Homework 3

1 Gibbs sampling and mean field VB for the probit model

1.1 Q1: Data scaling

- We add a constant column in order to take into account the bias parameter in the model.
- We need to scale data because if we use distances, the prediction will be incorrect otherwise.

1.2 Q2: Variance of ϵ_i

The probit model is the following

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

with $\epsilon_i \sim (0, 1)$.

If one considers $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, then one can write $\epsilon_i = \sigma\epsilon'_i$ where $\epsilon'_i \sim (0, 1)$, and we obtain,

$$y_i = sgn(\beta^T x_i + \epsilon_i) = sgn(\beta^T x_i + (\sigma\epsilon'_i)) = sgn(\frac{1}{\sigma}\beta^T x_i + \epsilon'_i) = sgn(\beta'^T x_i + \epsilon'_i),$$

where $\beta' = \beta/\sigma$

1.3 Q3. Gibbs sampling

1.3.1 Distributions computation

$$y_i = sign(z_i)$$
$$z_i = \beta^T x_i + \epsilon_i$$
$$P(y_i = 1/\beta) = \Phi(\beta^T x_i)$$

with $\epsilon_i \sim N(0, 1)$ and $\beta \sim N_p(\mu, I_p/\tau)$

We have then the joint distribution of β, z, y :

$$p(\beta, z, y) = p(\beta)p(z/\beta)p(y/\beta, z)$$

with $z_i/\beta \sim N(\beta^T x_i, 1)$ and $p(y_i/\beta, z_i) = \prod_{i=1}^{n} \mathbf{1}_{\{} y_i z_i > 0\}$

By applying the Bayes rule we have:

$$p(\beta/z_i)\infty p(\beta)p(z_i/\beta) = \exp(-\frac{1}{2}[\frac{||\beta||^2}{\tau} + \sum_{i=1}^{n}(z_i - \beta^T x_i)^2])$$

Similarly,

$$p(z_i/\beta, y_i)\infty p(z_i/\beta)p(y_i, z_i/\beta) = \exp(-\frac{\sum_{i=1}^{n}(z_i-\beta^T x_i)^2}{2})\prod_{i=1}^{n}\mathbf{1}_{\{y_i z_i} > 0\}$$
$$= \exp(-\frac{\sum_{i=1}^{n}(||Z-\beta^T X||^2}{2})\prod_{i=1}^{n}\mathbf{1}_{\{y_i z_i} > 0\} \quad \text{with} \quad Z = (z_1, z_2, .., z_n)^T \quad \text{and} \quad X = (x_1, x_2, .., x_n)^T$$

Then, we can deduce that, $\boxed{\beta/z \sim N(\hat{\mu}, \hat{\Sigma})}$ with $\hat{\mu} = \hat{\Sigma}X^T z$ and $\hat{\Sigma} = (\frac{1}{\tau}I_p + X^T X)^{-1}$

Furthermore, $\forall i, (z_i|\beta, y_i)$ is sampled from **truncated normal distribution** with $mean = \beta^T x_i$ and variance=1 and a $support = \{z \in R^p, y_i z_i > 0\}$

