\sum_i q_{(k)i}\right)}{f_0(x_i)\exp\Psi\left(\alpha+\sum_i q_{(k)i}\right)+f_1(x_i)\exp\Psi\left(\beta+n-\sum_i q_{(k)i}\right)}$$

The variational algorithm iterates this step until convergence, and the variational approximation to the posterior is

$$P \mid X^n = x^n \sim \text{Beta}\left(\alpha+\sum_i q_i, \beta+n-\sum_i q_i\right)$$

2

(a)



(b)

We estimate

$$\hat{\delta} = \hat{\mu}_1 - \hat{\mu}_2$$

where

$$\begin{aligned} \hat{\mu}_1 &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ \hat{\mu}_2 &= \frac{1}{m}\sum_{i=1}^{m} Y_i \end{aligned}$$

Observe that $\hat{\delta}$ is normally distributed, with

$$\begin{aligned}\mathbb{E}\left[\hat{\delta}\right] &= \mathbb{E}\left[\hat{\mu}_1 - \hat{\mu}_2\right] \\ &= \mu_1 - \mu_2\end{aligned}$$

and

$$\begin{aligned}\mathbb{V}\left[\hat{\delta}\right] &= \mathbb{V}\left[\hat{\mu}_1 - \hat{\mu}_2\right] \\ &= \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{m}\sum_{i=1}^{m} Y_i\right] \\ &= \frac{1}{n}\mathbb{V}\left[X\right] + \frac{1}{m}\mathbb{V}\left[Y\right] \\ &= \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\end{aligned}$$

which we estimate using $\alpha = \hat{\mu}_1 - \hat{\mu}_2$ and $\beta^2 = \frac{s_1^2}{n} + \frac{s_2^2}{m}$ respectively.

From the data, a 95% confidence interval for δ is

$$\alpha \pm \beta z_{0.05/2} = 0.431 \pm 0.364$$

(c)

The posterior density is

$$\begin{aligned}\pi\left(\mu_1, \mu_2, \sigma_1, \sigma_2 \mid X^n, Y^m\right) &\propto L_n\left(X^n \mid \mu_1, \sigma_1\right) L_m\left(Y^m \mid \mu_2, \sigma_2\right) \pi\left(\mu_1, \mu_2, \sigma_1, \sigma_2\right) \\ &\propto \frac{1}{\sigma_1\sigma_2}\prod_{i=1}^{n}\frac{1}{\sigma_1}\exp\left\{-\frac{\left(X_i-\mu_1\right)^2}{2\sigma_1^2}\right\}\prod_{i=1}^{m}\frac{1}{\sigma_2}\exp\left\{-\frac{\left(Y_i-\mu_2\right)^2}{2\sigma_2^2}\right\}\end{aligned}$$

(d)

We decompose the joint posterior such that

$$\begin{aligned}\pi\left(\mu_1, \mu_2, \sigma_1, \sigma_2 \mid D\right) &= p\left(\mu_1, \sigma_1 \mid X^n\right) p\left(\mu_2, \sigma_2 \mid Y^m\right) \\ &= p\left(\mu_1 \mid \sigma_1, X^n\right) p\left(\sigma_1 \mid X^n\right) p\left(\mu_2 \mid \sigma_2, Y^m\right) p\left(\sigma_2 \mid Y^m\right)\end{aligned}$$

where

$$\begin{aligned}p\left(\sigma_1 \mid X^n\right) = \int p\left(\mu_1, \sigma_1 \mid X^n\right) d\mu_1 &\propto \int \frac{1}{\sigma_1}\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left\{-\frac{\left(X_i-\mu_1\right)^2}{2\sigma_1^2}\right\} d\mu_1 \\ &= \int \frac{1}{\left(2\pi\right)^{n/2}\sigma_1^{n+1}}\exp\left\{-\sum_{i=1}^{n}\frac{\left(X_i-\mu_1\right)^2}{2\sigma_1^2}\right\} d\mu_1 \\ &= \int \frac{1}{\left(2\pi\right)^{n/2}\sigma_1^{n+1}}\exp\left\{-\frac{\left(\mu_1-\bar{X}\right)^2}{2\frac{\sigma_1^2}{n}}\right\}\exp\left\{-\frac{\sum_{i=1}^{n}\left(X_i^2-\bar{X}^2\right)}{2\sigma_1^2}\right\} d\mu_1 \\ &= \frac{\sqrt{2\pi}\sigma_1}{\left(2\pi\right)^{n/2}\sigma_1^{n+1}\sqrt{n}}\exp\left\{-\frac{\sum_{i=1}^{n}\left(X_i^2-\bar{X}^2\right)}{2\sigma_1^2}\right\} \\ &\propto \frac{1}{\sigma_1^n}\exp\left\{-\frac{\sum_{i=1}^{n}\left(X_i^2-\bar{X}^2\right)}{2\sigma_1^2}\right\}\end{aligned}$$

