Electronic Journal of Statistics
Vol. 6 (2012) 2463–2485
ISSN: 1935-7524
DOI: 10.1214/12-EJS756

Convergence rates for MCMC algorithms for a robust Bayesian binary regression model

Vivekananda Roy

Department of Statistics
Iowa State University
e-mail: vroy@iastate.edu

Abstract: Most common regression models for analyzing binary random variables are logistic and probit regression models. However it is well known that the estimates of regression coefficients for these models are not robust to outliers [26]. The robit regression model [1, 16] is a robust alternative to the probit and logistic models. The robit model is obtained by replacing the normal (logistic) distribution underlying the probit (logistic) regression model with the Student's $t-$distribution. We consider a Bayesian analysis of binary data with the robit link function. We construct a data augmentation (DA) algorithm that can be used to explore the corresponding posterior distribution. Following [10] we further improve the DA algorithm by adding a simple extra step to each iteration. Though the two algorithms are basically equivalent in terms of computational complexity, the second algorithm is theoretically more efficient than the DA algorithm. Moreover, we analyze the convergence rates of these Markov chain Monte Carlo (MCMC) algorithms. We prove that, under certain conditions, both algorithms converge at a geometric rate. The geometric convergence rate has important theoretical and practical ramifications. Indeed, the geometric ergodicity guarantees that the ergodic averages used to approximate posterior expectations satisfy central limit theorems, which in turn allows for the construction of asymptotically valid standard errors. These standard errors can be used to choose an appropriate (Markov chain) Monte Carlo sample size and allow one to use the MCMC algorithms developed in this paper with the same level of confidence that one would have using classical (iid) Monte Carlo. The results are illustrated using a simple numerical example.

AMS 2000 subject classifications: Primary 60J27; secondary 62F15.
Keywords and phrases: Convergence rate, data augmentation algorithm, geometric ergodicity, Markov chain, robit regression, robust regression.

Received December 2011.

Contents

1. Introduction

The logistic and probit regression models are commonly used in practice to analyze binary data. However, the estimators of the regression coefficients for these popular models are not robust to outliers [26]. [16] proposed a robust alternative to the logistic and probit models which he called the robit regression model. In order to describe the robit model, suppose that $Y = (Y_1, Y_2, \ldots, Y_n)$ is a vector of n independent binary random variables such that $P(Y_i = 1) = F_\nu(x_i^T \beta)$, where $F_\nu(\cdot)$ is the cdf of the univariate Student's t distribution with known and fixed degrees of freedom ν, the x_i's, $i = 1, 2, \ldots, n$ are $p \times 1$ known vector of covariates associated with Y_i and β is the $p \times 1$ vector of unknown regression coefficients. The joint probability mass function (pmf) of Y is given by

$$p(y|\beta) = \prod_{i=1}^{n} \Big(F_\nu(x_i^T \beta)\Big)^{y_i} \Big(1 - F_\nu(x_i^T \beta)\Big)^{1-y_i}, \tag{1.1}$$

where $y = (y_1, y_2, \ldots, y_n)$. The above model, as an alternative to the logistic model, has been previously suggested by [24] and [1]. Both probit and logistic regression models can be well approximated by the robit model. In fact, a robit link with about seven degrees of freedom provides an excellent approximation to the logit link, and the probit link can be well approximated by a robit link with large degrees of freedom. Gelman and Hill [7, chap. 6] showed that in the presence of outliers, the robit model, unlike the logistic and probit models, can effectively downweight the discordant data points for a better model fitting. On the other hand, in the absence of any discrepancy in the data set, if the data actually come from say, a logistic model, then the estimated response curve obtained by fitting the robit model is close to the true logistic model.

We consider a Bayesian analysis of binary data with the pmf of Y as defined in (1.1) and a normal prior on β. The posterior density $\pi(\beta|y)$ is given by

$$\pi(\beta|y) = \frac{1}{m(y)} \prod_{i=1}^{n} \Big(F_\nu(x_i'\beta)\Big)^{y_i} \Big(1 - F_\nu(x_i'\beta)\Big)^{1-y_i} \times \phi_p(\beta; \beta_a, \Sigma_a^{-1}), \tag{1.2}$$

where $\phi_p(\beta; \beta_a, \Sigma_a^{-1})$ is the density of the p-dimensional normal distribution with mean β_a, dispersion matrix Σ_a^{-1}, evaluated at β, $m(y)$ is the normalizing constant, that is, $m(y) = \int_{\mathbb{R}^p} p(y|\beta)\phi_p(\beta; \beta_a, \Sigma_a^{-1})d\beta$. Proper choice of the dispersion matrix, Σ_a^{-1} is important. One choice is to take $\Sigma_a = cX^TX$ for some constant c, where X is the $n \times p$ design matrix [see e.g. 38]. The simplicity of this choice of Σ_a is that only one hyper-parameter c needs to be specified.

