MVA – Probabilistic Graphical Models
Homework 3: Gibbs Sampling and VB

Wilson Jallet*

December 21, 2019

Question 1

This operation puts all the data on the same scale – this is especially useful because the prior on β assigns the same variance in each direction.

Question 2

If we supposed that ε_i had a variance of σ^2, we could write $\varepsilon_i = \sigma\varepsilon_i'$ where $\varepsilon_i' \sim \mathcal{N}(0,1)$, and we'd have

$$y_i = \operatorname{sgn}(\beta^T x_i + \varepsilon_i) = \operatorname{sgn}({\beta'}^T x_i + \varepsilon_i')$$

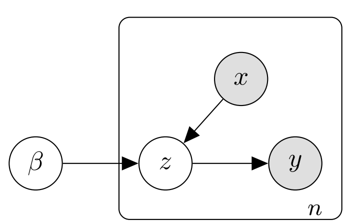
where $\beta' = \beta/\sigma$.

Question 3

We define the following graphical model:

- observed features $x_i \in \mathbb{R}^p$, $i \in \{1,\dots,n\}$
- random variable $\beta \sim \mathcal{N}(0, \tau I_p)$
- latent variables $z_i = \beta^T x_i + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0,1)$
- observed labels $y_i = \operatorname{sgn}(z_i) \in \{-1, 1\}$

It has the following representation:



To perform inference on the model, we need the posterior distribution of β, z given the data X, y.

*wilson.jallet@polytechnique.org

Our first approach is to use Gibbs sampling. To use it, we need to derive the conditional posteriors of the variables. Evidently,

$$
\begin{aligned}
p(y_i|\beta) &= \Phi(y_i \beta^T x_i) \\
p(z_i|\beta) &\sim \mathcal{N}(\beta^T x_i, 1) \\
p(y_i, z_i|\beta) &= \mathbb{1}_{\{y_i z_i > 0\}}
\end{aligned}
$$

By Bayes' theorem we have the posteriors

$$
\begin{aligned}
p(\beta|z) \propto p(\beta)p(z|\beta) &\propto \exp\left(-\frac{1}{2\tau}\|\beta\|^2 - \frac{1}{2}\sum_{i=1}^{n}(z_i - \beta^T x_i)^2\right) \\
&= \exp\left(-\frac{1}{2\tau}\|\beta\|^2 - \frac{1}{2}\|z - X\beta\|^2\right)
\end{aligned} \tag{1}
$$

and

$$
p(z|\beta, y) \propto p(z|\beta)p(y, z|\beta) \propto \exp\left(-\frac{1}{2}\|z - X\beta\|^2\right)\prod_{i=1}^{n}\mathbb{1}_{\{y_i z_i > 0\}} \tag{2}
$$

where $X = (x_1|\dots|x_n)^T \in \mathbb{R}^{n\times p}$ is the design matrix. By identification $\beta|z \sim \mathcal{N}(\mu_p, \Sigma_p)$ where

$$
\Sigma_p^{-1} = \frac{1}{\tau}I_p + X^T X, \quad \mu_p = \Sigma_p X^T z \tag{3}
$$

and for all i, $\boxed{z_i|\beta, y_i \sim \mathrm{T}\mathcal{N}(x_i^T\beta, 1; y_i)}$ where $\mathrm{T}\mathcal{N}(\cdot; y_i)$ is the truncated Gaussian with support in the orthant $\{z \in \mathbb{R} : y_i z_i > 0\}$.

With all this in place, we use Gibbs sampling to sample from the posterior distribution of β, z given the data (X, y). For inference and testing, we **split the dataset up as 2/3rds for training and 1/3rd for testing**. Figure 1 shows approximate posterior marginals for $\beta, z|X, y$ in the form of histograms made from samples.

The testing accuracy (predicting using MAP) is of about $\approx 75\%$, using 4000 samples of $\beta|X_{\text{train}}, y_{\text{train}}$.

