20592 - Statistics and Probability
Bayesian Estimation of a Probit Regression Model

Stefano Cortinovis
stefano.cortinovis@studbocconi.it

Daniele Micheletti
daniele.micheletti@studbocconi.it

Andrea Teruzzi
andrea.teruzzi@studbocconi.it

Leonardo Yang
leonardo.yang@studbocconi.it

1 Project Introduction

This report aims at illustrating the use of Bayesian methods for estimating the coefficient of a probit regression model for binary outcomes. Section 2 succinctly describes the probit regression model and the Bayesian approach to estimating its coefficients using instances of the Metropolis-Hastings algorithm. Next, section 3 and 4 briefly explain two such instances, namely the Metropolis algorithm and the auxiliary variable Gibbs sampler. Then, section 5 compares the performance of the two algorithms on a dataset provided by Finney [2] to study the relationship between the occurrence of transient skin vasorestriction and the rate and volume of the air inspired by the individuals undergoing the test. Performance is assessed by means of basic diagnostics, such as trace plots and acceptance rate computation. As prescribed by the project's description, this report is heavily inspired by Albert and Chib [1].

2 Probit Model

Let $Y_1, ..., Y_n$ be a sample of n independent binary random variables, where $Y_i \sim B(p_i)$. The probit model assumes that $p_i = \Phi(\mathbf{x}_i^T \beta), i = 1, ..., n$, where $\mathbf{x}_i^T = (x_{i1} \quad \dots \quad x_{1k})$ is a k-dimensional vector of observed covariates, β is a $k \times 1$ vector of unknown coefficients, and Φ denotes the cumulative distribution function of a standard normal random variable. Given a sample of observations $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$, the standard estimation procedure for the probit model employs the maximum likelihood estimation of coefficient β. However, the Bayesian approach to the estimation of β constitutes an interesting alternative. In particular, suppose the posterior density of β, denoted by $\pi(\beta|\mathbf{y})$, was known and tractable. Then, the value of the posterior first moment, $\mathbb{E}[\beta|\mathbf{y}]$, would constitute a reasonable estimate for β. However, if we denote the proper or improper prior density of β by $\pi(\beta)$, the posterior density of β has the form:

$$\pi(\beta|\mathbf{y}) \propto \pi(\beta) \prod_{i=1}^{n} \Phi(\mathbf{x}_i^T \beta)^{y_i} (1 - \Phi(\mathbf{x}_i^T \beta))^{1-y_i},$$

which is intractable.

To address the issue of sampling from intractable posteriors, the Metropolis-Hastings algorithm is often employed in Bayesian statistics. By means of a proposal density used to sample candidates and a probabilistic criterion to accept or reject them, this method is able to generate a Markov chain $\{\beta_t\}_{t \geq 1}$ having the target posterior as stationary distribution. This means that, if such a Markov chain of length T is generated and $t_0 < T$ denotes the step at which the chain can be considered to have converged to the stationary distribution $\pi(\beta|\mathbf{y})$, the estimator

$$\hat{\beta} = \frac{1}{T - t_0} \sum_{t=t_0}^{T} \beta_t$$

is unbiased and consistent for $\mathbb{E}[\beta|\mathbf{y}]$.

Instances of the MH algorithm differ in terms of the proposal density used. In particular, the methods employed in this report are the Metropolis algorithm and the auxiliary variable Gibbs sampler.

3 Metropolis Algorithm

The Metropolis algorithm is an instance of the MH algorithm with proposal density $q(\beta^*|\beta_t)$ symmetric around the current value of the chain, namely β_t. When $q(y|x)$ has such a property, the candidate acceptance probability becomes:

$$\alpha(\beta_t, \beta^*) = \min\left\{1, \frac{\pi(\beta^*|\mathbf{y})}{\pi(\beta_t|\mathbf{y})}\right\} = \min\left\{1, \frac{\pi(\beta^*)}{\pi(\beta_t)}\frac{\mathcal{L}(\beta^*)}{\mathcal{L}(\beta_t)}\right\}$$

where $\mathcal{L}(\beta)$ denotes the likelihood of the model at β.