Inference based on (1.2) often reduces to calculation of (posterior) expectations

$$E_\pi h := \int_{\mathbb{R}^p} h(\beta)\pi(\beta|y)d\beta.$$

Unfortunately, $E_\pi h$ is a ratio of two intractable integrals which are not available in closed form. Moreover, classical Monte Carlo methods based on independent and identically distributed (iid) samples are problematic when p is large. In this case we may resort to MCMC methods. Here we construct a data augmentation (DA) algorithm that can be used to explore the posterior density $\pi(\beta|y)$. DA algorithms, like its deterministic counterpart the EM algorithms, often suffer from slow convergence [see e.g. 36]. In the case when the prior mean $\beta_a = 0$, following [10] we present an alternative MCMC algorithm, called the *sandwich algorithm*, that is computationally equivalent to the DA algorithm, but converges faster to the stationary distribution and is more efficient (See Section 3 for details).

Let $\{\beta_m\}_{m=0}^\infty$ be the Markov chain associated with either the DA or the sandwich algorithm. Let

$$L^1(\pi) = \left\{ h : \mathbb{R}^p \to \mathbb{R} \;\middle|\; \int_{\mathbb{R}^p} |h(\beta)|\, \pi(\beta|y)\, d\beta < \infty \right\}.$$

Similarly, let $L^2(\pi)$ denote the vector space of real-valued functions that are square integrable with respect to the target density $\pi(\beta|y)$. If $h \in L^1(\pi)$ and the Markov chain $\{\beta_m\}_{m=0}^\infty$ is Harris ergodic, then $\bar{h}_m := (1/m)\sum_{j=0}^{m-1} h(\beta_j)$ is a consistent estimator of $E_\pi h$ since by ergodic theorem $\bar{h}_m \to E_\pi h$ with probability 1 as $m \to \infty$. However, in practice we need to choose the sample size m. The sample size m must be large enough to ensure that the Monte Carlo error $\bar{h}_m - E_\pi h$ is sufficiently small and this is where an approximate distribution of $\bar{h}_m - E_\pi h$ can be used. In fact if there is a central limit theorem (CLT) for h, that is,

$$\sqrt{m}(\bar{h}_m - E_\pi h) \xrightarrow{\mathrm{d}} N(0, \sigma_h^2), \quad \text{as } m \to \infty, \tag{1.3}$$

and if $\hat{\sigma}_h^2$ is a consistent estimator of the asymptotic variance σ_h^2, then an asymptotic 95% confidence interval (CI) for $E_\pi h$ based on m iterations of the chains can be constructed as $\bar{h}_m \pm 2\hat{\sigma}_h/\sqrt{m}$. If we are satisfied with the width of the CI, we stop, otherwise, we increase the sample size m to achieve the desired level of precision. Unfortunately, unlike in the case of classical Monte Carlo methods based on iid draws from the target distribution, the second moment condition ($h \in L^2(\pi)$) is no longer enough to guarantee a Markov chain CLT (1.3). The most common method of establishing the CLT in (1.3) is to show that the Markov chain $\{\beta_m\}_{m=0}^\infty$ is *geometrically ergodic* (see Section 2 for the definition), and this requires a deep theoretical convergence analysis of the chain. Moreover, due to the serial correlation in the Markov chain, the asymptotic variance σ_h^2 has a complicated form, and consistent estimation of σ_h^2 is a challenging problem. On the other hand, if $\{\beta_m\}_{m=0}^\infty$ is geometrically ergodic then results in [9, 12, 2], and [6] show that specialized techniques such as the regenerative

simulation and the batch means can be used to construct consistent estimators of σ_h^2. (For more on standard errors of MCMC based estimates, see [13] and [5].) The main result in this paper is a proof that, under certain easily verifiable conditions on n, ν, c, and the design matrix X, the Markov chains underlying the DA and sandwich algorithms presented here converge at a geometric rate. As described above, our convergence rate results allow one to use the MCMC algorithms developed in this paper with the same level of confidence that one would have using classical (iid) Monte Carlo.

