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

Lecture 13: February 8

Lecturer: Jonathan Taylor *Scribes: Kenneth Tay*

13.1 (Agresti 7.2) Bayesian Methods for Logistic Regression

In the Bayesian set-up, we have the model

- $X \in \mathbb{R}^{n \times p}$ fixed,
- $\beta \mid X$ distributed according to some prior density $g(\beta)$ (a common prior being $g(\beta) = \mathcal{N}(0, \Sigma)$),
- $Y_i \mid \beta, X_i \overset{iid}{\sim} \text{Bernoulli}(\pi_\beta X_i)$ for $i = 1, \ldots, n$. For the logistic model, $\pi_\beta(x) = \dfrac{e^{X_i^T \beta}}{1 + e^{X_i^T \beta}}$.

We can compute the log likelihood:

$$\log L(\beta \mid Y) = \sum_{i=1}^{n} (X_i^T \beta) Y_i - \log(1 + e^{X_i^T \beta}),$$

which in turn allows us to compute the log posterior density:

$$\begin{aligned} \log h(\beta \mid Y) &\propto \text{log likelihood } + \text{ log prior} \\ &= \log L(\beta \mid Y) + \log g(\beta) \\ &= \sum_{i=1}^{n} \left[(X_i^T \beta) Y_i - \log(1 + e^{X_i^T \beta}) \right] - \frac{\beta^T \Sigma^{-1} \beta}{2} + c(\Sigma), \end{aligned}$$

where $c(\Sigma)$ is some constant depending only on Σ, and which gets wiped out when the posterior is normalized.

13.2 Sampling from Posterior $g(\beta \mid Y)$

Often times, we will want to compute functions w.r.t. to the posterior density, i.e. $\int_{\mathbb{R}^p} f(\beta) h(\beta \mid Y) d\beta$ for some function f. In some cases, the posterior as calculated above is "nice enough" that we can use it directly. However, it is often difficult to do so.

Instead of trying to compute $h(\beta \mid Y)$ directly, we can try to get a sample distribution $\{\beta^1, \ldots, \beta^T\}$ which approximates the true posterior distribution of β. Once we have this sample, we have the approximation

$$\int_{\mathbb{R}^p} f(\beta) h(\beta \mid Y) d\beta \approx \frac{1}{T} \sum_{t=1}^{T} f(\beta^t).$$

MCMC methods allow us to obtain such a sample of β. We will discuss 2 of these methods: Metropolis-Hastings and Gibbs Sampling.

With MCMC methods, we may choose to use the sample only after some point ("burning in") so that the estimate is not affected by our initialization of the algorithm:

$$\int_{\mathbb{R}^p} f(\beta)h(\beta \mid Y)d\beta \approx \frac{1}{T-K}\sum_{t=K+1}^{T} f(\beta^t).$$

We may also use "thinning", i.e. using only every r^{th} entry as there may be strong correlation between successive samples:

$$\int_{\mathbb{R}^p} f(\beta)h(\beta \mid Y)d\beta \approx \frac{1}{(T-K)/r}\sum_{j=1}^{(T-K)/r} f(\beta^{K+rj}).$$

13.2.1 Metropolis-Hastings

The basic version of the Metropolis-Hastings algorithm is as follows:

1. Initialize at some β^0, set $\beta_{cur} = \beta^0$.
2. (Loop) For as many samples as you want:
   (a) "Propose" a new β:
   $$\beta_{proposed} = \beta_{cur} + \varepsilon,$$
   where ε follows some distribution, e.g. $\varepsilon \sim \mathcal{N}(0, \tilde{\Sigma})$. (Here, $\tilde{\Sigma}$ is some fixed covariance matrix.)
   (b) Compute the acceptance probability
   $$\alpha = \min\left(\frac{h(\beta_{proposed} \mid Y)}{h(\beta_{cur} \mid Y)}, 1\right).$$
   (c) Generate a random variable $X = \text{Bernoulli}(\alpha)$. If $X = 1$, add $\beta_{proposed}$ to the sample and set $\beta_{cur} = \beta_{proposed}$. If $X = 0$, add β_{cur} to the sample and leave β_{cur} as is.

The theorem is that for suitable proposals, the stationary distribution of this algorithm is the posterior $g(\beta \mid Y)$.

13.2.2 Gibbs Sampling

The basic version of the Gibbs sampler is as follows:

1. Initialize at some β^0, set $\beta_{cur} = \beta^0$. Write $\beta_{cur} = (\beta_{1,cur}, \ldots, \beta_{p,cur})$.
2. (Loop) For as many samples as you want:
   (a) For $1 \leq j \leq p$, define the density
   $$g_j(v_j) := g\left((\beta_{1,cur}, \ldots, \beta_{j-1,cur}, v_j, \beta_{j+1,cur}, \ldots, \beta_p) \mid Y\right).$$
   Draw X from g_j and set $\beta_{j,cur} = X$.
   (b) Add β_{cur} to the sample.

Note that there is no need for a proposal in this set-up.

13.2.2.1 Data Augmentation

For logistic regression, data augmentation can make the Gibbs sample simpler.

Recall that we have an equivalent model for logistic regression:

- $X \in \mathbb{R}^{n \times p}$ fixed,
- $\beta \mid X$ distributed according to some prior density $g(\beta)$,
- $\varepsilon_i \mid \beta, X_i \overset{iid}{\sim}$ Logistic with ε_i independent of β and X_i, and
- $Y_i = I\{\varepsilon_i \leq X_i^T \beta\}$.

With this set-up,

$$\begin{aligned} \text{Posterior } g(\beta \mid Y) &= g(\beta \mid \{\varepsilon \leq X^T \beta\}) \\ &\propto \left[\int_{\mathbb{R}^n} \prod_i I\{\varepsilon_i \leq X_i^T \beta\} f(\varepsilon_i) d\varepsilon_i \right] g(\beta), \end{aligned}$$

where f is the density for the logistic distribution. Thus, the Gibbs sampler for (ε, β) becomes easier: each β step amounts to picking from the prior distribution truncated for some interval.