Markov Chain Monte Carlo for Classification

1. Introduction

We are now given a 150×10 design matrix X where each row of X corresponds to one single training data and we have 150 training data in total. Besides, we're also given the actual 10×1 weight parameter β, whose true value is 10 equally spaced values from -2 to 2, for the associated logistic and cauchit classification model.

2. Sampling for Cauchit Model

To draw 150 $Y_i's$ from the cauchit model, we need to sample 150 $\epsilon_i \sim$ Cauchy(0, 1) which is equivalent to sampling from t(1) distribution. We can use the rt(150, 1) in R to do so. After the sampling, I first calculate $Y^* = X\beta + \epsilon$, where ϵ is the vector that contains our 150 sampled ϵ_i. Then I simply use the latent variable interpretation $Y = \mathbb{I}(Y^* > 0)$ for drawing 150 $Y_i's$ from the cauchit model.

3. Rejection Sampling for Logistic Model

Now I use rejection sampling for drawing 150 $\epsilon_i \sim$ Logistic(0, 1) where Cauchy(0, 1) is used as the candidate distribution. In rejection sampling, we first draw a sample x_i from the candidate distribution $p(.)$ ($p(.)$ is Cauchy(0, 1) in this section), then draw $u_i \sim \mathcal{U}[0, 1]$. Accept x_i if $u_i \leqslant \pi(x_i)/Mp(x_i)$, where $\pi(.)$ is the target distribution we hope to sampling from ($\pi(.)$ is Logistic(0, 1) in this section). Theoretically speaking, M satisfies $1 \leqslant M$ and also $(\pi(x)/p(x)) \leqslant M$. Also, the probability of each x_i being accepted is $1/M$, so M needs to be as small as possible for efficiency. The ideal situation is that $M = \sup(\pi(x)/p(x))$. However, in the Logistic/Cauchy case, it is not possible to find a closed form expression of $\sup(\pi(x)/p(x))$ by first order derivative. Given that Cauchy distribution has heavy tails than logistic distributions, the value of $\pi(x)/p(x)$ will gradually decrease to 0 as x goes to positive and negative infinity. I visualise the Cauchy and Logistic density together with the ratio of their density between (-6, 6):

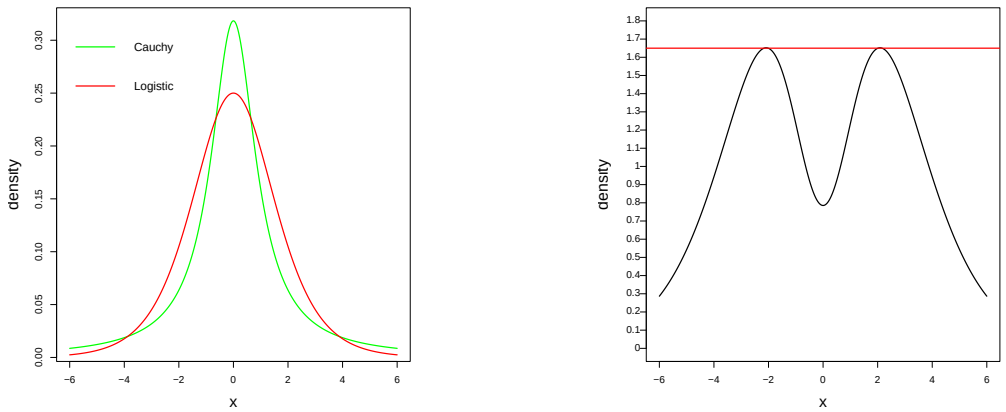

Figure 1: Left: Cauchy VS Logistic. Right: Ratio of Logistic/Cauchy Density.

From Figure 1 the ratio $\pi(x)/p(x)$ goes towards 0 when x goes to Inf or -Inf. Note that $\pi(x)/p(x)$ is symmetric around 0. Spotted from Figure 1, the maximum value of $\pi(x)/p(x)$ is between 1.6 and 1.7, around 1.65. Without losing too much efficiency, I choose $M = 1.7$ which is only slightly above $\sup(\pi(x)/p(x))$. Using $M = 1.7$, the acceptance rate of the sampling is about 58.82%, which agrees with the theoretical acceptance rate $1/1.7 \approx 58.82\%$.