The remainder of this paper is organized as follows. Section 2, contains a description of the DA algorithm as well as the statement of our results regarding convergence rates of the DA algorithm. In Section 3, we show how the group action recipe of [10] can be used to improve the DA algorithm. Our results are illustrated using a simple example in Section 4. A short discussion appears in Section 5. Technical results and proofs are relegated to the Appendices.

2. The geometric ergodicity of the Data Augmentation algorithm

2.1. The Data Augmentation algorithm

We mentioned in the introduction that sampling directly from (1.2) is rather difficult. But, as we explain now, it is easy to construct a data augmentation algorithm for (1.2) by introducing two sets of new (latent) random variables. In particular, let $z_1, \ldots, z_n$ be n independent variables with $z_i \sim t_\nu(x_i^T\beta, 1)$ where $t_\nu(\mu, 1)$ denotes the univariate Student's t distribution with location μ, scale 1 and degrees of freedom ν. If we define $Y_i = I_{\mathbb{R}_+}(z_i)$, then $Y_1, \ldots, Y_n$ are n independent Bernoulli random variables with $P(Y_i = 1) = P(z_i > 0) = F_\nu(x_i^T\beta)$. Thus $z_1, \ldots, z_n$ can be thought of as latent variables underlying the binary data y. Now we use the fact that t distribution can be expressed as a scale mixture of normal distributions, that is, if $z_i|\lambda_i \sim N(\mu, 1/\lambda_i)$ and $\lambda_i \sim$ Gamma $(\nu/2, \nu/2)$, then the marginal distribution of z_i is $t_\nu(\mu, 1)$. The joint posterior density of (β, λ, z) given y is

$$\begin{aligned}\pi(\beta, (\lambda, z)|y) &= \frac{1}{m(y)}\left[\prod_{i=1}^{n}\left\{I_{\mathbb{R}_+}(z_i)I_{\{1\}}(y_i) + I_{\mathbb{R}_-}(z_i)I_{\{0\}}(y_i)\right\}\phi\Big(z_i; x_i^T\beta, \frac{1}{\lambda_i}\Big)q\Big(\lambda_i, \frac{\nu}{2}, \frac{\nu}{2}\Big)\right] \\ &\quad \times \phi_p\Big(\beta; \beta_a, \Sigma_a^{-1}\Big); \qquad \lambda_i \in \mathbb{R}_+, z_i \in \mathbb{R}, \beta \in \mathbb{R}^p, \end{aligned} \tag{2.1}$$

where $\mathbb{R}_+ = (0, \infty)$, $\mathbb{R}_- = (-\infty, 0]$, $I_A(\cdot)$ is the indicator function of the set A, $z = (z_1, \ldots, z_n)$, $\lambda = (\lambda_1, \ldots, \lambda_n)$, $\phi(x; a, b)$ is the density of the univariate normal distribution with mean a, variance b, evaluated at the point x, that is $\phi \equiv \phi_1$, and $q(\omega; a, b)$ is the gamma density with shape parameter a, scale parameter b, evaluated at ω (i.e., $q(\omega; a, b) = b^a\omega^{a-1}e^{-b\omega}/\Gamma(a)$).

The β marginal density of (2.1) is $\pi(\beta|y)$, that is,

$$\int_{\mathbb{R}^n}\int_{\mathbb{R}_+^n} \pi(\beta, (\lambda, z)|y)d\lambda dz = \pi(\beta|y).$$

Thus we can use $\pi(\beta,(\lambda,z)|y)$ to construct a DA algorithm with stationary density $\pi(\beta|y)$ if we can sample from the two conditional densities $\pi(\lambda,z|\beta,y)$ and $\pi(\beta|z,\lambda,y)$ ([35, 8]). In fact, straightforward calculations show that $\beta|\lambda,z,y\sim N_p\big(\hat{\beta},(X^T\Lambda X+\Sigma_a)^{-1}\big)$, where

$$\hat{\beta}=(X^T\Lambda X+\Sigma_a)^{-1}(X^T\Lambda z+\Sigma_a\beta_a),$$

and Λ is the $n\times n$ diagonal matrix with diagonal elements $\lambda_1,\ldots,\lambda_n$. We can draw from $\pi(\lambda,z|\beta,y)$ by first drawing from $\pi(z|\beta,y)$ and then from $\pi(\lambda|z,\beta,y)$. It can be shown that conditional on (β,y), $z_1,\ldots,z_n$ are independent with $z_i|\beta,y\sim Tt_\nu(x_i^T\beta,y_i)$, where $Tt_\nu(x_i^T\beta,y_i)$ denote the truncated t distribution with mean $x_i^T\beta$, variance 1 and degrees of freedom ν that is truncated left at 0 if $y_i=1$ and truncated right at 0 if $y_i=0$. Sampling from the truncated t distribution can be done by the inversion method. Lastly, conditional on (z,β,y), λ_i's are independent with $\lambda_i|z,\beta,y\sim$ Gamma $((\nu+1)/2,(\nu+(z_i-x_i^T\beta)^2)/2)$ for $i=1,\ldots,n$. A single iteration of the DA algorithm uses the current state β to produce the new state β' through the following two steps:

