Efficient simulation of Bayesian logistic regression models

by Chris C Holmes

Imperial College London

Leonhard Knorr-Held

Ludwig-Maximilians-University Munich

Summary

In this paper we highlight a data augmentation approach to inference in the Bayesian logistic regression model. We demonstrate that the resulting conditional likelihood of the regression coefficients is multivariate normal, equivalent to a standard Bayesian linear regression, which allows for efficient simulation using a block Gibbs sampler. We illustrate that the method is particularly suited to problems in covariate set uncertainty and random effects models.

Some Key Words: Auxiliary variables, Bayesian logistic regression, Data augmentation, Markov chain Monte Carlo, Model averaging, Random effects, Scale mixture of normals, Variable selection.

1. Introduction

Binary regression using Generalised Linear Models (GLMs) is a widely used technique in applied statistics and the Bayesian approach to this subject is well documented, e.g. Dey, Gosh and Mallick (1999). However, inference in Bayesian GLMs is complicated by the fact that no conjugate prior exists for the parameters in the model, other than for normal regression, and this makes simulation difficult. In a seminal paper, with over 200 citations at time of writing, Albert & Chib (1993) demonstrated a data augmentation approach for binary probit regression models which renders the conditional distributions of the model parameters equivalent to those under the Bayesian linear regression model with Gaussian noise. Hence, conjugate priors are available to the conditional likelihood and the block Gibbs sampler can then be used to great effect. In this paper we demonstrate that this is also possible for logistic regression, by using a scale mixture of normals representation and additional auxiliary variables. This is an important discovery as typically the logit link is the method of choice for most statistical applications, due to the strong interpretation of the regression coefficients in terms of the change to the log-odds of one class over another for unit change in the associated covariate. In addition, the logit link avoids the need for a table look up, as in the cumulative normal (probit link) which is known to be sensitive to evaluation in the tails of the link function.

In §2 we present the method and algorithms for sampling from a Bayesian logistic regression model. The approach is also well suited to generalisations of the standard logistic model and in §3, §4 we describe two such applications, namely, in covariate set uncertainty and random effect models. Finally, in §5 we offer a brief discussion, contrasting the approach to existing methods and pointing to possible extensions.

2. A data augmentation approach to the logistic regression model

To begin, consider the Bayesian logistic regression model,

$$\begin{aligned} y_i &\sim \text{Bernoulli } (g^{-1}(\eta_i)) \\ \eta_i &= \boldsymbol{x}_i \boldsymbol{\beta} \\ \boldsymbol{\beta} &\sim \pi(\boldsymbol{\beta}) \end{aligned} \tag{1}$$

where $y_i \in \{0, 1\}, i = 1, \ldots, n$ is a binary response variable for a collection of n objects with associated p covariate measurements $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$, $g(u) = \log(u/(1-u))$ is the logistic link function, η_i is the linear predictor and $\boldsymbol{\beta}$ represents a $(p \times 1)$ column vector of regression coefficients which *a priori* are from some distribution $\pi(\cdot)$.

The logistic model in (1) has an equivalent representation using auxiliary variables,

$$\begin{aligned} y_i &= \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{otherwise} \end{cases} \\ z_i &= \boldsymbol{x}_i \boldsymbol{\beta} + \epsilon_i \\ \epsilon_i &\sim \pi(\epsilon_i) \\ \boldsymbol{\beta} &\sim \pi(\boldsymbol{\beta}) \end{aligned} \tag{2}$$

where y_i is now deterministic conditional on the sign of the stochastic auxiliary variable z_i and $\pi(\epsilon_i)$ is the standard logistic distribution. Under independence of ϵ_i, $i = 1, \ldots, n$, the marginal distribution of y in model (2), having integrated out z and ϵ, is the same as in (1). In what follows, we shall introduce a further set of variables, λ_i, $i = 1, \ldots, n$, and note the additional representation

