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

Wilson Jallet[*]

January 14, 2020

## Question 1

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

## Question 2

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

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

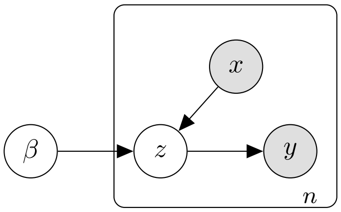
where $\beta' = \beta/\sigma$.

## 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 $\beta, z$ 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 $\beta, z$ 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

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

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

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

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

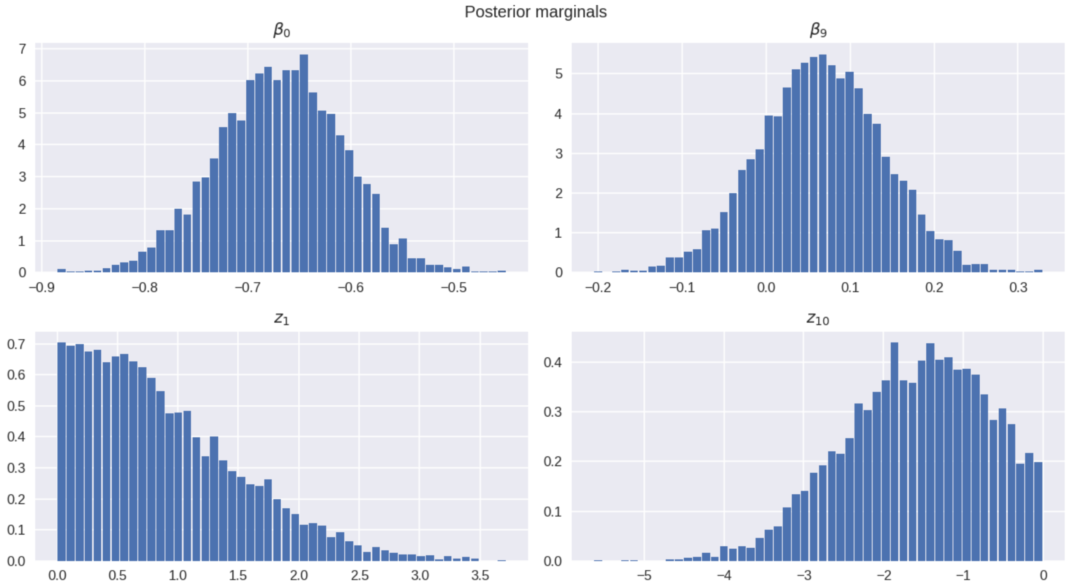


Figure 1: Some approximate posterior marginals for $\beta$ and $z$ given the data $X, y$.

- **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 $\Sigma_p$ 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 $\beta$ and are added to the constants.