The proposal used in this report is a multivariate normal distribution. In particular, at each sampling step performed by the algorithm, a candidate is drawn from $q(\beta^*|\beta_t) = N_k(\beta_t, V)$ where $V = \mathcal{I}^{-1}(\beta_t)$ is the inverse of the Fisher information matrix computed at the current value of β_t. Moreover, recall that, in the case of a generalized linear model with link function g, it can be shown that $\mathcal{I}(\beta_t) = X^T W X$ where $X = [\mathbf{x}_1 \dots \mathbf{x}_n]$ and W is a diagonal matrix such that

$$w_{ii} = \frac{1}{\mathbb{V}(Y_i)}\left(\frac{\partial \mathbf{x}_i^T\beta}{\partial g(\mathbf{x}_i^T\beta)}\right)^{-2}.$$

This means that, in the case of the probit model, $w_{ii} = \phi^2(\mathbf{x}_i^T\beta)/(p_i(1-p_i))$.

Algorithm 1 summarizes the steps performed by the Metropolis algorithm presented in this section.

Algorithm 1 Metropolis Algorithm for Probit Estimation

procedure METROPOLIS(β_0)
  $\beta \leftarrow \beta_0$
  $m \leftarrow [\beta_0]$
  repeat
    $\beta^* \leftarrow \beta^* \sim N_k(\beta, \mathcal{I}^{-1}(\beta))$
    $u \leftarrow U \sim \text{Unif}(0,1)$
    $\alpha \leftarrow \alpha(\beta, \beta^*) = \min\left\{1, \frac{\pi(\beta^*)}{\pi(\beta)}\frac{\mathcal{L}(\beta^*)}{\mathcal{L}(\beta)}\right\}$
    if $u \leq \alpha$ **then**
      $\beta \leftarrow \beta^*$
    end if
    append β to m
  until stopping criterion is reached
  return m
end procedure

4 Auxiliary Variable Gibbs Sampler

Sometimes, when a Gibbs sampler cannot be easily devised for a target distribution, it is possible to exploit a vector of auxiliary random variables to successfully complete the task. This straightforward extension of the simple Gibbs sampler is called auxiliary variable Gibbs sampler. In particular, if the target distribution is $\pi(\beta|\mathbf{y})$ and we are able to find a random vector $\mathbf{Z}$ such that the full conditionals $\pi(\beta|\mathbf{y}, \mathbf{Z})$ and $\pi(\mathbf{Z}|\mathbf{y}, \beta)$ are known and can be sampled from, then a Gibbs sampler can be used to generate the Markov chain $\{(\beta_t, \mathbf{Z}_t)\}_{t\geq 1}$. Crucially, it is possible to show that the sequence $\{\beta_t\}_{t\geq 1}$ is also a Markov chain and has the target distribution $\pi(\beta|\mathbf{y})$ as its stationary distribution.

In regards to the scenario considered in this report, it turns out that any probit model can be expressed in terms of a normal linear model. In particular, introduce auxiliary vector $\mathbf{Z} = (Z_1 \dots Z_n)$ such that the Z_i are independent $N(\mathbf{x}_i^T\beta, 1)$ and define $Y_i = \mathbb{1}_{(0,\infty)}(Z_i)$, i.e. $Y_i = 1$ if $Z_i > 0$ and $Y_i = 0$

otherwise. Then, it is possible to show that the Y_i are independent Bernoulli random variables with $p_i = \Phi(\mathbf{x}_i^T \beta)$, as prescribed by the probit model.