$$\begin{aligned}
y_i &= \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{otherwise} \end{cases} \\
z_i &= \boldsymbol{x}_i \boldsymbol{\beta} + \epsilon_i \\
\epsilon_i &\sim N(0, \lambda_i) \\
\lambda_i &= (2\psi_i)^2 \\
\psi_i &\sim KS \\
\boldsymbol{\beta} &\sim \pi(\boldsymbol{\beta})
\end{aligned} \tag{3}$$

where $N(0, \lambda_i)$ is a mean zero normal distribution with variance λ_i and ψ_i, $i = 1, \ldots, n$, are independent random variables following the Kolmogorov-Smirnov (KS) distribution, e.g. Devroye (1986). In this case, ϵ_i has a scale mixture of normal form with a marginal logistic distribution (Andrews & Mallows, 1974), so that the marginal distributions $\pi(\boldsymbol{\beta}|y)$ for models (3), (2) and (1) are equivalent.

The advantage of working with representation (3) is that, for judicious choice of $\pi(\boldsymbol{\beta})$, it lends itself to efficient simulation using the block Gibbs sampler. In particular, in the case of a normal prior on $\boldsymbol{\beta}$, $\pi(\boldsymbol{\beta}) = N(m, v)$, the full conditional distribution of $\boldsymbol{\beta}$ is still normal,

$$\begin{aligned}
\boldsymbol{\beta}|\boldsymbol{z}, \boldsymbol{\lambda}, \boldsymbol{y} &\sim N(\hat{\boldsymbol{\beta}}, V) \\
\hat{\boldsymbol{\beta}} &= V(v^{-1}m + \boldsymbol{x}'Wz) \\
V &= (v^{-1} + \boldsymbol{x}'W\boldsymbol{x})^{-1}, \\
W &= diag(\lambda_1^{-1}, \ldots, \lambda_n^{-1}),
\end{aligned} \tag{4}$$

here $\boldsymbol{x} = (\boldsymbol{x}_1', \boldsymbol{x}_2', \ldots, \boldsymbol{x}_n')'$, while the full conditional for z_i is truncated normal,

$$z_i|\boldsymbol{\beta}, \boldsymbol{x}_i, y_i, \lambda_i \;\propto\; \begin{cases} N(\boldsymbol{x}_i\boldsymbol{\beta}, \lambda_i)\; I(z_i > 0) & \text{if } y_i = 1 \\ N(\boldsymbol{x}_i\boldsymbol{\beta}, \lambda_i)\; I(z_i \le 0) & \text{otherwise,} \end{cases} \tag{5}$$

which is simple to sample from, see for example Robert (1995).

The conditional distribution of the variance parameter λ_i does not have a standard form though updating is conveniently achieved through a Metropolis-Hastings proposal from the prior distribution $\pi(\lambda_i)$. This involves sampling from the KS distribution for which efficient and exact algorithms exist (Devroye, 1986). The Metropolis-Hastings acceptance ratio will then be

$$\alpha = \left(\frac{\lambda_i}{\lambda_i^*}\right)^{\frac{1}{2}} \exp\left((z_i - \eta_i)^2 \left(\frac{1}{2\lambda_i} - \frac{1}{2\lambda_i^*}\right)\right), \tag{6}$$

where λ_i is the current and λ_i^* the proposed new value. Note that sampling from the prior avoids the evaluation of the KS density using a table look up, as it is known only as an infinite series.

Alternatively we can easily construct a joint proposal for λ_i and z_i by first sampling λ_i^* from the prior, as above, then generating z_i^* from the truncated normal (5), conditional on λ_i^*, and then accept-reject λ_i^* and z_i^* jointly. The acceptance ratio is in this case

$$\alpha = \begin{cases} \frac{1-\Phi(\eta_i/\sqrt{\lambda_i^*})}{1-\Phi(\eta_i/\sqrt{\lambda_i})} & \text{if } y_i = 1 \\ \frac{\Phi(\eta_i/\sqrt{\lambda_i^*})}{\Phi(\eta_i/\sqrt{\lambda_i})} & \text{if } y_i = 0, \end{cases} \tag{7}$$

here Φ denotes the cumulative distribution function of the standard normal distribution.

