Gibbs Sampling for the Probit Regression Model with Gaussian Markov Random Field Latent Variables

Mohammad Emtiyaz Khan
Department of Computer Science
University of British Columbia

May 8, 2007

Abstract

We consider a binary probit model where the latent variable follow a Gaussian Markov Random Field (GMRF). Our main objective is to derive an efficient Gibbs sampler for the above model. For this purpose, we first review two Gibbs samplers available for the classical probit model with one latent variable. We find that the joint update of variables increases the rate of convergence. We use these results to derive Gibbs samplers for the probit model with GMRF latent variables. We discuss three different approaches to Gibbs sampling for the above model. The first two approaches are direct extensions of the Gibbs sampler for the classical probit model. The third approach involves a slight modification in the probit model and suggests that it may be possible to block sample all its variables at once.

1 Introduction

In this report, we derive Gibbs samplers for the probit regression model with Gaussian Markov Random Field Latent variables. The probit models are very useful techniques in statistics, and has found many applications. Various sampling methods exist in the literature for inference using this model [4] [3]. The classical probit model assumes only one latent variable associated with the measurements, and hence doesn't take the spatial correlation into account. We consider a more general case where there are multiple latent variables and are correlated with each other. There are standard samplers available in literature (for example in [6]), however there is still a lot of scope to improve on these samplers.

2 Probit Model

In this section, we consider the classical version of probit model. Given data $\{y_t\}_{t=1}^T$ with each $y_t \in \{0, 1\}$, we have the following binary regression model:

$$Pr(y_t = 1|\beta) = \Phi(x_t\beta) \tag{1}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function, $\beta \in \mathcal{R}^d$ is the (unknown) latent variable, and $x_t \in \mathcal{R}^{1\times d}$ is a (known) row vector. Further, we define $X = [x_1', \ldots, x_T']'$ and $y = [y_1, \ldots, y_T]$. The task is to infer β given the data $\{y, X\}$.

The likelihood is given as follows:

$$\mathcal{L}(\beta|y, X) = \prod_{t=1}^{T} \Big[\Phi(x_t\beta)\Big]^{y_t} \Big[1 - \Phi(x_t\beta)\Big]^{(1-y_t)} \tag{2}$$

We can see that the form of the likelihood is quite complicated because of the presence of the non-linear function Φ. A simpler formulation can be obtained by using auxiliary variables as described in [3]. We introduce the auxiliary variables $\{z_t\}_{t=1}^T$, and get the following equivalent model:

$$z_t = x_t\beta + \epsilon_t \tag{3}$$

$$y_t = \begin{cases} 1 & \text{if } z_t > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where each $z_t \in \mathcal{R}$ and $\epsilon_t \sim \mathcal{N}(0, 1)$ (i.i.d. in t)[1]. Fig. 2 shows the graphs for these two models. The advantage of using the above form is that it allows us to perform Gibbs sampling. We now describe and compare the two Gibbs samplers for the above model. We will now discuss the Gibbs samplers described in [3] and [4]. Our purpose is to compare these samplers, which will helps us to understand the modifications necessary for GMRF latent variable model. There are methods based on the Metropolis-Hastings algorithm. While implementing a random walk Metropolis-Hastings algorithm, we found that it is sensitive to the variance increments which can be estimated by using methods like maximum-likelihood. However this will create a problem for high dimensional model as a good estimate of variance is difficult to obtain in that case. For this reason, in this study we focus only on Gibbs sampling. Also from here on, we denote a whole time-series by the variable itself, for example, $z = \{z_1, \ldots, z_T\}$.

2.1 Gibbs sampler I - Albert and Chib (A&C)

The Gibbs sampler described in [3] involves iterative sampling of $\beta|z$ and $z|\beta$. Assuming a normal prior on the β, $\beta \sim \mathcal{N}(\mu_\beta, \Sigma_\beta)$, the conditional distribution of β given z is a normal distribution (derivation in Appendix A:

$$p(\beta|z) = \mathcal{N}(\beta; \tilde{\mu}_\beta, \tilde{\Sigma}_\beta) \tag{5}$$

[1] To denote a normal random variable z, we will use $\mathcal{N}(z; \mu, \sigma^2)$ or $\mathcal{N}(\mu, \sigma^2)$ interchangeably.

