DM3 : Probabilistic Graphical models 2019/2020

AL HOUCINE KILANI
SALMANE NAOUMI

January 9, 2020

EMAILS: al_houceine.kilani@ens-paris-saclay.fr salmane.naoumi@ens-paris-saclay.fr

1 Gibbs sampling and mean field VB for the probit model

1.1 Question 1 :

Since we are choosing $\beta \sim \mathcal{N}(0, \tau I_p)$, we are assigning a variance τ on each one of the predictors. But the data are not on the same scale and the predictors do not necessarily have the same units. Thus, normalizing the data force them to fall on the same scale.

1.2 Question 2 :

This does not change the problem formulation because we can always rewrite it this way :
Let

$$\epsilon'_i \sim \mathcal{N}(0, \sigma^2)$$

Then

$$\epsilon'_i \sim \sigma^2 \mathcal{N}(0, 1)$$

$$\epsilon'_i \sim \sigma^2 \epsilon_i$$

Thus

$$y_i = sgn(\beta'^T x_i + \epsilon'_i)$$

$$y_i = sgn(\beta'^T x_i + \sigma^2 \epsilon_i)$$

Since $\sigma^2 > 0$, we then have :

$$y_i = sgn(\frac{\beta'}{\sigma^2}^T x_i + \epsilon_i)$$

Thus, rewriting $\beta = \frac{\beta'}{\sigma^2}$:

$$y_i = sgn(\beta^T x_i + \epsilon_i)$$

1.3 Question 3 :

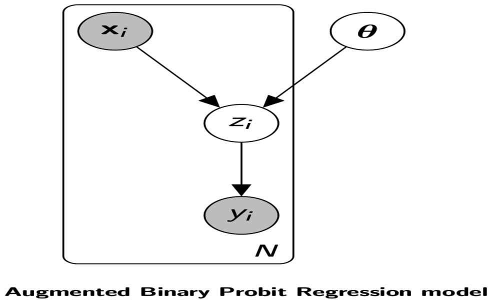

Figure 1: Graphical model of the problem ($\theta = \beta$ in our case), source = [1]

We are interested in computing the joint posterior distribution of $(z, \beta|Y, X)$. Since it is hard to do so, we are going to approximate the sampling from this distribution using Gibbs Sampling.

For this we need the conditional distributions : $(z|\beta, Y, X)$ and $(\beta|z, Y, X)$. But given this graphical model, we have that $(\beta|z, Y, X) = (\beta|z, X)$ (once we know z we know y so there is no need for y)

Since

$$\epsilon_i \sim \mathcal{N}(0, 1)$$

we have

$$(z_i|\beta, x_i) \sim \mathcal{N}(\beta^T x_i, 1)$$

and we have :

$$p(y_i|z_i) = 1(y_i = 1)1(z_i > 0) + 1(y_i = 0)1(z_i <= 0)$$

Thus

$$\begin{aligned} p(\beta|z, Y, X) &= p(\beta|z, X) \\ &= p(\beta|z, X) \\ &\propto p(\beta|X)p(z|\beta, X) \\ &= p(\beta)p(z|\beta, X) \\ &\propto e^{\frac{-||\beta||^2}{2\tau}} \prod_i e^{\frac{-(z_i-\beta^T x_i)^2}{2}} \\ &= e^{\frac{-||\beta||^2}{2\tau}} e^{\sum_i \frac{-(z_i-\beta^T x_i)^2}{2}} \\ &= e^{\frac{-||\beta||^2}{2\tau}} e^{\sum_i \frac{-(z_i-\beta^T x_i)^2}{2}} \\ &= exp(\frac{-1}{2}(\frac{\beta^T\beta}{\tau} + (z - X\beta)^T(z - X\beta))) \\ &= exp(\frac{-1}{2}(\beta^T \frac{I}{\tau}\beta + z^T z - z^T X\beta - \beta^T X^T z + \beta^T X^T X\beta)) \\ &= exp(\frac{-1}{2}(\beta^T(\frac{I}{\tau} + X^T X)\beta + z^T z - z^T X\beta - \beta^T X^T z)) \end{aligned} \tag{1}$$