The end result is

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

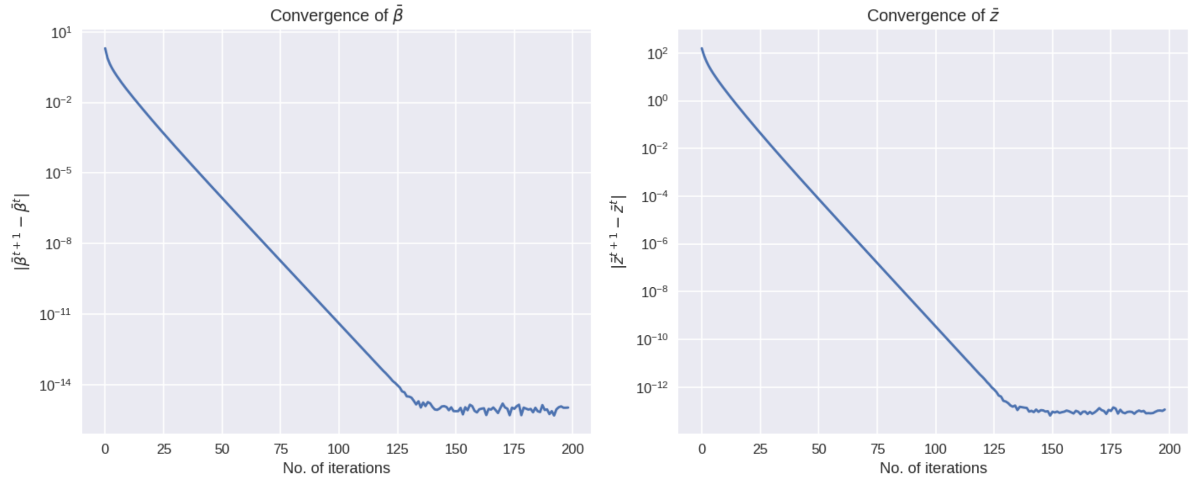


Figure 2: $L_1$-loss between consecutive iterations of the VI algorithm.

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

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

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 $\approx$ 10.4 seconds. Variational approximation converges in 200 iterations (see fig. 2) in $\approx$ 0.22 seconds, and sampling $M = 5000$ times took $\approx$ 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 $\beta$ than Gibbs. For prediction, we obtain similar MAP prediction accuracy on the test set – around 75%.

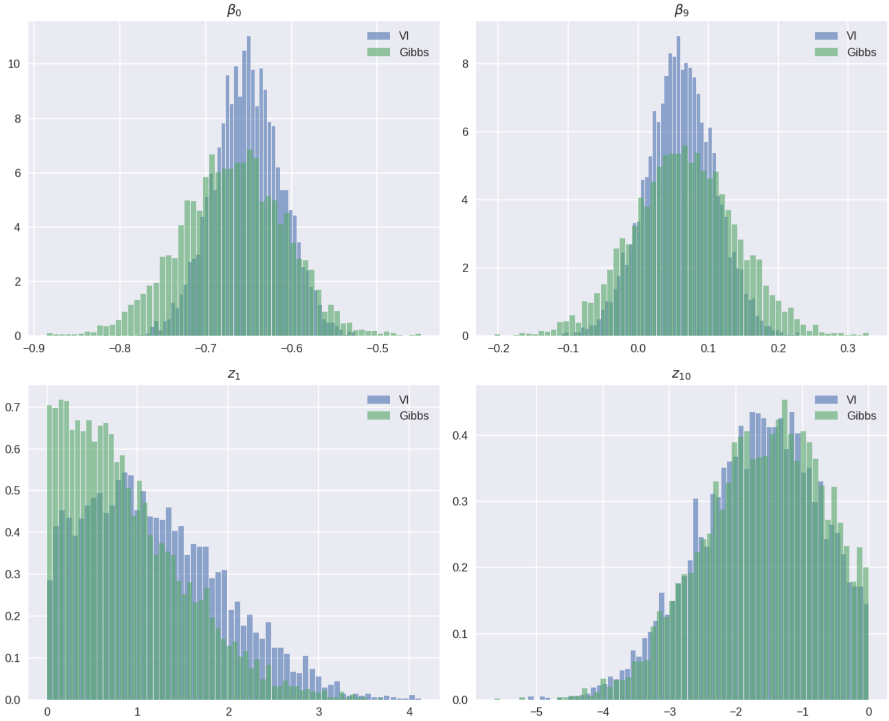


Figure 3: Comparison of the posterior marginals between Variational Bayes and Gibbs sampling. Histograms built with $M = 5000$ samples.

