Bayesian Logistic Regression with Polya-Gamma latent variables

Kaspar Märtens Sherman Ip

October 19, 2015

Abstract

Your abstract.

1 Introduction

Motivation for Bayesian approach etc

2 Data augmentation scheme

In Polson et al. (2013), the Polya-Gamma family of distributions is carefully constructed so that introducing latent variables from this family yields a simple Gibbs sampler for the Bayesian logistic regression model.

Let

$$y_i \sim \text{Bernoulli}\left(\frac{1}{1+\exp(-x_i^T \boldsymbol{\beta})}\right)$$

for data points $i = 1, ..., N$, with x_i the vector of covariates, and $\boldsymbol{\beta}$ the parameter vector with a prior distribution $\boldsymbol{\beta} \sim N(b, B)$. Sampling from the posterior distribution of $\boldsymbol{\beta}$ can be achieved by introducing the auxiliary random variables $\omega_i, i = 1, ..., N$, and iterating the following two-step Gibbs sampling scheme:

1. $(\omega_i|\boldsymbol{\beta}) \sim PG(1, x_i^T \boldsymbol{\beta})$
2. $(\boldsymbol{\beta}|y, \omega) \sim N(m_\omega, V_\omega)$

where the first conditional distribution is a Polya-Gamma $PG(1, z)$ with some real number z, and the second one is a multivariate normal with the mean and covariance specified in Polson et al. (2013). Note that there is a latent variable ω_i for each data point, i.e. the first step needs to be repeated N times, whereas the parameters β are sampled jointly.

One way for constructing a random variable X from a Polya-Gamma distribution $PG(1, z)$ with $z \in \mathbb{R}$ is according to the definition, i.e.

$$X \leftarrow \sum_{k=1}^{\infty} \frac{g_k}{(k-0.5)^2 + \frac{z^2}{4\pi^2}}$$

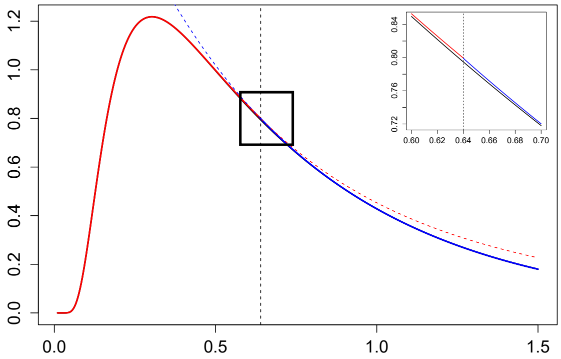


Figure 1: Visualisation of the accept-reject algorithm for the target PG(1, 1) distribution (density in black). The proposal distribution is defined in two pieces: for $x \in (0, 0.64]$ (density in red) and $x \in (0.64, \infty)$ (blue). The middle portion of the figure has been zoomed in (top right corner). The dashed lines (red and blue) extend the densities of proposal distributions outside their defined range.

where $g_k \sim \Gamma(1,1)$ are i.i.d. random variables. The definition contains an infinite sum and it is not clear how its truncation to a finite number of terms will affect the results.

Instead, an accept-reject sampling procedure is proposed to sample from $PG(1, z)$.

3 Implementation

3.1 Gibbs sampling

3.2 Sampling from the Polya-Gamma distribution

4 Experiments and results

4.1 Tests on simulated data

4.1.1 ???

```
devtools::load_all()

## Loading PolyaGamma

data = generate_from_simple_logistic_model(1000)
obj = gibbs_sampler(data$y, data$X, lambda = 0.01, n_iter = 100)
plot(obj)
```

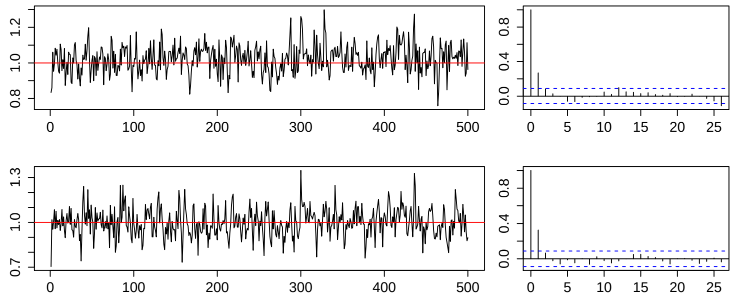
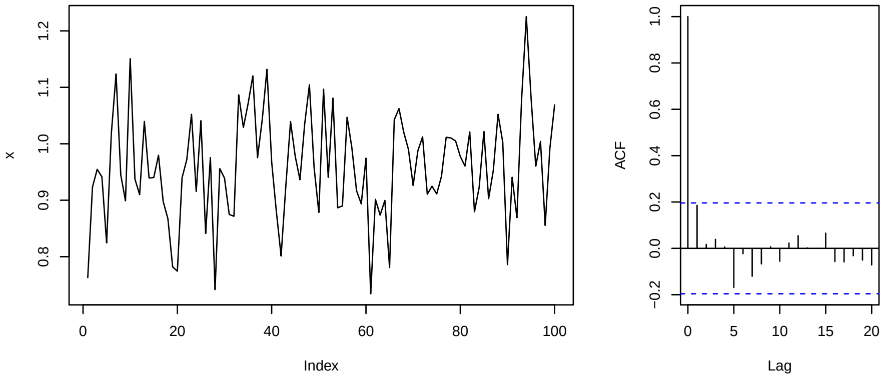
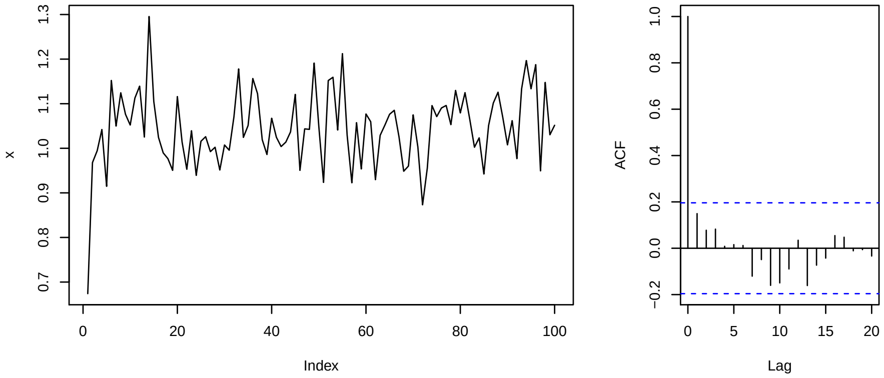


Figure 2: Traceplot of β_1 on simulated data





say something about the posterior distribution of beta
effective sample size??

4.1.2 Efficient sampling from Polya-Gamma distribution

Compare the naive approach versus accept-reject algorithm (autocorrelations, computation time)

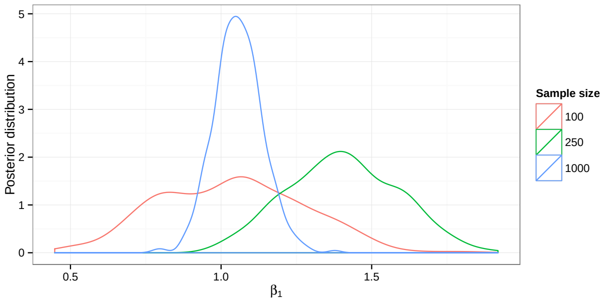


Figure 3: Posterior distribution (smoothed histograms) of β_1 for different sample sizes.

4.1.3 Comparison with BayesLogit package

check that the results are the same, compare computation time

4.2 Tests on real data

5 Future work?

References

Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013.