1.3.2 Results discussion

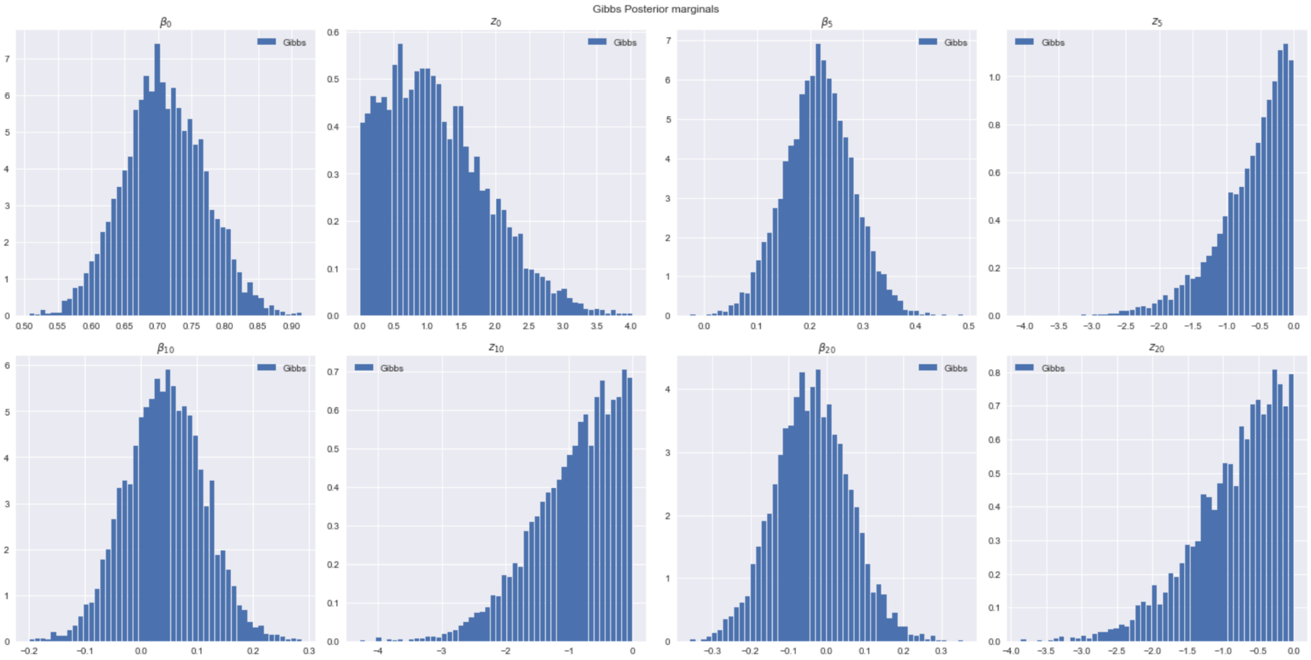

Figure 1: Gibbs Posterior Marginals

The figure above show the results of Gibbs sampling that generates 1000 samples and projects these results by showing some components of the estimate β and of the latent variable z. One can notice that β coefficient has a gaussian shape, proved in the previous section, and the components of z have a gaussian shape with hidde/truncated part. These results are then validated by the theory.

1.4 Q4. Mean Field Variational algorithm

1.4.1 Distributions computation

In the Mean Field algorithm we want to approximate the prior distribution $p(\beta, z|X, y)$ from the set

$$Q = \{q, q(\beta, z) = q_b(\beta)q_z(z)\}$$

According the Variational theory, we need to

$$q_b^* = argmax_{q \in Q} KL(q||q_z)$$

and

$$q_z^* = argmax_{q \in Q} KL(q||q_b)$$

with

$$KL(q||p) = E_q[\log(p)]$$

Therefore the optimal solutions are the following:

$$\log(q_b^*) = E_{q_z^*}[\log(p(z, \beta, y)] + const$$

$$\log(q_z^*) = E_{q_b^*}[\log(p(z, \beta, y)] + const$$

1. Compute $\log(p(z, \beta, y))$:

   $$\log(p(z, \beta, y) = \log(p(\beta)) + \log(z|\beta) + \log(p(y|z)) = -\frac{1}{2\tau}||\beta||^2 - \frac{1}{2}||z - \beta^T X||^2 + const_0$$

2. Compute $\log(q_b^*)$:

   $\log(q_b^*(\beta)) = E_{q_z^*}[\log(p(z, \beta, y)] + const$

   $=-E_{q_z^*}[\frac{1}{2\tau}||\beta||^2] - E_{q_z^*}[\frac{1}{2}||z - \beta^T X||^2] + E_{q_z^*}[\log(p(y|z))] + const_1$

   $= -E_{q_z^*}[\frac{1}{2\tau}||\beta||^2] - E_{q_z^*}[\frac{1}{2}||z - \beta^T X||^2] + const_2$

   However $E_{q_z^*}[\frac{1}{2\tau}||\beta||^2] = \frac{1}{2\tau}||\beta||^2$

   and $E_{q_z^*}[\frac{1}{2}||z - \beta^T X||^2] = \frac{1}{2}[E_{q_z^*}[||z||^2] + E_{q_z^*}[||\beta^T X||^2] - 2E_{q_z^*}[zX^T\beta]]$

   $= \frac{1}{2}(||\beta^T X||^2 - 2E_{q_z^*}[z]X^T\beta) + const_3$

Let's note $\hat{z} = E_{q_z^*}[z]$

Therefore,

$$\log(q_b^*(\beta)) = \hat{z}X^T\beta - \frac{1}{2}\beta(X^TX + \frac{1}{\tau}\mathbf{I})\beta^T + const_4$$

And by introducing a new variable $\tilde{\beta} = \hat{(}X^TX + \frac{1}{\tau}\mathbf{I})^{-1}zX^T = \hat{\Sigma}zX^T$