4. Metropolis-Hastings Sampling for Posterior β

To sample from the posterior distribution of β for the two models under both the $\mathcal{N}(\mathbf{0}, \mathbf{I}_{10})$ prior where $\mathbf{I}_{10}$ is the 10×10 identity matrix and also the unit information prior $\mathcal{N}(\mathbf{0}, 150(X^\top X)^{-1})$. So there will be 4 sets of output in total.

We will use the Random Walk Metropolis-Hastings (RWMH) algorithm to do the sampling. Assume we want to sample n high dimensional samples $\boldsymbol{x}_i's$ from the target distribution $\pi(.)$. There are 4 main steps generally for RWMH sampling:

1. For $i \in \{0, 1, \ldots, n-1\}$, draw $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, V)$, where V is a 10×10 covariance matrix.
2. Set $\boldsymbol{y} = \boldsymbol{x}_i + h\boldsymbol{\varepsilon}_i$
3. Set $u \sim \mathcal{U}[0, 1]$
4. Set $\boldsymbol{x}_{i+1} = \boldsymbol{y}$, if $\log(u) \leqslant \log(\pi(\boldsymbol{y})) - \log(\pi(\boldsymbol{x}_i))$ and set $\boldsymbol{x}_{i+1} = \boldsymbol{x}_i$, if otherwise

I will use pre-conditioning in our Random Walk Metropolis-Hastings algorithm to improve the performance. Before actually do the MCMC sampling, there is actually an issue in calculating the posterior likelihood value of logistic regression model. The problem comes when we are calculating the log likelihood of the logistic regression model. The log likelihood of the logistic model given that $y_i = 0$, $\log(p(y_i = 0|\boldsymbol{x}_i, \beta))$, is $\log\left(1 - 1/(1 + \exp(-\boldsymbol{x}_i^\top \beta))\right)$. Imagine when $\boldsymbol{x}_i^\top \beta$ goes really large, then $\exp(-\boldsymbol{x}_i^\top \beta)$ will tend to 0 and hence $1/(1 + \exp(-\boldsymbol{x}_i^\top \beta))$ will be calculated as 1 in R, which leads to $\log(p(y_i = 0|\boldsymbol{x}_i, \beta)) = \log(0) = -\text{Inf}$. This will cause a problem in our RWMH sampling. The solution to this is to use the logSumExp() function in the matrixStats package in R. logSumExp(c$(x_1, x_2, \ldots, x_n)$)=log$(\exp(x_1) + \exp(x_2) + \ldots + \exp(x_n))$, which is an excellent way to avoid numerical underflow in R. For example, $\log(1 + \exp(1888)) = \text{Inf}$ while logSumExp(c(0,1888))=1888 in R. Therefore, we can rewrite:

$$\log\left(p(y_i = 1|\boldsymbol{x}_i, \beta)\right) = \log\left(\frac{1}{1 + \exp(-\boldsymbol{x}_i^\top \beta)}\right) = -\log\left(1 + \exp(-\boldsymbol{x}_i^\top \beta)\right) = -\text{logSumExp}(c(0, -\boldsymbol{x}_i^\top \beta))$$

$$\log\left(p(y_i = 0|\boldsymbol{x}_i, \beta)\right) = \log\left(\frac{1}{1 + \exp(\boldsymbol{x}_i^\top \beta)}\right) = -\log\left(1 + \exp(\boldsymbol{x}_i^\top \beta)\right) = -\text{logSumExp}(c(0, \boldsymbol{x}_i^\top \beta))$$

The RWMH sampling strategy for all the four settings (logistic with i.i.d standard normal prior, logistic with unit information prior, cauchit with i.i.d standard normal prior and cauchit with unit information prior) are the same. In all rounds of RWMH