If we say that $(\beta|z,Y,X) \sim \mathcal{N}(\mu,\Sigma)$ then by identification with :

$$(\beta-\mu)^T\Sigma^{-1}(\beta-\mu) = \beta^T\Sigma^{-1}\beta - \beta^T\Sigma\mu - \mu^T\Sigma^{-1}\beta + \mu^T\Sigma^{-1}\mu$$

We get that :

$$\begin{gathered} \Sigma^{-1} = \frac{I}{\tau} + X^TX \\ \beta^T\Sigma^{-1}\mu = \beta^TX^Tz \\ \\ \mu = \Sigma X^Tz \end{gathered} \tag{2}$$

And we have :

$$\begin{aligned} p(z|\beta,Y,X) &\propto p(z|\beta,X)p(Y|z,X,\beta) \\ &= p(z|\beta,X)p(Y|z) \\ &= \prod_i p(z_i|\beta,x_i)p(y_i|z_i) \\ &= \prod_i e^{\frac{-(z_i-\beta^Tx_i)^2}{2}}[1(y_i=1)1(z_i>0)+1(y_i=0)1(z_i<=0)] \\ &= e^{\sum_i \frac{-(z_i-\beta^Tx_i)^2}{2}} \prod_i 1_{y_iz_i>0} \end{aligned} \tag{3}$$

This distribution is a truncated version of the normal distribution according to the value of y_i.
Now we can perform Gibbs sampling of $(z,\beta|Y,X)$ using $(z|\beta,Y,X)$, $(\beta|z,Y,X)$ following this scheme :

- Initialize z_0, compute Σ
- for t = 0..T
  - Compute μ_t
  - Sample $\beta_t \sim p(\beta|z_t,Y,X) = p(\beta|z_t,X) = \mathcal{N}(\mu_t,\Sigma)$
  - Sample $z_{t+1} \sim p(z|\beta_t,Y,X) = \mathcal{N}(X\beta_t,I)$ truncated given the sign of Y
  - If we are in the predicting phase : We take the values of β we sampled, we predict y given the sign of z (where we use β.

We split the data using a ratio train/test of 75%/25%.

We then perform a Maximum a posteriori prediction for the y_i using the samples of β we got after the burn-in. We achieve a score of 77% of accuracy.

