Chapter 24. Markov chain Monte Carlo (MCMC) inference

24.1.

$$p(x_1|x_2) = \mathcal{N}(x_1|-\frac{1}{2}x_2 + \frac{3}{2}, \frac{3}{4})$$
$$p(x_2|x_1) = \mathcal{N}(x_2|-\frac{1}{2}x_1 + \frac{3}{2}, \frac{3}{4})$$

24.2.
For class belongings, we have:

$$p(z_i = k|x_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\pi_k \mathcal{N}(x_i|\mu_k, \sigma_k^2)}{\sum_l \pi_l \mathcal{N}(x_i|\mu_l, \sigma_l^2)}$$

For class proportions, we have:

$$p(\boldsymbol{\pi}|\mathbf{z}) = \mathrm{Dir}(\{\alpha_k + \sum_{i=1}^{N} \mathbf{1}_{z_i=k}\}_{k=1}^{K})$$

For the means, we have:

$$p(\mu_k|\sigma_k^2, \mathbf{z}, \mathbf{x}) = \mathcal{N}(\mu_k|\frac{\sigma_0^2 \sum_{i:z_i=k} x_i + \sigma_k^2 \mu_0}{\sigma_k^2 + N_k \sigma_0^2}, \frac{\sigma_0^2 \sigma_k^2}{\sigma_k^2 + N_k \sigma_0^2})$$

For the variances, we have:

$$p(\sigma_k^2|\mu_k, \mathbf{z}, \mathbf{x}) = \mathrm{IG}(\sigma_k^2|a_0 + \sum_{i:z_i=k} (x_i - \mu_k)^2, b_0 + N_k).$$

24.4.

$$p(\mu|\theta_{1:D},\tau^2) \propto p(\mu)\prod_{j=1}^{D} p(\theta_j|\mu,\tau^2) = \mathcal{N}(\mu|\mu_0,\gamma_0^2)\prod_{j=1}^{D}\mathcal{N}(\theta_j|\mu,\tau^2)$$

$$\propto e^{-\frac{(\mu-\mu_0)^2}{2\gamma_0^2}-\sum_{j=1}^{D}\frac{(\mu-\theta_j)^2}{2\tau^2}} \propto \mathcal{N}(\mu|\frac{D\bar{\theta}\gamma_0^2+\mu_0\tau^2}{D\gamma_0^2+\tau^2},\frac{\gamma_0^2\tau^2}{D\gamma_0^2+\tau^2}).$$

$$p(\theta_j|\mu,\tau^2,\mathcal{D}_j,\sigma^2) \propto p(\theta_j|\mu,\tau^2)\prod_{i=1}^{N_j} p(x_{i,j}|\theta_j,\sigma^2) = \mathcal{N}(\theta_j|\mu,\tau^2)\prod_{i=1}^{N_j}\mathcal{N}(x_{i,j}|\theta_j,\sigma^2)$$

$$\propto e^{-\frac{(\theta_j-\mu)^2}{2\tau^2}-\sum_{i=1}^{N_j}\frac{(\theta_j-x_{i,j})^2}{2\sigma^2}} = \mathcal{N}(\theta_j|\frac{N_j\bar{x}_j\tau^2+\sigma^2}{N_j\tau^2+\sigma^2},\frac{\sigma^2\tau^2}{N_j\tau^2+\sigma^2}).$$

$$p(\tau^2|\theta_{1:D},\mu) \propto p(\tau^2)\prod_{j=1}^{D} p(\theta_j|\mu,\sigma^2) = \mathrm{IG}(\tau^2|\frac{\eta_0}{2},\frac{\eta_0\tau_0^2}{2})\prod_{j=1}^{D}\mathcal{N}(\theta_j|\mu,\tau^2)$$

$$\propto (\tau^2)^{-\frac{\eta_0}{2}-1-\frac{D}{2}}e^{-\frac{\eta_0\tau_0^2}{2\tau^2}-\sum_{j=1}^{D}\frac{(\mu-\theta_j)^2}{2\tau^2}} \propto \mathrm{IG}(\tau^2|\frac{\eta_0+D}{2},\frac{\eta_0\tau_0^2+\sum_j(\mu-\theta_j)^2}{2}).$$

$$p(\sigma^2|\theta_{1:D},\mathcal{D}) \propto p(\sigma^2)\prod_{j=1}^{D}\prod_{i=1}^{N_j} p(x_{i,j}|\theta_j,\sigma^2)$$

$$\propto (\sigma^2)^{-\frac{\nu_0}{2}-1}e^{-\frac{\nu_0\sigma_0^2}{2\sigma^2}}(\sigma^2)^{-\sum_{j=1}^{D}\frac{N_j}{2}}e^{-\frac{\sum_{j=1}^{D}\sum_{i=1}^{N_j}(x_{i,j}-\theta_j)^2}{2\sigma^2}}$$

$$\propto \mathrm{IG}(\sigma^2|\frac{1}{2}(\nu_0+\sum_{j=1}^{D}N_j),\frac{1}{2}(\nu_0\sigma_0^2+\sum_{j=1}^{D}\sum_{i=1}^{N_j}(x_{i,j}-\theta_j)^2)).$$