1. Draw $\{(\lambda_i,z_i),i=1,2,\ldots,n\}$ by first drawing $z_i\sim Tt_\nu(x_i^T\beta,y_i)$ and then draw $\lambda_i\sim$ Gamma $\big(\frac{\nu+1}{2},\frac{\nu+(z_i-x_i^T\beta)^2}{2}\big)$.
2. Then draw $\beta'\sim N_p\big(\hat{\beta},(X^T\Lambda X+\Sigma_a)^{-1}\big)$.

The Markov transition density of the DA algorithm is given by

$$k(\beta'|\beta)=\int_{\mathbb{R}_+^n}\int_{\mathbb{R}^n}\pi(\beta'|\lambda,z,y)\pi(\lambda,z|\beta y)\,dz\,d\lambda\ .$$

Note that, while the Markov transition density does depend on the data, y, and the design matrix X, these quantities are fixed, so this dependence is suppressed in the notation. The basic theory of DA algorithms implies that $k(\beta'|\beta)$ is reversible with respect to the posterior density $\pi(\beta|y)$; i.e., we have

$$k(\beta'|\beta)\pi(\beta|y)=k(\beta|\beta')\pi(\beta'|y),$$

for all $\beta,\beta'\in\mathbb{R}^p$. It follows immediately that the posterior density is invariant for the chain; i.e.,

$$\int_{\mathbb{R}^p}k(\beta'|\beta)\pi(\beta|y)\,d\beta'=\pi(\beta'|y),$$

for all $\beta\in\mathbb{R}^p$. Let $\mathcal{Z}$ denote the subset of $\mathbb{R}^n$ where z lives, that is, $\mathcal{Z}$ is the Cartesian product of n positive and negative half lines ($\mathbb{R}_+$ and $\mathbb{R}_-$), where the ith component is either $\mathbb{R}_+$ (if $y_i=1$) or $\mathbb{R}_-$ (if $y_i=0$). Note that, the joint posterior density $\pi(\beta,(\lambda,z)|y)$ is strictly positive on $\mathbb{R}^p\times(\mathbb{R}_+^n\times\mathcal{Z})$. So the Markov chain $\{\beta_m\}_{m=0}^\infty$ driven by $k(\beta'|\beta)$ is Harris ergodic; that is, irreducible, aperiodic and Harris recurrent (Tan and Hobert [34, Lemma 1], Hobert [8]). See [22] for definitions of irreducibility, aperiodicity and Harris recurrence. Harris ergodicity implies that, no matter how the chain is started, the chain converges to its stationary distribution and that ergodic average based on the DA algorithm, $\bar{h}_m$ converges almost surely to its population counterpart, which is, of course, the (posterior) expectation $E_\pi h$.

2.2. Geometric convergence of the DA algorithm

We begin with defining what it means for the DA algorithm to converge at a geometric rate. Let $K(\cdot,\cdot)$ denote the Markov transition function (Mtf) of the DA Markov chain; that is, $K(\beta, A) = \int_A k(\beta'|\beta)d\beta'$ for $\beta \in \mathbb{R}^p$ and a measurable set $A \subset \mathbb{R}^p$. Also for $m = 2, 3, \ldots$ the m-step Markov transition function of $\{\beta_m\}_{m=0}^{\infty}$ is defined inductively as

$$K^m\big(\beta, A\big) = \Pr\big(\beta_m \in A|\beta_0 = \beta\big) = \int_{\mathbb{R}^p} K^{m-1}\big(\beta', A\big)K\big(\beta, d\beta'\big) \ ,$$

where $K^1 \equiv K$. Let $\Pi(\cdot|y)$ denote the probability measure corresponding to the posterior density $\pi(\beta|y)$. Harris ergodicity implies that the total variation distance between the probability measures $K^m\big(\beta,\cdot\big)$ and $\Pi(\cdot|y)$ decreases to zero as m gets large; that is, for any starting value, $\beta \in \mathbb{R}^p$, we have

$$\big\|K^m\big(\beta,\cdot\big) - \Pi(\cdot|y)\big\|_{\mathrm{TV}} \downarrow 0 \quad \text{as } m \to \infty \ .$$

