Lucas Elbert
lucas.elbert@polytechnique.edu

1 Gibbs sampling and mean field VB for the probit model

We study the Gibbs sampling algorithm and the mean field variational Bayes for a probit model $y_i = sign(\beta^\top x_i + \epsilon_i), \epsilon_i \sim N(0,1)$ with a gaussian prior $\beta \sim N(0, \tau I_p)$ and test it on the German credit dataset to classify good and bad credits.

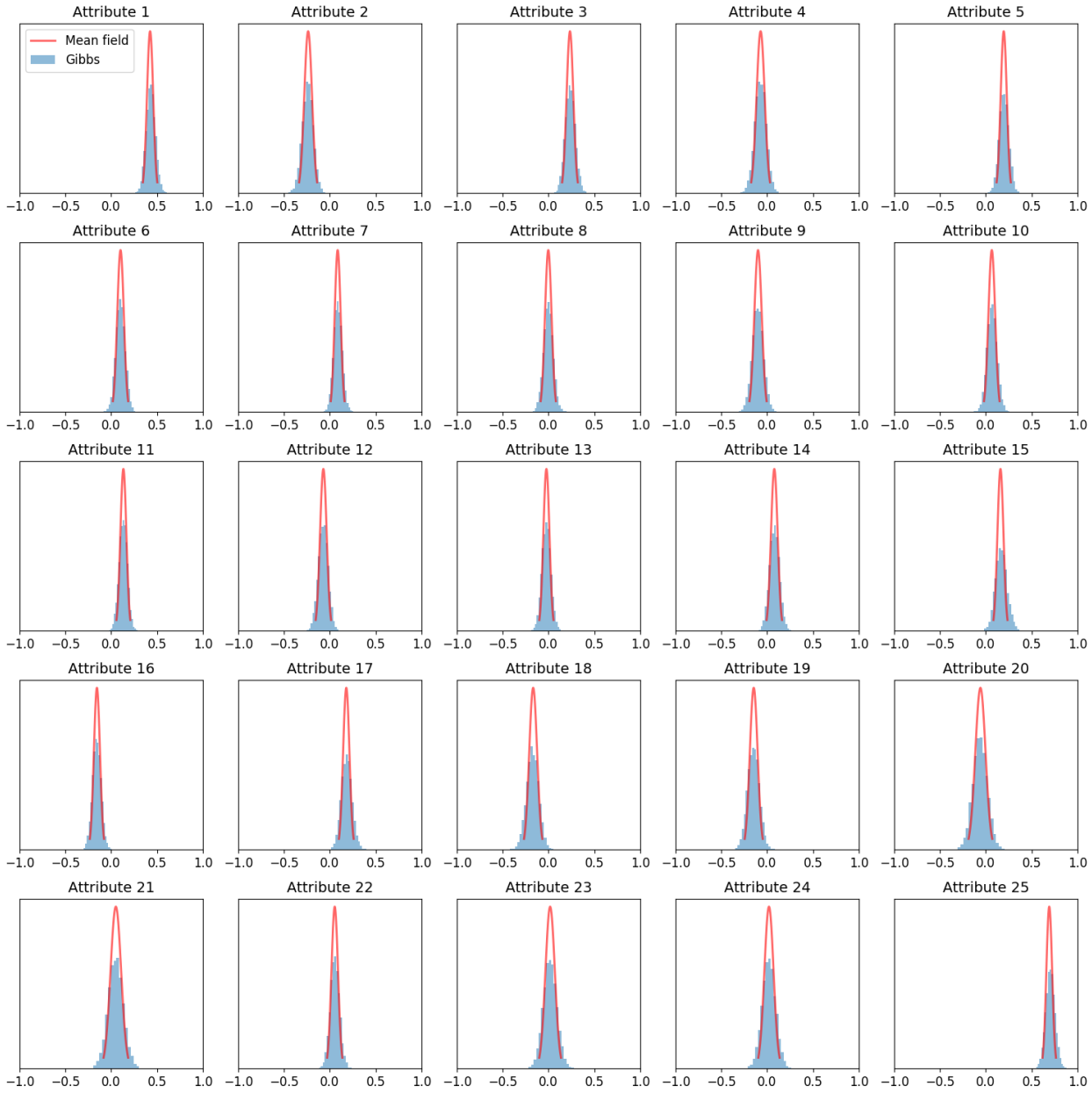

Figure 1: β marginals estimations of the Gibbs sampler and the mean field VB

1.1 Preprocessing

It is important to normalize the predictors to give all predictors a priori the same chance of influence. As the prior distributes all β_i according to $N(0, \tau)$ we generally favour smaller β_is over larger but we do not favor any predictor over another. Without rescaling predictors with numerically larger (absolute) values would have