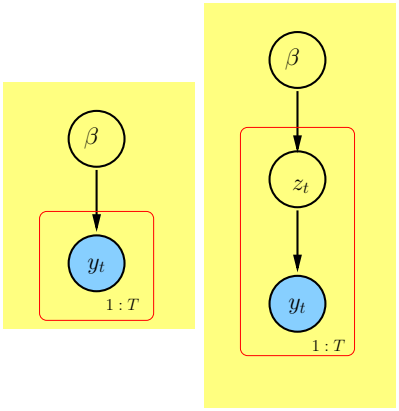

Figure 1: Graphs for the Probit model with and without auxiliary variables

$$\tilde{\Sigma}_\beta = (\Sigma_\beta^{-1} + X'X)^{-1} \tag{6}$$

$$\tilde{\mu}_\beta = \tilde{\Sigma}_\beta(\Sigma_\beta^{-1}\mu_\beta + Xz) \tag{7}$$

The conditional distribution of z given β is a truncated normal (derivation in Appendix A:

$$p(z_t|\beta, y_t, x_t) \propto \begin{cases} \mathcal{N}(z_t; x_t\beta, 1)\mathcal{I}(z_t > 0, y_t = 1) \\ \mathcal{N}(z_t; x_t\beta, 1)\mathcal{I}(z_t \leq 0, y_t = 0) \end{cases} \tag{8}$$

where $\mathcal{I}(A)$ is an indicator variable for an event A. Sampling from a truncated normal distribution is quite easy in the case when the mean of the parent normal distribution is close to zero. Various efficient methods are discussed in [5].

For some models, β and z could be highly correlated, which causes the algorithm to converge slowly. This is in fact a serious issue for high-dimensional model as we will see in Section 2.3. It is well known that block sampling increases the mixing of the variables and hence the rate of convergence [4]. We next describe the a Gibbs sampler which may be useful in such situations.

2.2 Gibbs sampler II - Holmes and Held (H&H)

The Gibbs sampler described in [4] makes use of the following factorization:

$$p(\beta, z|y) = p(\beta|z)p(z|y) \tag{9}$$

The conditional distribution of $\beta|z$ (first term in above equation) is given by Eq.(5) derived in previous section. To obtain $p(z|y)$, we will have to integrate β from the joint of z and β:

$$p(z|y) = p(y|z)p(z) = p(y|z)\int_\beta p(z|\beta)p(\beta)d\beta \tag{10}$$

The term under integral is derived in Appendix A, and given in Eq.(42). Substituting in the above equation and integrating over β, we get:

$$p(z|y) = p(y|z)\mathcal{N}(z; 0, P^{-1}) = \mathcal{I}_v(y, z)\mathcal{N}(z; 0, P^{-1}) \tag{11}$$

where $P = I_T - X\tilde{\Sigma}_\beta X'$, and $\mathcal{I}_v(y, z)$ is the *vectored* version of $\mathcal{I}$ defined earlier. Hence $p(z|y)$ is a multivariate truncated normal distribution of dimension T. It is now possible to first sample $z|y$ using the above distribution, then sample β given z using Eq. (5).

There is a serious issue here. It is usually difficult to sample from a high-dimensional truncated normal distribution. An efficient method to do so is described in [4]. However the derivation of the method is not[2]. Hence we skip details of the procedure. The method works in practice and we have used it for our implementation.

2.3 Comparison of Gibbs samplers I and II

In this section, we compare the two Gibbs samplers discussed earlier. Our main criteria for comparison is the rate of convergence. We present results on two data sets: Pima Indians Diabetes Database [2] and the MNIST digit recognition data [1]. The Pima dataset is of lower dimensionality than the MNIST data, and this will allow us to compare the rate of convergence in two different situations. We expect that for high-dimensional data there should be a significant difference in the rate of convergence for the two methods.

For the Pima dataset, we have $T = 768$ measurements (i.e y_t), with 8 attributes (i.e. x_t). The measurement $y_t = 1$ is interpreted as "tested positive for diabetes." The problem is to find a "common cause" of the disease (which is β in the case of a probit model). We run the two samplers for 1000 iterations with the prior $\mathcal{N}(0, I_8)$ on β. Fig. 2 shows the mean of β and samples of β_1 found using the two Gibbs samplers. We can see that the means are almost the same for both the methods. There is also not much difference in the mixing of the two variables, however H&H shows slightly better mixing than A&C. Histograms of β_1 and β_7 are also shown in Fig. 2 which are almost the same. We see that for this low-dimensional data, there is not much difference in the performance of the two samplers. However it seems that the H&H is better in mixing. Next we present result for the high dimensional data and we will see that H&H is indeed better in mixing.

