Computational Statistics: Final Project

Bottani Sophie, Chiarini Emanuele,
Gatteschi Giulia, Giannelli Enrico

December 2021

1 Introduction

In this report we explore two different approaches for the estimation of the coefficients of a Bayesian Probit model. In particular, we describe a Metropolis algorithm and an Auxiliary Variables Gibbs sampler. In section 2 we present the theory behind the models and in section 3 the derivation of the full conditionals of the Auxiliary Variables Gibbs sampler. In section 4 we run the two models both on a simulated and on a real dataset with different parameter values. In section 5 we explain one of the diagnostics we ran, the Gelman Rubin statistics while in section 6 we conclude.

2 The models

Consider $Y_1, ..., Y_n$ so that $Y_i \overset{\text{iid}}{\sim} Bernoulli(\pi_i)$, a set of covariates $X_2, ..., X_p$, and a vector of ones X_1:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{12} & \dots & x_{1p} \\ 1 & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n2} & \dots & x_{np} \end{bmatrix}$$

We use the vectors $X_1, ..., X_p$ to form the linear predictor:

$$\eta_i = X_i^T \beta$$

We now link this to π_i. Since we are considering a Probit model we choose as link function the probit link:

$$\phi^{-1}(\pi_i) = \eta_i$$

It is then useful to compute the likelihood function:

$$f(Y|\beta) = \prod_{i=1}^{n} (\phi(\eta_i))^{y_i} (1 - \phi(\eta_i))^{1-y_i}$$

where ϕ is the Standard Normal cdf.

Being in a Bayesian setting we then compute the posterior distribution which is what we will use to learn about our parameters.

$$\pi(\beta|Y) \propto \pi(\beta) \prod_{i=1}^{n} (\phi(\eta_i))^{y_i} (1 - \phi(\eta_i))^{1-y_i}$$

where $\pi(\beta)$ is a prior distribution assigned to the vector of parameters β.

This posterior distribution is not solvable analytically. Thus, we resort to a solution that consists in simulating $\pi(\beta|Y)$ through two different MCMC processes: a Metropolis algorithm and an Auxiliary Variables Gibbs Sampler.

2.1 Metropolis algorithm

In a Metropolis algorithm we sample from a proposal distribution and accept or reject the proposed β_t depending on the acceptance probability it yields.

Here we approximate the posterior distribution using as proposal distribution a multivariate normal with the current update of β, β_t as mean and with the inverse of the Fisher information evaluated at the current update as covariance matrix. Thus we have:

$$q(\beta^*|\beta_t) \sim N(\beta_t, \tau V)$$
$$V = (-\mathcal{I}''(\beta_t))^{-1}$$

It easily follows that the acceptance probability is:

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

2.2 Auxiliary Variables Gibbs sampler

In this implementation of the Auxiliary Variables Gibbs sampler we take advantage of the representation of the probit model in terms of latent normal variables.

Let $Z_1, Z_2, ..., Z_n$ be n auxiliary variables defined as

$$Y_i = \begin{cases} 1 & Z_i > 0 \\ 0 & Z_i \leq 0 \end{cases} \qquad \forall i = 1, ..., n$$

with $Z_i \overset{\text{iid}}{\sim} N(X_i^T\beta, 1)$.

Equivalently, we can write:

$$Z_i = X_i^T\beta + \epsilon_i$$

where $\epsilon_i \overset{\text{iid}}{\sim} N(0, 1)$.

In order to implement the algorithm we need to work out the following full conditionals from which the sampling will be carried out sequentially: $\pi(\beta|Z, Y)$ and $\pi(Z|\beta, Y)$. The full derivations are included in the following section.

