STATS 305B: Methods for Applied Statistics I **Winter 2016/17**

Lecture 14: February 10

Lecturer: Jonathan Taylor *Scribes: Kenneth Tay*

14.1 Gibbs Sampling for Probit Model

We work through the Gibbs sampler for the probit model to show how data augmentation can make Gibbs sampling easier.

We first specify the model:

- Prior $g(\beta) \propto \exp\left(-\beta^T \Sigma^{-1} \beta / 2\right)$ (i.e. normal),
- Likelihood $P(Y_i = 1 \mid X_i, \beta) = \Phi(X_i^T \beta)$. We can also think of Y_i in terms of a latent variable ε_i:

$$Y_i = \begin{cases} 1 & \text{if } \varepsilon_i \leq X_i^T \beta, \\ 0 & \text{otherwise.} \end{cases}$$

  Here, $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,1)$.

With these two pieces, we can compute the posterior density for β:

$$\begin{aligned} h(\beta \mid Y) &= h(\beta \mid \varepsilon \leq X\beta) \\ &\propto \left[\int_{\{\varepsilon \leq X\beta\}} e^{-\|\varepsilon\|^2/2} d\varepsilon \right] \cdot g(\beta), \\ h(\varepsilon, \beta \mid \varepsilon \leq X\beta) &\propto e^{-\|\varepsilon\|^2/2} \cdot e^{-\beta^T \Sigma^{-1} \beta / 2} 1_{\{\varepsilon \leq X\beta\}}. \end{aligned}$$

Note that if we ignore the indicator variable, the remainder corresponds to the distribution $\mathbb{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} I & 0 \\ 0 & \Sigma \end{pmatrix} \right)$.

Goal: Sample $(\varepsilon, \beta) \mid \varepsilon \leq X\beta$**.**

Recall the Gibbs sampler from the previous lecture. WLOG, assume that $\Sigma = I$. (If not, we can write $\beta = \Sigma^{-1/2}\gamma$ with $\gamma \sim \mathcal{N}(0, I)$.)

Consider what happens to each step of choosing a new ε_i or β_i.

- When moving a particular ε_i, fix ε_{-i} and β. Then the density is $\propto e^{-\varepsilon_i^2/2} 1_{\{\varepsilon_i \leq X_i^T \beta\}}$. We can draw ε_i^{new} from this density, i.e. $\sim \mathcal{N}(0,1) \mid (-\infty, X_i^T \beta)$.

  Note that all the ε_i^{new}'s can be drawn simultaneously (no need to be sequential as in the usual Gibbs sampler)!

- When moving a particular β_i, fix ε and β_{-i}. The constraints from the indicator variable become

$$
\begin{aligned}
\varepsilon_j &\le X_j^T \beta && \text{for } 1 \le j \le n,\\
\varepsilon_j &\le \sum_{l \ne i} X_{jl}\beta_l + X_{ji}\beta_i && \text{for } 1 \le j \le n,\\
X_{ji}\beta_i &\ge \varepsilon_j - \sum_{l \ne i} X_{jl}\beta_l. && \text{for } 1 \le j \le n,
\end{aligned}
$$

These constraints bound β_i on either side, depending on the sign of X_{ji}:

$$
\max_{j:X_{ji}>0} \frac{1}{X_{ji}} \left[\varepsilon_j - \sum_{l \ne i} X_{jl}\beta_l\right] \le \beta_i \le \min_{j:X_{ji}<0} \frac{1}{X_{ji}} \left[\varepsilon_j - \sum_{l \ne i} X_{jl}\beta_l\right],
$$
$$
L(\beta_{-i}, \varepsilon) \le \beta_i \le U(\beta_{-i}, \varepsilon).
$$

We draw β_i^{new} from $\mathcal{N}(0,1) \mid [L, U]$.