As mentioned above, to use a Gibbs sampler to sample from target $\pi(\beta|\mathbf{y})$ using $\mathbf{Z}$ as the auxiliary variable, the full conditionals $\pi(\beta|\mathbf{y}, \mathbf{Z})$ and $\pi(\mathbf{Z}|\mathbf{y}, \beta)$ have to be known distributions. First, the full conditional of β is given by:

$$\pi(\beta|\mathbf{y}, \mathbf{Z}) = \pi(\beta|\mathbf{Z}) \propto \pi(\beta)\pi(\mathbf{Z}|\beta) = \pi(\beta) \prod_{i=1}^{n} \phi(Z_i; \mathbf{x}_i^T \beta, 1),$$

which depends on the choice of the prior $\pi(\beta)$. In particular, if $\pi(\beta)$ is chosen to be non-informative, we have that

$$\begin{aligned}
\pi(\beta|\mathbf{y}, \mathbf{Z}) &\propto \prod_{i=1}^{n} \phi(Z_i; \mathbf{x}_i^T \beta, 1) \propto \exp\left\{ \sum_{i=1}^{n} (Z_i - \mathbf{x}_i^T \beta)^2 \right\} \\
&\propto \exp\Big\{ (\mathbf{X}\beta)^T (\mathbf{X}\beta) - 2(\mathbf{X}\beta)^T \mathbf{Z} \Big\} \\
&= \exp\Big\{ (\mathbf{X}\beta)^T (\mathbf{X}\beta) - 2(\mathbf{X}\beta)^T \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \mathbf{Z} \Big\}.
\end{aligned}$$

By completing the square with the product of the transpose of $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}$ with itself,

$$\begin{aligned}
\pi(\beta|\mathbf{y}, \mathbf{Z}) &\propto \exp\Big\{ \big(\mathbf{X}\beta - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}\big)^T \big(\mathbf{X}\beta - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}\big) \Big\} \\
&= \exp\Big\{ \big(\beta - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}\big)^T (X^T X) \big(\beta - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}\big) \Big\}
\end{aligned}$$

which is the kernel of a multivariate normal random variable with with mean $\hat{\beta}_Z = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}$ and variance $(\mathbf{X}^T\mathbf{X})^{-1}$. That is, if $\pi(\beta)$ is chosen to be non-informative, $\beta|\mathbf{y}, \mathbf{Z} \sim N_k(\hat{\beta}_Z, (\mathbf{X}^T\mathbf{X})^{-1})$. On the other hand, if the conjugate prior $\beta \sim N_k(\beta^*, \mathbf{B}^*)$ is chosen, we it can be shown that $\beta|\mathbf{y}, \mathbf{Z} \sim N_k(\tilde{\beta}, \tilde{\mathbf{B}})$ where $\tilde{\beta} = (\mathbf{B}^{*-1} + \mathbf{X}^T\mathbf{X})^{-1}(\mathbf{B}^{*-1}\beta^* + \mathbf{X}^T\mathbf{Z})$ and $\tilde{\mathbf{B}} = (\mathbf{B}^{*-1} + \mathbf{X}^T\mathbf{X})^{-1}$.

Next, for each i, the full conditional of Z_i is given by:

$$\pi(Z_i|\mathbf{y}, \beta) \propto \begin{cases} \phi(Z_i; \mathbf{x}_i^T\beta, 1)\mathbb{1}_{(0,+\infty)}(Z_i) & \text{if } y_i = 1 \\ \phi(Z_i; \mathbf{x}_i^T\beta, 1)\mathbb{1}_{(-\infty,0]}(Z_i) & \text{if } y_i = 0 \end{cases}$$

that is, Z_i has a $N(\mathbf{x}_i^T\beta, 1)$ truncated by 0 at the left if $y_i = 1$ and at the right if $y_i = 0$.

Notice that, if we choose the prior $\pi(\beta)$ in one of the two ways mentioned above, both full conditionals have a known distribution and can be easily sampled from. This means that we can indeed exploit $\mathbf{Z}$ to use an auxiliary variable Gibbs sampler to sample from $\pi(\beta|\mathbf{y})$. Finally, Algorithm 2 summarizes the steps performed by the auxiliary variable Gibbs sampler presented in this section.