We consider the 189 images of digits 2 and 3 in the MNIST dataset, each of size 16×16. We vectorize the images into vectors of length 256 (i.e. $d = 256$). Our goal is to find β which is of size 256. We run the two Gibbs samplers for 300 iterations with the prior $\mathcal{N}(0, I_{256})$ on β. Fig. 3 shows the mean of β after 300 iterations. Both of the methods converge to the same image, but results for H&H are better than A&C[3]. This is further confirmed in Fig. 4 which shows

[2]The technical report quoted in [6] page 159, for the derivation is not accessible to the author till the date this report is written

[3]The reader may argue whether this is a "good" feature image, however classification with

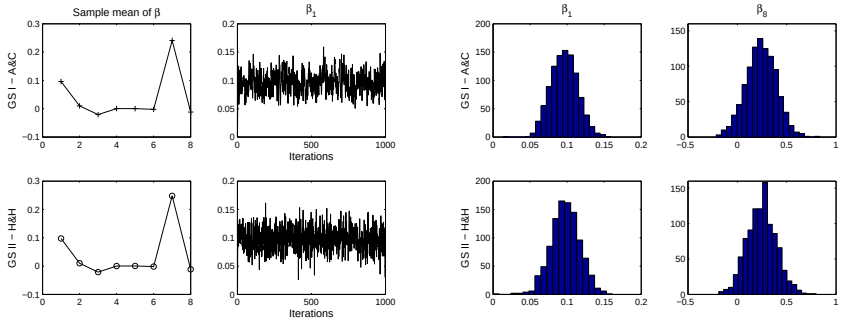

Figure 2: (Right) Mean of β and samples for β_1 with A&C and H&H. (Left) Histogram of β_1 and β_7 with A&C and H&H

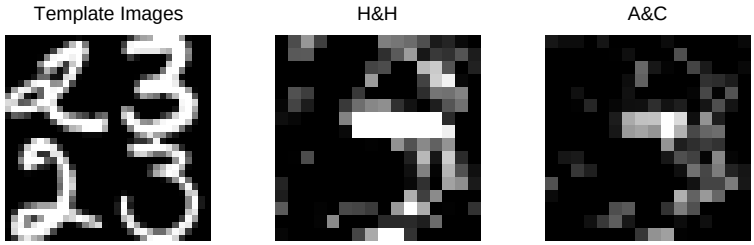

Figure 3: Template images and mean of β with A&C and H&H

the convergence of the mean with iterations. It is clear that H&H converges much faster than A&C. Finally Fig. 4 shows the histogram of some instances of β. We can clearly see that mixing in H&H is better than A&C.

3 A probit model with GMRF latent variables

We now consider an extension of the probit model such that when β is a Gaussian Markov Random Field (GMRF). Let $\{\beta_t\}_{t=1}^T$ be a GMRF with the following distribution:

$$p(\beta_1, \ldots, \beta_T|\alpha) \propto \exp\Big[-\frac{1}{2} \sum_{t=2}^{T} \frac{(\beta_t - \beta_{t-1})^2}{\alpha} + \frac{1}{\alpha}\beta_1'\beta_1 \Big] \tag{12}$$

test results (not presented in this report) show that the this β gives satisfactory classification results