24.5.
Using the fact that a Student distribution can be written as a Gaussian scale mixture, a linear regreession with Student's t errors can be written as:

$$y_i = \mathbf{w}^T\mathbf{x}_i + \epsilon_i$$

$$\epsilon_i|z_i \sim \mathcal{N}(0,\sigma^2 z_i)$$

$$z_i \sim \mathrm{IG}(\frac{\nu}{2},\frac{\nu}{2})$$

Now, consider the following independent, conjugate priors:

$$\mathbf{w} \sim \mathcal{N}(\mathbf{w}_0,\mathbf{V}_{\mathbf{w}}),\sigma^2 \sim \mathrm{IG}(\nu_0,S_0)$$

To do Gibbs sampling, first we compute $p(\mathbf{w}|\mathbf{z}, \mathcal{D}, \sigma^2, \nu)$:
Since $\mathbf{y} = \mathbf{w}^T\mathbf{X} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathrm{diag}(z_i))$, we obtain:

$$p(\mathbf{w}|\mathbf{z}, \mathcal{D}, \sigma^2, \nu) = \mathcal{N}(\mathbf{w}|\hat{\mathbf{w}}, \mathbf{D}_\mathbf{w})$$

where

$$\hat{\mathbf{w}} = \mathbf{D}_\mathbf{w}(\mathbf{V}_\mathbf{w}^{-1}\mathbf{w}_0 + \frac{1}{\sigma^2}\mathbf{X}^T \mathrm{diag}(z_i^{-1})\mathbf{y})$$

and

$$\mathbf{D}_\mathbf{w} = (\mathbf{V}_\mathbf{w}^{-1} + \frac{1}{\sigma^2}\mathbf{X}^T \mathrm{diag}(z_i^{-1})\mathbf{X})^{-1}.$$

Next, to sample from $p(\mathbf{z}|\mathbf{w}, \mathcal{D}, \sigma^2, \nu)$, note that each of z_i are conditionally independent given the data and $(\mathbf{w}, \sigma^2)$, i.e. $p(\mathbf{z}|\mathbf{w}, \mathcal{D}, \sigma^2, \nu)$ is a product of univariate densities. Moreover, each of these is an inverse-gamma density. To see why, recall that $z_i \sim \mathrm{IG}(\frac{\nu}{2}, \frac{\nu}{2})$ and

$$\begin{aligned} p(\mathbf{z}|\mathbf{w}, \mathcal{D}, \sigma^2, \nu) &\propto p(\mathbf{y}|\mathbf{w}, \mathbf{z}, \sigma^2)p(\mathbf{z}|\nu)p(\mathbf{w})p(\sigma^2) \\ &\propto p(\mathbf{y}|\mathbf{w}, \mathbf{z}, \sigma^2)p(\mathbf{z}|\nu) \\ &\propto \prod[(z_i)^{-\frac{\nu+1}{2}-1}e^{-\frac{1}{2z_i}(\nu+\frac{(y_i-\mathbf{w}^T\mathbf{x}_i)^2}{\sigma^2})}]. \end{aligned}$$

Therefore, we conclude:

$$p(z_i|\mathcal{D}, \mathbf{w}, \sigma^2, \nu) \sim \mathrm{IG}(\frac{\nu+1}{2}, \frac{1}{2}(\nu + \frac{(y_i - \mathbf{w}^T\mathbf{x}_i)^2}{\sigma^2})).$$

Lastly, to sample from $p(\sigma^2|\mathbf{w}, \mathbf{z}, \mathcal{D}, \nu)$, we compute:

$$\begin{aligned} p(\sigma^2|\mathbf{w}, \mathbf{z}, \mathcal{D}, \nu) &\propto p(\sigma^2)\prod p(y_i|\mathbf{w}, \mathbf{z}, \sigma^2, \nu) \\ &\propto \mathrm{IG}(\sigma^2|\nu_0, S_0)\prod \mathcal{N}(y_i|\mathbf{w}^T\mathbf{x}_i, \sigma^2 z_i)\mathrm{IG}(z_i|\frac{\nu}{2}, \frac{\nu}{2}) \\ &\propto \mathrm{IG}(\nu_0 + \frac{N}{2}, S_0 + \frac{1}{2}(\mathbf{y} - \mathbf{w}^T\mathbf{X})^T \mathrm{diag}(z_i^{-1})(\mathbf{y} - \mathbf{w}^T\mathbf{X})). \end{aligned}$$

