MVA - Probabilistic Graphical Models

DM3

Ariane ALIX, Hassen MIRI

January 15th, 2020

Gibbs sampling and mean field VB for the probit model

We would like to classify the last column (good vs bad credit) based on the previous columns (see the UCI repository for more information on each variable). To do so, we consider the probit model, where $y_i = \text{sgn}(\beta^T x_i + \epsilon_i), \epsilon_i \sim N(0,1)$ and a Gaussian prior $\beta \sim N(0, \tau I_p)$, with $p = \dim\beta$, and $\tau = 102$.

1. The prior on β has the same variance in all directions, whereas the predictors are not homogeneous (different mean and covariance). As a consequence, if a predictor has values much larger than the others, it will have the biggest impact on $\beta^T x_i$ and therefore y_i. With the pre-processing, when we normalize the different predictors, they are all put on a same scale and thus can impact the classification with the same weight.

2. Since we only look at the sign of $\beta^T x_i + \epsilon_i$ to compute y_i, the variance of ϵ_i has no impact. Indeed:

$$\text{sgn}(\beta^T x_i + \epsilon_i) = \text{sgn}(\frac{\beta}{\sigma}^T x_i + \frac{\epsilon_i}{\sigma})$$

Hence our problem is equivalent to any problem of finding $\beta' = \frac{\beta}{\sigma}$ whith a noise of variance σ. Therefore we can take a constant such as 1 as the variance of the noise.

3. We define the latent variables $z_i = \beta^T xi + \epsilon_i$, which means that $z_i|\beta \sim N(\beta^T x_i; 1)$.

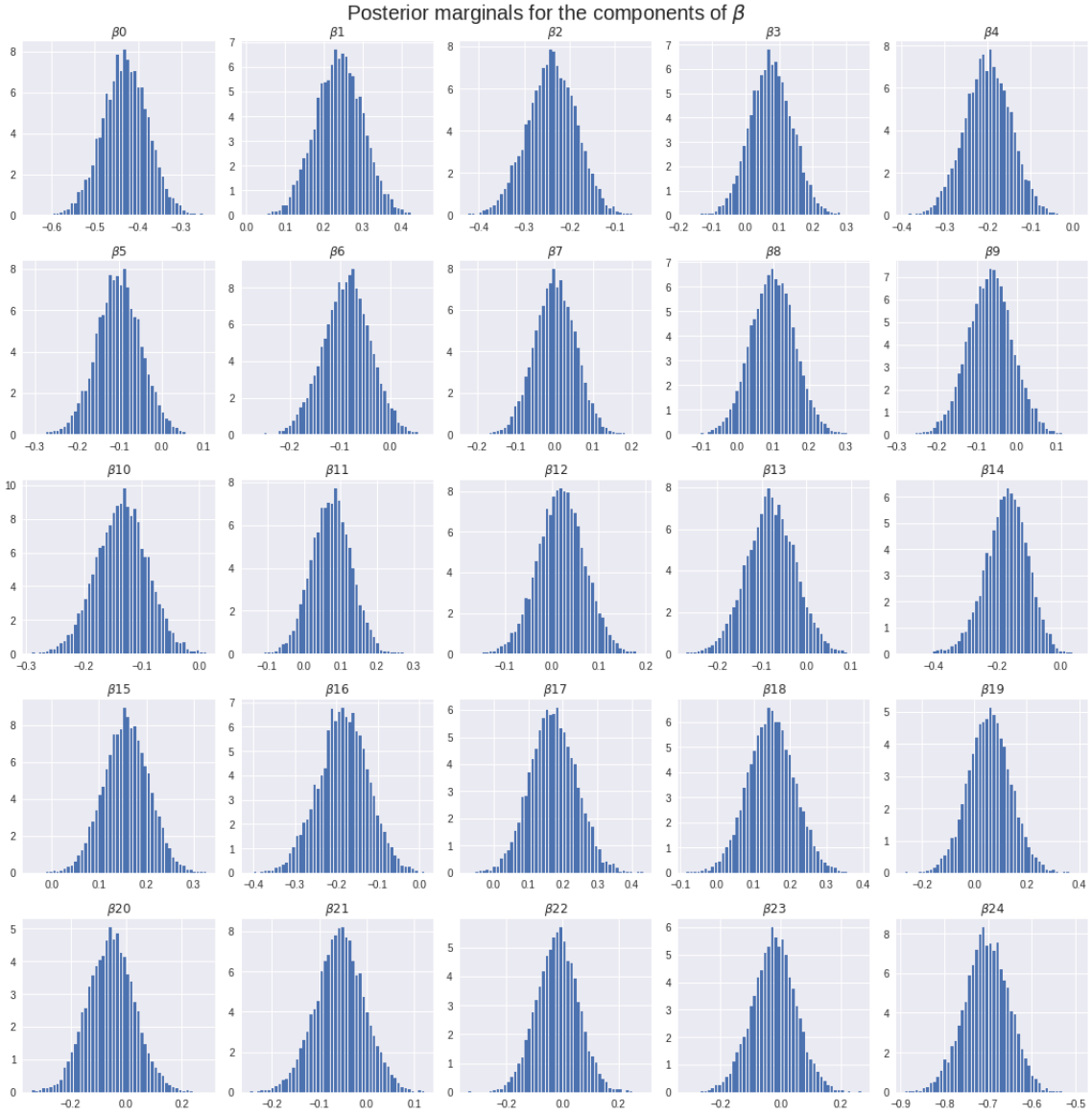
We are looking for the posterior of β:

$$
\begin{aligned}
p(\beta|z;y) &= p(\beta|z) \\
&\propto p(z|\beta)p(\beta) \quad \text{using Bayes' law} \\
&= \exp(-\frac{1}{2\tau}\beta^T\beta)\exp(-\frac{1}{2}\sum_{i=1}^{n}(z_i-\beta^T x_i)^2) \\
&= \exp(-\frac{1}{2}\beta^T\frac{1}{\tau}I_p\beta - \frac{1}{2}\beta^T x^T x\beta + \beta^T x^T z) \\
&= \exp(-\frac{1}{2}(\beta-\mu)^T\Sigma^{-1}(\beta-\mu))
\end{aligned}
$$

With

$$\Sigma^{-1} = \frac{1}{\tau}I_p + x^T x \quad \text{and} \quad \mu = \Sigma x^T z$$

Here are the posteriors obtained with the Gibbs sampling algorithm for all components of β, with 10,000 samples including a burn-in phase of 500:

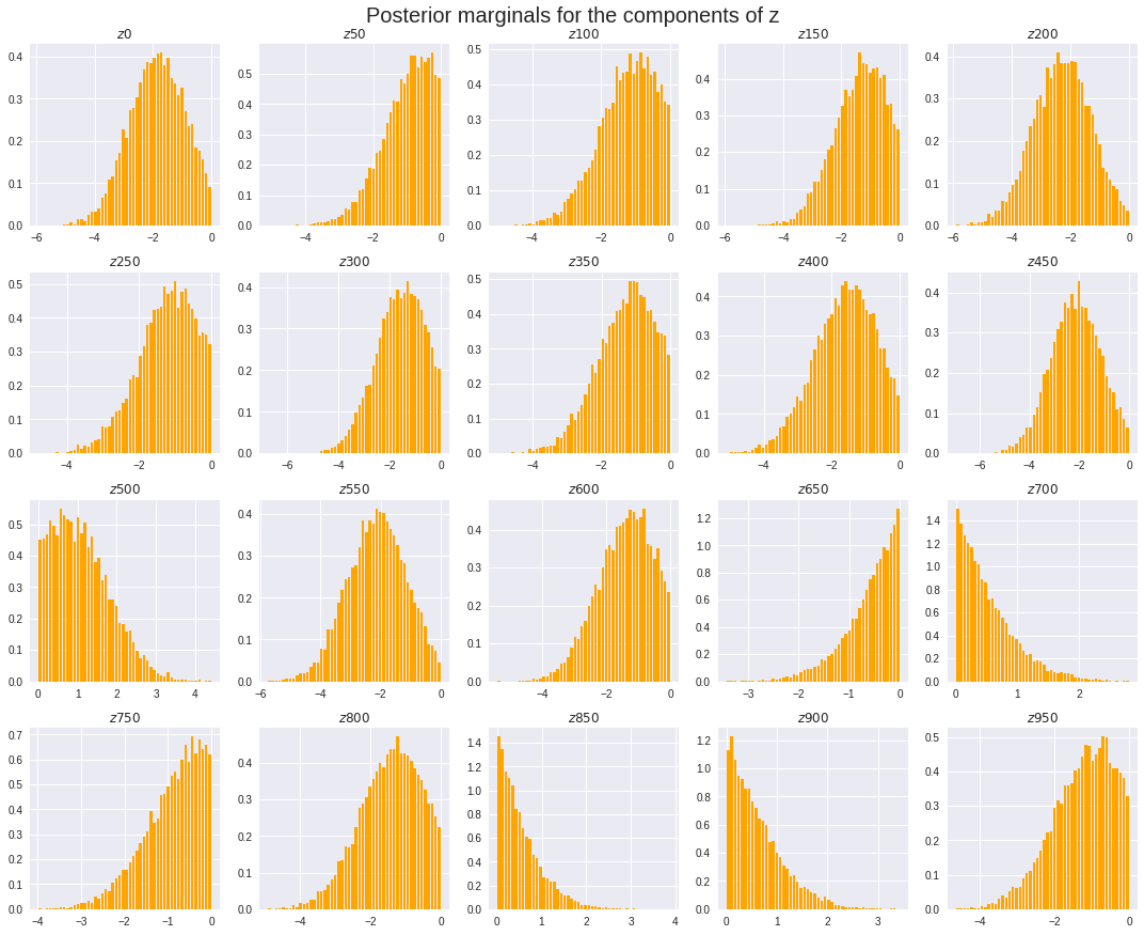


And the posterior of z is:

$$\begin{aligned} p(z_i|\beta, y_i) &\propto p(y_i, z_i|\beta)p(z_i|\beta) \\ &= \exp(-\frac{1}{2}(z_i - \beta^T x_i)^2)\mathbf{1}_{\{\text{sgn}(y_i)=\text{sgn}(z_i)\}} \end{aligned}$$