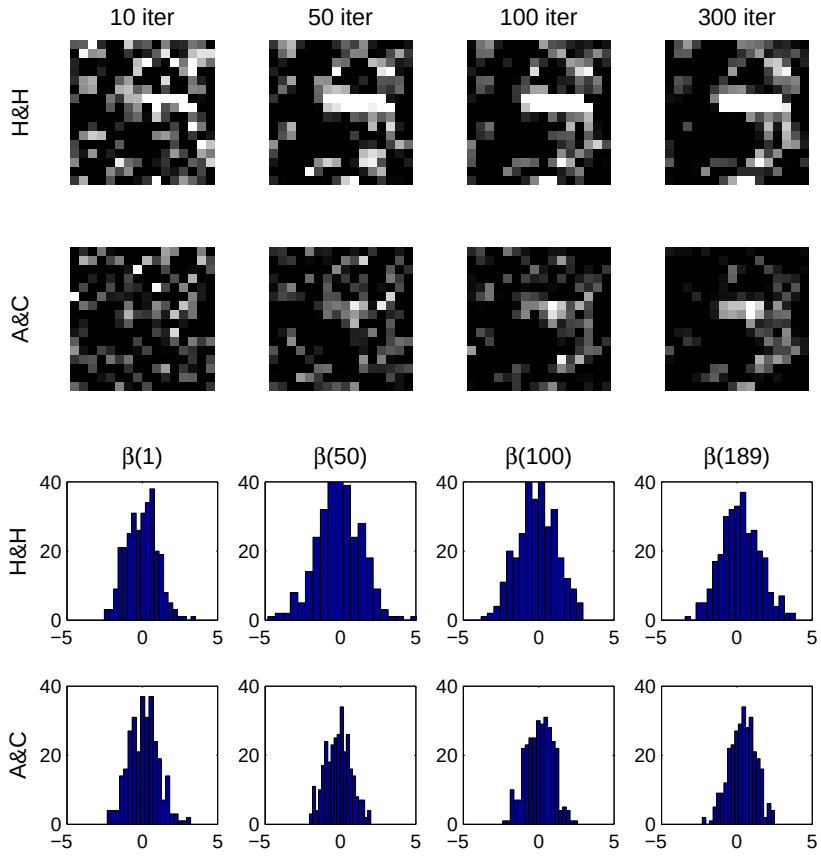

Figure 4: (Top) Convergence of mean (Bottom) Histogram of few instances of β

where each $\beta_t \in \mathcal{R}^d$ and $\alpha > 0$ is a parameter for the covariance of the conditional distribution $\beta_t|\beta_{t-1}$ (we will refer to it as the variance of the GMRF). There are other extensions of the above model with higher order dependence or cyclic dependence (for example see [6], page 109). These distributions can be obtained with slight modifications in the above equation. The distribution can be more conveniently expressed in the form of a Gaussian distribution :

$$p(\beta) \propto \mathcal{N}(0, \alpha Q^{-1}) \tag{13}$$

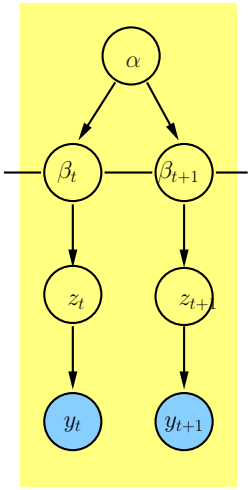

Figure 5: Graph of an auxiliary variable probit model with GMRF latent variables

where Q is a tridiagonal matrix of size $Td \times Td$,

$$Q = \begin{pmatrix} 2I & -I & 0 & \dots & 0 & 0 \\ -I & 2I & -I & \dots & 0 & 0 \\ 0 & -I & 2I & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 2I & -I \\ 0 & 0 & 0 & \dots & -I & I \end{pmatrix} \tag{14}$$

where I is an identity matrix of size $d \times d$. Given β, we define an auxiliary variable probit model as follows:

$$z_t = x_t\beta_t + \epsilon_t \quad \text{where } \epsilon_t \sim \mathcal{N}(0,1) \tag{15}$$

$$y_t = \begin{cases} 1 & \text{if } z_t > 0 \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

We also assume an inverse gamma[4] prior on variance $\alpha \sim \mathcal{IG}(a,b)$. Fig. 3 shows the graph of the above model. β in the classical probit model is now replaced with a process correlated in t. Each of the β_t is correlated with an auxiliary variable. As dimensionality is increased, (compared to the probit model discussed in Section 2), we are likely to face more trouble with mixing in Gibbs sampling. Hence we would like to have more block sampling than single site sampling.

