Bayesian Analysis (2018) **13**, Number 2, pp. 359–383

On the Use of Cauchy Prior Distributions for Bayesian Logistic Regression

Joyee Ghosh[*][§], Yingbo Li[†][§], and Robin Mitra[‡]

Abstract. In logistic regression, separation occurs when a linear combination of the predictors can perfectly classify part or all of the observations in the sample, and as a result, finite maximum likelihood estimates of the regression coefficients do not exist. Gelman et al. (2008) recommended independent Cauchy distributions as default priors for the regression coefficients in logistic regression, even in the case of separation, and reported posterior modes in their analyses. As the mean does not exist for the Cauchy prior, a natural question is whether the posterior means of the regression coefficients exist under separation. We prove theorems that provide necessary and sufficient conditions for the existence of posterior means under independent Cauchy priors for the logit link and a general family of link functions, including the probit link. We also study the existence of posterior means under multivariate Cauchy priors. For full Bayesian inference, we develop a Gibbs sampler based on Pólya-Gamma data augmentation to sample from the posterior distribution under independent Student-t priors including Cauchy priors, and provide a companion R package `tglm`, available at CRAN. We demonstrate empirically that even when the posterior means of the regression coefficients exist under separation, the magnitude of the posterior samples for Cauchy priors may be unusually large, and the corresponding Gibbs sampler shows extremely slow mixing. While alternative algorithms such as the No-U-Turn Sampler (NUTS) in Stan can greatly improve mixing, in order to resolve the issue of extremely heavy tailed posteriors for Cauchy priors under separation, one would need to consider lighter tailed priors such as normal priors or Student-t priors with degrees of freedom larger than one.

Keywords: binary regression, existence of posterior mean, Markov chain Monte Carlo, probit regression, separation, slow mixing.

1 Introduction

In Bayesian linear regression, the choice of prior distribution for the regression coefficients is a key component of the analysis. Noninformative priors are convenient when the analyst does not have much prior information, but these prior distributions are often improper which can lead to improper posterior distributions in certain situations. Fernández and Steel (2000) investigated the propriety of the posterior distribution and the existence of posterior moments of regression and scale parameters for a linear regression model, with errors distributed as scale mixtures of normals, under the independence

[*] The University of Iowa, Iowa City, IA, joyee-ghosh@uiowa.edu

[†] Clemson University, Clemson, SC, ybli@clemson.edu

[‡] University of Southampton, Southampton, UK, R.Mitra@soton.ac.uk

[§] These authors contributed equally.

https://doi.org/10.1214/17-BA1051

Jeffreys prior. For a design matrix of full column rank, they showed that posterior propriety holds under mild conditions on the sample size; however, the existence of posterior moments is affected by the design matrix and the mixing distribution. Further, there is not always a unique choice of noninformative prior (Yang and Berger, 1996). On the other hand, proper prior distributions for the regression coefficients guarantee the propriety of posterior distributions. Among them, normal priors are commonly used in normal linear regression models, as conjugacy permits efficient posterior computation. The normal priors are informative because the prior mean and covariance can be specified to reflect the analyst's prior information, and the posterior mean of the regression coefficients is the weighted average of the maximum likelihood estimator and the prior mean, with the weight on the latter decreasing as the prior variance increases.

A natural alternative to the normal prior is the Student-t prior distribution, which can be viewed as a scale mixture of normals. The Student-t prior has tails heavier than the normal prior, and hence is more appealing in the case where weakly informative priors are desirable. The Student-t prior is considered robust, because when it is used for location parameters, outliers have vanishing influence on posterior distributions (Dawid, 1973). The Cauchy distribution is a special case of the Student-t distribution with 1 degree of freedom. It has been recommended as a prior for normal mean parameters in a point null hypothesis testing (Jeffreys, 1961), because if the observations are overwhelmingly far from zero (the value of the mean specified under the point null hypothesis), the Bayes factor favoring the alternative hypothesis tends to infinity. Multivariate Cauchy priors have also been proposed for regression coefficients (Zellner and Siow, 1980).

While the choice of prior distributions has been extensively studied for normal linear regression, there has been comparatively less work for generalized linear models. Propriety of the posterior distribution and the existence of posterior moments for binary response models under different noninformative prior choices have been considered (Ibrahim and Laud, 1991; Chen and Shao, 2001).

Regression models for binary response variables may suffer from a particular problem known as separation, which is the focus of this paper. For example, complete separation occurs if there exists a linear function of the covariates for which positive values of the function correspond to those units with response values of 1, while negative values of the function correspond to units with response values of 0. Formal definitions of separation (Albert and Anderson, 1984), including complete separation and its closely related counterpart quasicomplete separation, are reviewed in Section 2. Separation is not a rare problem in practice, and has the potential to become increasingly common in the era of big data, with analysis often being made on data with a modest sample size but a large number of covariates. When separation is present in the data, Albert and Anderson (1984) showed that the maximum likelihood estimates (MLEs) of the regression coefficients do not exist (i.e., are infinite). Removing certain covariates from the regression model may appear to be an easy remedy for the problem of separation, but this ad-hoc strategy has been shown to often result in the removal of covariates with strong relationships with the response (Zorn, 2005).

In the frequentist literature, various solutions based on penalized or modified likelihoods have been proposed to obtain finite parameter estimates (Firth, 1993; Heinze and

Schemper, 2002; Heinze, 2006; Rousseeuw and Christmann, 2003). The problem has also been noted when fitting Bayesian logistic regression models (Clogg et al., 1991), where posterior inferences would be similarly affected by the problem of separation if using improper priors, with the possibility of improper posterior distributions (Speckman et al., 2009).

Gelman et al. (2008) recommended using independent Cauchy prior distributions as a default weakly informative choice for the regression coefficients in a logistic regression model, because these heavy tailed priors avoid over-shrinking large coefficients, but provide shrinkage (unlike improper uniform priors) that enables inferences even in the presence of complete separation. Gelman et al. (2008) developed an approximate EM algorithm to obtain the posterior mode of regression coefficients with Cauchy priors. While inferences based on the posterior mode are convenient, often other summaries of the posterior distribution are also of interest. For example, posterior means under Cauchy priors estimated via Monte Carlo and other approximations have been reported in Bardenet et al. (2014); Chopin and Ridgway (2015). It is well-known that the mean does not exist for the Cauchy distribution, so clearly the prior means of the regression coefficients do not exist. In the presence of separation, where the maximum likelihood estimates are not finite, it is not clear whether the posterior means will exist. To the best of our knowledge, there has been no investigation considering the existence of the posterior mean under Cauchy priors and our research is filling this gap. We find a necessary and sufficient condition where the use of independent Cauchy priors will result in finite posterior means here. In doing so we provide further theoretical underpinning of the approach recommended by Gelman et al. (2008), and additionally provide further insights on their suggestion of centering the covariates before fitting the regression model, which can have an impact on the existence of posterior means.

When the conditions for existence of the posterior mean are satisfied, we also empirically compare different prior choices (including the Cauchy prior) through various simulated and real data examples. In general, posterior computation for logistic regression is known to be more challenging than probit regression. Several MCMC algorithms for logistic regression have been proposed (O'Brien and Dunson, 2004; Holmes and Held, 2006; Gramacy and Polson, 2012), while the most recent Pólya-Gamma data augmentation scheme of Polson et al. (2013) emerged superior to the other methods. Thus we extend this Pólya-Gamma Gibbs sampler for normal priors to accommodate independent Student-t priors and provide an R package to implement the corresponding Gibbs sampler.

The remainder of this article is organized as follows. In Section 2 we derive the theoretical results: a necessary and sufficient condition for the existence of posterior means for coefficients under independent Cauchy priors in a logistic regression model in the presence of separation, and extend our investigation to binary regression models with other link functions such as the probit link, and multivariate Cauchy priors. In Section 3 we develop a Gibbs sampler for the logistic regression model under independent Student-t prior distributions (of which the Cauchy distribution is a special case) and briefly describe the NUTS algorithm of Hoffman and Gelman (2014) which forms the basis of the software Stan. In Section 4 we illustrate via simulated data that Cauchy priors may lead to coefficients of extremely large magnitude under separation, accompanied by slow mixing Gibbs samplers, compared to lighter tailed priors such as Student-t priors with degrees of freedom 7 (t_7) or normal priors. In Section 5 we compare Cauchy, t_7, and normal priors based on two real datasets, the Single Proton Emission Computed Tomography (SPECT) data with quasicomplete separation and the Pima Indian Diabetes data without separation. Overall, Cauchy priors exhibit slow mixing under the Gibbs sampler compared to the other two priors. Although mixing can be improved by the NUTS algorithm in Stan, normal priors seem to be the most preferable in terms of producing more reasonable scales for posterior samples of the regression coefficients accompanied by competitive predictive performance, under separation. In Section 6 we conclude with a discussion and our recommendations.

2 Existence of Posterior Means Under Cauchy Priors

In this section, we begin with a review of the concepts of complete and quasicomplete separation proposed by Albert and Anderson (1984). Then based on a new concept of solitary separators, we introduce the main theoretical result of this paper, a necessary and sufficient condition for the existence of posterior means of regression coefficients under independent Cauchy priors in the case of separation. Finally, we extend our investigation to binary regression models with other link functions, and Cauchy priors with different scale parameter structures.

Let $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ denote a vector of independent Bernoulli response variables with success probabilities $\pi_1, \pi_2, \ldots, \pi_n$. For each of the observations, $i = 1, 2, \ldots, n$, let $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ denote a vector of p covariates, whose first component is assumed to accommodate the intercept, i.e., $x_{i,1} = 1$. Let $\mathbf{X}$ denote the $n \times p$ design matrix with $\mathbf{x}_i^T$ as its ith row. We assume that the column rank of $\mathbf{X}$ is greater than 1. In this paper, we mainly focus on the logistic regression model, which is expressed as:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^T\boldsymbol{\beta}, \quad i = 1, 2, \ldots, n, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is the vector of regression coefficients.

