Probabilistic Graphical Models **MVA Fall 2019**

Homework 3 — January 15

Instructor: Pierre Latouche *Student: Dimitri Theveneau, Antoine Moulin*

1 Gibbs sampling and mean field VB for the probit model

Download the German credit dataset available in e.g. the UCI repository:

https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/german.data-numeric

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 $y_i = \text{sign}\left(\beta^\top x_i + \epsilon_i\right), \epsilon_i \sim \mathcal{N}(0,1)$ and a Gaussian prior $\beta \sim \mathcal{N}(0, \tau I_p)$, with $p = \dim \beta$ and $\tau = 10^2$.

Note: all the code is avaible in the notebook associated to this report.

1. Start by adding a constant column, and normalising all the predictors (so that their average is 0 and their standard deviation is one). Why is it important to pre-process the predictors in this way? (Think in particular about the choice of the prior distribution.)

   Solution. Adding the first constant column allows to create an affine model from a linear one. This column is used to create a bias that will be estimated with the parameters. We chose a prior in which the variance of β is the same on all coordinates. This means that we incite the coordinates of β to be close to zero and close from each others. If we don't normalise the predictors, one of the features might have values way smaller than the other ones and imply a larger value associated in β, which will be badly estimated because of the prior. Thus, in order to have the same variance in the prior for each coordinate, we need to normalise the features.

2. Explain why the variance of the ϵ_i is one, and not σ^2, with σ^2 an extra parameter that must be estimated.

   Solution. β's values will be influenced directly by ϵ_i in order to have the best separation. If ϵ_i has large values, β will tend to have higher values too (in absolute). Thus, the choice of ϵ_i is compensated by β in the results. We can without restriction take it to one.

3. From the calculations in the last lecture, implement a Gibbs sampler for this model (based on the introduction of latent variables $z_i = \beta^\top x_i + \epsilon_i$, as explained during the course). Run your algorithm on the German dataset, and plot the (approximated) marginal posterior distribution of each parameter.

   Solution. We denote $X \in \mathbb{R}^{n \times p}$ the design matrix formed by the x_i's, $z \in \mathbb{R}^n$ the vector formed by the z_i's. Let us compute

$$
\begin{aligned}
p\left(\beta | z, X\right) &\propto p(z|\beta, X)p(\beta) \\
&\propto \exp\left(-\frac{1}{2}\left[(z - X\beta)^\top (z - X\beta) + \frac{1}{\tau}\beta^\top I_p \beta\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left[z^\top z - 2z^\top X\beta + \beta^\top X^\top X\beta + \frac{1}{\tau}\beta^\top\beta\right]\right) \\
&= \exp\left(-\frac{1}{2}\left[z^\top z - 2z^\top X\beta + \beta^\top\left(X^\top X + \frac{1}{\tau}I\right)\beta\right]\right)
\end{aligned}
$$

Then, denoting $\tilde{\Sigma} = \left(X^\top X + \frac{1}{\tau} I\right)^{-1}$,

$$\begin{aligned}
p\left(\beta | z, X\right) &= \exp\left(-\frac{1}{2}\left[z^\top z - 2 z^\top X\beta + \beta^\top \tilde{\Sigma}^{-1}\beta \right.\right.\\
&\qquad \left.\left. - z^\top X \tilde{\Sigma} X^\top z + z^\top X\tilde{\Sigma}X^\top z\right]\right)\\
&\propto \exp\left(-\frac{1}{2}\left[z^\top P z + \left(\beta - \widetilde{\mu}\right)^\top \tilde{\Sigma}^{-1}\left(\beta - \widetilde{\mu}\right)\right]\right)\\
&\propto \mathcal{N}\left(\beta; \widetilde{\mu}, \tilde{\Sigma}\right)\mathcal{N}\left(z; 0, P^{-1}\right)
\end{aligned}$$

with

$$\begin{aligned}
\widetilde{\mu} &= \tilde{\Sigma} X^\top z\\
P &= I - X\tilde{\Sigma}X^\top
\end{aligned}$$

From this

$$\boxed{\begin{aligned}
z_i | y_i = 0, x_i, \beta &\sim \mathcal{N}\left(x_i^\top \beta, 1\right)\mathcal{I}\left(z_i < 0\right)\\
z_i | y_i = 1, x_i, \beta &\sim \mathcal{N}\left(x_i^\top \beta, 1\right)\mathcal{I}\left(z_i \geq 0\right)
\end{aligned}}$$

4. Now implement the mean field variational algorithm seen during the course, and compare the results with those of Gibbs; both in terms of speed, and accuracy. Comment. (Give details about your derivations of the distributions computed at every iteration.)