We now perform a variance to precision transformation $\tau_1 = \frac{1}{\sigma_1^2}$, so that

$$\begin{aligned} p(\tau_1 \mid X^n) = p(\sigma_1(\tau_1) \mid X^n) &\propto \tau_1^{n/2} \exp\left\{-\tau_1\left(\frac{1}{2}\right)\sum_{i=1}^{n}\left(X_i^2 - \bar{X}^2\right)\right\}\left|\frac{d\sigma_1(\tau_1)}{d\tau_1}\right| \\ &= \tau_1^{n/2} \exp\left\{-\tau_1 / \left(\frac{1}{2}\sum_{i=1}^{n}\left(X_i^2 - \bar{X}^2\right)\right)^{-1}\right\}\left|-\frac{1}{2}\tau^{-\frac{3}{2}}\right| \\ &\propto \tau_1^{\frac{n-1}{2}-1} \exp\left\{-\tau_1 / \left(\frac{1}{2}\sum_{i=1}^{n}\left(X_i^2 - \bar{X}^2\right)\right)^{-1}\right\} \end{aligned}$$

which is a Gamma $\left(\frac{n-1}{2}, \frac{2}{\sum_{i=1}^{n}\left(X_i^2 - \bar{X}^2\right)}\right)$ distribution. Similarly,

$$p(\tau_2 \mid Y^m) = p(\sigma_2(\tau_2) \mid Y^m) \propto \tau_2^{\frac{m-1}{2}-1} \exp\left\{-\tau_2 / \left(\frac{1}{2}\sum_{i=1}^{m}\left(Y_i^2 - \bar{Y}^2\right)\right)^{-1}\right\}$$

which is a Gamma $\left(\frac{m-1}{2}, \frac{2}{\sum_{i=1}^{m}\left(Y_i^2 - \bar{Y}^2\right)}\right)$ distribution. Finally,

$$\begin{aligned} p(\mu_1 \mid \sigma_1, X^n) = \frac{p(\mu_1, \sigma_1 \mid X^n)}{p(\sigma_1 \mid X^n)} &\propto \frac{\frac{1}{\sigma_1}\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left\{-\frac{(X_i-\mu_1)^2}{2\sigma_1^2}\right\}}{\frac{1}{\sigma_1^n}\exp\left\{-\frac{\sum_{i=1}^{n}\left(X_i^2-\bar{X}^2\right)}{2\sigma_1^2}\right\}} \\ &\propto \prod_{i=1}^{n}\exp\left\{-\frac{(X_i-\mu_1)^2}{2\sigma_1^2}\right\} \\ &\propto \exp\left\{-\frac{(\mu_1-\bar{X})^2}{2\frac{\sigma_1^2}{n}}\right\} \\ p(\mu_2 \mid \sigma_2, Y^m) = \frac{p(\mu_2, \sigma_2 \mid Y^m)}{p(\sigma_2 \mid Y^m)} &\propto \exp\left\{-\frac{(\mu_2-\bar{Y})^2}{2\frac{\sigma_2^2}{m}}\right\} \end{aligned}$$

which are Normal $\left(\bar{X}, \frac{\sigma_1^2}{n}\right)$ and Normal $\left(\bar{Y}, \frac{\sigma_2^2}{m}\right)$ distributions, respectively.

An estimated 95% PI using 10000 samples is 0.432 ± 0.376. The estimated PI's center is nearly identical to the CI center, while its width is slightly larger. Assuming that the true PI is close to the CI, we conclude that direct sampling is a good way to estimate this posterior distribution.