14.2 (Agresti 8) Multinomial Regression

Recall the multinomial distribution: If $Y \sim \text{Multinom}(N, \pi)$, where $\pi \in \mathbb{R}_0^k$, $\sum \pi_i = 1$, then it has mass function

$$
f(y_1, \dots, y_k \mid \pi) = \binom{N}{y_1, \dots, y_k} \prod_{i=1}^k \pi_i^{y_i},
$$

supported on $y_1, \dots, y_k \in \mathbb{Z}_0$, $\sum y_i = N$. Note that there are actually only $k-1$ free parameters in this model. If we take the k^{th} category as the baseline category, we can write the mass function as

$$
f(y_1, \dots, y_{k-1} \mid \pi) = \binom{N}{y_1, \dots, y_k} \prod_{i=1}^k \pi_i^{y_i},
$$

where $y_k = N - \sum_{i=1}^{k-1} y_i$, $\pi_k = 1 - \sum_{i=1}^{k-1} \pi_i$. This mass function is supported on $y_1, \dots, y_{j-1} \in \mathbb{Z}_0$, $\sum_{i=1}^{k-1} y_i \le N$.

We can compute the log likelihood for this model:

$$
\begin{aligned}
\log L(\pi_1, \dots, \pi_{k-1} \mid Y) &= C + \sum_{i=1}^{k-1} y_i \log \pi_i + \left(N - \sum_{i=1}^{k-1} y_i\right) \log\left(1 - \sum_{i=1}^{k-1} \pi_i\right)\\
&= C + \sum_{i=1}^{k-1} y_i \log\left(\frac{\pi_i}{1 - \sum_{i=1}^{k-1} \pi_i}\right) + N \log\left(1 - \sum_{j=1}^{k-1} \pi_j\right),
\end{aligned}
$$

where C is some constant that does not depend on π. The identity above shows that this model is an exponential family with sufficient statistic $(y_1, \dots, y_{k-1})$ and natural parameters $\eta_j = \log\left(\dfrac{\pi_j}{1 - \sum_{l=1}^{k-1} \pi_l}\right)$.

We can invert the relationship between η and π to obtain $\pi_j = \dfrac{e^{\eta_j}}{1+\sum_{l=1}^{k-1} e^{\eta_l}}$. This allows us to rewrite the log likelihood in terms of the natural parameters:

$$\log L(\eta_1, \ldots, \eta_{k-1} \mid Y) = \sum_{i=1}^{k-1} \eta_i y_i - N \log \left(1 + \sum_{i=1}^{k-1} e^{\eta_i}\right).$$

14.2.1 Baseline Multinomial Logit

In this model, we have $Y_i \overset{iid}{\sim} \text{Multinom}(N_i, \pi_\beta(X_i))$, where for $1 \le j \le k-1$,

$$[\pi_\beta(X_i)]_j = \frac{\exp(X_i^T \beta_j)}{1+\sum_{l=1}^{k-1} \exp(X_i^T \beta_l)}, \qquad \text{or equivalently,} \qquad \eta_i = X_i^T \beta.$$

Note here that β is a $p \times (k-1)$ matrix. (In logistic regression, β was just a $p \times 1$ matrix.) In the above expression for $[\pi_\beta(X_i)]_j$, β_j refers to the j^{th} column of β.

We can compute the log likelihood and its gradient for this model:

$$\log L(\beta \mid Y) = \sum_{i=1}^{n} \left[\sum_{l=1}^{k-1} Y_{il}(X_i^T \beta_l) - N_i \log \left(1 + \sum_{l=1}^{k-1} e^{X_i^T \beta_l}\right)\right],$$
$$\nabla \log L(\beta \mid Y) = X^T [Y - \mathbb{E}_\beta(Y)] \in \mathbb{R}^{p \times (k-1)}.$$

Note that the gradient of the log likelihood has the exact same form that we had for logistic and loglinear regression.

The Hessian $\nabla^2 \log L(\beta \mid Y)$ is some kind of tensor. It can also be thought of as Cov $(\nabla \log L(\beta \mid Y))$. It is given by

$$[\nabla^2 \log L(\beta \mid Y)]_{ijkl} = \sum_{c=1}^{n} X_{ci} X_{ck} \text{Cov}_\beta(Y_c)_{jl}.$$