$\pi(\beta|Z, Y)$ is given by:

$$\pi(\beta|Z, Y) = \pi(\beta|Z) \propto \pi(\beta)\pi(Z|\beta)$$

which follows a Normal density. The parameters of this Normal density differ depending on the choice of the prior $\pi(\beta)$ as follows:

- In the case of a non-informative prior

$$\mu = (X^TX)^{-1}X^TZ \quad \text{and} \quad \Sigma = (X^TX)^{-1}$$

- While in the case of a normal prior with mean β_0 and variance V_0

$$\mu = (V_0^{-1} + X^TX)^{-1}(V_0^{-1}\beta_0 + X^TZ) \quad \text{and} \quad \Sigma = (V_0^{-1} + X^TX)^{-1}$$

$\pi(Z|\beta, Y)$ is,instead, given by:

$$\pi(Z|\beta, Y) \propto \pi(Y|Z)\pi(Z|\beta)$$

Which we will see implies that conditionally on β and on Y, $Z_1, ...Z_n$ are independent variables so that Z_i has a normal distribution centered at $X_i^T\beta$, scaled by 1 and truncated by 0: at the left if the corresponding observation is $Y_i = 1$ or at the right otherwise.

3 Full conditional Derivations

3.1 Derivation of $\pi(\beta|Z, Y)$

We now derive the full conditionals. We start with $\pi(\beta|Z, Y)$ in the non-informative prior case. As we said before:

$$\pi(\beta|Z, Y) = \pi(\beta|Z) \propto \pi(Z|\beta)\pi(\beta)$$

Having $Z_i|\beta \sim N\left(X_i^T\beta, 1\right)$ by construction and $\pi(\beta) \propto 1$ by assumption we get:

$$\begin{aligned}
&\propto \exp\left[-\frac{1}{2}\left(Z - X\beta\right)^T\left(Z - X\beta\right)\right] \\
&= \exp\left[-\frac{1}{2}\left(Z^TZ - 2\beta^TX^TZ + \beta^TX^TX\beta\right)\right]
\end{aligned}$$

Which simply by multiplying (and dividing) by $(X^TX)^{-1}$:

$$\begin{aligned}
&= \exp\left[-\frac{\left(X^TX\right)^{-1}}{2\left(X^TX\right)^{-1}}\left(Z^TZ - 2\beta^TX^TZ + \beta^TX^TX\beta\right)\right] \\
&= \exp\left[-\frac{1}{2\left(X^TX\right)^{-1}}\left(\left(Z^TZ\right)\left(X^TX\right)^{-1} - 2\left(\beta^TX^TZ\right)\left(X^TX\right)^{-1} + \beta^T\beta\right)\right] \\
&= \exp\left[-\frac{1}{2\left(X^TX\right)^{-1}}\left(\beta - \left(X^TX\right)^{-1}\left(X^TZ\right)\right)^T\left(\beta - \left(X^TX\right)^{-1}\left(X^TZ\right)\right)\right]
\end{aligned}$$

Which implies as wanted:

$$\beta|Z, Y \sim N\left(\left(X^TX\right)^{-1}\left(X^TZ\right), \left(X^TX\right)^{-1}\right)$$

We can similarly perform this derivation in the case in which $\beta \sim N(\beta_0, V_0)$:

$$\begin{aligned}
\pi(\beta|Z,Y) = \pi(\beta|Z) &\propto \pi(Z|\beta)\pi(\beta) \\
&\propto \exp\left[-\frac{1}{2}\left(Z - X\beta\right)^T\left(Z - X\beta\right)\right]\exp\left[-\frac{1}{2}\left(\beta - \beta_0\right)^TV_0^-1\left(\beta - \beta_0\right)\right] \\
&= \exp\left[-\frac{1}{2}\left(Z^TZ - 2\beta^TX^TZ + \beta^TX^TX\beta + \beta^TV_0^{-1}\beta - 2\beta^TV_0^{-1}\beta_0 + \beta_0^TV_0^{-1}\beta_0\right)\right] \\
&= \exp\left[-\frac{1}{2}\left(\beta^T\left(V_0^{-1} + X^TX\right)\beta - 2\beta^T\left(V_0^{-1}\beta_0 + X^TZ\right) + \left(Z^TZ + \beta_0^TV_0^{-1}\beta_0\right)\right)\right]
\end{aligned}$$