Note that the above expression gives no information about the *rate* at which the total variation distance converges to zero. The Markov chain is called *geometrically ergodic* if there exists a function $M : \mathbb{R}^p \to [0, \infty)$ and a constant $r \in (0, 1)$ such that, for all m,

$$\big\|K^m\big(\beta,\cdot\big) - \Pi(\cdot|y)\big\|_{\mathrm{TV}} \leq M(\beta)r^m \ .$$

It is known that if a reversible Markov chain is geometrically ergodic, then there is a CLT (1.3) for every function that is square integrable with respect to the stationary distribution [29]. Unfortunately, Harris ergodicity of a Markov chain, which is generally easy to verify, does not imply geometric ergodicity. We will establish geometric ergodicity of the DA algorithm by establishing the so-called *drift condition*, which we now describe. (See Jones and Hobert [13] for a gentle introduction to these ideas.)

A function $V : \mathbb{R}^p \to \mathbb{R}_+$ is said to be *unbounded off compact sets* if, for each $\alpha > 0$, the level set $\{\beta : V(\beta) \leq \alpha\}$ is compact. We say that a geometric drift condition holds if there exists a $V : \mathbb{R}^p \to \mathbb{R}_+$ that is unbounded off compact sets and constants $\rho \in [0, 1)$, and $L \in \mathbb{R}$ such that, for all $\beta \in \mathbb{R}^p$,

$$(KV)(\beta) \leq \rho V(\beta) + L$$

where

$$(KV)(\beta) = \int_{\mathbb{R}^p} V(\beta')\, k(\beta'|\beta)\, d\beta' \ .$$

The function V is called the *drift function*. Since $\pi(\beta, (\lambda, z)|y) > 0$ for all $(\beta, \lambda, z) \in \mathbb{R}^p \times \mathbb{R}_+^n \times \mathcal{Z}$, a geometric drift condition implies that the DA algorithm is geometrically ergodic (Meyn and Tweedie [22, chap. 15.], Hobert [8]).

Let W be an $n \times p$ matrix whose ith row is w_i^T where $w_i = x_i I_{\{0\}}(y_i) - x_i I_{\{1\}}(y_i)$. We now define two conditions on y and X which are used to prove the geometric ergodicity of the DA algorithm.

A1 The design matrix X has full rank, and
A2 there exists a vector $a = (a_1, ..., a_n)^T$ with $a_i > 0$ for all $i = 1, 2, \ldots, n$ such that $W^T a = 0$.

[32] provide a simple way to check the condition A2 that can be easily implemented in most statistical software languages. By establishing a drift condition for the DA algorithm we prove the following theorem in Appendix B.

Theorem 1. *Assume that A1 and A2 hold. The DA algorithm is geometrically ergodic if* $\Sigma_a = cX^TX$, $\nu > 2$ *and*

$$n < \frac{c\nu}{(\nu+1)(1+2\sqrt{\beta_a^T X^T X \beta_a})}. \tag{2.2}$$

Ideally, we would like to be able to say that the DA algorithm is geometrically ergodic for any n, ν, y, X, and Σ_a. As mentioned in the Introduction, the assumption $\Sigma_a = cX^TX$ is made in the literature and this simple choice requires only one hyperparameter c to be specified. Several inequalities used in the proof of Theorem 1 heavily depends on this assumption. Similarly the conditions A1 and A2 are crucial in our proof. The condition $\nu > 2$ guarantees the existence of finite second moment of the latent t random variables. In our opinion, the most restrictive of all conditions is (2.2), which is a direct consequence of the several inequalities we have used in the proof. Note that (2.2) implies that $n < c$. Thus when n is large, the elements of the prior covariance matrix becomes small. We prove Theorem 1 by establishing a drift condition using the drift function $V(\beta) = \beta^T X^T X \beta$. We believe that a substantial reduction of the conditions in Theorem 1 would require the use of a different drift function $V(\beta)$, which means starting over from square one.