## Question 5

We know that the (true) posterior variance of $\beta$ given $X$ is by the law of total variance

$$\mathbb{V}(\beta) = \mathbb{E}[\mathbb{V}(\beta|z)] + \mathbb{V}(\mathbb{E}[\beta|z]) = \Sigma_p + \mathbb{V}(\Sigma_p X^T z) = \Sigma_p + \Sigma_p X \mathbb{V}(z) X^T \Sigma_p \succeq \Sigma_p$$

while under $q_1^*$, $\beta$ exactly has variance $\Sigma_p$. So we know that asymptotically VB under-estimates the true posterior variance. It also has lesser variance than the Gibbs iterates $\beta^{(t)}$ by using the same of argument total variance:

$$\mathbb{V}(\beta^{(t)}) = \Sigma_p + \mathbb{V}(\Sigma_p X^T z^{(t-1)})$$

## Question 6

Figure 4 shows a completely separated dataset in $\mathbb{R}^2$ – the separation line $x_1 + 1.5x_2 + 0.1 > 0$ is supported by a normal vector $\beta_{\text{true}} = (1, 1.5)^T$, and bias 0.1. A logistic regression model with maximum likelihood estimation would break down and not converge.

For this dataset, the Gibbs sampler might not converge: if we have bias (an additional column in $X$ with ones), we get the trace plot of Figure 5.

If we introduce instead a smaller constant column (of value $c = 10^{-3}$), we get Figure 6 where we can see the trace plot converge fast and that the posterior histograms resemble actual normal distributions. This is the same thing as introducing a different, much higher *prior* for the bias variable, which in the supervised regression setting is equivalent to introducing penalization for the bias. This is still not perfect, as we can see in the histogram that the posterior distribution of $\beta_0$ lies in the range $\sim 400 - 600$: the corresponding bias (renormalizing by $c$) is estimated to be in $\sim 0.4 - 0.6$, which is wrong. In this setting MAP prediction accuracy on a test set of 1/5th of the total data yields an accuracy of 89.5%.

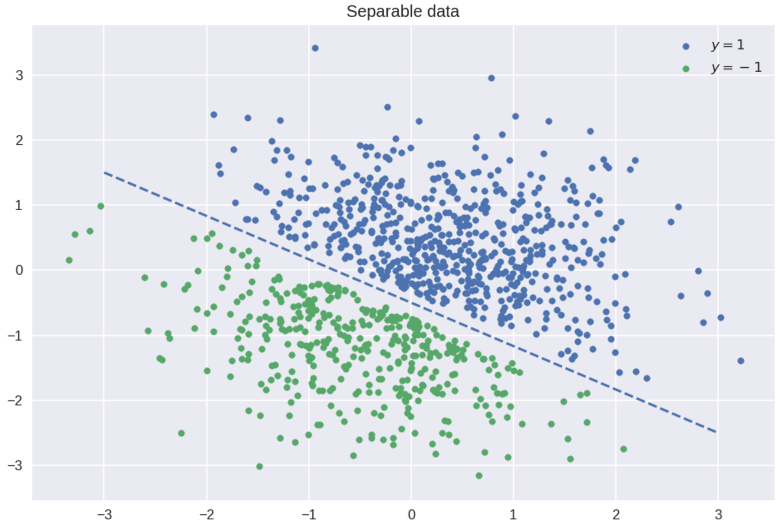
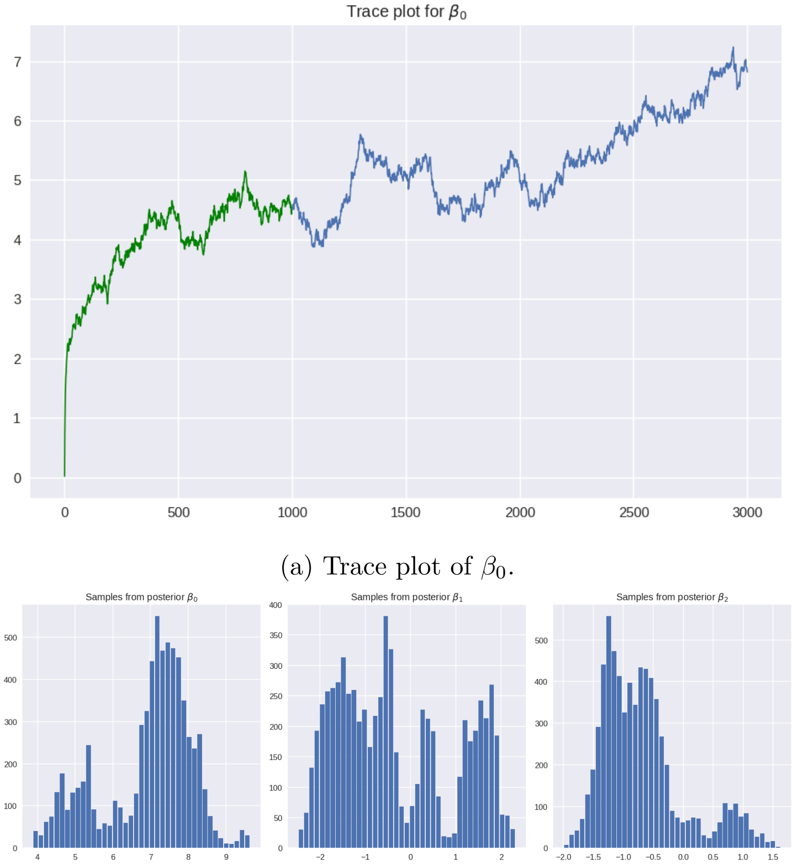