2.1 A Brief Review of Separation

We denote two disjoint subsets of sample points based on their response values: $A_0 = \{i : y_i = 0\}$ and $A_1 = \{i : y_i = 1\}$. According to the definition of Albert and Anderson (1984), complete separation occurs in the sample if there exists a vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_p)^T$, such that for all $i = 1, 2, \ldots, n$,

$$\mathbf{x}_i^T\boldsymbol{\alpha} > 0 \text{ if } i \in A_1, \quad \mathbf{x}_i^T\boldsymbol{\alpha} < 0 \text{ if } i \in A_0. \tag{2}$$

Consider a simple example in which we wish to predict whether subjects in a study have a certain kind of infection based on model (1). Let $y_i = 1$ if the ith subject is infected and 0 otherwise. The model includes an intercept ($x_{i1} = 1$) and the other covariates are age (x_{i2}), gender (x_{i3}), and previous records of being infected (x_{i4}). Suppose in the sample, all infected subjects are older than 25 ($x_{i2} > 25$), and all subjects who are not infected are younger than 25 ($x_{i2} < 25$). This is an example of complete separation because (2) is satisfied for $\boldsymbol{\alpha} = (-25, 1, 0, 0)^T$.

If the sample points cannot be completely separated, Albert and Anderson (1984) introduced another notion of separation called quasicomplete separation. There is quasicomplete separation in the sample if there exists a non-null vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_p)^T$, such that for all $i = 1, 2, \dots, n$,

$$\mathbf{x}_i^T \boldsymbol{\alpha} \geq 0 \ \text{ if } i \in A_1, \quad \mathbf{x}_i^T \boldsymbol{\alpha} \leq 0 \ \text{ if } i \in A_0, \tag{3}$$

and equality holds for at least one i. Consider the set up of the previous example where the goal is to predict whether a person is infected or not. Suppose we have the same model but there is a slight modification in the dataset: all infected subjects are at least 25 years old ($x_{i2} \geq 25$), all uninfected subjects are no more than 25 years old ($x_{i2} \leq 25$), and there are two subjects aged exactly 25, of whom one is infected but not the other. This is an example of quasicomplete separation because (2) is satisfied for $\boldsymbol{\alpha} = (-25, 1, 0, 0)^T$ and the equality holds for two observations with age exactly 25.

Let $\mathcal{C}$ and $\mathcal{Q}$ denote the set of all vectors $\boldsymbol{\alpha}$ that satisfy (2) and (3), respectively. For any $\boldsymbol{\alpha} \in \mathcal{C}$, all sample points must satisfy (2), so $\boldsymbol{\alpha}$ cannot lead to quasicomplete separation which requires at least one equality in (3). This implies that $\mathcal{C}$ and $\mathcal{Q}$ are disjoint sets, while both can be non-empty for a certain dataset. Note that Albert and Anderson (1984) define quasicomplete separation only when the sample points cannot be separated using complete separation. Thus according to their definition, only one of $\mathcal{C}$ and $\mathcal{Q}$ can be non-empty for a certain dataset. However, in our slightly modified definition of quasicomplete separation, the absence of complete separation is not required. This permits both $\mathcal{C}$ and $\mathcal{Q}$ to be non-empty for a dataset. In the remainder of the paper, for simplicity we use the term "separation" to refer to either complete or quasicomplete separation, so that $\mathcal{C} \cup \mathcal{Q}$ is non-empty.

2.2 Existence of Posterior Means Under Independent Cauchy Priors

When Markov chain Monte Carlo (MCMC) is applied to sample from the posterior distribution, the posterior mean is a commonly used summary statistic. We aim to study whether the marginal posterior mean $E(\beta_j \mid \mathbf{y})$ exists under the independent Cauchy priors suggested by Gelman et al. (2008). Let $C(\mu, \sigma)$ denote a Cauchy distribution with location parameter μ and scale parameter σ. The default prior suggested by Gelman et al. (2008) corresponds to $\beta_j \overset{\text{ind}}{\sim} C(0, \sigma_j)$, for $j = 1, 2, \dots, p$.

For a design matrix with full column rank, Albert and Anderson (1984) showed that a finite maximum likelihood estimate of $\boldsymbol{\beta}$ does not exist when there is separation in the data. However, even in the case of separation and/or a rank deficient design matrix, the posterior means for some or all β_j's may exist because they incorporate the information from the prior distribution. Following Definition 2.2.1 of Casella and Berger (1990, pp. 55), we say $E(\beta_j \mid \mathbf{y})$ exists if $E(|\beta_j| \mid \mathbf{y}) < \infty$, and in this case, $E(\beta_j \mid \mathbf{y})$ is given by

$$E(\beta_j \mid \mathbf{y}) = \int_0^\infty \beta_j \; p(\beta_j \mid \mathbf{y}) \; d\beta_j + \int_{-\infty}^0 \beta_j \; p(\beta_j \mid \mathbf{y}) \; d\beta_j. \tag{4}$$

Note that alternative definitions may require only one of the integrals in (4) to be finite for the mean to exist, e.g., Bickel and Doksum (2001, pp. 455). However, according to the definition used in this paper, both integrals in (4) have to be finite for the posterior mean to exist. Our main result shows that for each $j = 1, 2, \dots, p$, the existence of $E(\beta_j \mid \mathbf{y})$ depends on whether the predictor $\mathbf{X}_j$ is a solitary separator or not, which is defined as follows:

Definition 1. *The predictor* $\mathbf{X}_j$ *is a solitary separator, if there exists an* $\boldsymbol{\alpha} \in (\mathcal{C} \cup \mathcal{Q})$ *such that*

$$\alpha_j \neq 0, \quad \alpha_r = 0 \textit{ for all } r \neq j. \tag{5}$$

This definition implies that for a solitary separator $\mathbf{X}_j$, if $\alpha_j > 0$, then $x_{i,j} \geq 0$ for all $i \in A_1$, and $x_{i,j} \leq 0$ for all $i \in A_0$; if $\alpha_j < 0$, then $x_{i,j} \leq 0$ for all $i \in A_1$, and $x_{i,j} \geq 0$ for all $i \in A_0$. Therefore, the hyperplane $\{\mathbf{x} \in \mathbb{R}^p : x_j = 0\}$ in the predictor space separates the data into two groups A_1 and A_0 (except for the points located on the hyperplane). The following theorem provides a necessary and sufficient condition for the existence of marginal posterior means of regression coefficients in a logistic regression model.

Theorem 1. *In a logistic regression model, suppose the regression coefficients (including the intercept)* $\beta_j \overset{ind}{\sim} C(0, \sigma_j)$ *with* $\sigma_j > 0$ *for* $j = 1, 2, \dots, p$, *so that*

$$p(\boldsymbol{\beta}) = \prod_{j=1}^p p(\beta_j) = \prod_{j=1}^p \frac{1}{\pi\sigma_j(1 + \beta_j^2/\sigma_j^2)}. \tag{6}$$

Then for each $j = 1, 2, \dots, p$, *the posterior mean* $E(\beta_j \mid \mathbf{y})$ *exists if and only if* $\mathbf{X}_j$ *is not a solitary separator.*

A proof of Theorem 1 is available in Appendices A and B (Ghosh et al., 2017).

Remark 1. *Theorem 1 implies that under independent Cauchy priors in logistic regression, the posterior means of all coefficients exist if there is no separation, or if there is separation with no solitary separators.*

Remark 2. *Gelman et al. (2008) suggested centering all predictors (except interaction terms) in the pre-processing step. A consequence of Theorem 1 is that centering may have a crucial role in the existence of the posterior mean* $E(\beta_j \mid \mathbf{y})$.

We expand on the second remark with a toy example where a predictor is a solitary separator before centering but not after centering. Consider a dataset with $n = 100$, $\mathbf{y} = (\underbrace{0, \dots 0}_{25}, \underbrace{1, \dots, 1}_{75})^T$ and a binary predictor $\mathbf{X}_j = (\underbrace{0, \dots 0}_{50}, \underbrace{1, \dots, 1}_{50})^T$. Here $\mathbf{X}_j$ is a solitary separator which leads to quasicomplete separation before centering. However, the centered predictor $\mathbf{X}_j = (\underbrace{-0.5, \cdots -0.5}_{50}, \underbrace{0.5, \dots, 0.5}_{50})^T$ is no longer a solitary separator because after centering the hyperplane $\{\mathbf{x} : x_j = -0.5\}$ separates the data but

$\{\mathbf{x} : x_j = 0\}$ does not. Consequently, the posterior mean $E(\beta_j \mid \mathbf{y})$ does not exist before centering but it exists after centering.

2.3 Extensions of the Theoretical Result

So far we have mainly focused on the logistic regression model, which is one of the most widely used binary regression models because of the interpretability of its regression coefficients in terms of odds ratios. We now generalize Theorem 1 to binary regression models with link functions other than the logit. Following the definition in McCullagh and Nelder (1989, pp. 27), we assume that for $i = 1, 2, \ldots, n$, the linear predictor $\mathbf{x}_i^T\boldsymbol{\beta}$ and the success probability π_i are connected by a monotonic and differentiable link function $g(\cdot)$ such that $g(\pi_i) = \mathbf{x}_i^T\boldsymbol{\beta}$. We further assume that $g(.)$ is a one-to-one function, which means that $g(.)$ is strictly monotonic. This is satisfied by many commonly used link functions including the probit. Without loss of generality, we assume that $g(\cdot)$ is a strictly increasing function.