sampling, I drew 30000 samples and each time h was carefully tuned such that the acceptance rate of the RWMH sampling is between 22.5% and 25.5% which follows the Goldilocks principle. During the sampling procedure, I first use the vanilla RWMH method by setting $V = \mathbf{I}_{10}$ with the starting point being the actual β (10 evenly spaced real number between -2 and 2) and draw 30000 points. Using the first 30000 samples, I calculate an estimated covariance matrix $\hat{C}_1$ of the posterior distribution of β via cov() in R and set $V = \hat{C}_1$ to do the second pre-conditioned RWMH sampling with starting point being the actual β. Then use the samples from the second pre-conditioned RWMH to compute a second estimated covariance matrix $\hat{C}_2$ and set $V = \hat{C}_2$ to do the final round of pre-conditioned RWMH sampling with true β being the starting point. Meanwhile, during the final round of pre-conditioned RWMH sampling I also start three other RWMH sampling with very different starting point from the final pre-conditioned RWMH chain with true β as the starting point (with all other parameters unchanged). Then I use the four chains to calculate the Gelman-Rubin diagnostic $\hat{R}$ and also to generate Gelman-Rubin diagnostic plots for each component of β. The Gelman-Rubin diagnostic will help us determine whether the final round of pre-conditioned RWMH chain with true β as the starting point is mixing well and where is its burn-in position. After that, I discard every sampled point in the final round RWMH sampling (the one with true β as its starting point) before the burn-in position and calculate the sample mean $\hat{\beta}$ of the rest points and use $\hat{\beta}$ as the point estimator of the posterior distribution of β.

For all the four settings, the final round of the pre-conditioned RWMH is mixing well. This can be seen from the traceplot and the autocorrelation plot of the first component of the sampled β final pre-conditioned RWMH in each of the four settings, which I think is representative of how all the 10 components of the sampled β behave. I include these plots as Figure 4-7 in Appendix.

I also list a table of the Gelman-Rubin diagnostic $\hat{R}$ as below:

	cauchit-normal	cauchit-info	logistic-normal	logistic-info
$\hat{R}$ 1st component	1.00	1.00	1.00	1.00
$\hat{R}$ 2nd component	1.00	1.00	1.00	1.00
$\hat{R}$ 3rd component	1.00	1.00	1.00	1.00
$\hat{R}$ 4th component	1.00	1.00	1.00	1.01
$\hat{R}$ 5th component	1.01	1.00	1.00	1.00
$\hat{R}$ 6th component	1.00	1.00	1.00	1.00
$\hat{R}$ 7th component	1.00	1.00	1.00	1.00
$\hat{R}$ 8th component	1.00	1.00	1.00	1.00
$\hat{R}$ 9th component	1.00	1.00	1.00	1.00
$\hat{R}$ 10th component	1.00	1.00	1.00	1.00
multivariate $\hat{R}$	1.01	1.01	1.01	1.01

Table 1: Table of estimated $\hat{R}$ for all 10 component of all four final pre-conditioned RWMH chains

From Table 1, we can see that the estimated $\hat{R}$ for all 10 components of the four final

sets of pre-conditioned RWMH chains are less than 1.1, which again indicates excellent mixing of the four final pre-conditioned RWMH chains with true β being the starting point.

Furthermore, the Gelman-Rubin diagnostic plot is also good for checking convergence of the chain. For cauchit model with i.i.d standard normal prior, the component whose $\hat{R}$ converges below 1.1 the slowest is the 6th component while for logistic model with i.i.d standard normal prior, the 9th component seems to converge the slowest:

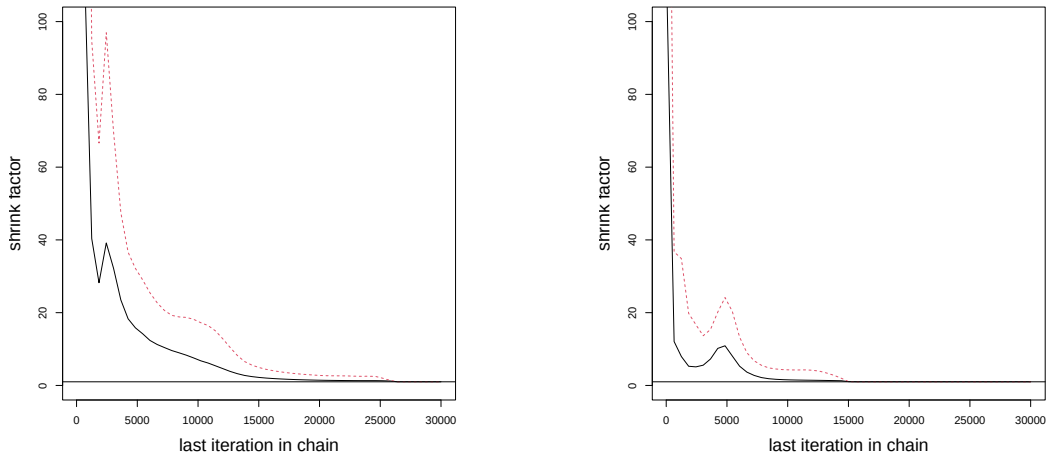

Figure 2: Left: Gelman-Rubin plot for the 6th component of the final pre-conditioned RWMH sampling for cauchit model with i.i.d standard normal prior. Right: Gelman-Rubin plot for the 9th component of the final pre-conditioned RWMH sampling for logistic model with i.i.d standard normal prior.