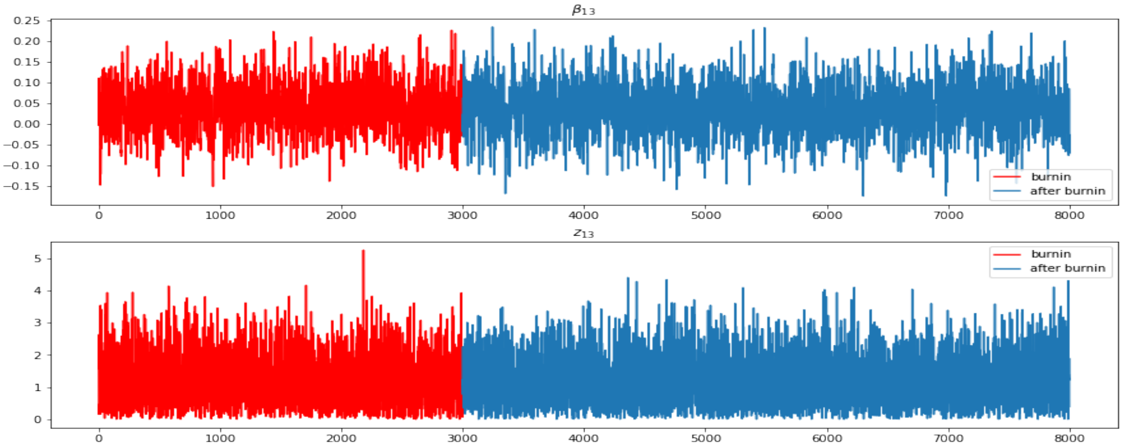

Figure 2: Values taken by the sampled parameters β and z

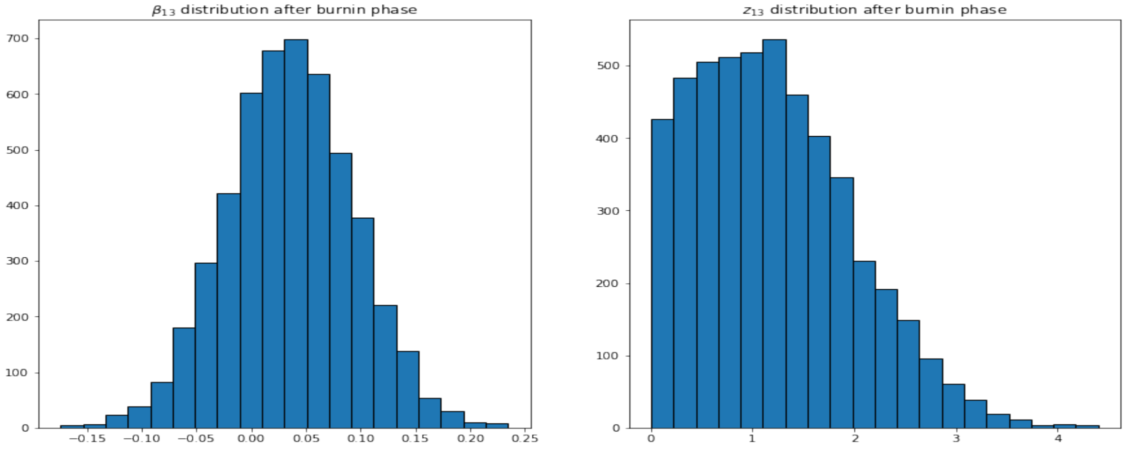

Figure 3: histogram of values taken by the sampled parameters β and z

1.4 Question 4 :

Same as before, we want to model the joint posterior probability $p(\beta, z|X, y)$ but now we will approximate it with a distribution $q(\beta, z)$. Furthermore, we are going to suppose the mean-field factorization :

$$q(\beta, z) = q_1(\beta)q_2(z)$$

The optimal form for each factor is [2]:

$$q_1^*(\beta) \propto \exp[E_{q_2(z)}(log(p(\beta, z, y|X)))]$$
$$q_2^*(\beta) \propto \exp[E_{q_1(\beta)}(log(p(\beta, z, y|X)))]$$

We have that :

$$\begin{aligned} log(p(\beta, z, y|X)) &= log(p(y|z, \beta, X) + log(p(\beta, z|y, X)) \\ &= log(p(y|z, \beta, X)) + log(p(z|\beta, y, X)) + log(p(\beta|y, X)) \\ &= log(p(y|z)) + log(p(z|\beta)) + log(p(\beta)) \\ &= log(p(y|z)) - \frac{||z - X\beta||^2}{2} - \frac{||\beta||^2}{2\tau} \end{aligned} \tag{4}$$

Thus we get that :

$$\begin{aligned} log[q_1^*(\beta)] &= E_{q_2(z)}(log(p(\beta, z, y|X))) + C_0 \\ &= E_{q_2(z)}[log(p(y|z)) - \frac{||z - X\beta||^2}{2} - \frac{||\beta||^2}{2\tau}] + C_0 \\ &= E_{q_2(z)}[log(p(y|z))] - E_{q_2(z)}[\frac{||z - X\beta||^2}{2}] - E_{q_2(z)}[\frac{||\beta||^2}{2\tau}] + C_0 \\ &= C_1 - \frac{1}{2}E_{q_2(z)}[||z - X\beta||^2] - \frac{||\beta||^2}{2\tau} \\ &= C_1 - \frac{1}{2}(E_{q_2(z)}[z^T z] - 2E_{q_2(z)}[z](X\beta)^T + (X\beta)^T(X\beta) - \frac{||\beta||^2}{2\tau} \\ &= C_2 + E_{q_2(z)}[z](X\beta)^T - \frac{1}{2}(X\beta)^T(X\beta) - \frac{||\beta||^2}{2\tau} \\ &= C_2 + E_{q_2(z)}[z]\beta^T X^T - \frac{1}{2}\beta^T X^T X\beta - \frac{1}{2\tau}\beta^T\beta \\ &= C_2 + E_{q_2(z)}[z]\beta^T X^T - \frac{1}{2}\beta^T(X^T X + \frac{I}{\tau})\beta \end{aligned} \tag{5}$$