a much higher influence on the prediction. Therefore the rescaling is important to bring them all into the same range.
Furthermore the mean centering gives the advantage that however we choose β, in expectation our prediction will only be influenced by the bias (β_i corresponding to the constant column). Both aspects are covered by our mean-0-std-1 preprocessing.

1.2 Why the noise is $N(0,1)$

Adding a parameter σ and assuming $\epsilon_i \sim N(0,\sigma^2)$ does not extend our model. In fact we have

$$y_i \sim sign(\beta^\top x_i + N(0,1)) = sign(\frac{\beta^\top x_i}{\sigma} + N(0,\sigma^2))$$

Therefore the new parameter gives only a re- and over-parameterization of the model. Because all β, σ pairs with the same ratio create the same model, the concrete values (apart from the ratio) would be determined by the prior only.

1.3 Gibbs sampler

We implement a Gibbs sampler of the posterior of β. As it is difficult to directly sample from $p(\beta|y)$ we introduce the hidden variables z_i and will sample from $p(\beta, z|y)$

$$z_i = x_i\beta + \epsilon_i$$

In class we have seen the following
Theorem 1: If $x \sim N(\mu,\Sigma)$ and $\mu \sim N(m,S)$ then

$$\begin{aligned} p(\mu|x) \propto p(x|\mu)p(\mu) &= N(x;\mu,\Sigma)N(\mu;m,S) \\ &= N(\mu;m_p,S_p) \\ \text{with } S_p &= (S^{-1}+\Sigma^{-1})^{-1} \\ m_p &= S_p(S^{-1}m + \Sigma^{-}1x) \end{aligned}$$

This can be applied here and we get

$$\begin{aligned} p(\beta|z,y) = p(\beta|z) \propto p(z|\beta)p(\beta) &\sim N(z;X\beta,I_n)N(\beta;0,\tau I_p) \\ &= N((X^\top X)^{-1}X^\top z;\beta,(X^\top X)^{-1})N(\beta;0,\tau I_p) \\ &= N(\beta;m_p,S_p) \end{aligned}$$

$$\begin{aligned} S_p &= (\frac{1}{\tau}I_p + X^\top X)^{-1} \\ m_p &= S_p X^\top z \end{aligned}$$

and

$$p(z|\beta,y) \propto \prod_{i=1}^{n} N(z_i;\beta^\top x_i,1)\mathbb{1}_{(z_i y_i>0)}$$

We can then iteratively sample β and z from those two distributions. This is a MCMC method, I used a burn in period of 5000, then took 10000 more samples and used them to create histograms for the marginals of each component β_i of β. Figure 1 shows in blue these histograms.

1.4 Mean field variational Bayes

We now implement the mean field variational Bayes algorithm to perform the same task, i.e. estimating the posterior $p(\beta|y)$. We again do this by introducing the hidden variables $z_i = x_i\beta + \epsilon_i$, try to estimate $p(z,\beta|y) \approx q(z,\beta|y)$ and finally consider its marginal. In the mean field VB algorithm we try to estimate p by a factorizing $q(z,\beta|y) = q_1(z)q_2(\beta)$.