In this way, the formulas (4), (5), (6) or (7), provide the basis of efficient sampling from the Bayesian logistic regression model. The approach is easy to incorporate into statistical software and we believe it to be highly efficient compared to current sampling techniques. Furthermore, in the next two sections we highlight two generalisations of the standard logistic regression model where the auxiliary variable representation (3) is especially useful, namely in situations of covariate set uncertainty and for random effects models.

3. Covariate set uncertainty

It is often the case that the statistical analyst may suspect that some of the available covariates are irrelevant to the regression task. A convenient approach to this problem is to adopt a prior distribution on the covariate matrix $\pi(\boldsymbol{x})$ that places mass on the 2^p possible sub-models made up of differing covariates or columns of $\boldsymbol{x}$. In particular, consider the covariate indicator vector $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_p\}$, $\gamma_i \in \{0, 1\}$, $i = 1, \ldots, p$, such that $\gamma_i = 1$ if the ith covariate is present in the model and $\gamma_i = 0$ if it is not. A prior on the model space can be specified via a prior on the covariate indicator, $\pi(\boldsymbol{\gamma})$. The parameter vector $\boldsymbol{\gamma}$ can then be included in the model specification and updated as part of the simulation.

Bayesian analysis of models of random dimension have become extremely popular following the introduction of sampling techniques such as Green (1995). However, simulation of variable dimensional models can be problematic as a change to the model structure typically causes a large change to the likelihood of the current parameter values in the model, see Brooks *et al.* (2002). A key advantage of using model (3) is that when updating the covariate set defined by $\boldsymbol{\gamma}$ we can condition on z and jointly update the $\boldsymbol{\beta}$'s as well, from their full conditional distribution given the new model structure. The vector z retains information about the likelihood which allows for optimal updates to be made to $\boldsymbol{\beta}$, given a change in the covariate set. Updating the $\boldsymbol{\beta}$ vector jointly with $\boldsymbol{\gamma}$ is extremely important as typically, when the covariates are non-orthogonal, there is strong linear dependence between the regression coefficients.

To sample from the posterior model space, we suggest a Metropolis-Hastings step to update the current covariate set, defined by $\boldsymbol{\gamma}$, with a joint update to $\boldsymbol{\beta}$ as well,

$$q(\boldsymbol{\gamma}^*, \boldsymbol{\beta}^*) = \pi(\boldsymbol{\beta}^* | \boldsymbol{\gamma}^*, z, W) \; q(\boldsymbol{\gamma}^*),$$

where $q()$ denotes a proposal distribution, $\pi(\boldsymbol{\beta}^* | \boldsymbol{\gamma}^*, z, W)$ is the conditional multivariate normal posterior distribution (4) given the covariate set defined by $\boldsymbol{\gamma}^*$, and $q(\boldsymbol{\gamma}^*)$ is a, possibly symmetric, Metropolis-Hastings proposal density that may, or may not, be based on the current covariate set $\boldsymbol{\gamma}$. In this case, some straightforward algebra leads to the acceptance probability of the move as,

$$\alpha = \min \left\{ 1, \frac{|V_{\gamma^*}|^{1/2} |v_{\gamma}|^{1/2}}{|V_{\gamma}|^{1/2} |v_{\gamma^*}|^{1/2}} \frac{\exp(-0.5 \hat{\boldsymbol{\beta}}'_{\gamma^*} V_{\gamma^*}^{-1} \hat{\boldsymbol{\beta}}_{\gamma^*})}{\exp(-0.5 \hat{\boldsymbol{\beta}}'_{\gamma} V_{\gamma}^{-1} \hat{\boldsymbol{\beta}}_{\gamma})} \frac{\pi(\boldsymbol{\gamma}^*) q(\boldsymbol{\gamma} | \boldsymbol{\gamma}^*)}{\pi(\boldsymbol{\gamma}) q(\boldsymbol{\gamma}^* | \boldsymbol{\gamma})} \right\} \tag{8}$$