Theorem 2. *In a binary regression model with link function $g(.)$ described above, suppose the regression coefficients have independent Cauchy priors in (6). Then for each $j = 1, 2, \ldots, p$,*

(1) a necessary condition for the existence of the posterior mean $E(\beta_j \mid \mathbf{y})$ is that $\mathbf{X}_j$ is not a solitary separator;

(2) a sufficient condition for the existence of $E(\beta_j \mid \mathbf{y})$ consists of the following:

(i) $\mathbf{X}_j$ is not a solitary separator, and

(ii) $\forall \epsilon > 0$,

$$\int_0^\infty \beta_j p(\beta_j) g^{-1}(-\epsilon\beta_j) d\beta_j < \infty, \quad \int_0^\infty \beta_j p(\beta_j) \left[1 - g^{-1}(\epsilon\beta_j)\right] d\beta_j < \infty. \tag{7}$$

Note that (7) in the sufficient condition of Theorem 2 imposes constraints on the link function $g(.)$, and hence the likelihood function. A proof of this theorem is given in Appendix C (Ghosh et al., 2017). Moreover, it is shown that condition (7) holds for the probit link function.

In certain applications, to incorporate available prior information, it may be desirable to use Cauchy priors with nonzero location parameters. The following corollary states that for both logistic and probit regression, the condition for existence of posterior means derived in Theorems 1 and 2 continues to hold under independent Cauchy priors with nonzero location parameters.

Corollary 1. *In logistic and probit regression models, suppose the regression coefficients $\beta_j \overset{ind}{\sim} C(\mu_j, \sigma_j)$, for $j = 1, 2, \ldots, p$. Then a necessary and sufficient condition for the existence of the posterior mean $E(\beta_j \mid \mathbf{y})$ is that $\mathbf{X}_j$ is not a solitary separator, for $j = 1, 2, \ldots, p$.*

A proof of Corollary 1 is available in Appendix D (Ghosh et al., 2017).

In some applications it could be more natural to allow the regression coefficients to be dependent, *a priori*. Thus in addition to independent Cauchy priors, we also study the existence of posterior means under a multivariate Cauchy prior, with the following density function:

$$p(\boldsymbol{\beta}) = \frac{\Gamma\left(\frac{1+p}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\pi^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}\left[1 + (\boldsymbol{\beta} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu})\right]^{\frac{1+p}{2}}}, \tag{8}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{\mu}$ is a $p \times 1$ location parameter and $\boldsymbol{\Sigma}$ is a $p \times p$ positive-definite scale matrix. A special case of the multivariate Cauchy prior is the Zellner–Siow prior (Zellner and Siow, 1980). It can be viewed as a scale mixture of g-priors, where conditional on g, $\boldsymbol{\beta}$ has a multivariate normal prior with a covariance matrix proportional to $g(\mathbf{X}^T\mathbf{X})^{-1}$, and the hyperparameter g has an inverse gamma prior, $\text{IG}(1/2, n/2)$. Based on generalizations of the g-prior to binary regression models (Fouskakis et al., 2009; Sabanés Bové and Held, 2011; Hanson et al., 2014), the Zellner–Siow prior, which has a density (8) with $\boldsymbol{\Sigma} \propto n(\mathbf{X}^T\mathbf{X})^{-1}$, can be a desirable objective prior as it preserves the covariance structure of the data and is free of tuning parameters.

Theorem 3. *In logistic and probit regression models, suppose the vector of regression coefficients $\boldsymbol{\beta}$ has a multivariate Cauchy prior as in (8). If there is no separation, then all posterior means $E(\beta_j \mid \mathbf{y})$ exist, for $j = 1, 2, \ldots, p$. If there is complete separation, then none of the posterior means $E(\beta_j \mid \mathbf{y})$ exist, for $j = 1, 2, \ldots, p$.*

A proof of Theorem 3 is available in Appendices E and F (Ghosh et al., 2017). The study of existence of posterior means under multivariate Cauchy priors in the presence of quasicomplete separation has proved to be more challenging. We hope to study this problem in future work. Note that although under (8), the induced marginal prior of β_j is a univariate Cauchy distribution for each $j = 1, 2, \ldots, p$, the multivariate Cauchy prior is different from independent Cauchy priors, even with a diagonal scale matrix $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2)$. In fact, as a rotation invariant distribution, the multivariate Cauchy prior places less probability mass along axes than the independent Cauchy priors (see Figure 1). Therefore, it is not surprising that solitary separators no longer play an important role for existence of posterior means under multivariate Cauchy priors, as evident from Theorem 3.

So far we have considered Cauchy priors, which are t distributions with 1 degree of freedom. We close this section with a remark on lighter tailed t priors (with degrees of freedom greater than 1) and normal priors, for which the prior means exist.

Remark 3. *In a binary regression model, suppose that the regression coefficients have independent Student-t priors with degrees of freedom greater than one, or independent normal priors. Then it is straightforward to show that the posterior means of the coefficients exist because the likelihood is bounded above by one and the prior means exist. The same result holds under multivariate t priors with degrees of freedom greater than one, and multivariate normal priors.*

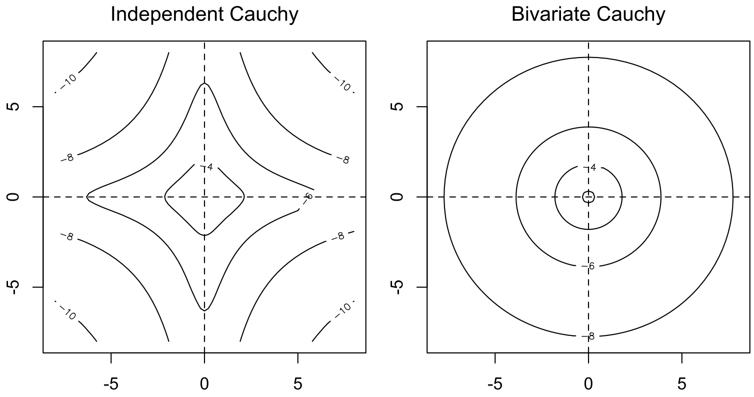

Figure 1: Contour plots of log-density functions of independent Cauchy distributions with both scale parameters being 1 (left) and a bivariate Cauchy distribution with scale matrix $\mathbf{I}_2$ (right). These plots suggest that independent Cauchy priors place more probability mass along axes than a multivariate Cauchy prior, and thus impose stronger shrinkage. Hence, if complete separation occurs, $E(\beta_j \mid \mathbf{Y})$ may exist under independent Cauchy priors for some or all $j = 1, 2, \ldots, p$ (Theorem 1), but does not exist under a multivariate Cauchy prior (Theorem 3).

3 MCMC Sampling for Logistic Regression

In this section we discuss two algorithms for sampling from the posterior distribution for logistic regression coefficients under independent Student-t priors. We first develop a Gibbs sampler and then briefly describe the No-U-Turn Sampler (NUTS) implemented in the freely available software Stan (Carpenter et al., 2016).

3.1 Pólya-Gamma Data Augmentation Gibbs Sampler

Polson et al. (2013) showed that the likelihood for logistic regression can be written as a mixture of normals with respect to a Pólya-Gamma (PG) distribution. Based on this result, they developed an efficient Gibbs sampler for logistic regression with a multivariate normal prior on $\boldsymbol{\beta}$. Choi and Hobert (2013) showed that their Gibbs sampler is uniformly ergodic. This guarantees the existence of central limit theorems for Monte Carlo averages of functions of $\boldsymbol{\beta}$ which are square integrable with respect to the posterior distribution $p(\boldsymbol{\beta} \mid \mathbf{y})$. Choi and Hobert (2013) developed a latent data model which also led to the Gibbs sampler of Polson et al. (2013). We adopt their latent data formulation to develop a Gibbs sampler for logistic regression with independent Student-t priors on $\boldsymbol{\beta}$.

Let $U = (2/\pi^2) \sum_{l=1}^{\infty} W_l/(2l-1)^2$, where $W_1, W_2, \ldots$ is a sequence of i.i.d. Exponential random variables with rate parameters equal to 1. The density of U is given by

$$h(u) = \sum_{l=0}^{\infty} (-1)^l \frac{(2l+1)}{\sqrt{2\pi u^3}} e^{-\frac{(2l+1)^2}{8u}}, \quad 0 < u < \infty. \tag{9}$$

Then for $k \geq 0$, the Pólya-Gamma (PG) distribution is constructed by exponential tilting of $h(u)$ as follows:

$$p(u; k) = \cosh\left(\frac{k}{2}\right) e^{-\frac{k^2 u}{2}} h(u), \quad 0 < u < \infty. \tag{10}$$

A random variable with density $p(u; k)$ has a $\mathrm{PG}(1, k)$ distribution.