Algorithm 2 Auxiliary Variable Gibbs Sampler for Probit Estimation

procedure AUXILIARYGIBBS(β_0)
  $\beta \leftarrow \beta_0$
  $m \leftarrow [\beta_0]$
  repeat
    $\mathbf{Z} \leftarrow \mathbf{Z}|\mathbf{y}, \beta \sim \pi(\mathbf{Z}|\mathbf{y}, \beta)$
    $\beta \leftarrow \beta|\mathbf{y}, \mathbf{Z} \sim \pi(\beta|\mathbf{y}, \mathbf{Z})$
    append β to m
  until stopping criterion is reached
  return m
end procedure

5 Diagnostics and Performance Comparison

Following the pseudo-code described in Algorithm 1 and 2 we implemented the Metropolis and auxiliary variable Gibbs sampler algorithms using the Python programming language. In both cases,

the user is allowed to specify whether to include an intercept in the estimation, as well as the initial value for β_0, with the default being the OLS estimate for β. Moreover, for the Metropolis algorithm, the user is allowed to specify a custom prior distribution, while the Gibbs sampler implemented the full conditional associated with the non-informative and conjugate multivariate normal prior mentioned in Section 4.

What follows is a short summary of the tests performed to assess the performance of the two algorithms. The dataset used for all tests is the one introduced by Finney [2] and briefly mentioned in Section 1.

In particular, both the Metropolis algorithm and the auxiliary variable Gibbs sampler were used to estimate a probit model using three different priors:

1. Non-informative prior;
2. Multivariate normal prior with high variance (diagonal covariance matrix with diagonal entries equal to 10) centered around the true values of the betas;
3. Multivariate normal prior with low variance (diagonal covariance matrix with diagonal entries equal to 1) centered around the true values of the betas.

Every model estimated included the intercept and the starting value β_0 was taken to be equal to $\hat{\beta}_{OLS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ for every estimation. For both algorithms, and for each of the priors mentioned above, the fitting procedure was run thrice with number of iterations equal to 200, 800, and 20000, respectively. Lastly, a *warmup*, i.e. number of iterations discarded from the output chain before computing the Bayesian estimator $\hat{\beta}$, of 200 was used for the runs with 20000 total iterations.

Below, distribution and trace plots of β for the simulations performed with 800 iterations are shown. Additionally, the acceptance rate behaviour across iterations for the Metropolis algorithm is also plotted.

First, Figure 1 compares the trace plots obtained using the two algorithms with a non-informative prior. From that, we can immediately notice how the Gibbs Sampler trace plot seems to be of higher quality in terms of the final chain's representativeness of the posterior of β. The main reason behind this is that, since the Gibbs Sampler can be interpreted of the MH framework as always accepting the proposal, it seems to better explore the sample space of β. The Metropolis algorithm, instead, accepts the proposal with a certain probability. This means that, sometimes, the proposal is rejected and the new state of the Markov chain is the same as in the previous iteration. From the trace plot, we can see that the output chain for the Metropolis algorithm explores the sample space of β based on "local movements", making the next observation more likely to be correlated with the previous one and increasing the number of iterations required for the chain to reach its stationary distribution.

Moreover, from the simulations performed using Metropolis, we also noticed that, as shown in Figure 2, the prior with high variance outperformed the one with low variance. Given that the true value of the coefficients is not very close to zero, this is expected. For priors centered at zero, lower variance means narrower exploration of the sample space of β. Regarding the Gibbs Sampler, as shown in figure 3, the simulation with multivariate normal prior produced a better trace plot then the one with non-informative prior.

Looking at the acceptance rate of the Metropolis with non-informative prior displayed in Figure 4, it is possible to notice a stabilization of the rate around 0.4 after nearly 500 iterations. This value can be considered quite satisfactory. On the one hand, it is not too high, which would lead to the risk of a local exploration. On the other hand, it is not too low, that would lead to the risk of slow convergence to the stationary distribution.

Finally the distribution plots of the the parameters obtained, as shown in Figure 5, are nearly the same as the ones discussed by Albert and Chib [1] in their paper.

References

[1] James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679, 1993.