From Figure 2, we can see that $\hat{R}$ for the worst component converges below 1.1 after the 25000th iteration in cauchit model with i.i.d standard normal prior and for logistic model with i.i.d standard normal prior it goes below 1.1 after the 15000th iteration. So to ensuring mixing, I discard the first 25000 samples and 15000 samples (from the final pre-conditioned RWMH with true β as the starting point) respectively and use the rest to calculate the sample mean in these two situations.

Similarly, for cauchit and logistic model with unit information prior, the $\hat{R}$ of the 9th component for both of these 2 settings in their final pre-conditioned RWMH converges the slowest below 1.1 as shown in Figure 3 in the next page.

As from Figure 3, both of the $\hat{R}$ for the worst components converge below 1.1 after the 20000th iteration. Again, I discard the first 20000 samples (from the final pre-conditioned RWMH chain with true β as the starting point) in both cases and use the rest to calculate the sample mean in these two situations.

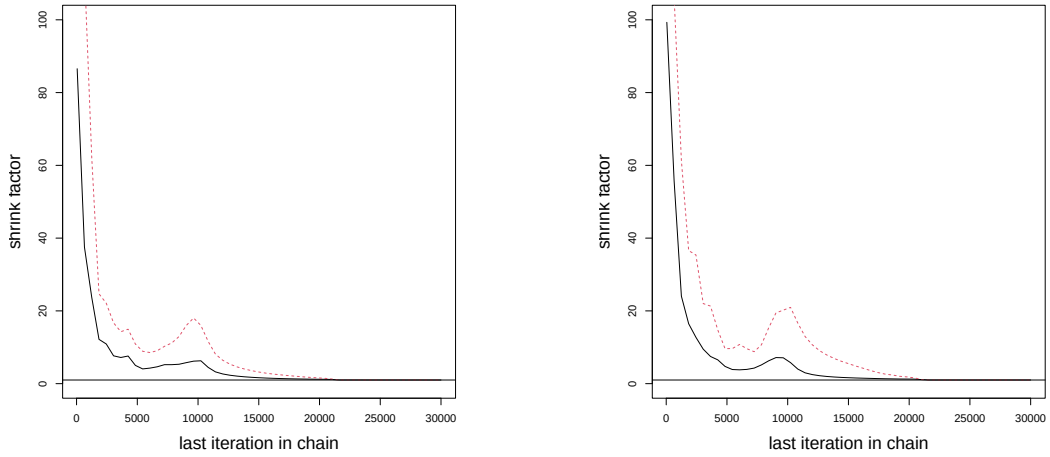

Figure 3: Left: Gelman-Rubin plot for the 9th component of the final pre-conditioned RWMH sampling for cauchit model with unit information prior. Right: Gelman-Rubin plot for the 9th component of the final pre-conditioned RWMH sampling for logistic model with unit information prior.

5. Prediction Assessment

To evaluate the performance of the model fitted with the four sets of estimated $\hat{\beta}$ by calculating sampled mean as in section 4, we can use the Brier Score using the formula of:

$$\mathcal{S} = \frac{1}{150}\sum_{i=1}^{150}(y_i - P(Y_i = 1))^2$$

The calculated Brier scores are summarised in Table 2 below:

	cauchit-normal	cauchit-info	logistic-normal	logistic-info
Brier Score	0.0707	0.0719	0.0279	0.0369

Table 2: Brier Score of the four sets of fitted classification model

We can see from Table 2 that all the four fitted models have pretty small prediction errors as indicated by relatively small Brier scores. The Brier scores of the logistic models are generally smaller than cauchit ones.