Let $t_v(0, \sigma_j)$ denote the Student-t distribution with v degrees of freedom, location parameter 0, and scale parameter σ_j. Since Student-t distributions can be expressed as inverse-gamma (IG) scale mixtures of normal distributions, for $j = 1, 2, \ldots, p$, we have:

$$\beta_j \sim t_v(0, \sigma_j) \Longleftrightarrow \begin{cases} \beta_j \mid \gamma_j \sim \mathrm{N}(0, \gamma_j), \\ \gamma_j \sim \mathrm{IG}\left(\frac{v}{2}, \frac{v\sigma_j^2}{2}\right). \end{cases}$$

Conditional on $\boldsymbol{\beta}$ and $\boldsymbol{\Gamma} = \mathrm{diag}(\gamma_1, \gamma_2, \ldots, \gamma_p)$, let $(y_1, z_1), (y_2, z_2), \ldots, (y_n, z_n)$ be n independent random vectors such that y_i has a Bernoulli distribution with success probability $\exp(\mathbf{x}_i^T \boldsymbol{\beta})/(1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta}))$, $z_i \sim PG(1, |\mathbf{x}_i^T \boldsymbol{\beta}|)$, and y_i and z_i are independent, for $i = 1, 2, \ldots, n$. Let $\boldsymbol{Z}_D = \mathrm{diag}(z_1, z_2, \ldots, z_n)$, then the augmented posterior density is $p(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{Z}_D \mid \mathbf{y})$. We develop a Gibbs sampler with target distribution $p(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{Z}_D \mid \mathbf{y})$, which cycles through the following sequence of distributions iteratively:

1. $\boldsymbol{\beta} \mid \boldsymbol{\Gamma}, \boldsymbol{Z}_D, \mathbf{y} \sim \mathrm{N}\left((\mathbf{X}^T \boldsymbol{Z}_D \mathbf{X} + \boldsymbol{\Gamma}^{-1})^{-1} \mathbf{X}^T \tilde{\mathbf{y}}, (\mathbf{X}^T \boldsymbol{Z}_D \mathbf{X} + \boldsymbol{\Gamma}^{-1})^{-1}\right)$, where $\tilde{y}_i = y_i - 1/2$ and $\tilde{\mathbf{y}} = (\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_n)^T$,
2. $\gamma_j \mid \boldsymbol{\beta}, \boldsymbol{Z}_D, \mathbf{y} \overset{\text{ind}}{\sim} \mathrm{IG}\left(\frac{v+1}{2}, \frac{\beta_j^2 + v\sigma_j^2}{2}\right)$, for $j = 1, 2, \ldots, p$,
3. $z_i \mid \boldsymbol{\Gamma}, \boldsymbol{\beta}, \mathbf{y} \overset{\text{ind}}{\sim} \mathrm{PG}(1, |\mathbf{x}_i^T \boldsymbol{\beta}|)$, for $i = 1, 2, \ldots, n$.

Steps 1 and 3 follow immediately from Choi and Hobert (2013); Polson et al. (2013) and step 2 follows from straightforward algebra. In the next section, for comparison of posterior distributions under Student-t priors with different degrees of freedom, we implement the above Gibbs sampler, and for normal priors we apply the Gibbs sampler of Polson et al. (2013). Both Gibbs samplers can be implemented using the R package `tglm`, available in the supplement.

3.2 Stan

Our empirical results in the next section suggest that the Gibbs sampler exhibits extremely slow mixing for posterior simulation under Cauchy priors for data with separation. Thus we consider alternative MCMC sampling algorithms in the hope of improving mixing. A random walk Metropolis algorithm shows some improvement over the Gibbs sampler in the $p = 2$ case. However, it is not efficient for exploring higher dimensional spaces. Thus we have been motivated to use the software Stan (Carpenter et al., 2016), which implements the No-U-Turn Sampler (NUTS) of Hoffman and Gelman (2014), a tuning free extension of the Hamiltonian Monte Carlo (HMC) algorithm (Neal, 2011).

It has been demonstrated that for continuous parameter spaces, HMC can improve over poorly mixing Gibbs samplers and random walk Metropolis algorithms. HMC is a Metropolis algorithm that generates proposals based on Hamiltonian dynamics, a concept borrowed from Physics. In HMC, the parameter of interest is referred to as the "position" variable, representing a particle's position in a p-dimensional space. A p-dimensional auxiliary parameter, the "momentum" variable, is introduced to represent the particle's momentum. In each iteration, the momentum variable is generated from a Gaussian distribution, and then a proposal of the position momentum pair is generated (approximately) along the trajectory of the Hamiltonian dynamics defined by the joint distribution of the position and momentum. Hamiltonian dynamics changing over time can be approximated by discretizing time via the "leapfrog" method. In practice, a proposal is generated by applying the leapfrog algorithm L times, with stepsize ϵ, to the current state. The proposed state is accepted or rejected according to a Metropolis acceptance probability. Section 5.3.3 of the review paper by Neal (2011) illustrates the practical benefits of HMC over random walk Metropolis algorithms. The examples in this section demonstrate that the momentum variable may change only slowly along certain directions during leapfrog steps, permitting the position variable to move consistently in this direction for many steps. In this way, proposed states using Hamiltonian dynamics can be far away from current states but still achieve high acceptance probabilities, making HMC more efficient than traditional algorithms such as random walk Metropolis.

In spite of its advantages, HMC has not been very widely used in the Statistics community until recently, because its performance can be sensitive to the choice of two tuning parameters: the leapfrog stepsize ϵ and the number of leapfrog steps L. Very small ϵ can lead to waste in computational power whereas large ϵ can yield large errors due to discretization. Regarding the number of leapfrog steps L, if it is too small, proposed states can be near current states and thus resemble random walk. On the other hand, if L is too large, the Hamiltonian trajectory can retrace its path so that the proposal is brought closer to the current value, which again is a waste of computational power.

The NUTS algorithm tunes these two parameters automatically. To select L, the main idea is to run the leapfrog steps until the trajectory starts to retrace its path. More specifically, NUTS builds a binary tree based on a recursive doubling procedure, that is similar in flavor to the doubling procedure used for slice sampling by Neal (2003), with nodes of the tree representing position momentum pairs visited by the leapfrog steps along the path. The doubling procedure is stopped if the trajectory starts retracing its path, that is making a "U-turn", or if there is a large simulation error accumulated due to many steps of leapfrog discretization. NUTS consists of a carefully constructed transition kernel that leaves the target joint distribution invariant. It also proposes a way for adaptive tuning of the stepsize ϵ.

We find that by implementing this tuning free NUTS algorithm, available in the freely available software Stan, substantially better mixing than the Gibbs sampler can be achieved in all of our examples in which posterior means exist. We still include the Gibbs sampler in this article for two main reasons. First, it illustrates that Stan can provide an incredible improvement in mixing over the Gibbs sampler in certain cases. Stan requires minimal coding effort, much less than developing a Gibbs sampler, which may be useful information for readers who are not yet familiar with Stan. Second, Stan currently works for continuous target distributions only, but discrete distributions for models and mixed distributions for regression coefficients frequently arise in Bayesian variable selection, for regression models with binary or categorical response variables (Holmes and Held, 2006; Mitra and Dunson, 2010; Ghosh and Clyde, 2011; Ghosh et al., 2011; Ghosh and Reiter, 2013; Li and Clyde, 2015). Unlike HMC algorithms, Gibbs samplers can typically be extended via data augmentation to incorporate mixtures of a point mass and a continuous distribution, as priors for the regression coefficients, without much additional effort.

4 Simulated Data

In this section, we use two simulation examples to empirically demonstrate that under independent Cauchy priors, the aforementioned MCMC algorithm for logistic regression may suffer from extremely slow mixing in the presence of separation in the dataset.

For each simulation scenario, we first standardize the predictors following the recommendation of Gelman et al. (2008). Binary predictors (with 0/1 denoting the two categories) are centered to have mean 0, and other predictors are centered and scaled to have mean 0 and standard deviation 0.5. Their rationale is that such standardizing makes the scale of a continuous predictor comparable to that of a symmetric binary predictor, in the sense that they have the same sample mean and sample standard deviation. Gelman et al. (2008) made a distinction between input variables and predictors, and they suggested standardizing the input variables only. For example, temperature and humidity may be input variables as well as predictors in a model; however, their interaction term is a predictor but not an input variable. In our examples, except for the constant term for the intercept, all other predictors are input variables and standardized appropriately.

We compare the posterior distributions under independent i) Cauchy, i.e., Student-t with 1 degree of freedom, ii) Student-t with 7 degrees of freedom (t_7), and iii) normal priors for the regression coefficients. In binary regression models, while the inverse cumulative distribution function (CDF) of the logistic distribution yields the logit link function, the inverse CDF of the Student-t distribution yields the robit link function. Liu (2004) showed that the logistic link can be well approximated by a robit link with 7 degrees of freedom. So a t_7 prior approximately matches the tail heaviness of the logistic likelihood underlying logistic regression. For Cauchy priors we use the default choice recommended by Gelman et al. (2008): all location parameters are set to 0 and scale parameters are set to 10 and 2.5 for the intercept and other coefficients, respectively. To be consistent we use the same location and scale parameters for the other two priors. Gelman et al. (2008) adopted a similar strategy in one of their analyses, to study the effect of tail heaviness of the priors. Among the priors considered here, the normal prior has the lightest tails, the Cauchy prior the heaviest, and the t_7 prior offers a compromise between the two extremes. For each simulated dataset, we run both the Gibbs sampler developed in Section 3.1 and Stan, for 1,000,000 iterations after a burn-in of 100,000 samples, under each of the three priors.

4.1 Complete Separation with a Solitary Separator

First, we generate a dataset with $p = 2$ (including the intercept) and $n = 30$. The continuous predictor $\mathbf{X}_2$ is chosen to be a solitary separator (after standardizing), which leads to complete separation, whereas the constant term $\mathbf{X}_1$ contains all one's and is not a solitary separator. A plot of $\mathbf{y}$ versus $\mathbf{X}_2$ in Figure 2 demonstrates this graphically. So by Theorem 1, under independent Cauchy priors, $E(\beta_1 \mid \mathbf{y})$ exists but $E(\beta_2 \mid \mathbf{y})$ does not.