(e)

The Gibbs sampling equations are

$$p(\mu_1 \mid \mu_2, \sigma_1, \sigma_2, D) = p(\mu_1 \mid \sigma_1, X^n) \propto p(\mu_1, \sigma_1 \mid X^n) \propto \exp\left\{-\frac{(\mu_1-\bar{X})^2}{2\frac{\sigma_1^2}{n}}\right\} \quad (\sigma_1 \text{ constant})$$

$$p(\mu_2 \mid \mu_1, \sigma_1, \sigma_2, D) = p(\mu_2 \mid \sigma_2, Y^n) \propto p(\mu_2, \sigma_2 \mid X^n) \propto \exp\left\{-\frac{(\mu_2-\bar{Y})^2}{2\frac{\sigma_2^2}{m}}\right\} \quad (\sigma_2 \text{ constant})$$

which are Normal $\left(\bar{X}, \frac{\sigma_1^2}{n}\right)$ and Normal $\left(\bar{Y}, \frac{\sigma_2^2}{m}\right)$ distributions respectively, and

$$p(\sigma_1 \mid \mu_1, \mu_2, \sigma_2, D) = p(\sigma_1 \mid \mu_1, X^n) \propto p(\mu_1, \sigma_1 \mid X^n)$$

$$\begin{aligned} & \propto \left(\frac{1}{\sigma_1^2}\right)^{\frac{n+1}{2}} \exp\left\{-\frac{\frac{1}{2}\sum_{i=1}^{n}(\mu_1 - X_i)^2}{2\sigma_1^2}\right\} \\ & = \left(\frac{1}{\sigma_1^2}\right)^{\frac{n+1}{2}} \exp\left\{-\frac{1}{\sigma_1^2}\Big/\left(\frac{1}{2}\sum_{i=1}^{n}(\mu_1 - X_i)^2\right)^{-1}\right\} \\ p(\sigma_2 \mid \mu_1, \mu_2, \sigma_1, D) &= p(\sigma_2 \mid \mu_1, D) \\ & \propto \left(\frac{1}{\sigma_2^2}\right)^{\frac{m+1}{2}} \exp\left\{-\frac{1}{\sigma_2^2}\Big/\left(\frac{1}{2}\sum_{i=1}^{m}(\mu_2 - Y_i)^2\right)^{-1}\right\} \end{aligned}$$

Transforming the variances σ_1^2, σ_2^2 to precisions,

$$\begin{aligned} p(\tau_1 \mid \mu_1, D) &\propto \tau_1^{\frac{n}{2}-1} \exp\left\{-\tau_1\Big/\left(\frac{1}{2}\sum_{i=1}^{n}(X_i - \mu_1)^2\right)^{-1}\right\} \\ p(\tau_2 \mid \mu_2, D) &\propto \tau_2^{\frac{m}{2}-1} \exp\left\{-\tau_2\Big/\left(\frac{1}{2}\sum_{i=1}^{m}(Y_i - \mu_2)^2\right)^{-1}\right\} \end{aligned}$$

which are Gamma $\left(\frac{n}{2}, \frac{2}{\sum_{i=1}^{n}(\mu_1 - X_i)^2}\right)$ and Gamma $\left(\frac{m}{2}, \frac{2}{\sum_{i=1}^{m}(\mu_2 - Y_i)^2}\right)$ distributions, respectively.

An estimated 95% PI using the first 10000 samples is 0.434 ± 0.539. The PI center is close to the CI center, but its width is significantly larger. Assuming that the true PI is close to the CI, this means that the samples have high variance compared to the true posterior distribution. Perhaps a burn-in period is necessary, so as to allow the Gibbs sampler to converge to the true posterior.

(f)

We sample from $\pi(\mu_1, \mu_2, \sigma_1, \sigma_2 \mid X^n, Y^m)$ using Metropolis-Hastings with proposal distribution

$$Q(\mu_1' \mid \mu_1)\, Q(\mu_2' \mid \mu_2)\, Q(\sigma_1' \mid \sigma_1)\, Q(\sigma_2' \mid \sigma_2)$$

where

$$\begin{aligned} Q(\mu_1' \mid \mu_1) &\sim \text{Normal}(\mu_1, 1) \\ Q(\mu_2' \mid \mu_2) &\sim \text{Normal}(\mu_2, 1) \\ Q(\sigma_1' \mid \sigma_1) &\sim \text{Gamma}(\sigma_1, 1) \\ Q(\sigma_2' \mid \sigma_2) &\sim \text{Gamma}(\sigma_2, 1) \end{aligned}$$

The acceptance probability is

$$r(\mu_1, \mu_2, \sigma_1, \sigma_2, \mu_1', \mu_2', \sigma_1', \sigma_2') = \min\left\{\frac{\pi(\mu_1', \mu_2', \sigma_1', \sigma_2' \mid X^n, Y^m)}{\pi(\mu_1, \mu_2, \sigma_1, \sigma_2 \mid X^n, Y^m)} \frac{Q(\sigma_1 \mid \sigma_1')\, Q(\sigma_2 \mid \sigma_2')}{Q(\sigma_1' \mid \sigma_1)\, Q(\sigma_2' \mid \sigma_2)}, 1\right\}$$

noting that the normal distribution is symmetric.