[4]We assume the form $\mathcal{IG}(\alpha,\beta)$ for Inverse Gamma

3.1 Gibbs sampler I

In this section, we describe a Gibbs sampler similar to A&C. This is discussed in [6]. We first sample $z|\beta, \alpha, y$ and then block sample $\beta, \alpha|z, y$. The conditional distribution of $z|\beta$ is same as before (Eq. (8)) except for the fact that the mean of the truncated normal random variable at time t depends on β_t:

$$p(z_t|\beta_t, y_t) \propto \begin{cases} \mathcal{N}(x_t\beta_t, 1)\mathcal{I}(z_t > 0, y_t = 1) \\ \mathcal{N}(x_t\beta_t, 1)I(z_t \leq 0, y_t = 0) \end{cases} \tag{17}$$

To derive the conditional distribution of $\beta, \alpha|z, y$, we first define:

$$\tilde{X} = \begin{pmatrix} x_1 & 0 & \dots & 0 \\ 0 & x_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & x_T \end{pmatrix} \tag{18}$$

The size of $\tilde{X}$ is $T \times Td$. Then $p(z|\beta, y) = \mathcal{N}(\tilde{X}\beta, 1)\mathcal{I}_v(z, y)$.

$$\begin{aligned} p(\beta, \alpha|z, y) &\propto p(z|y, \beta, \alpha)p(\beta|\alpha)p(\alpha) && (19) \\ &\propto p(y|z)\exp\Big[-\frac{1}{2}\Big\{(z - \tilde{X}\beta)'(z - \tilde{X}\beta) + \frac{1}{\alpha}\beta' Q\beta\Big\}\Big]p(\alpha) && (20) \\ &= \exp\Big[-\frac{1}{2}\Big\{(\beta - \mu)'\Sigma(\beta - \mu) + z'\tilde{P}z\Big\}\Big]p(\alpha)p(y|z) && (21) \\ &\propto \mathcal{N}(\beta; \mu, \Sigma) \cdot \mathcal{IG}(\alpha; a, b) \cdot \mathcal{N}(z; 0, \tilde{P}^{-1})\mathcal{I}_v(z, y) && (22) \end{aligned}$$

where $\Sigma^{-1} = \alpha^{-1}Q + \tilde{X}'\tilde{X}$, $\mu = \Sigma\tilde{X}'z$, $\tilde{P} = I_T - \tilde{X}\Sigma\tilde{X}'$. The third step is in the above derivation is obtained by completing squares for β. We sample from α and then β using the following joint distribution:

$$p(\beta, \alpha|z, y) = \mathcal{N}(\beta; \mu, \Sigma) \cdot \mathcal{IG}(\alpha; a, b) \tag{23}$$

We can see that the sampling β using this distribution is not efficient as it involves the inversion of Σ which will be very high dimensional. A better way is to use Kalman smoother to find the mean and the variance of marginal $p(\beta_t|z, \alpha)$.

We apply the above algorithm to Tokyo rainfall data. The data set is a record of daily rainfall during the years 1983 and 1984, and we wish to infer the the underlying probability of rainfall for each day (so β is of size 366). Also for $t = 60$ (February 29) there is only one measurement available. So we have total $T = 366 + 365 = 731$ measurements. We use the following GMRF for represent the β:

$$p(\beta|\alpha) \propto \exp\Big[-\frac{1}{2}\sum_{t=3}^{T}\frac{(\beta_t - 2\beta_{t-1} + \beta_{t-2})^2}{\alpha} + \frac{1}{\alpha}(\beta_2'\beta_2 + \beta_1'\beta_1)\Big] \tag{24}$$

The two measurements for each day are denote by y_{i1} and y_{i2} for which we have two auxiliary variables z_{i1} and z_{i2}, and conditional distribution of these

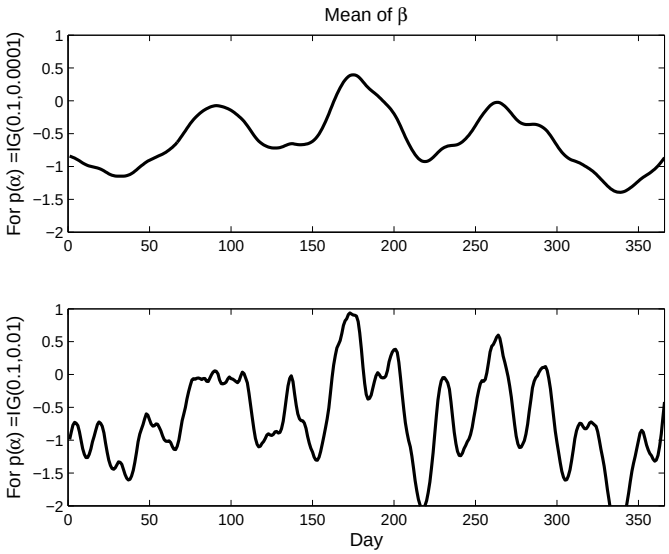

Figure 6: Mean of β for different priors on α.