24.6.
Assuming a normal prior on $\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{w}, \boldsymbol{\Sigma}_\mathbf{w})$, we can derive the conditional $p(\mathbf{w}|\mathbf{z}, \mathcal{D})$ as follows:

$$p(\mathbf{w}|\mathbf{z}, \mathcal{D}) = p(\mathbf{w}|\mathbf{z}) \propto p(\mathbf{z}|\mathbf{w})p(\mathbf{w})$$

$$\propto e^{-\frac{1}{2}(\mathbf{z}-\mathbf{w}^T\mathbf{X})^T(\mathbf{z}-\mathbf{w}^T\mathbf{X})+(\mathbf{w}-\boldsymbol{\mu}_{\mathbf{w}})^T\boldsymbol{\Sigma}_{\mathbf{w}}(\mathbf{w}-\boldsymbol{\mu}_{\mathbf{w}})}$$
$$\propto \mathcal{N}(\mathbf{w}|\tilde{\boldsymbol{\mu}}_{\mathbf{w}}, \tilde{\boldsymbol{\Sigma}}_{\mathbf{w}})$$

where

$$\tilde{\boldsymbol{\mu}}_{\mathbf{w}} = \tilde{\boldsymbol{\Sigma}}_{\mathbf{w}}(\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\boldsymbol{\mu}_{\mathbf{w}} + \mathbf{X}\mathbf{z})$$
$$\tilde{\boldsymbol{\Sigma}}_{\mathbf{w}} = (\boldsymbol{\Sigma}_{\mathbf{w}}^{-1} + \mathbf{X}^T\mathbf{X})^{-1}.$$

For $p(\mathbf{z}|\mathbf{w}, \mathcal{D})$, we first note that:

$$p(\mathbf{z}|\mathbf{w}, \mathcal{D}) \propto p(\mathbf{y}|\mathbf{z})p(\mathbf{z}|\mathbf{w}, \mathbf{X}) = \prod p(y_i|z_i)p(z_i|\mathbf{w}, \mathbf{x}_i)$$

Therefore, we get that each $p(z_i|\mathbf{w}, \mathbf{x}_i)$ is a truncated normal, given by:
$\mathcal{N}(z_i|\mathbf{w}^T\mathbf{x}_i, 1)\mathbf{1}_{z_i>0}$ if $y_i = 1$,
$\mathcal{N}(z_i|\mathbf{w}^T\mathbf{x}_i, 1)\mathbf{1}_{z_i\leq 0}$ if $y_i = 0$.

24.7.
The posterior density for $\mathbf{z}, \boldsymbol{\lambda}, \mathbf{w}$ and ν is given by:

$$p(\mathbf{z}, \boldsymbol{\lambda}, \mathbf{w}, \nu|\mathbf{y}) \propto p(\nu)\prod_i(\mathbf{1}_{z_i>0}\mathbf{1}_{y_i=1}+\mathbf{1}_{z_i\leq 0}\mathbf{1}_{y_i=0})\lambda_i e^{-\frac{\lambda_i(z_i-\mathbf{w}^T\mathbf{x}_i)^2}{2}}\frac{1}{\Gamma(\frac{\nu}{2})\frac{\nu}{2}^{\frac{\nu}{2}}}\lambda_i^{\frac{\nu}{2}-1}e^{-\frac{\nu\lambda_i}{2}}.$$

Decomposing this distribution, first we get:

$$p(\mathbf{w}|\mathbf{y}, \mathbf{z}, \boldsymbol{\lambda}, \nu) = \mathcal{N}(\mathbf{w}|(\mathbf{X}^T\boldsymbol{\Lambda}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Lambda}\mathbf{z}, (\mathbf{X}^T\boldsymbol{\Lambda}\mathbf{X})^{-1})$$

where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_i)$.
For z_i's, we observe that each of them are conditionally independent, to get:

$$p(z_i|y_i, \mathbf{w}, \lambda_i, \nu) \sim \mathcal{N}(z_i|\mathbf{w}^T\mathbf{x}_i, \frac{1}{\lambda_i})\mathbf{1}_{z_i>0}$$

if $y_i = 1$,

$$p(z_i|y_i, \mathbf{w}, \lambda_i, \nu) \sim \mathcal{N}(z_i|\mathbf{w}^T\mathbf{x}_i, \frac{1}{\lambda_i})\mathbf{1}_{z_i\leq 0}$$

if $y_i = 0$.
For λ_i's, we observe that each of them are conditionally independent, to get:

$$p(\lambda_i|y_i, z_i, \mathbf{w}, \nu) \sim \mathrm{Ga}(\lambda_i|\frac{\nu+1}{2}, \frac{2}{\nu + (z_i - \mathbf{w}^T\mathbf{x}_i)^2}).$$

For ν, we gather remaining components to get:

$$p(\nu|\mathbf{y}, \mathbf{z}, \mathbf{w}, \boldsymbol{\lambda}) \sim p(\nu)\prod_i(\frac{1}{\Gamma(\frac{\nu}{2})\frac{\nu}{2}^{\frac{\nu}{2}}}\lambda_i^{\frac{\nu}{2}-1}e^{-\frac{\nu\lambda_i}{2}}).$$