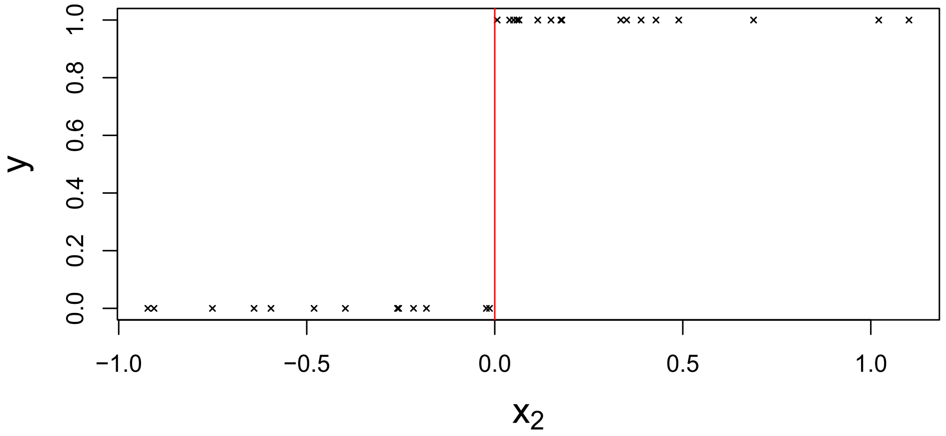

Figure 2: Scatter plot of $\mathbf{y}$ versus $\mathbf{X}_2$ in the first simulated dataset, where $\mathbf{X}_2$ is a solitary separator which completely separates the samples (the vertical line at zero separates the points corresponding to $y = 1$ and $y = 0$).

The results from the Gibbs sampler are reported in Figures 3 and 4. Figure 3 shows the posterior samples of $\boldsymbol{\beta}$ under the different priors. The scale of β_2, the coefficient corresponding to the solitary separator $\mathbf{X}_2$, is extremely large under Cauchy priors, less so under t_7 priors, and the smallest under normal priors. In particular, under Cauchy priors, the posterior distribution of β_2 seems to have an extremely long right tail. Moreover, although $\mathbf{X}_1$ is not a solitary separator, under Cauchy priors, the posterior samples of β_1 have a much larger spread. Figure 4 shows that the running means of both β_1 and β_2 converge rapidly under normal and t_7 priors, whereas under Cauchy priors, the running mean of β_1 does not converge after a million iterations and that of β_2 clearly diverges. We also ran Stan for this example but do not report the results here, because it gave warning messages about divergent transitions for Cauchy priors, after the burn-in period. Given that the posterior mean of β_2 does not exist in this case, the lack of convergence is not surprising.

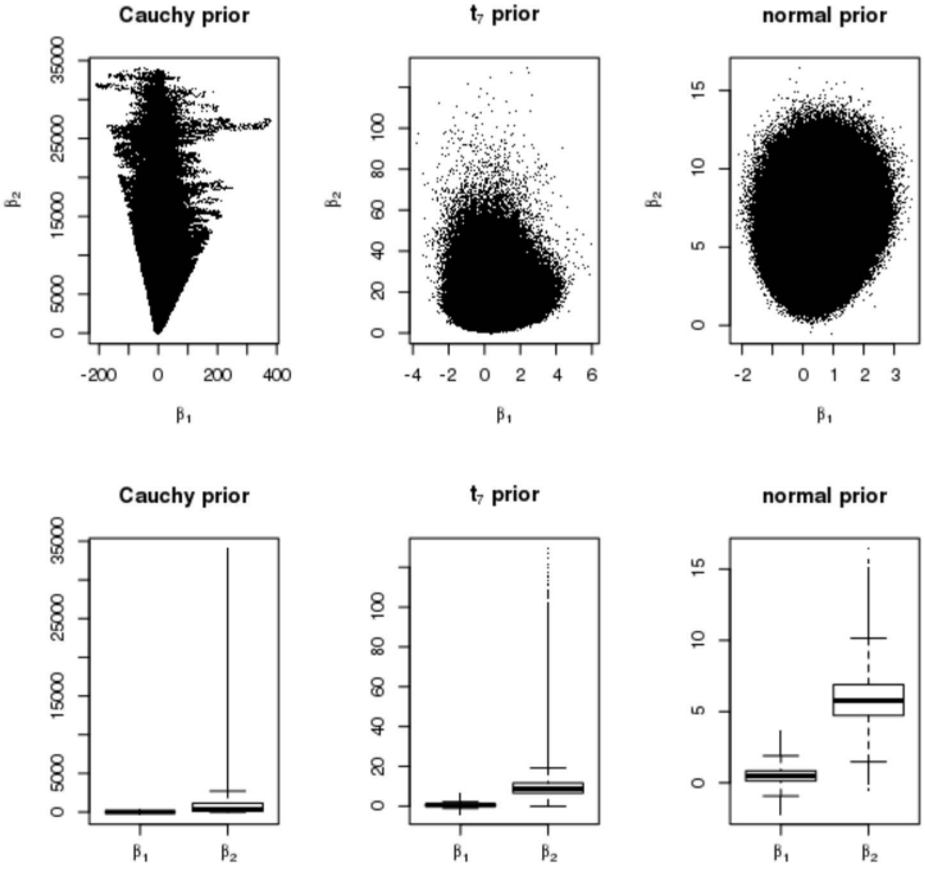

Figure 3: Scatter plots (top) and box plots (bottom) of posterior samples of β_1 and β_2 from the Gibbs sampler, under independent Cauchy, t_7, and normal priors for the first simulated dataset.

4.2 Complete Separation Without Solitary Separators

Now we generate a new dataset with $p = 2$ and $n = 30$ such that there is complete separation but there are no solitary separators (see Figure 5). This guarantees the existence of both $E(\beta_1 \mid \mathbf{y})$ and $E(\beta_2 \mid \mathbf{y})$ under independent Cauchy priors. The difference in the existence of $E(\beta_2 \mid \mathbf{y})$ for the two simulated datasets is reflected by the posterior samples from the Gibbs sampler: under Cauchy priors, the samples of β_2 in Figure 1 in the Appendix are more stabilized than those in Figure 3 in the manuscript. However, when comparing across prior distributions, we find that the posterior samples of neither β_1 nor β_2 are as stable as those under t_7 and normal priors, which is not surprising because among the three priors, Cauchy priors have the heaviest tails and thus yield the least shrinkage. Figure 2 in the Appendix shows that the convergence of the running means under Cauchy priors is slow. Although we have not verified the

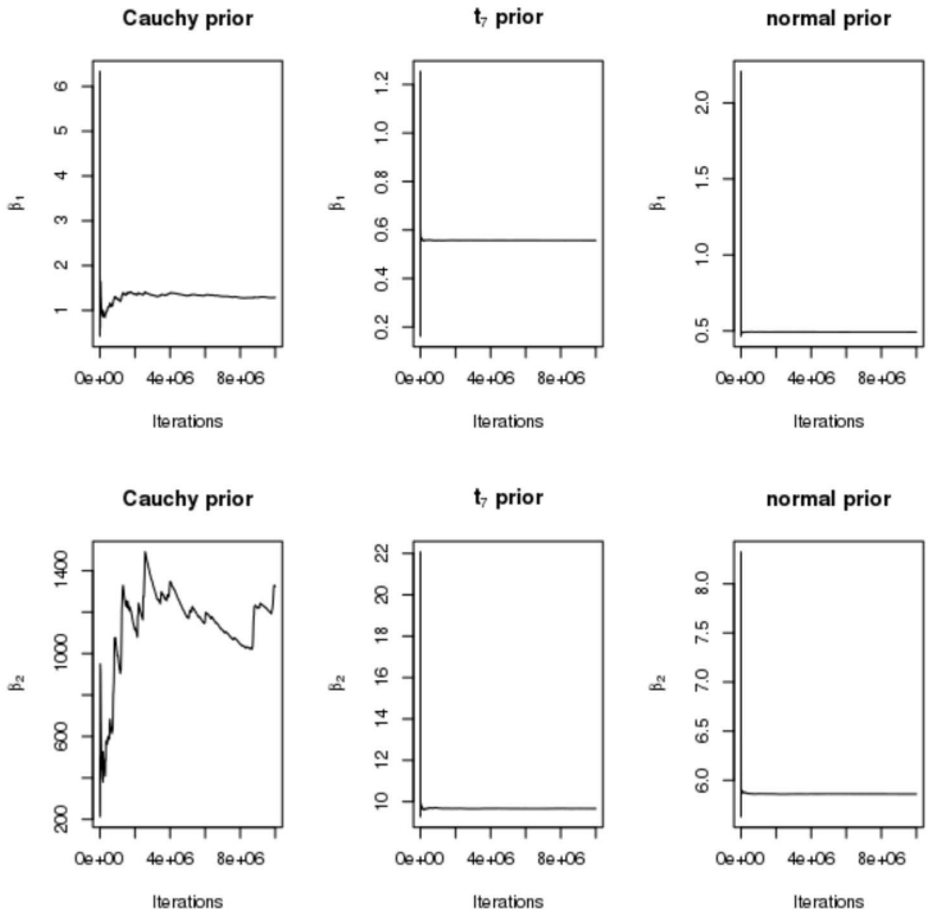

Figure 4: Plots of running means of β_1 (top) and β_2 (bottom) sampled from the posterior distributions via the Gibbs sampler, under independent Cauchy, t_7, and normal priors for the first simulated dataset. Here $E(\beta_1 \mid \mathbf{y})$ exists under independent Cauchy priors but $E(\beta_2 \mid \mathbf{y})$ does not.

existence of the second or higher order posterior moments under Cauchy priors, for exploratory purposes we examine sample autocorrelation plots of the draws from the Gibbs sampler. Figure 6 shows that the autocorrelation decays extremely slowly for Cauchy priors, reasonably fast for t_7 priors, and rapidly for normal priors.