Which simply by multiplying (and dividing) by $\left(V_0^{-1} + X^TX\right)^{-1}$:

$$\begin{aligned}
&= \exp\left[-\frac{\left(V_0^{-1} + X^TX\right)^{-1}}{2\left(V_0^{-1} + X^TX\right)^{-1}}\left(\beta^T\left(V_0^{-1} + X^TX\right)\beta - 2\beta^T\left(V_0^{-1}\beta_0 + X^TZ\right) + \left(Z^TZ + \beta_0^TV_0^{-1}\beta_0\right)\right)\right] \\
&= \exp\left[-\frac{1}{2\left(V_0^{-1} + X^TX\right)^{-1}}\left(\beta^T\beta - 2\beta^T\left(V_0^{-1}\beta_0 + X^TZ\right)\left(V_0^{-1} + X^TX\right)^{-1} + \left(Z^TZ + \beta_0^TV_0^{-1}\beta_0\right)\left(V_0^{-1} + X^TX\right)^{-1}\right)\right] \\
&\propto \exp\left[-\frac{1}{2\left(V_0^{-1} + X^TX\right)^{-1}}\left(\beta - \left(V_0^{-1} + X^TX\right)^{-1}\left(V_0^{-1}\beta_0 + X^TZ\right)\right)^T\left(\beta - \left(V_0^{-1} + X^TX\right)^{-1}\left(V_0^{-1}\beta_0 + X^TZ\right)\right)\right]
\end{aligned}$$

Which implies as wanted:

$$\beta|Z, Y \sim N\left(\left(V_0^{-1} + X^TX\right)^{-1}\left(V_0^{-1}\beta_0 + X^TZ\right), \left(V_0^{-1} + X^TX\right)^{-1}\right)$$

3.2 Derivation of $\pi(Z|\beta, Y)$

We now turn to $\pi(Z|\beta, Y)$.

We have:

$$\pi(Z|\beta, Y) = \frac{\pi(Z, \beta|Y)}{\pi(\beta)} \propto \frac{\pi(Y|Z, \beta)\pi(Z, \beta)}{\pi(\beta)} = \frac{\pi(Y|Z, \beta)\pi(Z|\beta)\pi(\beta)}{\pi(\beta)} = \pi(Y|Z)\pi(Z|\beta) = \prod_{i=1}^{N}\pi(Y_i|Z_i)\pi(Z_i|\beta)$$

Importantly, though: $Y_i|Z_i = \mathbb{1}_{[0,+\infty)}(Z_i)$

Thus, $Z_i|\beta, Y_i$ has a Normal distribution having as mean $(X_i^T\beta)$, as variance 1 and truncated by 0 at the left if $Y_i = 1$ and at the right if $Y_i = 0$.

4 Simulation

In the following section we are going to apply the Metropolis algorithm and the Auxiliary Variable Gibbs sampler to two datasets, one simulated and one real. For each dataset we are running both algorithms, starting from different prior distributions:

- **Uninformative** : no prior impacting the results.
- **Weak Normal Prior** : a vague standard multivariate Gaussian prior.
- **Strong Normal Prior** : a multivariate Gaussian distribution centered on the "true" β with greatly reduced variance.

Also, for each simulation we run three chains having three different starting points: vector of zeros, least squares estimates and MLE estimates for the parameters.

4.1 Simulated Data

The simulated dataset is composed of two regressors β_1 and β_2, together with a constant term β_0. The dataset is composed by $n = 1000$ realizations of Gaussian covariates that are then linearly combined into η. Using a normal CDF we convert η into probability p. Binary Y_i are then generated through a Bernoulli trial with probability p for each occurrence. The coefficients used in the linear combination that we seek to estimate are:

$$\beta_0 = 1 \quad \beta_1 = -1 \quad \beta_2 = 1.5$$

4.1.1 Metropolis Algorithm

We run our first simulation with an uninformative prior. The obtained estimates for the betas are quite precise. However, the plots of the autocorrelation function are showing serial correlation to be slightly too much persistent. Introducing thinning (discarding 50% of the observations) noticeably improves the situation, with autocorrelation "dying" much more quickly. Thinning also improves the estimation of β_2. Turning to the traceplots we can see that they explore the sample space quite effectively (Figure 1).