An estimated 95% PI using the first 10000 samples is 0.427 ± 0.458. The PI center is close to the CI center, but its width is somewhat larger. Assuming that the true PI is close to the CI, this means that the samples have high variance compared to the true posterior distribution. Similarly to the Gibbs sampler, a burn-in period may be necessary to allow for convergence to the true posterior.

3

(a)

Intuitively, the weights w_k get progressively smaller as k increases, because they are taken from the remainder of the unit length stick. In particular, w_k decreases faster as $\alpha \to 0$. Hence most of the probability mass of the Dirichlet

Process is found in the atoms $\delta_{(m_k,s_k)}$ such that k is small, and therefore the DP can be approximated with a finite number of atoms (mixture components) N.

More formally, Ishwaran and James (2002) show that the average pointwise ℓ_1 difference between the infinite DPM and the N-component DPM is approximately upper-bounded by $4n\exp\left(-\left(N-1\right)/\alpha\right)$, where n is the sample size and α is the concentration parameter. We can use this equation to choose N based on n, α and our desired average ℓ_1 error. As an example, when $n = 1000, N = 50, \alpha = 3$, the average ℓ_1 error bound is 3.2×10^{-4}, a very small number.

(b)

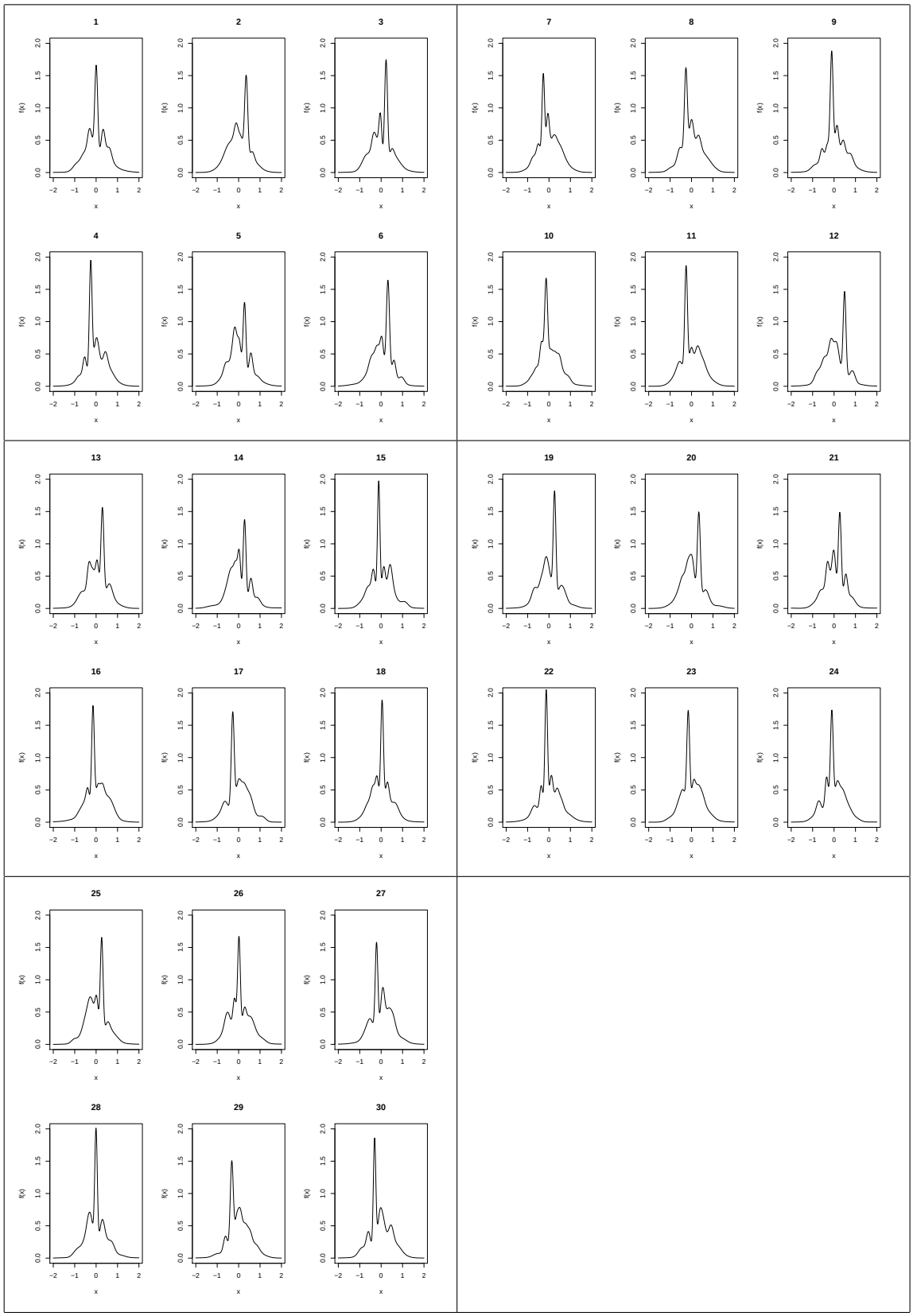
- θ is the prior mean for the mixture component means m_k. It controls the effective range of values m_k can take.
  - If θ was a fixed parameter, then m_k would be effectively restricted to a small subset of $\mathbb{R}$, because of its normal distribution $m_k \sim \mathcal{N}\left(\theta, \sigma_m\right)$.
  - To overcome this, a $\mathcal{N}\left(0, A = 1000\right)$ prior distribution is placed on θ. The large variance $A = 1000$ allows θ to take on a large range of values, in turn freeing m_k to take on a similarly large range of values.
- σ_m is the prior variance for the mixture mean components m_k.
  - Setting σ_m to 4 standard deviations of the data effectively restricts the mixture component means m_k to be close to the data samples, when θ is close to the sample mean.
- ν_1, ν_2 are hyperparameters for the Gamma prior distribution on the mixture component precisions $\left(s_k^2\right)^{-1}$. Hence their function is analogous to the bandwidth h in kernel density estimation.