In practice often it is assumed that $\beta_a = 0$. For the rest of this article, we assume that $\beta_a = 0$ and in this case the posterior density of β becomes

$$\tilde{\pi}(\beta|y) = \frac{1}{m(y)} \prod_{i=1}^{n} \Big(F_\nu(x_i^T\beta)\Big)^{y_i} \Big(1 - F_\nu(x_i^T\beta)\Big)^{1-y_i} \times \phi_p(\beta; 0, \Sigma_a^{-1}). \tag{2.3}$$

We can derive the corresponding complete posterior density $\tilde{\pi}(\beta, (\lambda, z)|y)$ just by replacing $\phi_p(\beta; \beta_a, \Sigma_a^{-1})$ with $\phi_p(\beta; 0, \Sigma_a^{-1})$ in (2.1), and use it to construct a data augmentation algorithm for $\tilde{\pi}(\beta|y)$. Obviously, a single iteration of this DA algorithm uses the current state β to produce the new state β' through the following two steps:

1. Draw $\{(\lambda_i, z_i), i = 1, 2, \ldots, n\}$ by first drawing $z_i \sim Tt_\nu(x_i^T\beta, y_i)$ and then draw $\lambda_i \sim \text{Gamma}\left(\frac{\nu+1}{2}, \frac{\nu+(z_i - x_i^T\beta)^2}{2}\right)$.
2. Then draw $\beta' \sim N_p\big((X^T\Lambda X + \Sigma_a)^{-1}X^T\Lambda z, (X^T\Lambda X + \Sigma_a)^{-1}\big)$.

Corollary 1. *Let* $\{\tilde{\beta}_m\}_{m=0}^{\infty}$ *be the Markov chain underlying the above DA algorithm for* $\tilde{\pi}(\beta|y)$. *From Theorem 1 it follows that under conditions A1 and A2,* $\{\tilde{\beta}_m\}_{m=0}^{\infty}$ *is geometrically ergodic if* $\Sigma_a = cX^TX$, $\nu > 2$ *and* $n < c\nu/(\nu+1)$.

As explained in [36], the standard DA algorithms often suffer from slow convergence. In the next section we present algorithms that have faster convergence than $\{\tilde{\beta}_m\}_{m=0}^{\infty}$.

3. Sandwich algorithms

Over the last decade, several authors have shown that it is possible to drastically improve the convergence behavior of DA algorithms by adding a simple, computationally inexpensive "extra step" to each iteration of the DA algorithms (see, e.g., [21, 19, 36, 10]). In this section, following [10], we construct improved DA algorithms for $\tilde{\pi}(\beta|y)$. Let $\tilde{\pi}(\lambda, z|y) = \int_{\mathbb{R}^p} \tilde{\pi}(\beta, (\lambda, z)|y)d\beta$, that is $\tilde{\pi}(\lambda, z|y)$ is the (λ, z) marginal density of the complete posterior density $\tilde{\pi}(\beta, (\lambda, z)|y)$. Straightforward calculations show that

$$\begin{aligned}\tilde{\pi}(\lambda, z|y) \quad &\propto \quad \frac{e^{-\frac{1}{2}[z^T\Lambda z - z^T\Lambda X(X^T\Lambda X+\Sigma_a)^{-1}X^T\Lambda z]}}{|X^T\Lambda X + \Sigma_a|^{\frac{1}{2}}}|\Lambda|^{\frac{\nu+1}{2}-1}\, e^{-\frac{\nu}{2}\Sigma\lambda_i} \\ &\times \quad \prod_{i=1}^{n}\Big[I_{\mathbb{R}_+}(z_i)I_{\{1\}}(y_i) + I_{\mathbb{R}_-}(z_i)I_{\{0\}}(y_i)\Big]\ .\end{aligned}$$

Suppose that $R((\lambda, z), (d\lambda', dz'))$ is a Markov transition function on $\mathbb{R}_+^n \times \mathcal{Z}$ that is reversible with respect to $\tilde{\pi}(\lambda, z \,|\, y)$. Consider adding an extra step to the DA algorithm where, after (λ, z) is drawn in the first step, we move to a new value, $(\lambda', z') \sim R((\lambda, z), \cdot)$, before drawing new value of β. To be more specific, let $\{\breve{\beta}_m\}_{m=0}^{\infty}$ be a new Markov chain that proceeds from the current state β to the next state β' via the following three steps

1. Draw $\{(\lambda_i, z_i), i = 1, 2, \ldots, n\}$ by first drawing $z_i \sim Tt_\nu(x_i^T\beta, y_i)$ and then draw $\lambda_i \sim$ Gamma $\big(\frac{\nu+1}{2}, \frac{\nu+(z_i - x_i^T\beta)^2}{2}\big)$.
2. Draw $(\lambda', z') \sim R((\lambda, z), \cdot)$
3. Then draw $\beta' \sim N_p\big((\Sigma_a + X^T\Lambda' X)^{-1}X^T\Lambda' z', (X^T\Lambda' X + \Sigma_a)^{-1}\big)$,

where Λ' is the $n \times n$ diagonal matrix with elements of λ' on the diagonal. Note that the first and third steps are the same as the DA algorithm. [37] call the above algorithm the "sandwich algorithm" since the draw from $R((\lambda, z), \cdot)$ is sandwiched between the two steps of the DA algorithm. We will provide a specific R later in this section.