Solution. We want to approximate the distribution $\pi \triangleq p\left(\beta, z | X, y\right)$ by $q\left(\beta, z\right) = q_1\left(\beta\right) q_2\left(z\right)$ solution of:

$$\arg\max_q KL\left(q || \pi\right)$$

An iteration of the algorithm consists in two steps: optimize $q_1\left(\beta\right)$ with $q_2\left(z\right)$ fixed and vice-versa. The optimal distribution satisfies:

$$\begin{aligned}
q_1\left(\beta\right) &\propto \exp\left(\mathbb{E}_{q_2}\left[\log p\left(\beta, y, z\right)\right]\right)\\
q_2\left(z\right) &\propto \exp\left(\mathbb{E}_{q_1}\left[\log p\left(\beta, y, z\right)\right]\right)
\end{aligned}$$

By construction, we have $p\left(\beta, y, z\right) = p\left(y|z\right) p\left(z|\beta\right) p\left(\beta\right)$. For all i, $z_i | \beta \sim \mathcal{N}\left(x_i^\top \beta, 1\right)$ and the z_i are independent, so $z|\beta \sim \mathcal{N}\left(X\beta, I\right)$. Hence,

$$\begin{aligned}
q_1\left(\beta\right) &\propto \exp\left(\mathbb{E}_{q_2}\left[\log p(y|z) - \frac{1}{2}\left|\left|z - X\beta\right|\right|_2^2 - \frac{1}{2\tau}\left|\left|\beta\right|\right|_2^2\right]\right)\\
&\propto \exp\left(\mathbb{E}_{q_2}\left[\log p(y|z) - \frac{1}{2}\left(z^\top z - 2 z^\top X\beta + \beta^\top X^\top X\beta\right) - \frac{1}{2\tau}\left|\left|\beta\right|\right|_2^2\right]\right)
\end{aligned}$$

The terms that do not depend on β are considered as a constant:

$$\begin{aligned}
q_1\left(\beta\right) &\propto \exp\left(\mathbb{E}_{q_2}(z)^\top X\beta - \frac{1}{2}\beta^\top X^\top X\beta - \frac{1}{2\tau}\beta^\top\beta\right)\\
&\propto \exp\left[-\frac{1}{2}\left(\beta - \bar{\beta}\right)^\top \tilde{\Sigma}^{-1}\left(\beta - \bar{\beta}\right)\right]
\end{aligned}$$

where $\bar{\beta} = \tilde{\Sigma} X^\top \bar{z}$, $\bar{z} = \mathbb{E}_{q_2}(z)$ and $\tilde{\Sigma}$ defined as previously. Thus, we have

$$\boxed{q_1\left(\beta\right) = \mathcal{N}\left(\beta; \bar{\beta}, \tilde{\Sigma}\right)}$$

For q_2, we have:

$$
\begin{aligned}
q_2(z) &\propto \exp\left(\mathbb{E}_{q_1}\left[\log p(y|z) - \frac{1}{2}\left|\left|z - X\beta\right|\right|_2^2 - \frac{1}{2\tau}\left|\left|\beta\right|\right|_2^2\right]\right)\\
&\propto \exp\left(\mathbb{E}_{q_1}\left[\log p(y|z)\right] - \frac{1}{2}\left[z^\top z - 2z^\top X\mathbb{E}_{q_1}(\beta)\right]\right)\\
&\propto \exp\left(\mathbb{E}_{q_1}\left[\log p(y|z)\right] - \frac{1}{2}\left|\left|z - X\mathbb{E}_{q_1}(\beta)\right|\right|_2^2\right)\\
&\propto \exp\left(\sum_{i=1}^{n}\log \mathbb{1}_{y_i z_i \geq 0} - \frac{1}{2}\left|\left|z - X\mathbb{E}_{q_1}(\beta)\right|\right|_2^2\right)
\end{aligned}
$$

Besides, $\mathbb{E}_{q_1}(\beta) = \bar{\beta}$ defined above. Thus,

$$\boxed{q_2(z) = \mathcal{N}^{\text{trunc}}\left(z; X\bar{\beta}, I, y\right)}$$

where $\mathcal{N}^{\text{trunc}}\left(z; X\bar{\beta}, I, y\right)$ denotes the truncated normal distribution, i.e. such that for all i, we have $q_2(z_i) = \mathcal{N}\left(z_i; x_i^\top\bar{\beta}, 1\right)\mathbb{1}_{y_i z_i \geq 0}$. Note that $\bar{z}_i = x_i^\top\beta + y_i\frac{\phi(x_i^\top\bar{\beta})}{\Phi\left(y_i x_i^\top\bar{\beta}\right)}$, where ϕ is the PDF of the standard normal distribution and Φ its CDF.

5. Bonus question: find formal arguments on why the posterior variance is under-estimated by the mean-field approach.

   Solution.

6. Complete separation occurs in a dataset if there exists β such that $y_i\beta^\top x_i > 0$ for all the datapoints. Represent graphically complete separation. Is maximum likelihood estimation possible in such a case? (Explain.) Construct a simple dataset with full separation, run your Gibbs sampler on this dataset, and comment.

   Solution. See the notebook.