  - Large values of ν_1 coupled with small values of ν_2 effectively restrict the precisions $\left(s_k^2\right)^{-1}$ to fall in a small interval near zero, analogous to large bandwidths h.
  - $\nu_1 = \nu_2 = 2$ implies that the prior distribution of $\left(s_k^2\right)^{-1}$ has mean $\nu_1\nu_2 = 4$, mode $\left(\nu_1 - 1\right)\nu_2 = 2$, and variance $\frac{2}{\sqrt{\nu_1}} = \sqrt{2}$. In other words, we expect the variance (bandwidth) of each mixture component to be less than 1.
- α is the prior to the stick-breaking construction $V_k \sim \text{Beta}\left(1, \alpha\right)$, therefore it controls the size of the stick fractions V_k to be broken off.
  - A small value of α results in large draws of V_k, therefore the first few weights w_k will be large and the rest small.
  - Conversely, a large value of α results in small draws of V_k, and the distribution of weights w_k becomes more even. As $\alpha \to \infty$, all weights $w_k \to 0$.
  - The prior on α is Gamma $\left(\eta_1 = 2, \eta_2 = 2\right)$. Hence the prior distribution of α has mean $\eta_1\eta_2 = 4$, mode $\left(\eta_1 - 1\right)\eta_2 = 2$, and variance $\frac{2}{\sqrt{\eta_1}} = \sqrt{2}$. Most draws of V_k will therefore be < 0.5, so F_0 will be moderately concentrated around the prior distribution of (m_k, s_k).

(e)

Compared to the optimal-bandwidth kernel density estimator in Chapter 26, the DPM predictive densities in part (d) are poor approximations of the true density. In particular:

1. The DPM predictive distributions are oversmoothed. This suggests that the posterior estimates of s_k^2 are too large, and therefore the prior parameters ν_1, ν_2 should be adjusted to generate larger precisions $\left(s_k^2\right)^{-1}$.
2. The DPM predictive distributions place unequal probability masses in each spike. This suggests that the mixture components are not distributed evenly throughout the spikes, which in turn suggests a lack of convergence. The low number of samples $B = 100$ may be to blame: since Gibbs sampling is an MCMC method, it requires some number of iterations to converge to its stationary distribution, the true posterior. Therefore the first $B = 100$ samples may not be representative of the true posterior.

(d)


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
f(x)
x