We conclude that:

$\boxed{q_b^*(\beta) \sim N(\tilde{\beta}, \hat{\Sigma})}$

3. Compute $\log(q_z^*)$:

   $\log(q_z^*(z)) = E_{q_b^*}[\log(p(z,\beta,y)] + const = -E_{q_b^*}[\frac{1}{2\tau}||\beta||^2|y,z] - E_{q_b^*}[\frac{1}{2}||z - \beta^TX||^2|y,z] + E_{q_b^*}[\log(p(y|z)|y,z] + const_0$

   However,

$$E_{q_b^*}[||\beta||^2|y,z] = const^*$$

$$E_{q_b^*}[||z - \beta^TX||^2|y,z] = E_{q_b^*}[||z||^2] - 2z^TX^TE_{q_b^*}[\beta] + const^{**}$$

$$E_{q_b^*}[\log(p(y|z)|y,z] = \sum_{i=1}^{n}\log(\mathbf{1}_{\{y_iz_i>0\}})$$

   Therefore,

$$\log(q_z^*(z)) = \frac{-1}{2}(E_{q_b^*}[||z||^2|z,y] - 2z^TX^TE_{q_b^*}[\beta|z,y]) + \sum_{i=1}^{n}\log(\mathbf{1}_{\{y_iz_i>0\}}) + const_1$$

   Let's denote $\beta^* = E_{q_b^*}[\beta|z,y]$

   then, $\boxed{q_z^*(z) \sim TN(X^T\beta^*, I)}$ the truncated normal distribution of the approximated law of z

1.4.2 Results discussion

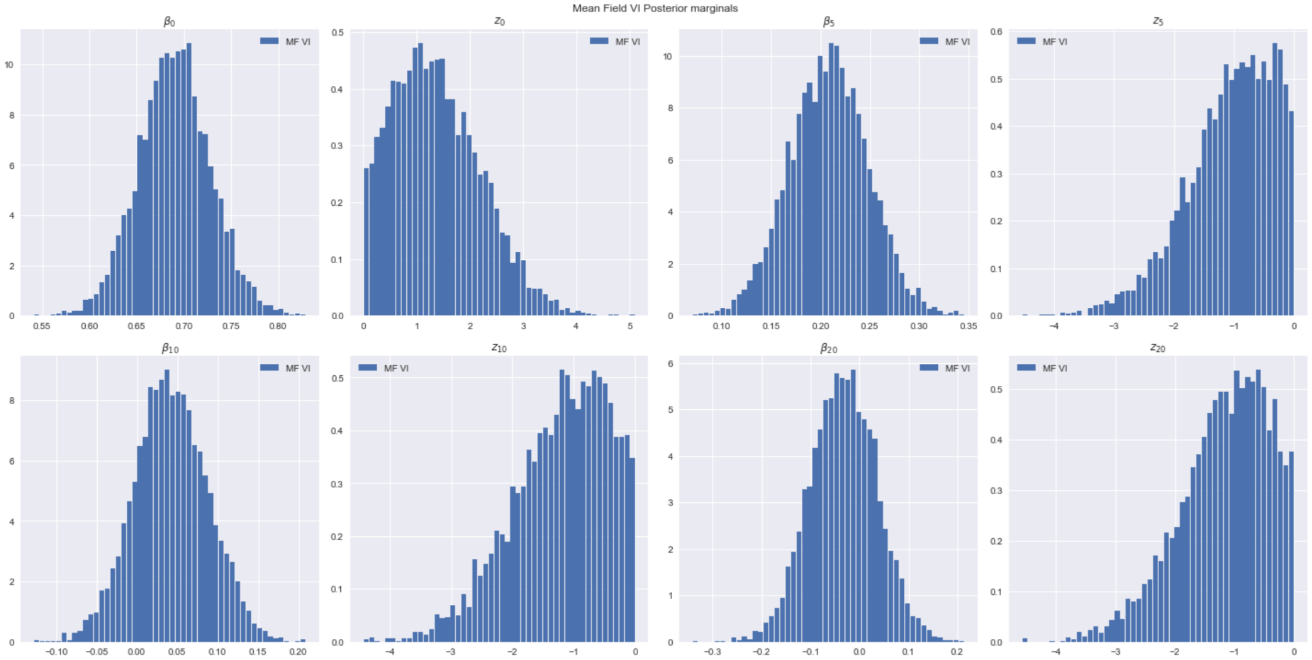

Figure 2: Mean Field Posterior Marginals

The figure above show the results of Mean Field Sampling that generates 5000 samples and projects these results by showing some components of the estimate β and of the latent variable z. Similarly, the results shown are validated by the previous computation.