where α denotes the acceptance probability of the proposal and $\{\hat{\boldsymbol{\beta}}_{\gamma}, V_{\gamma}\}$ are defined in (4), where the subscripts indicate that they are conditioned on the covariate set defined by $\boldsymbol{\gamma}$. Note that the realised (drawn) values of $\{\boldsymbol{\beta}, \boldsymbol{\beta}^*\}$ do not appear in the acceptance probability (8), which resembles the Bayes factor of a standard Bayesian linear model. This implicit marginalisation of $\boldsymbol{\beta}$ in the proposal step leads to efficient dimension sampling, as the $\boldsymbol{\beta}$'s are being updated from their full conditional distributions given the change to the covariate set.

To illustrate the approach we consider a binary classification problem taken from Ripley (1996). The regression task is to predict whether patients will test positive or negative for diabetes using a set of seven covariate measurements, observed on a group of adult females of Pima Indian heritage. There are 532 records, selected from a larger data set, with the following predictor variables: number of pregnancies (NP); plasma glucose concentration (Gl); distolic blood pressure (BP); triceps skin fold thickness (TST); body mass index (BMI); diabetes pedigree function (DP); and, age (Ag). We obtained the data from the web site `www.stats.ox.ac.uk/~ripley/PRbook/`. In Ripley (1996) they used a classical (non-Bayesian) logistic regression model and noted that some of

Covariates	NP	Gl	BP	TST	BMI	DP	Ag
$E[\gamma_i]$	0.925	0.998	0.009	0.034	0.992	0.946	0.131
MCMC Std	0.087	0.001	0.009	0.013	0.001	0.034	0.111

Table 1: Row 1, lists the covariate acronyms for the Pima Indian data set example in Section 3: (NP), number of pregnancies; (Gl), plasma glucose concentration; (BP), distolic blood pressure; (TST), triceps skin fold thickness; (BMI), body mass index; (DP), diabetes pedigree function; and, (Ag), age. Row 2, lists the posterior probabilities of covariate selection. In row 3, we report the MCMC standard deviations of the estimates $\pi(\gamma_i = 1|y)$, taken across nine consecutive post burn-in regions of size 1,000 MCMC samples.

the covariates appeared irrelevant. Ripley (1996) went on to perform stepwise variable selection using an AIC model choice criteria and found that the covariates blood pressure and skin thickness were dropped from the final model. We performed a Bayesian analysis using independent priors on the covariates and regression coefficients as, $\pi(\boldsymbol{\gamma}) = \prod_i \pi(\gamma_i)$, with $\pi(\gamma_i = 1) = 0.5$ for $i = 1, \ldots, p$ and $\pi(\boldsymbol{\beta}) = N(0, 100 I_p)$. Updates to λ_i were made using (6). Updates to the covariate set were made using a Metropolis proposal as follows. We select a covariate at random and propose $\gamma_i^* = 1$, if the current $\gamma_i = 0$, $\gamma_i^* = 0$ otherwise. This results in the final term, $\frac{\pi(\boldsymbol{\gamma}^*)q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^*)}{\pi(\boldsymbol{\gamma})q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma})}$, in (8) being one.

We performed a simulation of 10,000 iterations and discarded the first 1,000 as a burn-in. In Table 1, we show the estimates of the posterior probabilities, $\pi(\gamma_i = 1|y)$, for the seven covariates, along with the standard deviations in these MCMC estimates taken from nine consecutive regions of the post burn-in MCMC samples, $\{(1001, 3000), \ldots, (9001, 10000)\}$. The chain appears to be mixing well under the data augmentation approach. The overall acceptance rate of the covariate update proposals was around 4% which is good when considering the posterior probabilities $\pi(\gamma_i|y)$ shown in Table 1. The estimates of $\pi(\gamma_i = 1|y)$ are in accordance with the observations of Ripley (1996) though we find there also appears to be some doubt as to the relevance of age.