[2] D. J. Finney. The estimation from individual records of the relationship between dose and quantal response. *Biometrika*, 34(3/4):320–334, 1947.

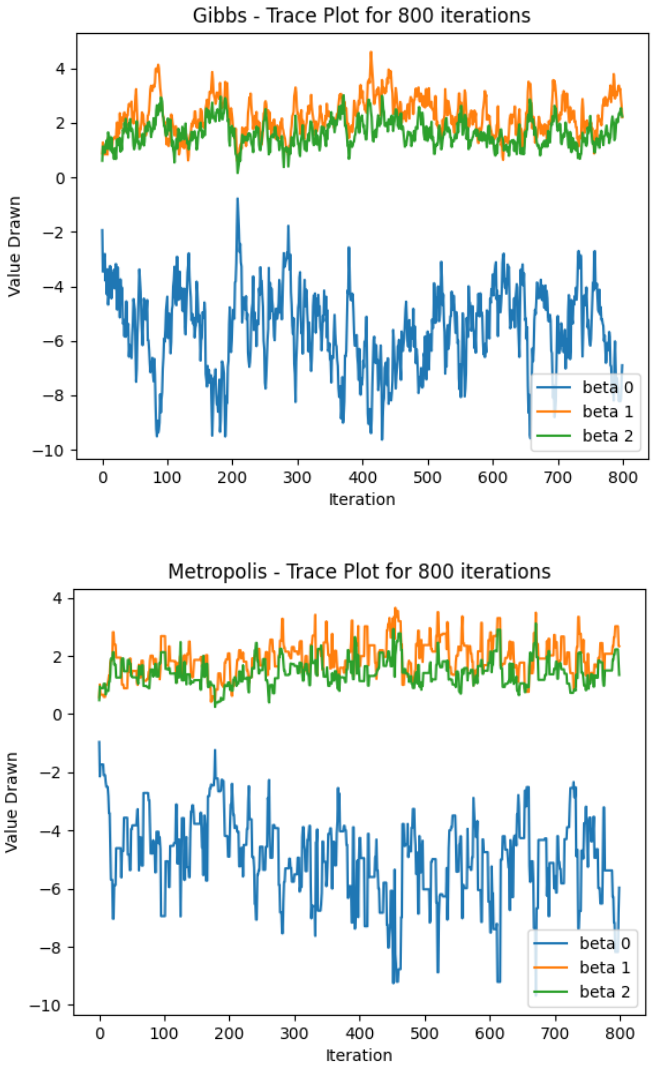


Figure 1: Trace plots for both algorithms with non informative prior.