1.4.3 Results comparison with Gibbs sampling

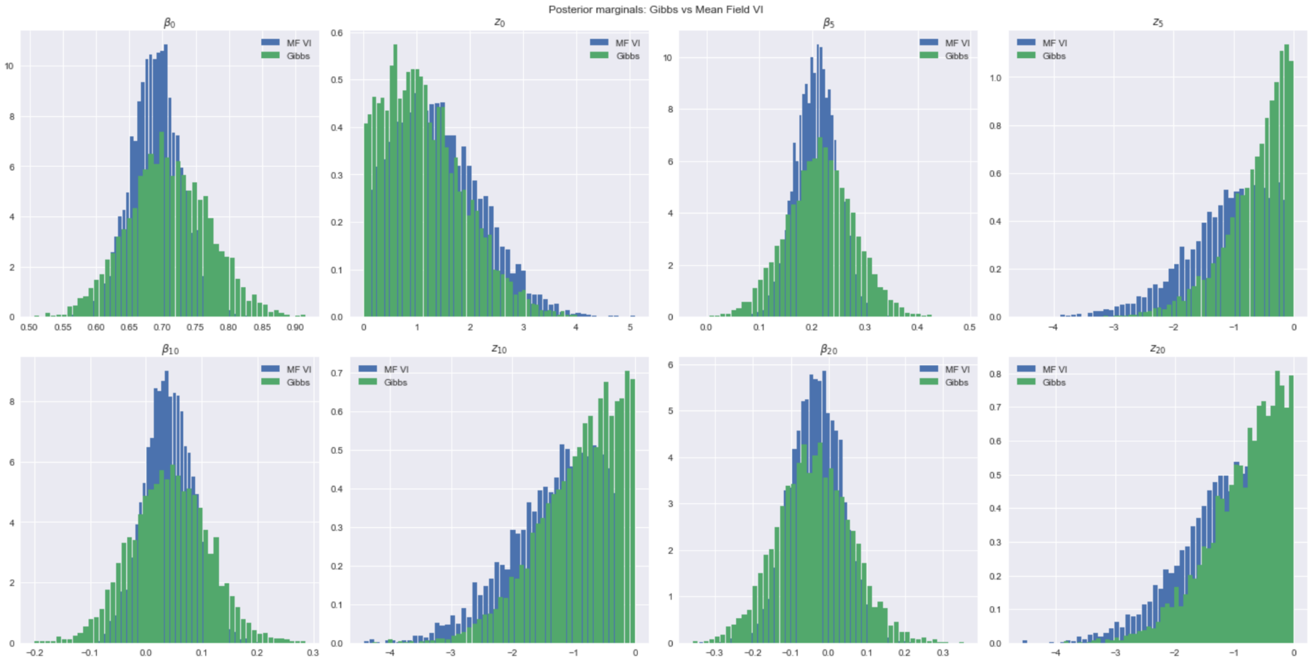

Figure 3: Comparison of Posterior Marginals: Mean Field vs Gibbs sampling

This figure shows the results of both Gibbs sampling and Mean Field for the same number of samples n= 5000. We notice first that both algorithms returns almost the same mean values for both β and z. However, the variance of β generated but Gibbs sampling seems to be lower that the variance of Mean Field. Furthermore, accuracy of Gibbs is 0.78 and the algorithm convergences in 2.985 s whereas Mean Field gives an accuracy of 0.77 and converges in 0.34. Mean Field Variational algorithm is more efficient than Gibbs sampling.