4. Random Effects Models

The proposed auxiliary variable approach is also well tailored to hierarchical logistic regression models, where latent random effects follow a Gaussian distribution. The effects may be conditionally independent, as in multilevel models, or dependent, as for example in in dynamic models (for a recent review see Fahrmeir and Knorr-Held, 2000) or in hierarchical models with latent Gaussian Markov random fields. In all these cases, the full conditional distributions for the random effects will follow multivariate Gaussian distributions, which are straightforward to sample from. If the random effects prior have a (spatial or temporal) Markov structure, the algorithm proposed in Rue (2001) provides a fast and efficient way to simulate from the full conditional distribution in one block.

As an illustration for a random effects model, we consider data on salamander mating taken from McCullagh and Nelder (1989). The data set is termed "challenging" by Karim and Zeger (1992) because of the binary response variable, (0= failure, 1=success) of the mating experiment between a female and a male, and the study design is crossed rather than nested, with two sets of random effects. The data have been collected in three experiments, one conducted in the summer and two in the fall where each male and female salamander has been taken from two different populations: rough-butt and whiteside population. The total number of experiments was 360. The data is available at `www.stat.uchicago.edu/~pmcc/glm/glm.html`.

We follow model B by Karim and Zeger (1992) closely, and include the following fixed effects in the model: an indicator of season, two indicators of the male and female salamander population and an indicator for the interaction term. We then add for each male and female salamander an additional random effect in the model, say b_j^f and b_k^m for the j-th female and the k-th male, which we assume to be independent realizations from a normal distribution with mean zero and variance σ_f^2 and σ_k^2, respectively. The linear predictor for the i-th experiment is hence

$$\eta_i = \boldsymbol{x}_i \boldsymbol{\beta} + b_j^f + b_k^m.$$

For the variance parameters σ_f^2 and σ_k^2 we assume independent inverse gamma distributions with parameters $a = 1.0$ and $b = 0.1$ respectively. Assuming the logistic regression model with our auxiliary variable approach, all full conditionals for the random effects follow normal distributions.

We performed a simulation of 21,000 iterations and discarded the first 1,000 as a burn-in, using either the separate update of λ_i, or the joint update of λ_i and z_i. Mixing of all parameters was very good with autocorrelations dropping quickly to zero. The results have been in good agreement with those obtained by Karim & Zeger (1992), despite the slightly different model specification (Karim & Zeger use improper priors for variance components and random effects of dimension 2, rather than 1). We focus here only on the acceptance rates for the separate and joint updates. Interestingly, the acceptance rates in the joint updates (min = 0.72, median of 0.97, max = 0.99) are consistently higher compared the ones obtained from the separate updates (min = 0.71, median of 0.89, max = 0.90). This suggests that the joint updates might be slightly more efficient than the separate ones, at virtually no additional computational cost. The acceptance rates are even higher in the corresponding model without the random effects, which suggests that sampling from the prior for λ_i is an efficient procedure. However, this might be different in other applications, and we will mention alternatives in the discussion.

5. Discussion

We have presented an auxiliary variable representation for the Bayesian logistic regression model that lends itself to efficient simulation using standard MCMC methods. In particular we highlighted two non-standard models where we believe that gains in efficiency will be marked.

Popular current alternatives for MCMC simulation in Bayesian logistic regression models are found in Albert & Chib (1993) and Gamerman (1997). In Albert & Chib (1993) it was noted that specifying a scale mixture for λ_i in (3) as $\lambda_i \sim Gamma(4, 4)$ induces a t-distribution for ϵ_i with 8 degrees of freedom which gives a good approximation to the logistic distribution (up to a change in scale). However, this remains an approximation and a qq-plot of the true logistic distribution against that found using the Student approximation reveals considerable departure in the tails, see Figure 1. In applications it will be difficult to assess the effect of this bias on the posterior distribution of the regression coefficients. Our approach, however, is exact and provides a fast and efficient algorithm for inference in logistic regression models. One current area of investigation is to use the method of Albert & Chib (1993) to construct an independence Metropolis kernel for updates to λ_i, where the accept-reject step corrects for the approximation.