where we put into constants C_x the terms that do not depend on β. By identification with $V \sim \mathcal{N}(\mu, \Sigma)$, we have :

$$log(p(V)) = \frac{-(V - \mu)^T\Sigma^{-1}(V - \mu)}{2} + ...$$
$$log(p(V)) = \frac{-1}{2}(V^T\Sigma^{-1}V - 2V^T\Sigma^{-1}\mu + \mu^T\Sigma^{-1}\mu) + ...$$

Thus :

$$\Sigma = (X^T X + \frac{I}{\tau})^{-1}$$
$$\mu = \Sigma X^T E_{q_2(z)}[z]$$

We thus conclude that :

$$q_1^*(\beta) \sim \mathcal{N}(\mu, \Sigma)$$

We secondly have :

$$\begin{aligned} log[q_2^*(z)] &= E_{q_1(\beta)}(log(p(\beta, z, y|X))) + C_0 \\ &= E_{q_1(\beta)}[log(p(y|z)) - \frac{||z - X\beta||^2}{2} - \frac{||\beta||^2}{2\tau}] + C_0 \\ &= log(p(y|z)) - \frac{1}{2}E_{q_1(\beta)}[z^T z] + XE_{q_1(\beta)}[\beta]z^T + E_{q_1(\beta)}[(X\beta)^T(X\beta)] + C_1 \\ &= \sum_{i=1}^{n} 1(y_i = 1)1(z_i > 0) + 1(y_i = 0)1(z_i <= 0) - \frac{z^T z}{2} + XE_{q_1(\beta)}[\beta]z^T + C_2 \\ &= \sum_{i=1}^{n} 1(y_i = 1)1(z_i > 0) + 1(y_i = 0)1(z_i <= 0) - \frac{1}{2}(||z - XE_{q_1(\beta)}[\beta]||^2) + C_3 \end{aligned} \tag{6}$$

We thus see that :

$$q_2^*(z) \sim T\mathcal{N}(XE_{q_1(\beta)}[\beta], I)$$

Finally, to perform the sampling we compute the following steps :

- Initialize β^0
- for t = 1..T :
  - Update $q_2^*(z)$ using β^{t-1}
  - Sample z^t using $q_2^*(z)$
  - Update $q_1^*(\beta)$ using z^t
  - Sample β^t using $q_1^*(\beta)$

In terms of speed Figure 4, this sampling is way faster than Gibbs sampling (it converges in less than 200 iterations and we got 5000 samples in less than a second whereas we need 17 seconds for Gibbs sampler). Furthermore, we got to see that Figure 5 the variance of the parameters we got with this approach is smaller than the one for Gibbs sampler (Cf Question 5). But this came with a slight trade-off in performance (71% MAP prediction accuracy).
In fact, we can see that the model predict all the instances as good credit which shows that the model does not take into account the imbalance of the dataset.

1.5 Question 5 :

The reason why mean-field inference underestimates the variance of the posterior is because it is designed to minimize $KL(q(z)||p(z|X)$ with respect to $q(z)$ [4].

The KL divergence is an expectation under q(z). This means that the objective penalizes regions of latent variable space where q(z) is high, but does not care what happens where q(z) is very low. This means that it has no problem fitting a low variance estimate to the posterior distribution.