The results in the simulation that uses weak normal priors are quite similar. In this case however, no thinning is required as the autocorrelation is not as persistent as in the previous case.

Turning to the case with a strong normal prior, we again can see little difference in the parameters estimates, that are similar to the weak uniform prior and slightly worse than the uninformative prior case. The traceplots point again towards a nice exploration of the state space and the ACF plots show serial correlation decreasing quite quickly.

In all three cases the Gelman-Rubin diagnostics (not presented here for conciseness) has remained in a neighbourhood of 1, suggesting that despite the different starting points the chains have all reached the same stationary distribution.

Mean of Beta Estimates Metropolis Algorithm			
	β_0	β_1	β_2
Uninformative Prior	1.0697	-0.9709	1.5942
Weak Normal Prior	1.0697	-0.9638	1.5883
Strong Normal Prior	1.0735	-0.9745	1.6019

4.1.2 Auxiliary Gibbs Sampler

As in the Metropolis Algorithm case, we run the first simulation with an uninformative prior. The estimated betas are slightly less precise than those of the Metropolis algorithm, but still well in the neighbourhood of the true betas. Looking at the autocorrelation function plots, we notice that there is some persistence in the serial correlation and we again introduce thinning and discard 50% of the observations. The post-thinning betas do not differ significantly from the previous one and the serial correlation problem is now solved. In both cases the traceplot are nicely exploring, more so than in the Metropolis Hastings case, the sample space. We also see that the estimates generated by chains starting at the LS and MLE estimates are more precise than the one starting at a vector of zeros.

Turning to the weak normal prior case, we can notice that the beta estimates are more precise compared to the uninformative prior run. In this case no thinning is required. The results are largely similar in the case with a strong normal prior, even if slightly less precise. In both runs, the traceplots pointed towards a very effective exploration of the sample space (Figure 2) and the autocorrelation plot showed that serial correlation went quickly to 0. In all three cases the Gelman Rubin statistics remained close to 1 suggesting that all the chains are managing to reach the stationary distribution.

Mean of Beta Estimates Auxiliary Gibbs Sampler			
	β_0	β_1	β_2
Uninformative Prior	1.1063	-1.0016	1.6518
Weak Normal Prior	1.0803	-0.9753	1.6125
Strong Normal Prior	1.1008	-0.9984	1.6485

4.2 Real Data

The real dataset on which we are focusing is the Pima Indians Diabetes Database, provided by the American National Institute of Diabetes. The focus of the dataset is on predicting whether a patient has diabetes, based on some diagnostical measurement, focusing on a population of Pima Indian women that are at least 21 years old. The target binary variable *Outcome* describes whether a patient has or not diabetes. The dataset includes 8 continuous regressors : Pregnancies, Glucose, Blood pressure, Skin thickness, Insulin, Body Mass Index, Diabetes Pedigree Function (DPF), Age. To get a sense of what the β of the regressor on this dataset are, we run a Probit regression with Statsmodel package. The results are as follows:

Probit estimate of the True Betas	
Constant	-0.5156
Pregnancies	0.2434
Glucose	0.6353
Blood Pressure	-0.1533
Skin thickness	0.0197
Insulin	-0.0854
Body Mass Index	0.4122
DPF	0.1650
Age	0.1198

4.2.1 Metropolis Algorithm

Our first simulation focuses again on the uninformative prior. In this run however we spot some problems, the acceptance rate is too low (around 15%) thus, signaling that the variability of the proposals is excessive. We therefore proceed to reduce τ, bringing the acceptance rate to about 40%. However, the variability in the traceplot remains too high, signalling that the chain may require more steps to fully explore the parameter space. As such, we increase the number of steps. Finally, we notice that the autocorrelation seems to be too persistent (Figure 5), therefore we introduce thinning and keep only 33% of the observations (Figure 6). After solving these problems, we found that the trace plots show an effective exploration.