Some results from Stan are reported in Figures 3 and 4 in the Appendix. Figure 3 in the Appendix shows posterior distributions with nearly identical shapes as those obtained using Gibbs sampling in Figure 1 in the Appendix, with the only difference being that more extreme values appear under Stan. This is most likely due to faster mixing in Stan. As Stan traverses the parameter space more rapidly, values in the tails appear more quickly than under the Gibbs sampler. Figures 2 and 4 in the Appendix

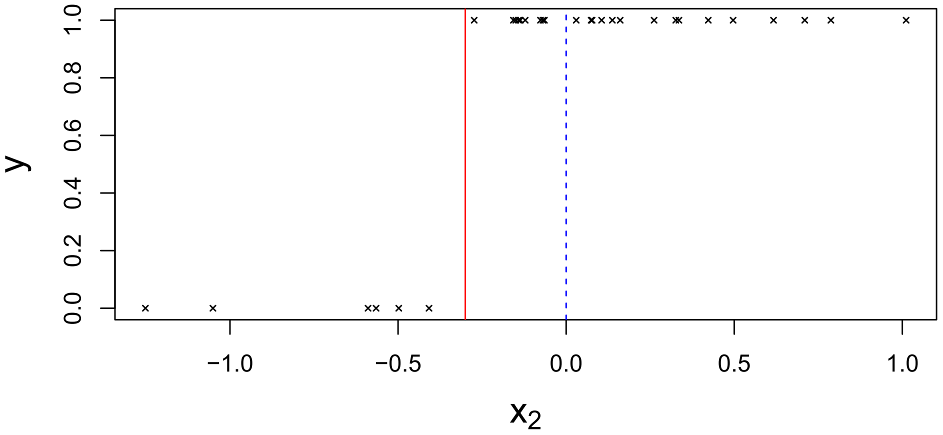

Figure 5: Scatter plot of $\mathbf{y}$ versus $\mathbf{X}_2$ for the second simulated dataset. The solid vertical line at -0.3 demonstrates complete separation of the samples. However, $\mathbf{X}_2$ is not a solitary separator, because the dashed vertical line at zero does not separate the points corresponding to $y = 1$ and $y = 0$. The other predictor $\mathbf{X}_1$ is a vector of ones corresponding to the intercept, which is not a solitary separator, either.

demonstrate that running means based on Stan are in good agreement with those based on the Gibbs sampler.

The autocorrelation plots for Stan in Figure 7 demonstrate a remarkable improvement over those for Gibbs in Figure 6 for all priors, and the difference in mixing is the most prominent for Cauchy priors.

To summarize, all the plots unequivocally suggest that Cauchy priors lead to an extremely slow mixing Gibbs sampler and unusually large scales for the regression coefficients, even when all the marginal posterior means are guaranteed to exist. While mixing can be improved tremendously with Stan, the heavy tailed posteriors under Stan are in agreement with those obtained from the Gibbs samplers. One may argue that in spite of the unnaturally large regression coefficients, Cauchy priors could lead to superior predictions. Thus in the next two sections we compare predictions based on posteriors under the three priors for two real datasets. As Stan generates nearly independent samples, we use Stan for MCMC simulations for the real datasets.

5 Real Data

5.1 SPECT Dataset

The “SPECT” dataset (Kurgan et al., 2001) is available from the UCI Machine Learning Repository[1]. The binary response variable is whether a patient’s cardiac image is normal

[1] https://archive.ics.uci.edu/ml/datasets/SPECT+Heart

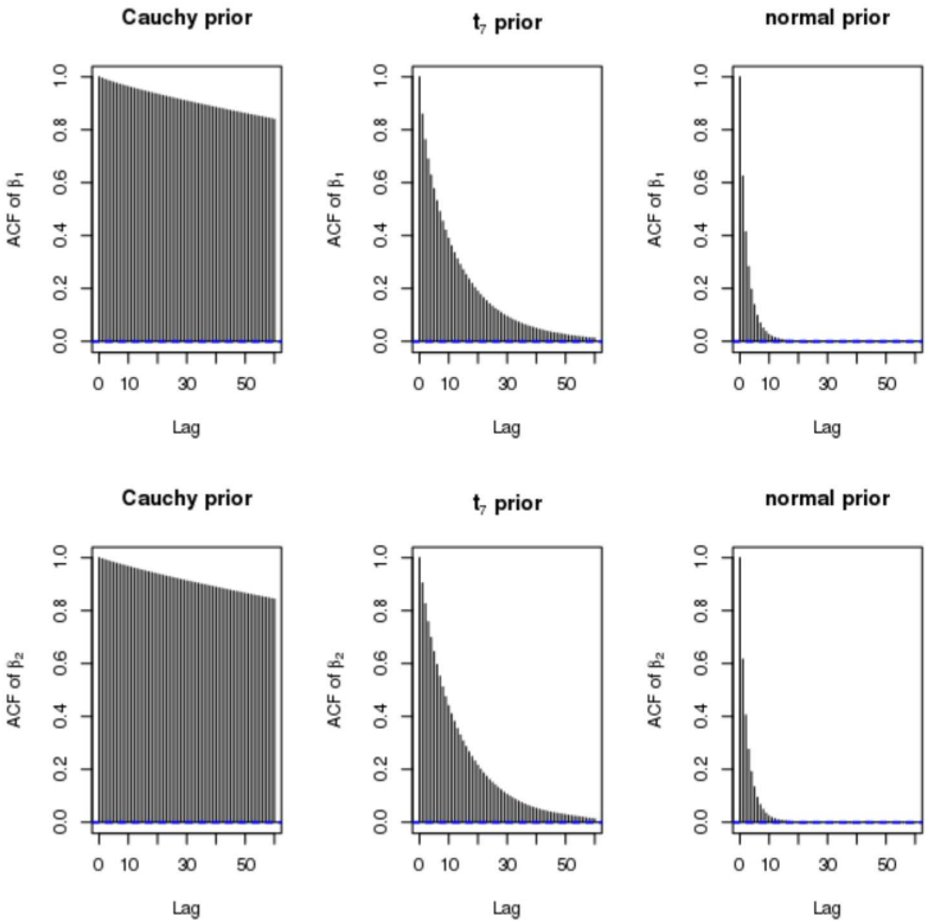

Figure 6: Autocorrelation plots of the posterior samples of β_1 (top) and β_2 (bottom) from the Gibbs sampler, under independent Cauchy, t_7, and normal priors for the second simulated dataset.

or abnormal, according to the diagnosis of cardiologists. The predictors are 22 binary features obtained from the cardiac images using a machine learning algorithm. The goal of the study is to determine if the predictors can correctly predict the diagnoses of cardiologists, so that the process could be automated to some extent.

Prior to centering, two of the binary predictors are solitary quasicomplete separators: $x_{i,j} = 0 \ \forall i \in A_0$ and $x_{i,j} \geq 0 \ \forall i \in A_1$, for $j = 18, 19$, with $\mathbf{X}_1$ denoting the column of ones. Ghosh and Reiter (2013) analyzed this dataset with a Bayesian probit regression model which incorporated variable selection. As some of their proposed methods relied on an approximation of the marginal likelihood based on the MLE of $\boldsymbol{\beta}$, they had to drop these potentially important predictors from the analysis. If one analyzed the dataset with the uncentered predictors, by Theorem 1, the posterior means $E(\beta_{18} \mid \mathbf{y})$ and

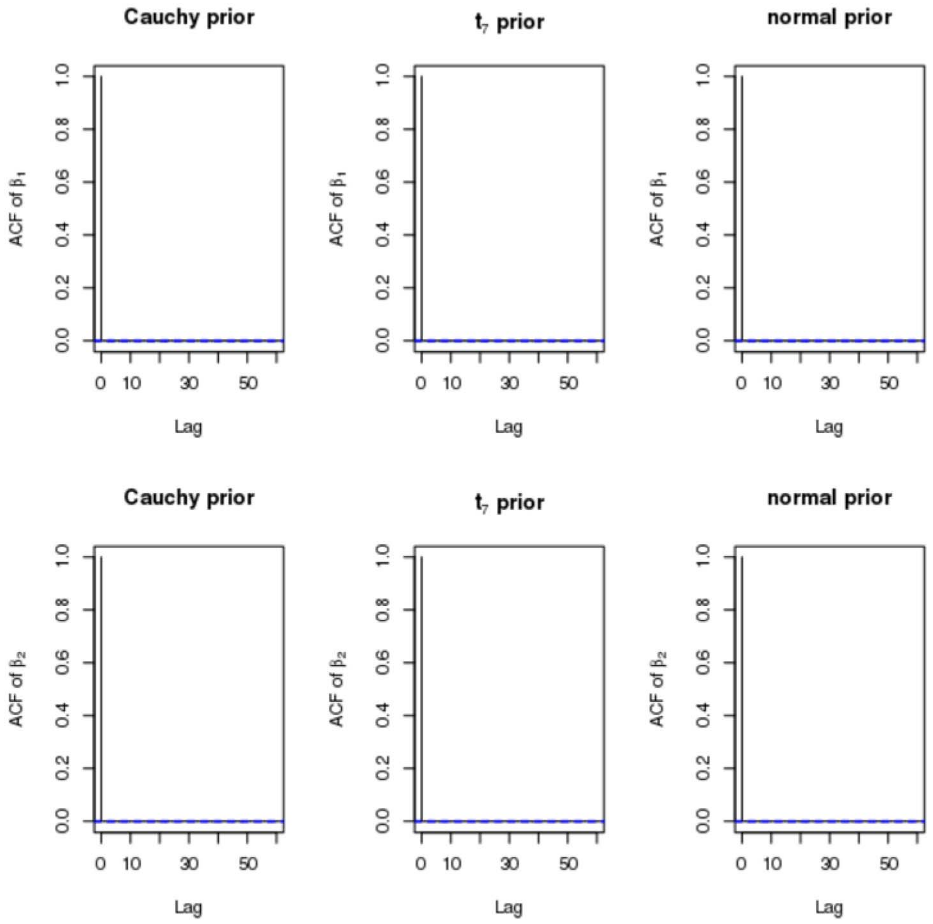

Figure 7: Autocorrelation plots of the posterior samples of β_1 (top) and β_2 (bottom) from Stan, under independent Cauchy, t_7, and normal priors for the second simulated dataset.