variables given β is given as follows:

$$z_{i1} = \beta_t + \epsilon_t^1 \quad (25)$$

$$z_{i2} = \beta_t + \epsilon_t^2 \quad (26)$$

where ϵ_t^k are standard Normal distribution.

Fig. 6 shows the results for 10000 iterations. We used two different priors for $\alpha\mathcal{IG}(0.1, 0.0001)$ $\mathcal{IG}(0.1, 0.0001)$ (The first prior is used in [6]). We see that given a good prior the algorithm converges, however it is sensitive to the prior. This is expected as sampling from prior is not a good idea. When β and α are not strongly correlated, it is more sensible to sample $\alpha|\beta$ and then block sample $\beta, z|\alpha, y$ using a method similar to H&H. In the next section we describe two other possibilities for Gibbs sampler.

3.2 Gibbs sampler II

A possible approach is to follow H&H approach. We block sample $z, \beta|y, \alpha$ and then sample $\alpha|\beta$. The conditional distributions are given as follows:

$$p(\alpha|\beta) \propto p(\beta|\alpha)p(\alpha) = \mathcal{IG}(\alpha; a, b + \frac{1}{2}\beta' Q\beta) \quad (27)$$

$$p(z, \beta|y, \alpha) \propto p(\beta|z, \alpha)p(y|z) = \mathcal{N}(\beta; \mu, \Sigma) \cdot \mathcal{N}(z; 0, \tilde{P}^{-1})\mathcal{I}_v(z, y) \quad (28)$$

where the last step is obtained using Eq. (22). Similar to Section 2.2, we have a high-dimensional truncated normal random variable, and we need an efficient

method to sample from this distribution (similar to the method discussed in [4]).

3.3 Gibbs sampler III

Another approach can be to sample from the joint distribution of $z, \beta, \alpha|y$. However sampling from that distribution is intractable because of introduction of α. One approach can be to modify the auxiliary variable model by making z_t dependent on α as follows :

$$z_t = x_t\beta_t + \sqrt{\alpha}\epsilon_t \tag{29}$$

This means that the variance of z_t is now α. Now the probit link corresponding to this model is $\Phi(X_t\beta_t/\sqrt{\alpha})$. This means that the probit link is a function of $\beta_t/\sqrt{\alpha}$. This definitely puts a serious problem on the model identification, as all β and α pair with same ratio will correspond to the same model. Although we are not clear on this issue, we outline the procedure here to demonstrate that this modification makes sampling from joint possible.

The joint of $(\beta, \alpha, z)|y$ is given by the following equation:

$$\begin{aligned}
p(\beta, \alpha, z|y) &\propto p(y|z)p(z|\beta, \alpha)p(\beta|\alpha)p(\alpha) && (30)\\
&= \mathcal{I}_v(y, z) \cdot \mathcal{N}(z; \tilde{X}\beta, \alpha I_T) \cdot \mathcal{N}(\beta; 0, \alpha Q^{-1}) \cdot \mathcal{IG}(\alpha; a, b) && (31)\\
&\propto \mathcal{I}_v(y, z) \cdot \mathcal{N}(z; 0, \alpha\bar{P}^{-1}) \cdot \mathcal{N}(\beta; \bar{\mu}, \alpha\Sigma) \cdot \mathcal{IG}(\alpha; a, b) && (32)
\end{aligned}$$