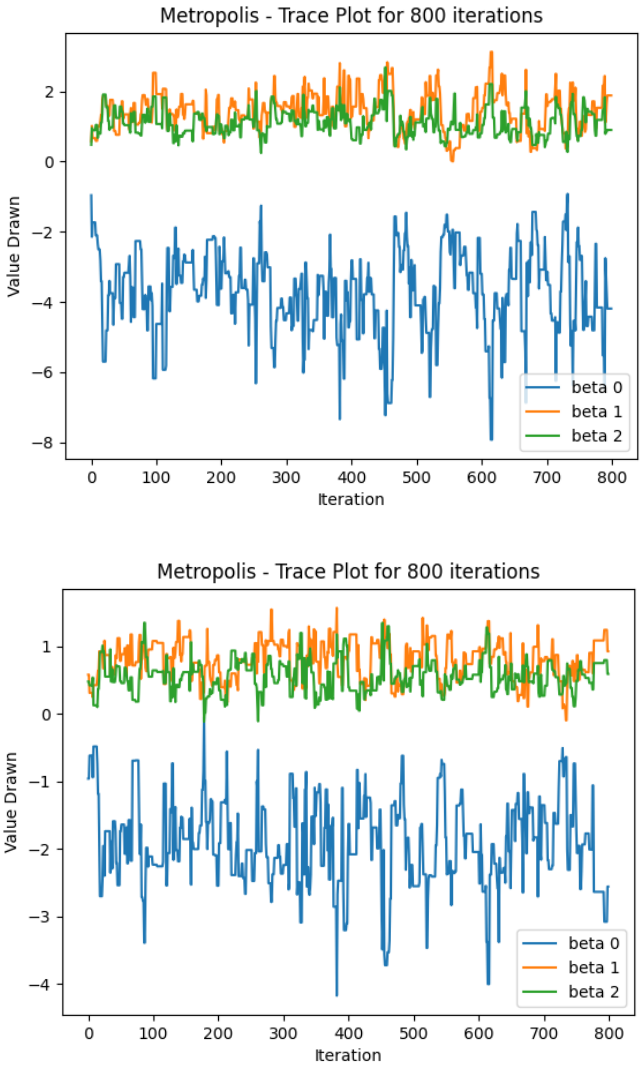


Figure 2: Starting from the top: trace plot for the coefficients sampled using high variance multivariate normal prior with Metropolis, trace plot for the coefficients sampled using low variance multivariate normal prior with Metropolis.

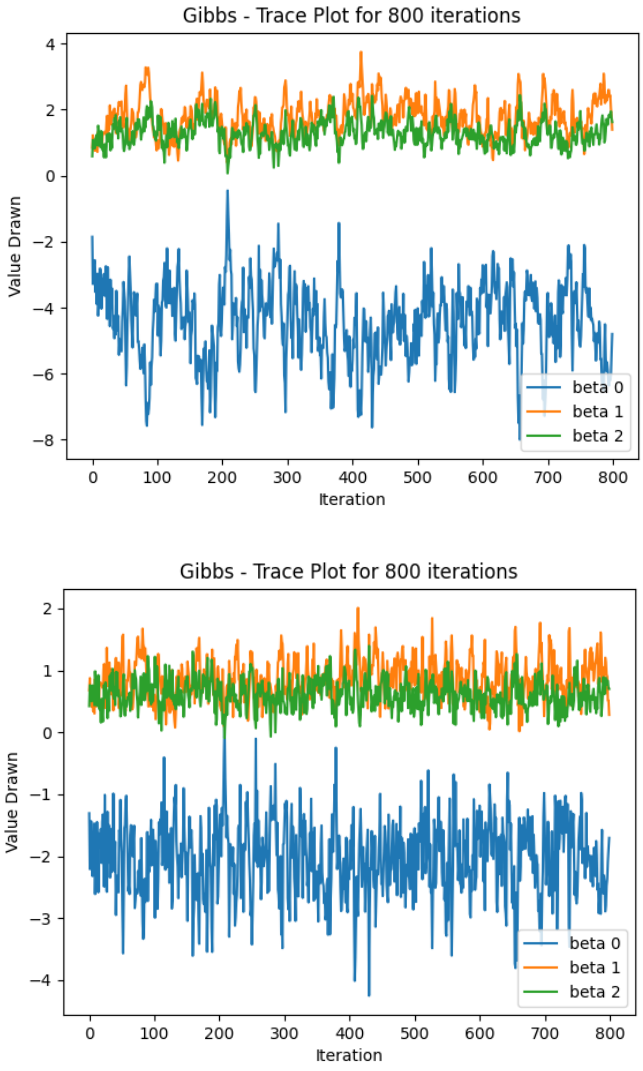
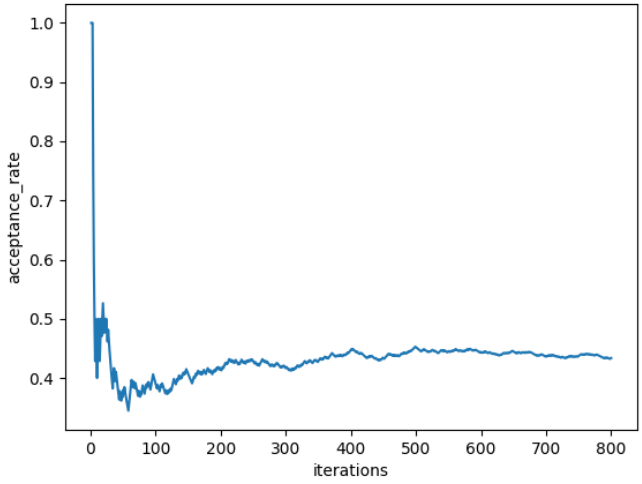


Figure 3: Starting from the top: trace plot for the coefficients sampled using high variance multivariate normal prior on Gibbs, trace plot for the coefficients sampled using low variance multivariate normal prior on Gibbs.