$E(\beta_{19} \mid \mathbf{y})$ would not exist under independent Cauchy priors. However, after centering there are no solitary separators, so the posterior means of all coefficients exist.

The SPECT dataset is split into a training set of 80 observations and a test set of 187 observations by Kurgan et al. (2001). We use the former for model fitting and the latter for prediction. First, for each of the three priors (Cauchy, t_7, and normal), we run Stan on the training dataset, for 1,000,000 iterations after discarding 100,000 samples as burn-in.

As in the simulation study, MCMC draws from Stan show excellent mixing for all priors. However, the posterior means of the regression coefficients involved in separation are rather large under Cauchy priors compared to the other priors. For example, the posterior means of (β_{18}, β_{19}) under Cauchy, t_7, and normal priors are $(10.02, 5.57)$, $(3.24, 1.68)$, and $(2.73, 1.43)$ respectively. These results suggest that Cauchy priors are too diffuse for datasets with separation.

Next for each $i = 1, 2, \ldots, n_{\text{test}}$ in the test set, we estimate the corresponding success probability π_i by the Monte Carlo average:

$$\widehat{\pi}_i^{\text{MC}} = \frac{1}{S}\sum_{s=1}^{S} \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}^{(s)}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}^{(s)}}}, \tag{11}$$

where $\boldsymbol{\beta}^{(s)}$ is the sampled value of $\boldsymbol{\beta}$ in iteration s, after burn-in. Recall that here $n_{\text{test}} = 187$ and $S = 10^6$. We calculate two different types of summary measures to assess predictive performance. We classify the ith observation in the test set as a success, if $\widehat{\pi}_i^{\text{MC}} \geq 0.5$ and as a failure otherwise, and compute the misclassification rates. Note that the misclassification rate does not fully take into account the magnitude of $\widehat{\pi}_i^{\text{MC}}$. For example, if $y_i = 1$ both $\widehat{\pi}_i^{\text{MC}} = 0.5$ and $\widehat{\pi}_i^{\text{MC}} = 0.9$ would correctly classify the observation, while the latter may be more preferable. So we also consider the average squared difference between y_i and $\widehat{\pi}_i^{\text{MC}}$:

$$MSE^{\text{MC}} = \frac{1}{n_{\text{test}}}\sum_{i=1}^{n_{\text{test}}} \left(\widehat{\pi}_i^{\text{MC}} - y_i\right)^2, \tag{12}$$

which is always between 0 and 1, with a value closer to 0 being more preferable. Note that the Brier score (Brier, 1950) equals $2MSE^{\text{MC}}$, according to its original definition. Since in some modified definitions (Blattenberger and Lad, 1985), it is the same as MSE^{MC}, we refer to MSE^{MC} as the Brier score.

	Cauchy	t_7	normal
MCMC	0.273	0.257	0.251
EM	0.278	0.262	0.262

Table 1: Misclassification rates based on $\widehat{\pi}_i^{\text{MC}}$ and $\widehat{\pi}_i^{\text{EM}}$, under Cauchy, t_7, and normal priors for the SPECT data. Small values are preferable.

	Cauchy	t_7	normal
MCMC	0.172	0.165	0.163
EM	0.179	0.178	0.178

Table 2: Brier scores MSE^{MC} and MSE^{EM}, under Cauchy, t_7, and normal priors for the SPECT data. Small values are preferable.

To compare the Monte Carlo estimates with those based on the EM algorithm of Gelman et al. (2008), we also estimate the posterior mode, denoted by $\widetilde{\boldsymbol{\beta}}$ under identical priors and hyperparameters, using the R package `arm` (Gelman et al., 2015). The EM estimator of π_i is given by:

$$\widehat{\pi}_i^{\text{EM}} = \frac{e^{\mathbf{x}_i^T \widetilde{\boldsymbol{\beta}}}}{1 + e^{\mathbf{x}_i^T \widetilde{\boldsymbol{\beta}}}}, \tag{13}$$

and MSE^{EM} is calculated by replacing $\widehat{\pi}_i^{\mathrm{MC}}$ by $\widehat{\pi}_i^{\mathrm{EM}}$ in (12).

We report the misclassification rates in Table 1 and the Brier scores in Table 2. MCMC achieves somewhat smaller misclassification rates and Brier scores than EM, especially under t_7 and normal priors. This suggests that a full Bayesian analysis using MCMC may produce estimates that are closer to the truth than modal estimates based on the EM algorithm. The predictions are similar across the three prior distributions with the normal and t_7 priors yielding slightly more accurate results than Cauchy priors.

5.2 Pima Indians Diabetes Dataset

We now analyze the "Pima Indians Diabetes" dataset in the R package `MASS`. This is a classic dataset without separation that has been analyzed by many authors in the past. Using this dataset we aim to compare predictions under different priors, when there is no separation. Using the training data provided in the package we predict the class labels of the test data. In this case the difference between different priors is practically nil. The Brier scores are same up to three decimal places, across all priors and all methods (EM and MCMC). The misclassification rates reported in Table 3 also show negligible difference between priors and methods. Here Cauchy priors have a slightly better misclassification rate compared to normal and t_7 priors, and MCMC provides slightly more accurate results compared to those obtained from EM. These results suggest that when there is no separation and maximum likelihood estimates exist, Cauchy priors may be preferable as default weakly informative priors in the absence of real prior information.

	Cauchy	t_7	normal
MCMC	0.196	0.199	0.199
EM	0.202	0.202	0.202

Table 3: Misclassification rates based on $\widehat{\pi}_i^{\mathrm{MC}}$ and $\widehat{\pi}_i^{\mathrm{EM}}$, under Cauchy, t_7, and normal priors for the Pima Indians data. Small values are preferable.

6 Discussion

We have proved that posterior means of regression coefficients in logistic regression are not always guaranteed to exist under the independent Cauchy priors recommended by Gelman et al. (2008), if there is complete or quasicomplete separation in the data. In particular, we have introduced the notion of a solitary separator, which is a predictor capable of separating the samples on its own. Note that a solitary separator needs to be able to separate without the aid of any other predictor, not even the constant term corresponding to the intercept. We have proved that for independent Cauchy priors, the absence of solitary separators is a necessary condition for the existence of posterior means of all coefficients, for a general family of link functions in binary regression models. For logistic and probit regression, this has been shown to be a sufficient condition as well. In general, the sufficient condition depends on the form of the link function.

We have also studied multivariate Cauchy priors, where the solitary separator no longer plays an important role. Instead, posterior means of all predictors exist if there is no separation, while none of them exist if there is complete separation. The result under quasicompelte separation is still unclear and will be studied in future work.

In practice, after centering the input variables it is straightforward to check if there are solitary separators in the dataset. The absence of solitary separators guarantees the existence of posterior means of all regression coefficients in logistic regression under independent Cauchy priors. However, our empirical results have shown that even when the posterior means for Cauchy priors exist under separation, the posterior samples of the regression coefficients may be extremely large in magnitude. Separation is usually considered to be a sample phenomenon, so even if the predictors involved in separation are potentially important, some shrinkage of their coefficients is desirable through the prior. Our empirical results based on real datasets have demonstrated that the default Cauchy priors can lead to posterior means as large as 10, which is considered to be unusually large on the logit scale. Our impression is that Cauchy priors are good default choices in general because they contain weak prior information and let the data speak.However, under separation, when there is little information in the data about the logistic regression coefficients (the MLE is not finite), it seems that lighter tailed priors, such as Student-t priors with larger degrees of freedom or even normal priors, are more desirable in terms of producing more plausible posterior distributions.

From a computational perspective, we have observed very slow convergence of the Gibbs sampler under Cauchy priors in the presence of separation. Note that if the design matrix is not of full column rank, for example when $p > n$, the p columns of $\mathbf{X}$ will be linearly dependent. This implies that the equation for quasicomplete separation (3) will be satisfied with equality for all observations. Empirical results (not reported here for brevity) demonstrated that independent Cauchy priors show convergence of the Gibbs sampler in this case also compared to other lighter tailed priors. Out-of-sample predictive performance based on a real dataset with separation did not show the default Cauchy priors to be superior to t_7 or normal priors.

In logistic regression, under a multivariate normal prior for $\boldsymbol{\beta}$, Choi and Hobert (2013) showed that the Pólya-Gamma data augmentation Gibbs sampler of Polson et al. (2013) is uniformly ergodic, and the moment generating function of the posterior distribution $p(\boldsymbol{\beta} \mid \mathbf{y})$ exists for all $\mathbf{X}, \mathbf{y}$. In our examples of datasets with separation, the normal priors led to the fastest convergence of the Gibbs sampler, reasonable scales for the posterior draws of $\boldsymbol{\beta}$, and comparable or even better predictive performance than other priors. The results from Stan show no problem in mixing under any of the priors. However, the problematic issue of posteriors with extremely heavy tails under Cauchy priors cannot be resolved without altering the prior. Thus, after taking into account all the above considerations, for a full Bayesian analysis we recommend the use of normal priors as a default, when there is separation. Alternatively, heavier tailed priors such as the t_7 could also be used if robustness is a concern. On the other hand, if the goal of the analysis is to obtain point estimates rather than the entire posterior distribution, the posterior mode obtained from the EM algorithm of Gelman et al. (2015) under default Cauchy priors (Gelman et al., 2008) is a fast viable alternative.