Moreover, the i.i.d standard normal prior performs better than unit information prior in both the cauchit and the logistic regression model. One possible reason for this is that the true β value is 10 evenly spaced values in (-2, 2), which is more likely to be drawn from an 10-dimensional i.i.d standard normal distribution compared to a unit information distribution. Indeed, the log density value of the 10-dimensional i.i.d standard normal distribution evaluated at true β is about -17.34 while the log density value of the unit information distribution at true β is about -137.22, which verifies my guess.

Appendix

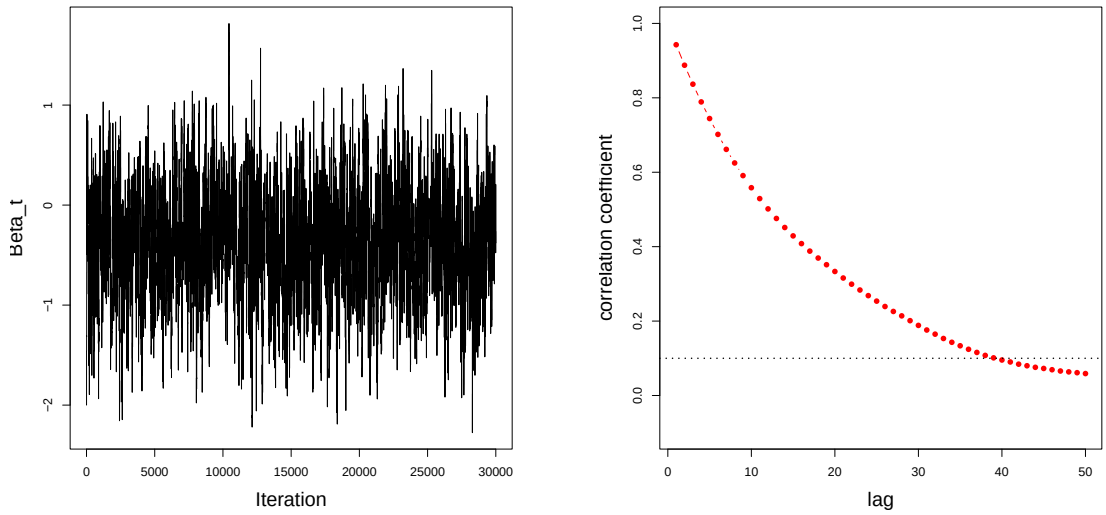

Figure 4: Traceplot and autocorrelation plot of the 1st component of the final preconditioned RWMH output for cauchit model with i.i.d standard normal prior

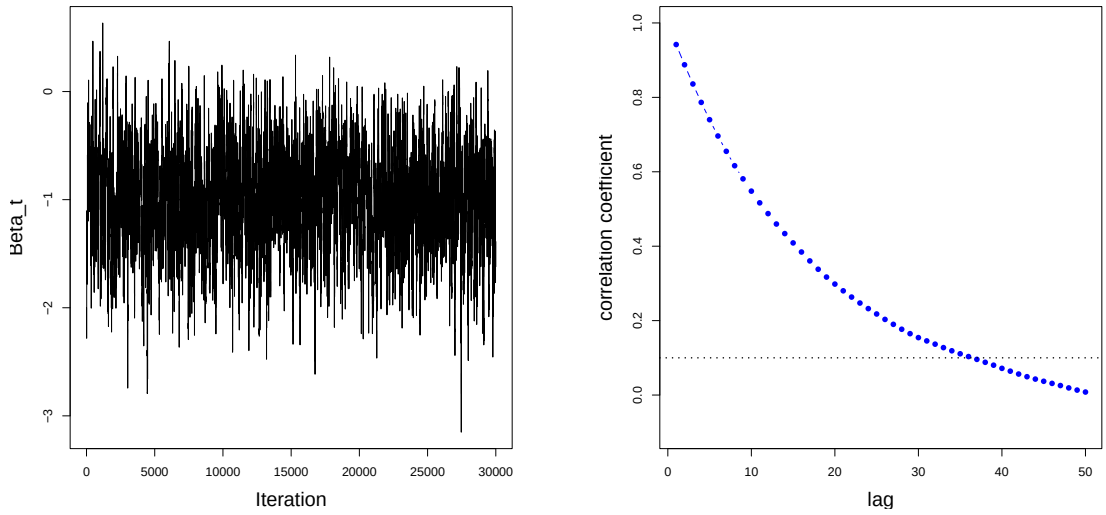

Figure 5: Traceplot and autocorrelation plot of the 1st component of the final preconditioned RWMH output for logistic model with i.i.d standard normal prior