Question 4

This time, we want to use variational inference, by approximating the true prior $p(\beta, z|X, y)$ by a distribution $q(\beta, z)$. Assume the mean-field factorization for q:

$$
q(\beta, z) = q_1(\beta)q_2(z) \tag{4}
$$

We denote by $\mathbb{E}^q$ the expectation operator under the distribution q. The optimal variational distribution satisfies

$$
\log q_1^*(\beta) = \mathbb{E}_{z\sim q_2^*}[\log p(\beta, z, y)|\beta, y] + \mathrm{cst} \tag{5a}
$$

$$
\log q_2^*(z) = \mathbb{E}_{\beta\sim q_1^*}[\log p(\beta, z, y)|z, y] + \mathrm{cst} \tag{5b}
$$

- **Joint probability.** The log-joint probability of (β, z, y) is written

$$
\begin{aligned}
\log p(\beta, z, y) &= \log p(y|z) + \log p(z|\beta) + \log p(\beta) \\
&= \log p(y|z) - \frac{1}{2}\|z - X\beta\|^2 - \frac{1}{2\tau}\|\beta\|^2 + \mathrm{cst}
\end{aligned} \tag{6}
$$

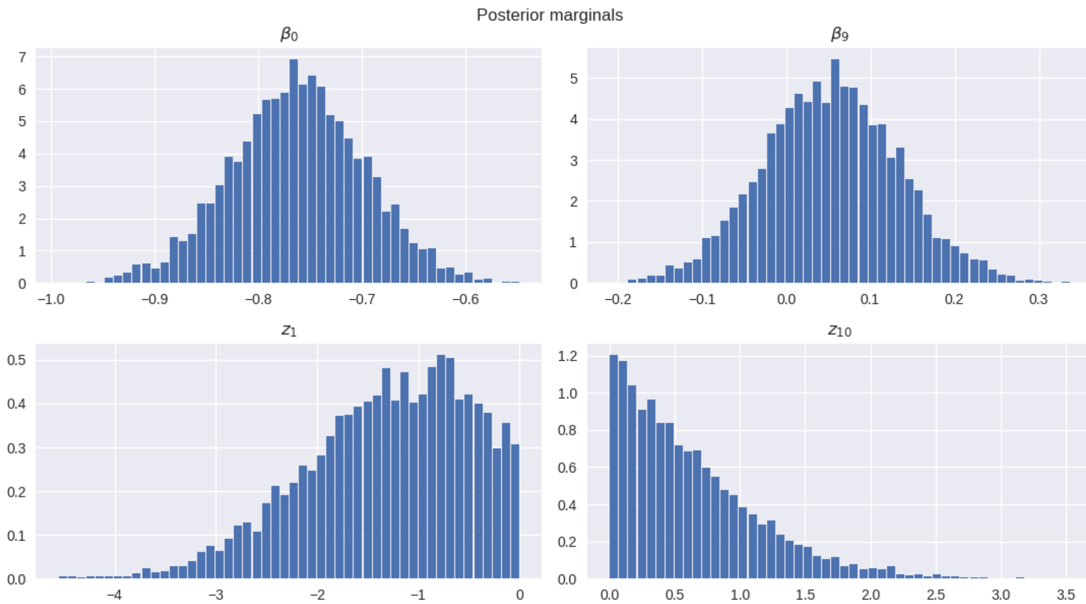


Figure 1: Some approximate posterior marginals for β and z given the data X, y.

- **Derivation of q_1.** The optimal form of the factor is

$$
\begin{aligned}
\log q_1^*(\beta) &= \mathbb{E}_z^q \left[\log p(y|z) - \frac{1}{2}\|z - X\beta\|^2 - \frac{1}{2\tau}\|\beta\|^2 \middle| \beta, y \right] + C_1 \\
&= \mathbb{E}_z^q[\log p(y|z)|\beta, y] - \frac{1}{2}\mathbb{E}_z^q \left[\|z - X\beta\|^2 \middle| \beta, y \right] - \frac{1}{2\tau}\|\beta\|^2 + C_1 \\
&= -\frac{1}{2}\mathbb{E}_z^q \left[\|z - X\beta\|^2 \middle| \beta, y \right] - \frac{1}{2\tau}\|\beta\|^2 + C_2 \\
&= -\frac{1}{2}\mathbb{E}_z^q \left[\|z\|^2 - 2z^T X\beta + \|X\beta\|^2 \middle| \beta, y \right] - \frac{1}{2\tau}\|\beta\|^2 + C_2 \\
&= \bar{z}^T X\beta - \frac{1}{2}\beta^T X^T X\beta - \frac{1}{2\tau}\|\beta\|^2 + C_3 \\
&= -\frac{1}{2}(\beta - \bar{\beta})\Sigma_p^{-1}(\beta - \bar{\beta}) + C_3
\end{aligned}
\tag{7}
$$

where Σ_p is defined as in eq. (3), and

$$\bar{z} = \mathbb{E}_{z \sim q_2^*}[z], \quad \bar{\beta} = \Sigma_p X^T \bar{z}. \tag{8}$$

The terms $\mathbb{E}_z^q[\log p(y|z)|\beta, y]$ and $\mathbb{E}_z^q \left[\|z\|^2 \right]$ do not depend on β and are added to the constants.