A routine calculation shows that the reversibility of R with respect to $\tilde{\pi}(\lambda, z|y)$ implies that the sandwich chain, $\{\breve{\beta}_m\}_{m=0}^{\infty}$, is reversible with respect to the target (posterior) density, $\tilde{\pi}(\beta|y)$. The sandwich algorithm is known to converge faster than the DA algorithm. In order to make precise comparisons of the DA and sandwich algorithms, we need to introduce some notations. Let

$$L_0^2(\tilde{\pi}) = \{h \in L^2(\tilde{\pi}) : \int_{\mathbb{R}^p} h(\beta)\,\tilde{\pi}(\beta|y)\,d\beta = 0\},$$

that is, $L_0^2(\tilde{\pi})$ is the subspace of $L^2(\tilde{\pi})$ of mean zero functions. Then $L_0^2(\tilde{\pi})$ is a Hilbert space with inner product defined as

$$\langle h_1, h_2 \rangle = \int_{\mathbb{R}^p} h_1(\beta)\, h_2(\beta)\, \tilde{\pi}(\beta|y)\, d\beta \ ,$$

and the norm is $\|h\| = \sqrt{\langle h, h \rangle}$. Let $\tilde{K}, \breve{K} : L_0^2(\tilde{\pi}) \to L_0^2(\tilde{\pi})$ denote the usual Markov operators defined by the DA chain $\{\tilde{\beta}_m\}_{m=0}^{\infty}$ and the sandwich chain $\{\breve{\beta}_m\}_{m=0}^{\infty}$, respectively. In particular, $\tilde{K}$ maps $h \in L_0^2(\tilde{\pi})$ to

$$(\tilde{K}h)(\beta) := \int_{\mathbb{R}^p} h(\beta')\, \tilde{k}(\beta'|\beta)\, d\beta' \ ,$$

where $\tilde{k}(\beta'|\beta)$ is the Markov transition density of the DA chain $\{\tilde{\beta}_m\}_{m=0}^{\infty}$ defined in Section 2.2. The Markov operator $\breve{K}$ is similarly defined. The reason that we define $\tilde{K}$ as an operator on $L_0^2(\tilde{\pi})$ (instead of $L^2(\tilde{\pi})$) is to eliminate the eigenvalue 1 associated with constant eigenfunction from its spectrum [See e.g. 23]. Let $\|\tilde{K}\|$ and $\|\breve{K}\|$ denote the (operator) norms of $\tilde{K}$ and $\breve{K}$, for example,

$$\|\tilde{K}\| = \sup_{h \in L_0^2(\tilde{\pi}), \|h\|=1} \|\tilde{K}h\| \ .$$

In general, the closer the norm of a Markov operator is to 0, the faster the corresponding Markov chain converges [see, e.g., 18]. There are close connections between convergence properties discussed in the previous section and the norm of a Markov operator. Indeed, a reversible Markov chain is geometrically ergodic if and only if the norm of the corresponding Markov operator is strictly less than 1 [29, 30]. From Roy [31] we know that, $\|\breve{K}\| \leq \|\tilde{K}\|$, and hence the sandwich chain converges at least as fast as the DA chain.

Another criterion that is used to compare Markov chains (with the same stationary distribution) is *efficiency ordering*. Let $h \in L^2(\tilde{\pi})$ and we want to estimate $E_{\tilde{\pi}}h$. Define $\tilde{\sigma}_h^2$ to be the asymptotic variance in the CLT for the ergodic averages $\overline{h}_m$ based on the DA algorithm if such a CLT exists, and ∞ otherwise. Similarly define $\breve{\sigma}_h^2$ for the sandwich algorithm. Sandwich chain is said to be at least as efficient as the DA chain if $\breve{\sigma}_h^2 \leq \tilde{\sigma}_h^2$ for all $h \in L^2(\tilde{\pi})$. In this case after running the two chains for equal number of iterations, we may expect that the sandwich algorithm results in a shorter CI for $E_{\tilde{\pi}}h$ than the DA algorithm. This is why if the two algorithms are similar in terms of simulation effort, we prefer the sandwich algorithm over the DA. In fact, the results in Hobert and Marchev [10] can be used to show that $\tilde{K} - \breve{K}$ is a positive operator, that is, $\langle (\tilde{K} - \breve{K})h, h \rangle \geq 0$ for all $h \in L_0^2(\tilde{\pi})$, and this implies that the sandwich chain is at least as efficient as the DA chain [23].