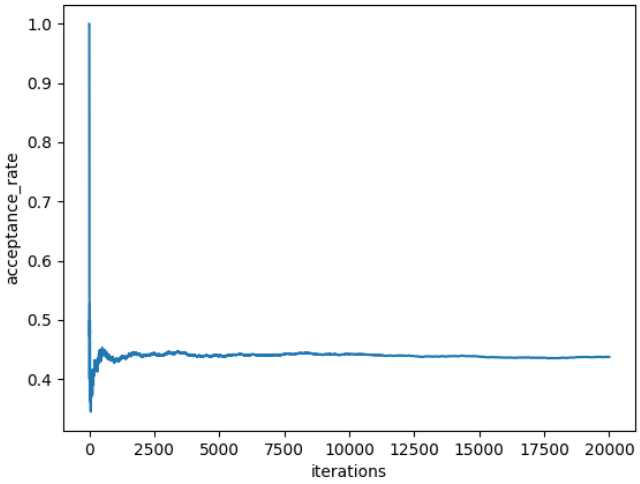
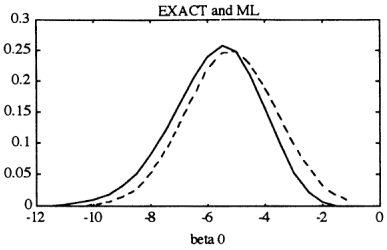
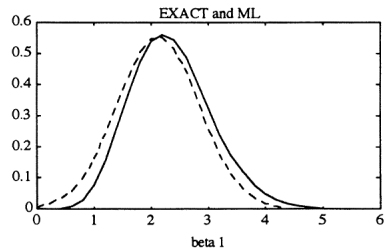
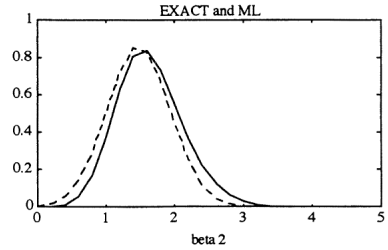


Figure 4: Starting from the top: acceptance rate of the Metropolis algorithm with non informative prior for 800 iterations, acceptance rate of the Metropolis algorithm with non informative prior for 20000 iterations.







Gibbs - Distribution Plot for Beta after 20000 iterations

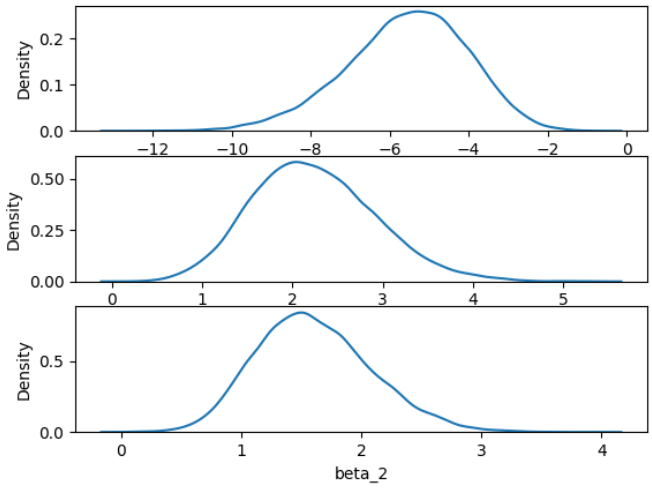


Figure 5: Starting from the top: the distribution plots from Albert and Chib, the distribution plots of the parameters with Gibbs and Metropolis, respectively.