Which is a truncated Gaussian of mean $\beta^T x_i$, variance 1 and support $\{y_i z_i > 0\}$

Below are some of the posteriors (20 out of 1000) obtained with the Gibbs sampling algorithm for z, with 10,000 samples including a burn-in phase of 500:



4. Using the mean field approach, we choose a distribution q to approximate $p(\beta, q|y)$ such as the variable β and z are independent, we can write then:

$$p(\beta, q|y) = q(\beta, z) = q_1(\beta)q_2(z)$$

The optimal variational distribution q_1^* and q_2^* verify the following equation:

$$\log q_1^*(\beta) = E_{z\sim q_2^*}[log(p(\beta, z, y))|\beta, y] + cst$$

$$\log q_2^*(z) = E_{\beta\sim q_1^*}[log(p(\beta, z, y))|z, y] + cst$$

We can write $log(p(\beta, z, y))$ as follow:

$$\log p(\beta, z, y) = \log p(y|z) + \log p(z|\beta) + \log p(\beta)$$

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

therefore we have

$$\log q_1^*(\beta) = E_{z\sim q_2^*}(\log p(y|z) - \frac{1}{2}\|z - X\beta\|^2 - \frac{1}{2\tau}\|\beta\|^2|\beta, y) + cst$$

$$= -\frac{1}{2}E_{z\sim q_2^*}(\|z - X\beta\|^2 - \frac{1}{2\tau}\|\beta\|^2|\beta, y) + cst$$

$$-\frac{1}{2}E_{z\sim q_2^*}(\|z\|^2 - 2z^T X\beta + \|X\beta\|^2 - \frac{1}{2\tau}\|\beta\|^2|\beta, y) + cst$$

$$= -\frac{1}{2}(\beta - \beta_{q_2})^T\Sigma^{-1}(\beta - \beta_{q_2}) + cst$$

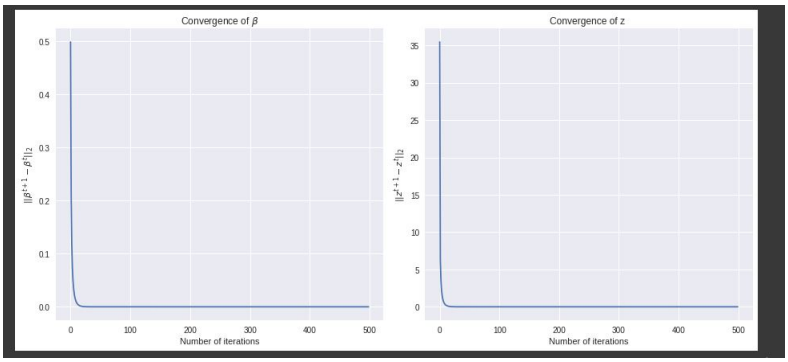
Where: $\beta_{q_2} = \Sigma X^t E_{z\sim q_2^*}(z)$ so we can conclude that $q_1^* \sim N(\beta_{q_2}, \Sigma)$
We have:

$$\log q_2^*(z) = E_{\beta\sim q_1^*}(log(p(y|z) + log(p(z|\beta) + log(p(\beta)|z, y) + cst$$

$$= \sum_{i=1}^{n} log(\mathbb{1}_{y_i z_i >= 0} + E_{\beta\sim q_1^*}(-\frac{1}{2}\|z - X\beta\|^2 - \frac{1}{2\tau}\|\beta\|^2) + cst$$

$$= \sum_{i=1}^{n} log(\mathbb{1}_{y_i z_i >= 0}\} + -\frac{1}{2}\|z - X\beta_{q_1}\|^2$$

where $\beta_{q_1} = E_{\beta\sim q_1^*}(\beta)$ so q_2^* follows a truncated gaussian distribution of mean $x\beta_{q_1}$ and variance Ip and support $\{y_i z_i > 0\}$



5. We compare the speed of the gibbs sampling algorithm and the mean field variational algorithm for 5000 samples with a burn-in of 500 iterations, we find the following results:

- gibbs time:98.74 seconds
- mean field variational time:50.29

We compare equally their accuracy on the test set:

- gibbs accuracy;0.715
- mean field variational accuracy:0.715

so we see that the mean field variational algorithm gives us the same accuracy for a shorter execution time

6. The maximum likelihood estimator for a logistic regression does not converge on this dataset. We test the gibbs algorithm on this dataset and calculate its accuracy.