Moving to the case with a weak normal prior, we keep all the tweaks to the algorithm developed over the previous run (as we notice again the same issues). Indeed, they again lead to a satisfactory result with traceplots and ACF plot showing no problems (Figure 3). Finally we move to the strong normal prior case. Again, our tweaked algorithm performance is adequate, with ACF showing no serial correlation issues and the traceplots nicely exploring the sample space. Indeed, in both cases, we found that the beta estimates' are close to the Probit ones.

Lastly, the Gelman Rubin statistics remained close to 1 in all three cases, as desired.

Mean of Beta Estimates Metropolis Algorithm			
	Uninformative Prior	Weak Normal Prior	Strong Normal Prior
Constant	-0.5130	-0.5163	-0.5033
Pregnancies	0.2451	0.2472	0.2442
Glucose	0.6274	0.6275	0.6374
Blood Pressure	-0.1515	-0.1522	-0.1547
Skin thickness	0.0128	0.0192	0.0315
Insulin	-0.0772	-0.0842	-0.0874
Body Mass Index	0.4070	0.4116	0.4064
DPF	0.1661	0.1675	0.1705
Age	0.1189	0.1187	0.1209

4.2.2 Auxiliary Gibbs Sampler

Starting from the uninformative prior case, we notice that the results are satisfactory without any change in the parameters, differently from the Metropolis algorithm case. The results are also quite similar in the weak and strong normal prior cases (Figure 4). The estimated betas are quite precise and traceplots and ACF plots point to an excellent convergence. Again, in all three cases the Gelman Rubin statistics remained close to 1.

Mean of Beta Estimates Auxiliary Gibbs Sampler			
	Uninformative Prior	Weak Normal Prior	Strong Normal Prior
Constant	-0.5165	-0.5158	-0.5163
Pregnancies	0.2411	0.2438	0.2443
Glucose	0.6394	0.6350	0.6386
Blood Pressure	-0.1545	-0.1548	-0.1556
Skin thickness	0.0228	0.0192	0.0249
Insulin	-0.0870	0.0858	0.0891
Body Mass Index	0.4148	0.4142	0.4169
DPF	0.1674	0.1636	0.1656
Age	0.1218	0.1214	0.1187

5 Gelman Rubin Diagnostics

Gelman and Rubin propose a general approach to monitoring convergence of Monte-Carlo Markov Chains in which more parallel chains are run. Starting with different initial values, we find convergence when the distinct chains have "forgotten" their starts and they are indistinguishable. Indeed, they should "look" like the stationary distribution. The diagnostic is based on a comparison of within-chain and between-chain variances. Generally, values in a neighborhood of 1 are considered good results.

6 Conclusions

As we saw from the previous section, we have different results when using a simulated dataset or a real one. In the case of the simulated dataset, the Metropolis Algorithm and the Auxiliary Gibbs Sampler behave in a fairly similar way, both giving quite precise estimates and exploring the sample space effectively. Moreover, both algorithms needed thinning when using an uninformative prior to solve the persistence in the serial correlation. When using a real dataset, instead, the Auxiliary Gibbs Sampler outperformed by far the Metropolis Algorithm. Indeed, the latter has an adequate performance only when some tweaks are applied to the algorithm and even in this case its performance is worse than the one of the Gibbs Sampler. The Auxiliary Gibbs Sampler instead, gives satisfactory results without the need of changing any of the parameters. Lastly, we noted how priors seem to play a very minor role in both cases and for both algorithms. This is probably due to the large number of observations we have in both the data sets.

Images

Note that we present only a selection of the plots in this report for conciseness purposes. The notebook includes all the images and results.