Last step is obtained by completing the squares for β. We have $\bar{\Sigma} = Q + \tilde{X}'\tilde{X}$, $\bar{mu} = \bar{\Sigma}\tilde{X}'z, \bar{P} = I_T - \tilde{X}\bar{\Sigma}\tilde{X}'$. Integrating out β, we get:

$$\begin{aligned}
p(\alpha, z|y) &= \mathcal{I}_v(y, z) \cdot \mathcal{N}(z; 0, \alpha\bar{P}^{-1}) \cdot \alpha^{-d/2} \cdot \mathcal{IG}(\alpha; a, b) && (33)\\
&= \mathcal{I}_v(y, z) \cdot \alpha^{-T/2} \exp\Big[-\frac{1}{2\alpha}z'Pz\Big] \cdot \mathcal{IG}(\alpha; a, b) && (34)\\
&= \mathcal{I}_v(y, z) \cdot \mathcal{IG}\Big(\alpha; a + \frac{T+d}{2}, b + \frac{1}{2}z'Pz\Big) && (35)
\end{aligned}$$

In the above steps we have just collected the terms for α and found the inverse gamma distribution. Integrating out α, we get $p(z|y)$. Hence we get the following conditional distributions:

$$\begin{aligned}
p(z|y) &= \mathcal{I}_v(y, z) \cdot (b + \frac{1}{2}z'Pz)^{-(a+(T+d)/2)} && (36)\\
p(\alpha|z) &= \mathcal{IG}\Big(\alpha; a + \frac{T+d}{2}, b + \frac{1}{2}z'Pz\Big) && (37)\\
p(\beta|\alpha, z) &= \mathcal{N}(\beta; \bar{\mu}, \alpha\Sigma) && (38)
\end{aligned}$$

Till now it is not clear to us if it is easy to sample from the $p(z|y)$, which is a multi-dimensional truncated distribution. However these distributions can be used to get a block update $(z, \beta, \alpha)|y$, which will increase the mixing in the algorithm. We hope to do more work on this in future.

4 Conclusions

In this report, we reviewed and derived Gibbs samplers for classical probit model. We found that joint update of variables increase the rate of convergence. We discussed three different approaches for Gibbs sampling of the probit model with GMRF latent variable. For the second approach, we need to find an efficient way of sample from the truncated multi-dimensional Normal random variable. Similarly, for the third approach we need to find how does the modification affect the probit model and whether it's possible to sample from the distribution given by Eq. 36.

Acknowledgment I would like to thank Kevin Murphy and Arnaud Doucet for their guidance, and Mark Schmidt for his code.

A Derivation of $p(\beta|z)$

This section describes the derivation done in Section 2.1. First we derive $p(\beta|z)$:

$$\begin{aligned} p(\beta|z) &\propto p(z|\beta)p(\beta) && (39)\\ &\propto \exp\Big[-\frac{1}{2}\Big\{(z-x_t\beta)^2+(\beta-\mu_\beta)'\Sigma_\beta(\beta-\mu_\beta)\Big\}\Big] && (40)\\ &= \exp\Big[-\frac{1}{2}\Big\{(\beta-\tilde{\mu}_\beta)'\tilde{\Sigma}_\beta(\beta-\tilde{\mu}_\beta)+z'Pz\Big\}\Big] && (41)\\ &\propto \mathcal{N}(\beta;\tilde{\mu}_\beta,\tilde{\Sigma}_\beta)\mathcal{N}(z;0,P^{-1}) && (42)\end{aligned}$$

where the third step is just completing the squares for β. Here $\tilde{\Sigma}_\beta$ and $\tilde{\mu}_\beta$ are as described in Eq. (6),(7) and $P = I_T - X\tilde{\Sigma}_\beta X'$. The distribution of z will be useful for the derivation of Gibbs sampler II described in Section 2.2.

For $p(z|\beta, y, X)$, we first note that :

$$p(z|\beta,y,X) \propto p(y|z)p(z|\beta,X) = \prod_{t=1}^{T} p(y_t|z_t)p(z_t|\beta,x_t) \quad (43)$$

Hence for each t, z_t a truncated version of standard Normal distribution with mean $x_t\beta$.

References

[1] MNIST digit recognition data. http://yann.lecun.com/exdb/mnist/.

[2] Pima Indians Diabetes Database. http://www.ics.uci.edu/ mlearn/databases/pima-indians-diabetes/pima-indians-diabetes.names.

[3] J.H. Albert and S. Chib. Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

[4] C.C. Holmes and L. Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1):145–168, 2006.

[5] C.P. Robert. Simulation of truncated normal variables. *Statistics and Computing*, 5(2):121–125, 1995.

[6] H. Rue and L. Held. *Gaussian Markov random fields*. Chapman & Hall/CRC, 2005.