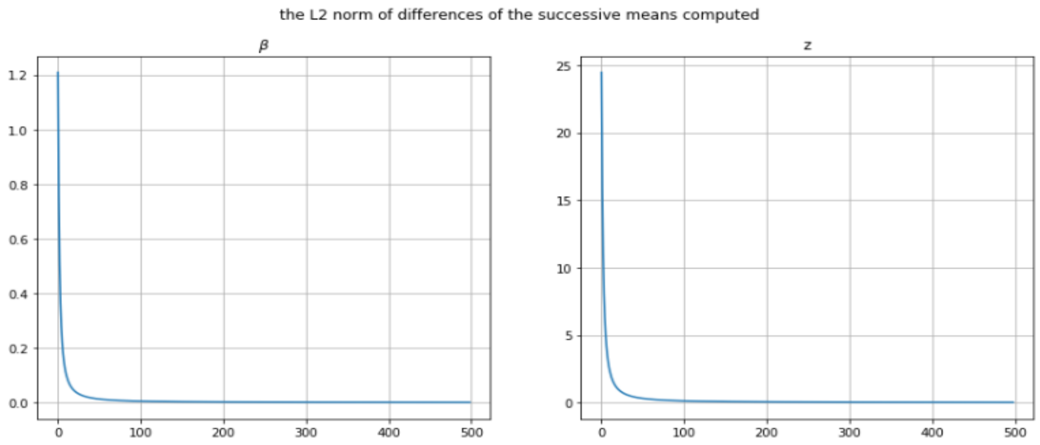

Figure 4: Convergence of the means : $||\beta^t - \beta^{t-1}||$ and $||z^t - z^{t-1}||$

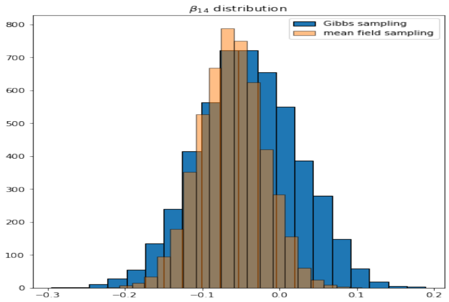

Figure 5: Comparison

1.6 Question 6:

If there is complete separation of the data points, the maximum likelihood estimation is not possible since a finite estimate of β does not exist as shown by Albert and Anderson [5]. In fact, there are infinite estimates, which causes a divergence of estimators and standard errors of iterative solutions such as Newton-Raphson method that will iterate until the bound on the number of iterations is reached.

For experiment, we run Gibbs Sampling on the following 2 blobs dataset with full separation and we got an accuracy of 100%. But even if we observe a unnaturally large regression coefficient for β_1, the posterior means for all coefficients are kind of stable because they potentially incorporate the information from the prior distribution.

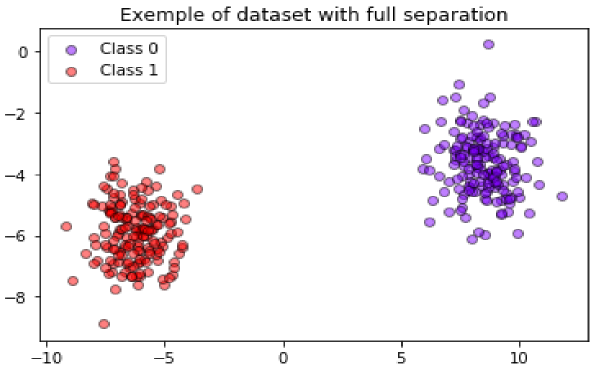

Figure 6: 2 blobs Dataset with complete separation

References

1] https://rstudio-pubs-static.s3.amazonaws.com/208180_b659633007eb45aa9c48e4c50b8afc07.html
[2] http://keyonvafa.com/variational-inference-probit-regression/
[3] https://rstudio-pubs-static.s3.amazonaws.com/348023_0eb459d58e2a494b8855a3b2fe36212c.html
[4] https://www.quora.com/Why-and-when-does-mean-field-variational-Bayes-underestimate-variance
[5] Albert, A. and Anderson, J. A. (1984). “On the Existence of Maximum Likelihood Estimates in Logistic Regression Models.” Biometrika, 71(1): 1–10. MR0738319. doi: https://doi.org/10.1093/biomet/71.1.1. 360, 362, 363