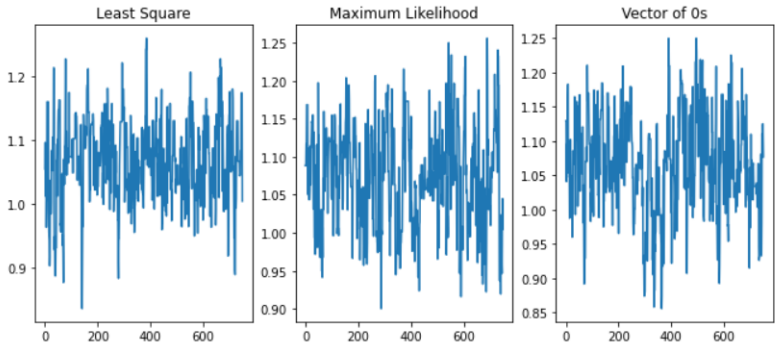

Figure 1: Traceplots Metropolis Algorithm Simulated Data with Uninformative Prior for β_0

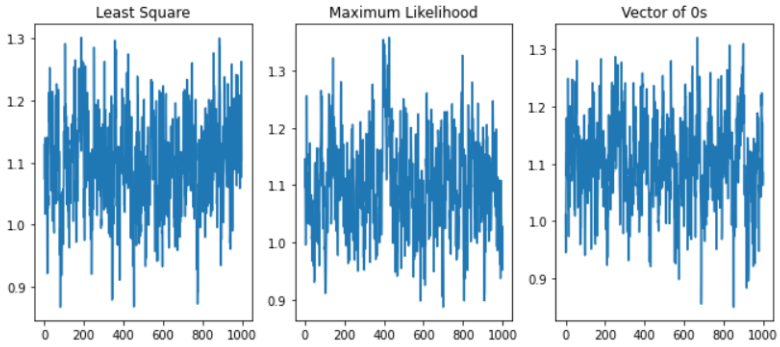

Figure 2: Traceplots Auxiliary Gibbs Sampler Simulated Data with Strong Normal Prior for β_0

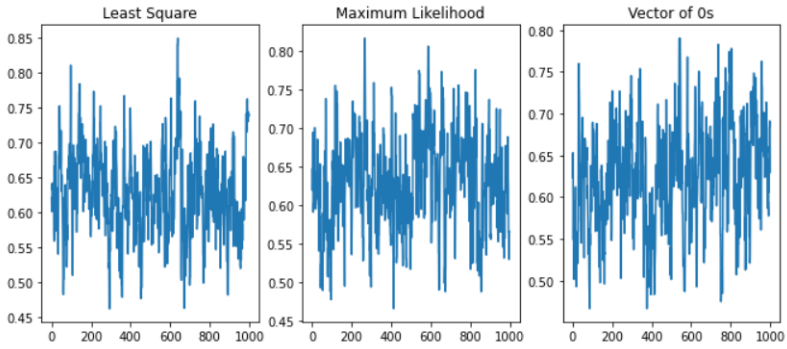

Figure 3: Traceplots Metropolis Algorithm Real Data with Weak Normal Prior for *Glucose*

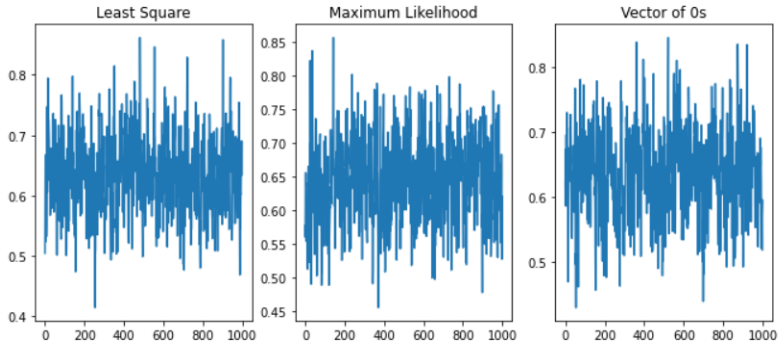

Figure 4: Traceplots Auxiliary Gibbs Sampler Real Data with Strong Normal Prior for *Glucose*