1.5 Q.6 Application

1.5.1 Data generation

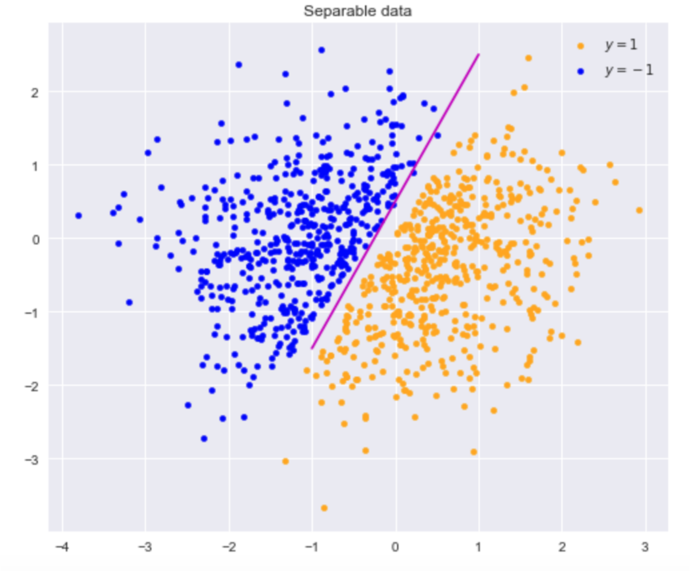

Figure 4: Generated Dataset: linearly seperable

We chose to simulate a separated dataset using 1000 data points ($n = 1000$) and in 2 dimensions ($p = 2$) and we labelled the dataset using the following line:

$$\beta^{*T}x + \beta_0 = 0$$

.

with $\beta_0 = 0.1$ and $\beta^* = [2.5, -1.5]$

If we use a Maximum likelihood estimation, under the logistic regression model, it might not converge as the first two components are large compared to the bias and the data is linearly separated, which mean that the prediction will give very high coefficients for the first two components that might cause them to go to infinity.

1.5.2 Gibbs sampling results

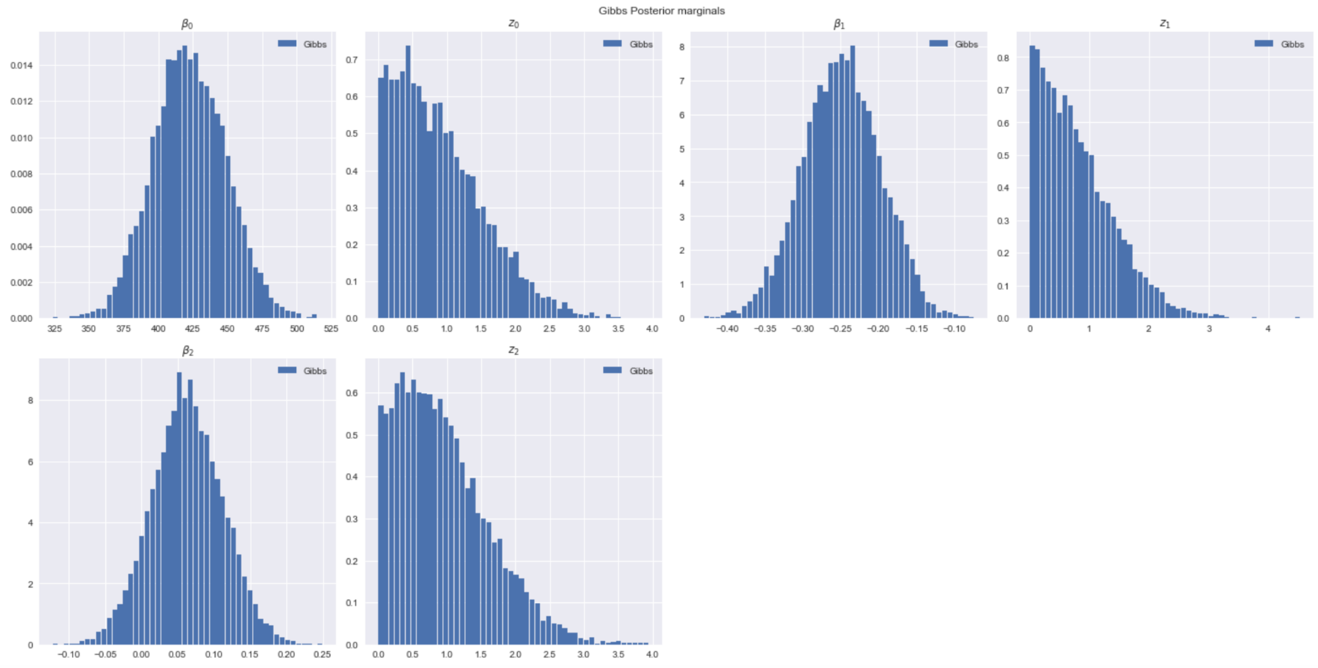

Figure 5: Gibbs Posterior Marginals for the new dataset

In our experiment, in order to get the algorithm to converge, we used a very large number of iterations 8000. As we can see the maximum value of β_0 is ~ 0.015 whereas for β_1 and β_2 it is ~ 8. To conclude, we can say the Gibbs sampling does not guarantee in some cases the converge, however in the specific case of our experiment, it did converge.