The end result is

$$\boxed{q_1^*(\beta) = \mathcal{N}(\Sigma_p X^T \bar{z}, \Sigma_p).} \tag{9}$$

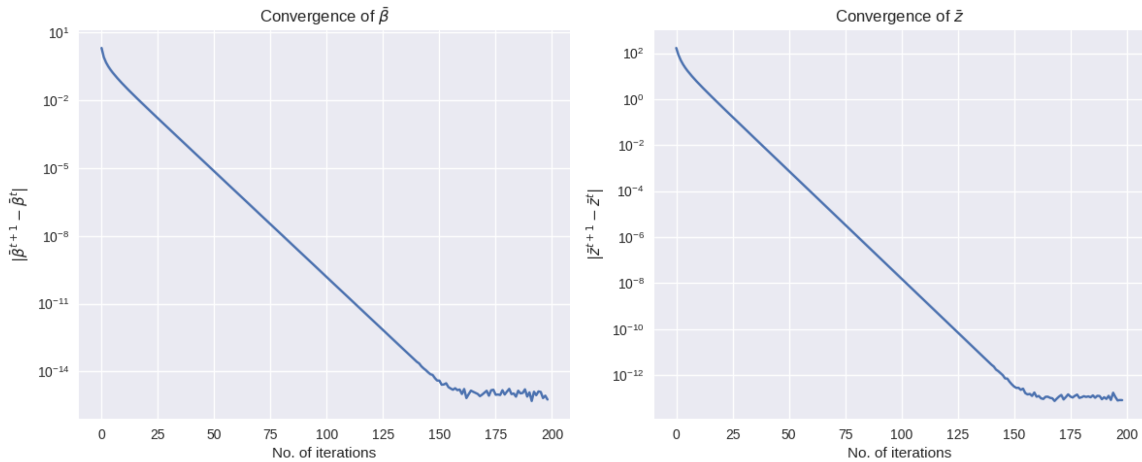


Figure 2: L_1-loss between consecutive iterations of the VI algorithm.

- **Derivation of q_2.** The optimal form of the factor is

$$\begin{aligned}
\log q_2^*(z) &= \mathbb{E}_{\beta\sim q_1^*}\left[\log p(y|z) - \frac{1}{2}\|z - X\beta\|^2 - \frac{1}{2\tau}\|\beta\|^2 \middle| z, y\right] + C_1 \\
&= \sum_{i=1}^{n}\{\mathbb{1}_{y_i=1}\ln(\mathbb{1}_{z_i>0}) + \mathbb{1}_{y_i=-1}\ln(\mathbb{1}_{z_i\leq 0})\} - \frac{1}{2}\left(\|z\|^2 - 2z^T X\bar{\beta}\right) + C_2 \\
&= \sum_{i=1}^{n}\{\mathbb{1}_{y_i=1}\ln(\mathbb{1}_{z_i>0}) + \mathbb{1}_{y_i=-1}\ln(\mathbb{1}_{z_i\leq 0})\} - \frac{1}{2}\left\|z - X\bar{\beta}\right\|^2 + C_3
\end{aligned} \tag{10}$$

where $\bar{\beta} = \mathbb{E}_{\beta\sim q_1^*}[\beta]$. Indeed, the expectation under q_1 of $\log p(y|z)$ conditionally on z, y is itself. This means that

$$\boxed{q_2^*(z) = \mathrm{T}\mathcal{N}(X\bar{\beta}, I_p; \mathcal{P}_y).} \tag{11}$$

- **Summary and algorithm.** The optimal mean-field distribution $q(\beta, z) = q_1(\beta)q_2(z)$ satisfies the fixed-point condition

$$q_1^*(\beta) = \mathcal{N}(\Sigma_p X^T\bar{z}, \Sigma_p) \tag{12a}$$
$$q_2^*(z) = \mathrm{T}\mathcal{N}(X\bar{\beta}, I_p; \mathcal{P}_y) \tag{12b}$$
$$\bar{\beta} = \mathbb{E}_{\beta\sim q_1^*}[\beta] = \Sigma_p X^T\bar{z} \tag{12c}$$
$$\bar{z} = \mathbb{E}_{z\sim q_2^*}[z] \tag{12d}$$

We can explicitly compute

$$\bar{z}_i = x_i^T\bar{\beta} + y_i\frac{\phi(x_i^T\bar{\beta})}{\Phi(y_i x_i^T\bar{\beta})}.$$

The coordinate ascent variational approximation algorithm now reduces to alternatively updating the means until convergence.