Supplementary Material

Supplementary Material for "On the Use of Cauchy Prior Distributions for Bayesian Logistic Regression" (DOI: 10.1214/17-BA1051SUPP; .pdf). In the supplementary material, we present additional simulation results for logistic and probit regression with complete separation, along with an appendix that contains the proofs of all theoretical results. The Gibbs sampler developed in the paper can be implemented with the R package `tglm`, available from the website: `https://cran.r-project.org/web/packages/tglm/index.html`.

References

Albert, A. and Anderson, J. A. (1984). "On the Existence of Maximum Likelihood Estimates in Logistic Regression Models." *Biometrika*, 71(1): 1–10. MR0738319. doi: `https://doi.org/10.1093/biomet/71.1.1`. 360, 362, 363

Bardenet, R., Doucet, A., and Holmes, C. (2014). "Towards Scaling up Markov Chain Monte Carlo: An Adaptive Subsampling Approach." *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 405–413. 361

Bickel, P. J. and Doksum, K. A. (2001). *Mathematical Statistics, volume I*. Prentice Hall Englewood Cliffs, NJ. 364

Blattenberger, G. and Lad, F. (1985). "Separating the Brier Score into Calibration and Refinement Components: A Graphical Exposition." *The American Statistician*, 39(1): 26–32. 377

Brier, G. W. (1950). "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review*, 78: 1–3. 377

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, A., Michael, Guo, J., Li, P., and Riddell, A. (2016). "Stan: A Probabilistic Programming Language." *Journal of Statistical Software*, in press. 367, 369

Casella, G. and Berger, R. L. (1990). *Statistical Inference*. Duxbury Press. MR1051420. 363

Chen, M.-H. and Shao, Q.-M. (2001). "Propriety of Posterior Distribution for Dichotomous Quantal Response Models." *Proceedings of the American Mathematical Society*, 129(1): 293–302. 360

Choi, H. M. and Hobert, J. P. (2013). "The Polya-Gamma Gibbs Sampler for Bayesian Logistic Regression is Uniformly Ergodic." *Electronic Journal of Statistics*, 7(2054–2064). 367, 368, 379

Chopin, N. and Ridgway, J. (2015). "Leave Pima Indians Alone: Binary Regression as a Benchmark for Bayesian Computation." *arxiv.org*. 361

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B., and Weidman, L. (1991). "Multiple Imputation of Industry and Occupation Codes in Census Public-Use Samples

Using Bayesian Logistic Regression." *Journal of the American Statistical Association*, 86(413): 68–78. 361

Dawid, A. P. (1973). "Posterior Expectations for Large Observations." *Biometrika*, 60: 664–666. 360

Fernández, C. and Steel, M. F. (2000). "Bayesian Regression Analysis with Scale Mixtures of Normals." *Econometric Theory*, 16(80–101). 359

Firth, D. (1993). "Bias Reduction of Maximum Likelihood Estimates." *Biometrika*, 80(1): 27–38. 360

Fouskakis, D., Ntzoufras, I., and Draper, D. (2009). "Bayesian Variable Selection Using Cost-Adjusted BIC, with Application to Cost-Effective Measurement of Quality of Health Care." *The Annals of Applied Statistics*, 3(2): 663–690. 366

Gelman, A., Jakulin, A., Pittau, M., and Su, Y. (2008). "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models." *The Annals of Applied Statistics*, 2(4): 1360–1383. MR2655663. doi: https://doi.org/10.1214/08-AOAS191. 359, 361, 363, 364, 370, 377, 378, 379

Gelman, A., Su, Y.-S., Yajima, M., Hill, J., Pittau, M. G., Kerman, J., Zheng, T., and Dorie, V. (2015). *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models*. R package version 1.8-5. URL http://CRAN.R-project.org/package=arm 377, 379

Ghosh, J. and Clyde, M. A. (2011). "Rao-Blackwellization for Bayesian Variable Selection and Model Averaging in Linear and Binary Regression: A Novel Data Augmentation Approach." *Journal of the American Statistical Association*, 106(495): 1041–1052. 370

Ghosh, J., Herring, A. H., and Siega-Riz, A. M. (2011). "Bayesian Variable Selection for Latent Class Models." *Biometrics*, 67: 917–925. MR2829266. doi: https://doi.org/10.1111/j.1541-0420.2010.01502.x. 370

Ghosh, J., Li, Y., and Mitra, R. (2017). "Supplementary Material for "On the Use of Cauchy Prior Distributions for Bayesian Logistic Regression"." *Bayesian Analysis*. doi: https://doi.org/10.1214/17-BA1051SUPP. 364, 365, 366

Ghosh, J. and Reiter, J. P. (2013). "Secure Bayesian Model Averaging for Horizontally Partitioned Data." *Statistics and Computing*, 23: 311–322. 370, 375

Gramacy, R. B. and Polson, N. G. (2012). "Simulation-Based Regularized Logistic Regression." *Bayesian Analysis*, 7(3): 567–590. 361

Hanson, T. E., Branscum, A. J., and Johnson, W. O. (2014). "Informative g-Priors for Logistic Regression." *Bayesian Analysis*, 9(3): 597–612. 366

Heinze, G. (2006). "A Comparative Investigation of Methods for Logistic Regression with Separated or Nearly Separated Data." *Statistics in Medicine*, 25: 4216–4226. 360

Heinze, G. and Schemper, M. (2002). "A Solution to the Problem of Separation in Logistic Regression." *Statistics in Medicine*, 21: 2409–2419. 360

Hoffman, M. D. and Gelman, A. (2014). "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *The Journal of Machine Learning Research*, 15(1): 1593–1623. 361, 369

Holmes, C. C. and Held, L. (2006). "Bayesian Auxiliary Variable Models for Binary and Multinomial Regression." *Bayesian Analysis*, 1(1): 145–168. 361, 370

Ibrahim, J. G. and Laud, P. W. (1991). "On Bayesian Analysis of Generalized Linear Models using Jeffreys's Prior." *Journal of the American Statistical Association*, 86(416): 981–986. 360

Jeffreys, H. (1961). *Theory of Probability*. Oxford Univ. Press. 360

Kurgan, L., Cios, K., Tadeusiewicz, R., Ogiela, M., and Goodenday, L. (2001). "Knowledge Discovery Approach to Automated Cardiac SPECT Diagnosis." *Artificial Intelligence in Medicine*, 23:2: 149–169. 374, 376

Li, Y. and Clyde, M. A. (2015). "Mixtures of g-Priors in Generalized Linear Models." *arxiv.org*. 370

Liu, C. (2004). "Robit Regression: A Simple Robust Alternative to Logistic and Probit Regression." In Gelman, A. and Meng, X. (eds.), *Applied Bayesian Modeling and Casual Inference from Incomplete-Data Perspectives*, 227–238. Wiley, London. 370

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall. 365

Mitra, R. and Dunson, D. B. (2010). "Two Level Stochastic Search Variable Selection in GLMs with Missing Predictors." *International Journal of Biostatistics*, 6(1): Article 33. MR2729579. doi: https://doi.org/10.2202/1557-4679.1173. 370

Neal, R. M. (2003). "Slice Samlping." *The Annals of Statistics*, 31(3): 705–767. MR1994729. doi: https://doi.org/10.1214/aos/1056562461. 369

Neal, R. M. (2011). "MCMC using Hamiltonian Dynamics." In Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (eds.), *Handbook of Markov Chain Monte Carlo*. Chapman & Hall / CRC Press. 369

O'Brien, S. M. and Dunson, D. B. (2004). "Bayesian Multivariate Logistic Regression." *Biometrics*, 60(3): 739–746. 361

Polson, N. G., Scott, J. G., and Windle, J. (2013). "Bayesian Inference for Logistic Models Using Pólya-Gamma Latent Variables." *Journal of the American Statistical Association*, 108(504): 1339–1349. 361, 367, 368, 379

Rousseeuw, P. J. and Christmann, A. (2003). "Robustness Against Separation and Outliers in Logistic Regression." *Computational Statistics and Data Analysis*, 42: 315–332. MR1996815. doi: https://doi.org/10.1016/S0167-9473(02)00304-3. 360

Sabanés Bové, D. and Held, L. (2011). "Hyper-g Priors for Generalized Linear Models." *Bayesian Analysis*, 6(3): 387–410. 366

Speckman, P. L., Lee, J., and Sun, D. (2009). "Existence of the MLE and Propriety of Posteriors for a General Multinomial Choice Model." *Statistica Sinica*, 19: 731–748. 361

Yang, R. and Berger, J. O. (1996). "A Catalog of Noninformative Priors." *Institute of Statistics and Decision Sciences, Duke University*. 360

Zellner, A. and Siow, A. (1980). "Posterior Odds Ratios for Selected Regression Hypotheses." In *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia (Spain)*, 585–603. Valencia, Spain: University of Valencia Press. 360, 366

Zorn, C. (2005). "A Solution to Separation in Binary Response Models." *Political Analysis*, 13(2): 157–170. 360

Acknowledgments

The authors thank the Editor-in-Chief, Editor, Associate Editor and the reviewer for suggestions that led to a greatly improved paper. The authors also thank Drs. David Banks, James Berger, William Bridges, Merlise Clyde, Jon Forster, Jayanta Ghosh, Aixin Tan, and Shouqiang Wang for helpful discussions. The research of Joyee Ghosh was partially supported by the NSA Young Investigator grant H98230-14-1-0126.