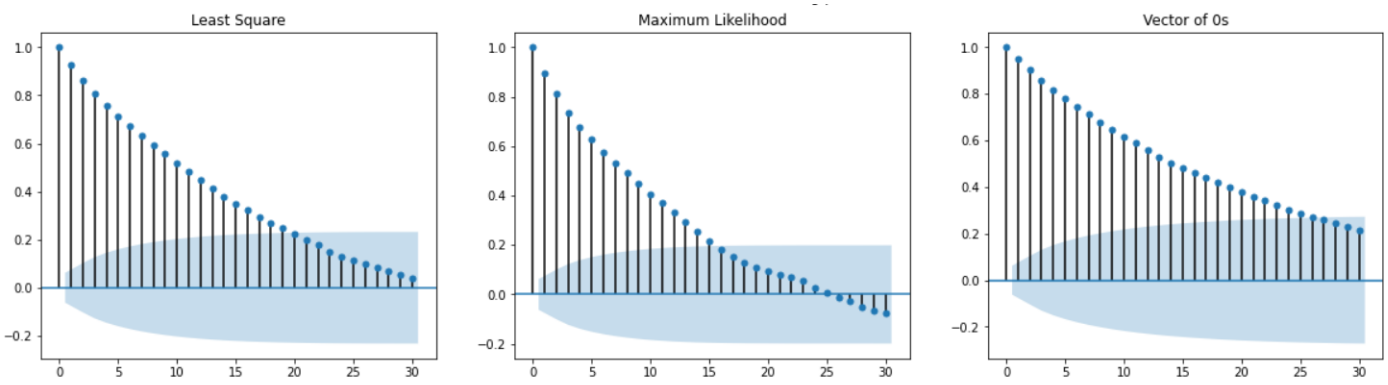

Figure 5: ACF Plot for the Metropolis Algorithm with Uninformative Prior on Real Data showing pre-thinning for *Glucose*

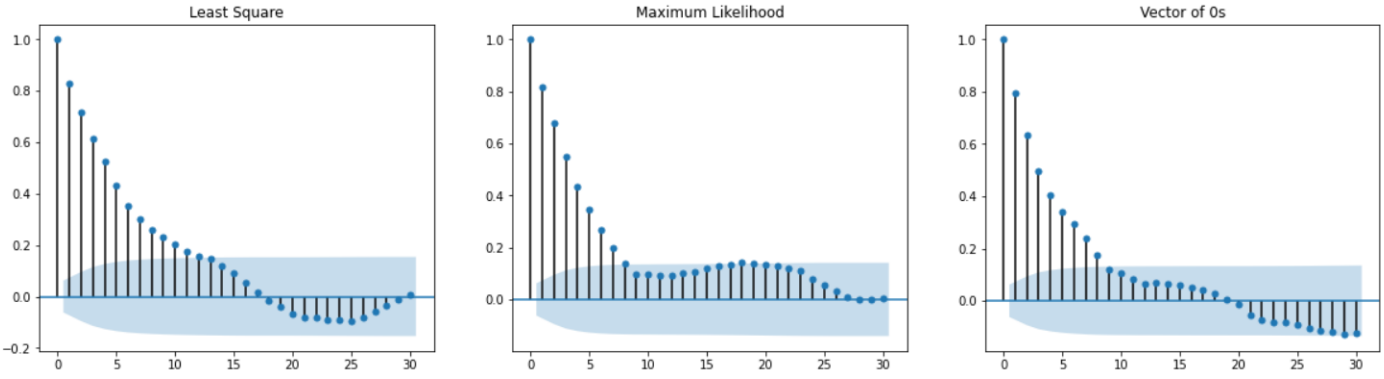

Figure 6: ACF Plot for the Metropolis Algorithm with Uninformative Prior on Real Data showing post-thinning for *Glucose*