▪ **Performance.** Inference with Gibbs sampling $M = 5000$ samples (and a burn-in of 100) takes ≈ 10.4 seconds. Variational approximation converges in 200 iterations (see fig. 2) in ≈ 0.22 seconds, and sampling $M = 5000$ times took ≈ 0.26 seconds. Figure 3 shows a comparison of the posterior marginals obtained with the two approaches: we can observe that the VI algorithm often returns lower posterior variance on β than Gibbs. For prediction, we obtain similar MAP prediction accuracy on the test set – around 75%.

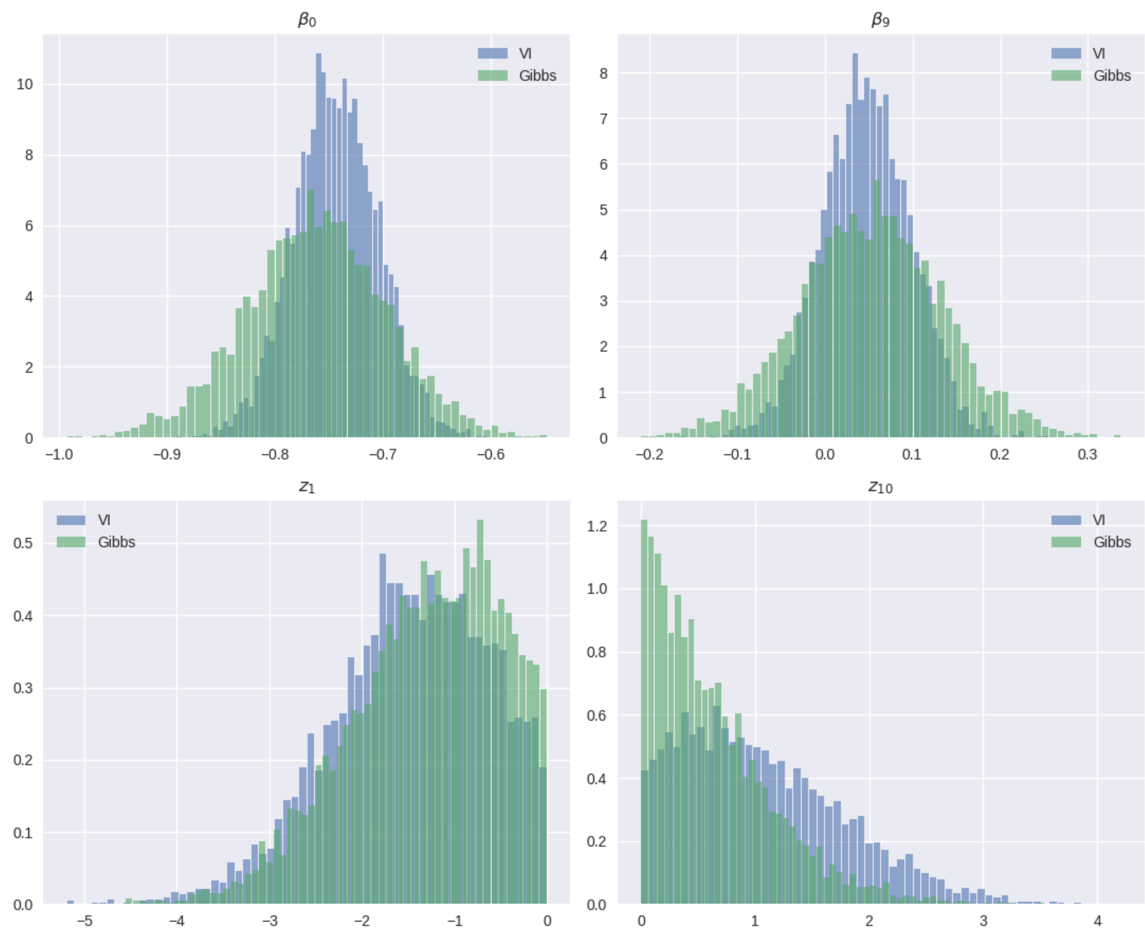


Figure 3: Comparison of the posterior marginals between Variational Bayes and Gibbs sampling. Histograms built with $M = 5000$ samples.

Question 5

Question 6