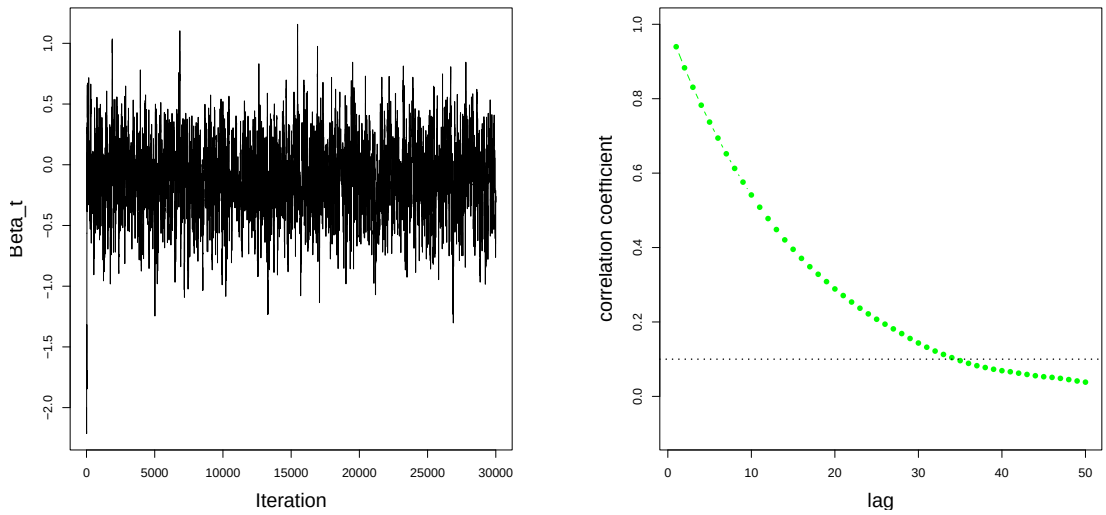

Figure 6: Traceplot and autocorrelation plot of the 1st component of the final preconditioned RWMH output for cauchit model with unit information prior

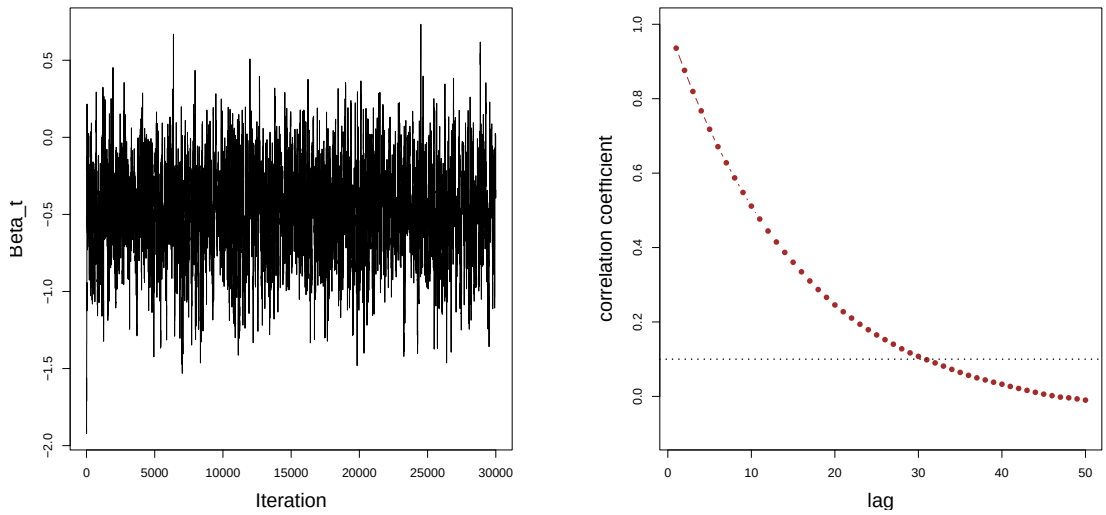

Figure 7: Traceplot and autocorrelation plot of the 1st component of the final preconditioned RWMH output for logistic model with unit information prior