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

We try to find q_1, q_2 in order to minimize the Kullback-Leibler distance $KL(q_1(.)q_2(.)||p(.,.|y))$. In class we have seen that a coordinate descent here in general leads to an iterative update scheme for q_1, q_2:

$$\begin{aligned} q_1(z) &\propto \exp(\mathbb{E}_{q_2(\beta)}[\log(p(z,y|\beta))]) \text{ with } z_i y_i > 0 \ \forall i \\ q_2(\beta) &\propto p(\beta)\exp(\mathbb{E}_{q_1(z)}[\log(p(z,y|\beta))]) \text{ with } z_i y_i > 0 \ \forall i \end{aligned}$$

We now derive forms of these update equations for our special case. By assumption we have with some normalization constant α

$$p(z, y|\beta) = \alpha \mathbb{1}_{zy>0} N(z; X\beta, I_n)$$

Furthermore for the calculations I will for the ease of calculation assume z, y are only a single observation. As our observations are independent the formulas easily generalize to to more observations.
$q_1(z)$:
Let q_2 be fixed. By ignoring factors that do not depend on z we get

$$\begin{aligned}
q_1(z) &\propto \mathbb{1}_{zy>0}\ exp(\mathbb{E}_{q_2(\beta)}[log(\alpha N(z; X\beta, I_n)]) \\
&\propto \mathbb{1}_{zy>0}\ \alpha\ exp(\mathbb{E}_{q_2(\beta)}[z^\top z - 2z^\top X\beta + \beta^\top X^\top X\beta]) \\
&\propto \mathbb{1}_{zy>0}\ exp(z^\top z - 2z^\top X\mathbb{E}_{q_2(\beta)}[\beta] + \beta^\top X^\top X\beta) \\
&\propto \mathbb{1}_{zy>0}\ exp(z^\top z - 2z^\top X\mathbb{E}_{q_2(\beta)}[\beta]) \\
&\propto \mathbb{1}_{zy>0}\ exp(z^\top z - 2z^\top X\mathbb{E}_{q_2(\beta)}[\beta] + (X\mathbb{E}_{q_2(\beta)}[\beta])^\top (X\mathbb{E}_{q_2(\beta)}[\beta])) \\
&\propto \mathbb{1}_{zy>0}\ N(z; X\mathbb{E}_{q_2(\beta)}[\beta], I_n)
\end{aligned}$$

Which is a truncated normal distribution.
$q_2(\beta)$:
Let now q_1 be fixed. By ignoring factors that do not depend on β we get

$$\begin{aligned}
q_2(\beta) &\propto p(\beta)\ exp(\mathbb{E}_{q_1(z)}[log(N(z; X\beta, I_n))]) \\
&\propto p(\beta)\ exp(\mathbb{E}_{q_1(z)}[z^\top z - 2z^\top X\beta + \beta^\top X^\top X\beta]) \\
&\propto p(\beta)\ exp(-2\mathbb{E}_{q_1(z)}[z]^\top X\beta] + \beta^\top X^\top X\beta) \\
&\propto p(\beta)\ exp(\mathbb{E}_{q_1(z)}[z]^\top \mathbb{E}_{q_1(z)}[z] - 2\mathbb{E}_{q_1(z)}[z]^\top X\beta + \beta^\top X^\top X\beta) \\
&\propto p(\beta)\ N(\mathbb{E}_{q_1(z)}[z]; X\beta, I_n) \\
&= N(\beta; 0, \tau I_p) N(\mathbb{E}_{q_1(z)}[z]; X\beta, I_n)
\end{aligned}$$

Here we apply again **Theorem 1** as already seen at the Gibbs sampling part and we get

$$\begin{aligned}
q_2(\beta) &= N(\beta; m_p, S_p) \\
S_p &= (\frac{1}{\tau} I_p + X^\top X)^{-1} \\
m_p &= S_p X^\top \mathbb{E}_{q_1(z)}[z]
\end{aligned}$$

We can initialize $q_2(\beta)$ to the prior and then use these update formulas to iteratively update q_1 and q_2. We then finally get the marginals of the posterior $p(\beta|y)$ from the marginals of q_2 which are easy to compute as q_2 is gaussian.

Comparison
Figure 1 shows in red the by mean-field VB estimated posterior marginal densities and in blue the histograms of the Gibbs sampler. The means of the two estimates always correspond almost perfectly. However the density estimate of mean-field-VB has a little too much mass on the mean, it systematically underestimates the variance. We could therefore say the estimation by Gibbs sampler is qualitatively better but this also comes to a runtime price. The 15000 Gibbs samples were computed in 25 seconds whereas the mean-field algorithm already converged with only 50 iterations in less than one second.