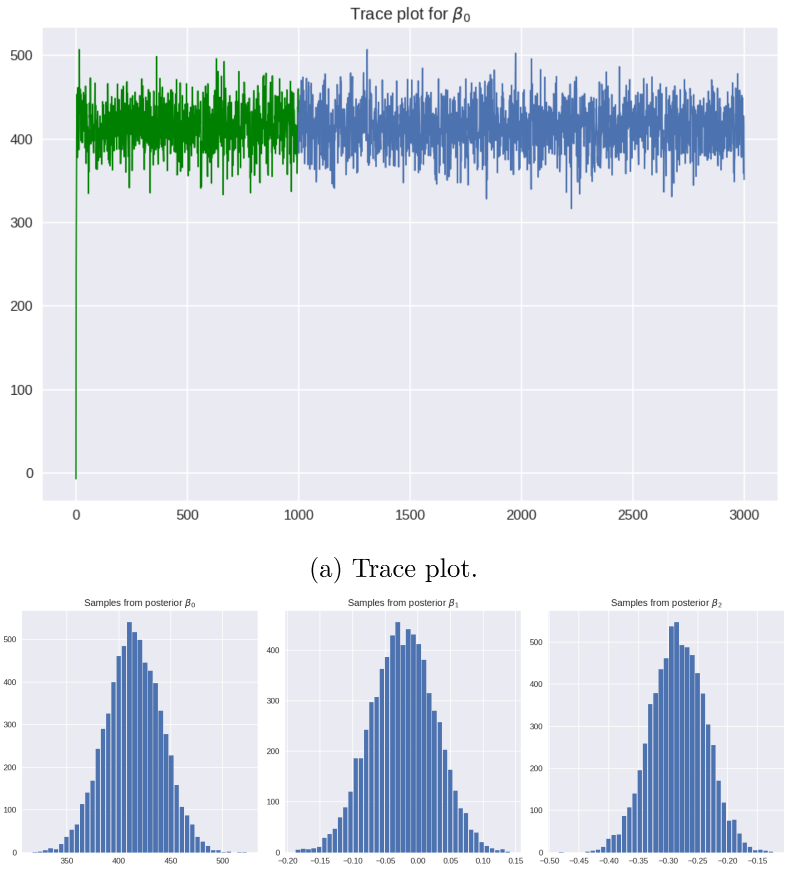
Figure 4: Linearly separable data in dimension $p = 2$.



(a) Trace plot of $\beta_0$.

(b) Histogram of the posterior $\beta$ marginals.

Figure 5: Failed convergence of Gibbs sampling for the separable dataset when a bias term is added. Increasing the amount of burn-in does not solve the problem.



(a) Trace plot.

(b) Histograms of the posterior $\beta$ marginals.

Figure 6: Gibbs sampling. Trace plot and posterior $\beta$ histograms using a smaller constant column for the separable data of fig. 4.