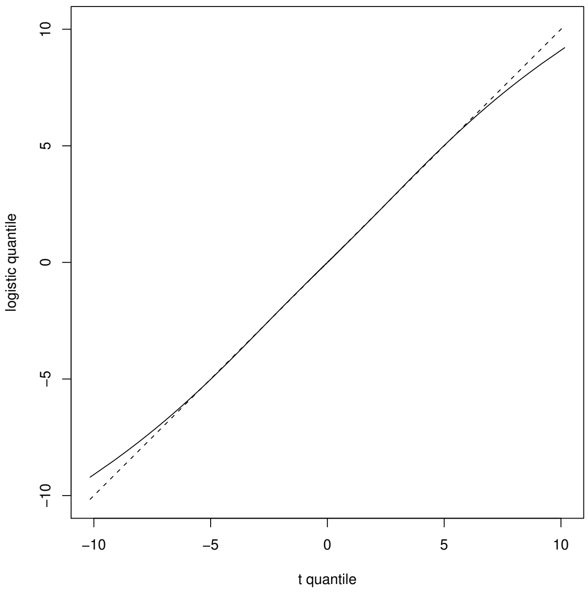

Figure 1: Plot of t-quantiles against logistic quantiles for probabilities between 0.0001 and 0.9999 (Solid line). The dashed line gives the reference line if the two distributions are identical.

An alternative algorithm without auxiliary variables is described in Gamerman (1997). Gamerman suggests a "weighted least squares" Metropolis-Hastings proposal based on a linear Taylor-approximation of the likelihood. This algorithm works well in practice, in particular if the number of parameters to be updated is not too large. However, acceptance rates will typically become too low in highly parameterized models, for example in dynamic logistic regression models. In contrast, the corresponding acceptance rates in our approach will always be unity due to the introduction of the auxiliary variables. Moreover, the extension of Gamerman's approach to variable dimension settings is non-trivial whereas we have shown in §3 this to be straightforward using auxiliary variables.

Finally we note that the approach proposed in this paper is straightforward to extend to the use of nonlinear regression splines (Denison *et al.*, 2002) and to logistic regression models for ordinal data, such as the cumulative (Albert & Chib, 1993) or the sequential model (Albert & Chib, 2001).

References

Andrews, D.F. & Mallows, C.L. (1974). Scale mixtures of normal distributions. *J. R. Statist. Soc.* B **36**, 99-102.

Albert, J. & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669-679.

Albert, J. & Chib, S. (2001). Sequential ordinal modeling with applications to survival data. *Biometrics*, **57**, 829-836.

Brooks, S.P., Giudici, P. & Roberts, G. O. (2003). Efficient construction of reversible jump MCMC proposal distributions (with discussion). To appear. *J. R. Statist. Soc.* B.

Denison, D.G.T., Holmes, C.C., Mallick, B.K. & Smith, A.F.M. (2002). *Bayesian methods for nonlinear classification and regression.* Chichester: Wiley.

Dey, D.P., Gosh, S. & Mallick, B. (1999). *Generalized Linear Models: A Bayesian Perspective.* New York: Marcel Dekker.

Devroye, L. (1986). *Non-Uniform Random Variate Generation.* New York: Springer.

Fahrmeir, L. and Knorr-Held, L. (2000). Dynamic and semiparametric models. *in*: M. Schimek (ed.), *Smoothing and Regression: Approaches, Computation and Applications*, Ch. 18, pp. 513-544, Wiley & Sons, New York.

Gamerman, D. (1997). Efficient sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing*, **7**, 57-68.

Green, P.J. (1995). Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711-732.

Karim, M.R. & Zeger, S.L. (1992). Generalized linear models with random effects; Salamander mating revisited. *Biometrics* **48**, 631-644.

Ripley, B.D. (1996). *Pattern Recognition and Neural Networks.* Cambridge University Press.

Robert, C.P. (1995). Simulation of truncated normal variables. *Statistics and Computing* **5**, 121-125.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *J. R. Statist. Soc.* B **63**, 325-338.