1.5 mean-field VB variance under-estimation

The variance estimate of mean-field algorithm is given by the diagonal elements of S_p that do never change during the algorithm.

$$\hat{Var}(\beta_i|y) = (S_p)_{i,i}$$

The law of total variance tells us

$$Var(\beta_i|y) = \mathbb{E}[Var(\beta_i|y, z)] + Var(\mathbb{E}[\beta_i|y, z])$$

Now if the mean-field assumption of $p(\beta|y, z)$ factorizing in q_1, q_2 is true then we can use our above calculations where we have shown that the variance estimate is given by the diagonal elements of S_p that never changes during the algorithm and write

$$Var(\beta_i|y) = ((S_p)_{i,i}) + Var(\mathbb{E}[\beta_i|y, z]) \geq (S_p)_{i,i} = \hat{Var}(\beta_i|y)$$

1.6 Complete separation

Complete separation as $y_i \beta_i^\top x_i > 0$ means that all y_i are separated from all x_i by a hyperplane. Figure 1.6 illustrates this in 2D. A maximum likelihood estimator for our model does not exist in this case. To see this let β be given such that it fulfills complete separation. Then we have

$$\begin{aligned} L(\beta) &= \prod_i p(y_i | x_i, \beta) \\ &= \prod_i \mathbb{P}(sign(y_i) = sign(N(\beta^\top x_i, 1))) \\ &= \prod_i \mathbb{P}(N(0,1) < |\beta^\top x_i|) \end{aligned}$$

therefore we have

$$\lim_{\lambda \to \infty} L(\lambda \beta) = 1$$

And thus the MLE does not exist. All entries of β would needed to be set to $\pm\infty$.

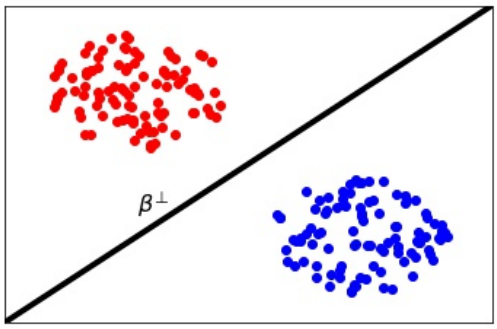

Figure 2: Two point clouds that are linearly separable. The black line is orthogonal to the separating β

I run the Gibbs sampler for this case. Figure 1.6 shows the histograms of the sampled β components. Even though the MLE does not exist we can still have meaningfull samples because the posterior distribution also depends on the chosen prior, which in our case limits the absolute value of the components of β. Anyhow the distributions still have a tail pointing to $\pm\infty$.

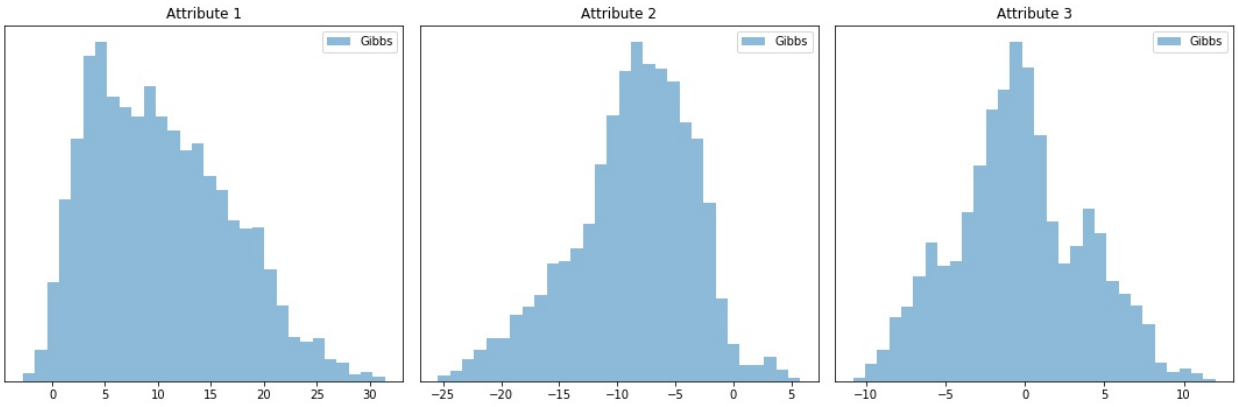

Figure 3: Histograms from the Gibbs sampler in a complete separation situation