Bayesian Probit Model[1]

Let there be I choices. For the purpose of the following discussion, it is useful to state the model from a latent variable perspective. Let y_i^* be an unobserved random continuous variable which is drawn from a normal density with a mean parameterized by a linear predictor $\mathbf{x}_i\boldsymbol{\beta}$ and an variance term σ. $\mathbf{x}_i$ is a vector of length K and $\boldsymbol{\beta}$ a vector of K parameters.

We only observe the latent variable in terms of a binary state: If y_i^* is strictly smaller than 0, we observe $y_i = 0$, for values larger than τ, we observe $y_i = 1$. The probability model of this can be written as:

$$\begin{aligned} y_i^* &\sim \mathbf{N}(\mathbf{x}_i\boldsymbol{\beta}, \sigma) \\ y_i &= \begin{cases} 0 & \text{if } y_i^* < 0 \\ 1 & \text{otherwise} \end{cases} . \end{aligned} \tag{1}$$

In order to identify the model σ is usually set to one. To complete the Bayesian model, it is necessary to specify a prior distribution for the coefficients. It is usually assumed that they are from a multivariate normal distribution:

$$\boldsymbol{\beta} \sim \mathbf{N}(\mathbf{b}_0, \mathbf{B}_0). \tag{2}$$

The posterior distribution can not be marginalized analytically. A Gibbs sampler was first suggested by Albert and Chib (1993) and can be motivated using data augmentation and the directed acyclic graph (DAG) of the model.

A data augmented DAG representation of the model appears in figure 1. Each node is a random variable. Rectangular nodes indicate observed variables, circle nodes represent unobserved variables. An arrow indicates the dependencies between these variables. The DAG helps to derive the full conditionals for the Gibbs sampler. Lauritzen et al. (1990) have shown that the conditional densities of any node $(\theta_1, \ldots .\theta_j, \ldots \theta_J)$ in a DAG $\mathcal{G}$ can be written as:

$$f(\theta_j|\theta_{\neg j}) \propto f(\theta_j|\text{parents}[\theta_j]) \times \prod_{w \in \text{chidren}[\theta_j]} f(w|\text{parents}[w]), \tag{3}$$

where $\theta_{\neg j}$ denotes all nodes in $\mathcal{G}$ other than θ_j.

[1] Notes by Moritz Marbach. Comments welcome.

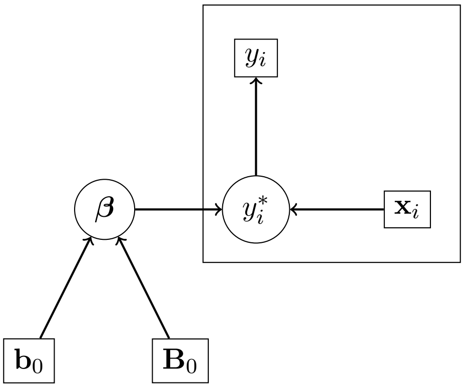

Figure 1: A directed acyclic graph representation of a probit model.

Given the DAG in figure 1, we can derive the full conditional densities for the two unobserved variables:

$$\begin{aligned} f(\boldsymbol{\beta}|\mathbf{b}_0,\mathbf{B}_0,\mathbf{y}^*,\mathbf{y},\mathbf{X}) &= f(\boldsymbol{\beta}|\mathbf{b}_0,\mathbf{B}_0) \times \prod_{i=1}^{n} f(y_i^*|\mathbf{x}_i,\boldsymbol{\beta}) \\ f(y_i^*|\mathbf{b}_0,\mathbf{B}_0,\boldsymbol{\beta},\mathbf{y},\mathbf{X}) &= f(y_i^*|\mathbf{x}_i,\boldsymbol{\beta}) \times f(y_i|y_i^*). \end{aligned} \tag{4}$$

The first density is a product of a normal prior density and the likelihood of n normal densities. It is known to have a closed form:

$$f(\boldsymbol{\beta}|\mathbf{b}_0,\mathbf{B}_0,\mathbf{X},\mathbf{y}^*) = \phi\Big((\mathbf{B}_0^{-1}+\mathbf{X}'\mathbf{X})^{-1}(\mathbf{B}_0^{-1}\mathbf{b}_0+\mathbf{X}'\mathbf{y}^*), (\mathbf{B}_0^{-1}+\mathbf{X}'\mathbf{X})^{-1}\Big), \tag{5}$$

where $\mathbf{X}$ is the matrix of all covariate vectors $\mathbf{x}_i$ and similar, $\mathbf{y}^*$ the vector of all y_i^*. $\phi(.,.)$ is a normal p.d.f. The proof is standard. The conditional density in the second equation is:

$$f(y_i^*|\mathbf{x}_i,\boldsymbol{\beta},y_i) = \begin{cases} \phi(\mathbf{x}_i\boldsymbol{\beta})\mathrm{T}(y_i^* \leq 0) & \text{if } y_i = 0 \\ \phi(\mathbf{x}_i\boldsymbol{\beta})\mathrm{T}(y_i^* > 0) & \text{if } y_i = 1 \end{cases}, \tag{6}$$

where $\phi(.)$ is a normal p.d.f. with a variance parameter $\sigma^2 = 1$ and $\mathrm{T}(.)$ is a truncation function. The Gibbs Sampler works from iterative sampling of the two densities.

The mixing of this sampler is not very good though. I better mixing can be obtained by marginal data augmentation. The alternative Gibbs sampler below corresponds to scheme 1 in Imai and van Dyk (2005, p. 317-318) that generalizes an earlier scheme in van Dyk and Meng (2001). See also Jackman (2009, 390) for a summary.

We define the following prior distributions:

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &\sim \mathrm{N}(0, \mathbf{B}_0) \\ \alpha^2 &\sim \alpha_0^2/\chi^2_{v_0}, \end{aligned} \tag{7}$$

where v_0 (degrees of freedom), α_0^2 (tuning parameter) and $\mathbf{B}_0$ (covariance) are hyper paramteres. χ^2_{df} is the χ^2_{df}-distribution with df deegres of freedom. Notice, that the prior mean for $\tilde{\boldsymbol{\beta}}$ has to be set to zero and the prior precision matrix is defined over the the unidentified version of $\boldsymbol{\beta}$ (I denote unidentified paramters with a tilde). The following MCMC scheme can be employed, denoting values from iteration t with superscripts '(t)'. van Dyk and Meng (2001) show that $v_0 = 0$ (which induces an improper prior on α^2) leads to the most efficient algorithm in a sense that the autocorrelation is reduced most. In practical applications, Imai and van Dyk (2005) use the sets ($v_0 = 3, \alpha_0^2 = 3$) and ($v_0 = 6, \alpha_0^2 = v_0$).

Define the following constant:

$$\mathbf{B}_1 = (\mathbf{B}_0^{-1} + \mathbf{X}'\mathbf{X})^{-1}$$

At each iteration t do:

1. Draw α^2 from its prior:

$$\begin{aligned}\alpha^2 &\sim \alpha_0^2/\chi^2_{v_0} \quad \text{or, equivalently}\\ \alpha^2 &\sim \text{invGamma}(v_0/2, \alpha_0^2/2)\end{aligned}$$

2. Draw $\tilde{y}_i$ for each i:

$$\tilde{y}_i^{*(t)} \sim \begin{cases} \text{N}(\alpha\mathbf{x}_i\boldsymbol{\beta}^{(t-1)}, \alpha^2)\text{T}(\tilde{y}_i^{*(t)} \leq 0) & \text{if } y_i = 0 \\ \text{N}(\alpha\mathbf{x}_i\boldsymbol{\beta}^{(t-1)}, \alpha^2)\text{T}(\tilde{y}_i^{*(t)} > 0) & \text{if } y_i = 1 \end{cases},$$

3. Discard the draw α^2 and draw a new α^2, this time from:

$$\begin{aligned}\alpha^2 &\sim \alpha_1^2/\chi^2_{n+v_0} \quad \text{or, equivalently}\\ \alpha^2 &\sim \text{invGamma}(v_1/2, \alpha_1^2/2),\end{aligned}$$

where:

$$\begin{aligned}v_1 &= v_0 + n\\ \alpha_1^2 &= \sum_{i=1}^{n}(\tilde{y}_i^{*(t)} - \mathbf{x}_i\tilde{\mathbf{b}}_1)^2 + \alpha_0^2 + \tilde{\mathbf{b}}_1'\mathbf{B}_0^{-1}\tilde{\mathbf{b}}_1\\ \tilde{\mathbf{b}}_1 &= \mathbf{B}_1\mathbf{X}'\tilde{\mathbf{y}}^{*(t)}\end{aligned}$$

4. Draw $\tilde{\boldsymbol{\beta}}$

$$\tilde{\boldsymbol{\beta}} \sim \text{N}(\tilde{\mathbf{b}}_1, \alpha^2\mathbf{B}_1)$$

Set $\boldsymbol{\beta}^{(t)} = \tilde{\boldsymbol{\beta}}/\alpha$ and discard α^2. Repeat T times until convergence.

References

Albert, J. H. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association 422*(88), 669–679.

Imai, K. and D. A. van Dyk (2005). A bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of Econometrics 124*(2), 311–334.

Jackman, S. (2009). *Bayesian Analysis for the Social Sciences*. New York: Wiley.

Lauritzen, S. L., A. P. Dawid, B. N. Larsen, and H.-G. Leimer (1990). Independence properties of directed markov fields. *Networks 20*(5), 491–505.

van Dyk, D. A. and X.-L. Meng (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics 10*(1), 1–50.