Recall that the only difference between a single iteration of the DA algorithm with that of the sandwich algorithm is that the sandwich algorithm has an extra step according to the Mtf R. Following [19] and [17], [10] gave a recipe of constructing R that involves group actions and (left) Haar measure. In order to use Hobert and Marchev's [2008] group action recipe, let G be the multiplicative group $\mathbb{R}_+$ where group composition is defined as multiplication. The

multiplicative group $\mathbb{R}_+$ is unimodular with Haar measure $\varrho(dg) = dg/g$, where dg denotes the Lebesgue measure on $\mathbb{R}_+$. We now define a (left) group action of G on $\mathbb{R}_+^n \times \mathcal{Z}$ (the support of the density $\tilde{\pi}(\lambda, z|y)$) as $g(\lambda, z) = (\lambda, gz)$ where $gz = (gz_1, \cdots, gz_n)$. (It is easy to verify that the above is a valid group action.) If $g \in G$ and $h : \mathbb{R}_+^n \times \mathcal{Z} \to \mathbb{R}$ is an integrable function (with respect to Lebesgue measure), straightforward calculations show that

$$\int_{\mathbb{R}_+^n} \int_{\mathcal{Z}} h(\lambda, z) dz d\lambda = \chi(g) \int_{\mathbb{R}^n} \int_{\mathcal{Z}} h(g(\lambda, z)) dz d\lambda, \tag{3.1}$$

where $\chi(g) = g^n$. Also it is easy to see that $\chi(g^{-1}) = 1/\chi(g)$ and $\chi(g_1, g_2) = \chi(g_1)\chi(g_2)$. Here $\chi(g)$ plays the role of the function j defined in Hobert and Marchev [10, page 543]. Consider a distribution on G with density function

$$\frac{\pi(\lambda, gz \,|\, y)\, \chi(g) \varrho(dg)}{\int_G \pi(\lambda, gz \,|\, y)\, \chi(g)\, \varrho(dg)} = (z^T \Lambda^{1/2} (I - Q) \Lambda^{1/2} z)^{\frac{n}{2}} g^{n-1} \frac{e^{-\frac{g^2}{2} z^T \Lambda^{1/2} (I-Q) \Lambda^{1/2} z}}{2^{(n-2)/2} \Gamma(n/2)} dg,$$

where $Q = \Lambda^{1/2} X (X^T \Lambda X + \Sigma_a)^{-1} X^T \Lambda^{1/2}$, or, equivalently

$$g^2 \sim \text{Gamma}\left(\frac{n}{2}, \frac{z^T \Lambda^{1/2} (I - Q) \Lambda^{1/2} z}{2}\right). \tag{3.2}$$

Note that $d(\lambda, z) := \int_G \pi(\lambda, gz \,|\, y)\, \chi(g)\, \varrho(dg)$ is positive for all $(\lambda, z) \in \mathbb{R}_+^n \times \mathcal{Z}$ and finite for almost all $(\lambda, z) \in \mathbb{R}_+^n \times \mathcal{Z}$.

We now provide an R for Step 2 of the sandwich algorithm. Given (λ, z), we make the transition $(\lambda, z) \to (\lambda', z')$ by drawing g^2 from (3.2) and setting $\lambda' = \lambda$ (that is, λ is left unaltered) and $z' = gz$. In other words, the corresponding Markov operator R on $L_0^2(\tilde{\pi}(\lambda, z \,|\, y))$ is defined as

$$(Rh)(\lambda, z) = \int_G \frac{h(\lambda, gz) \pi(\lambda, gz \,|\, y) \chi(g)}{d(\lambda, z)} \varrho(dg). \tag{3.3}$$

From Hobert and Marchev's [2008] Proposition 3, it follows that R is reversible with respect to $\tilde{\pi}(\lambda, z|y)$. Of course, the Markov chain driven by R is reducible, which is a common feature of efficient sandwich algorithms [14]. Note that reducibility of R does not stop the sandwich algorithm $\{\breve{\beta}_m\}_{m=0}^{\infty}$ from being Harris ergodic. From our discussion before, we know that the above sandwich algorithm is more efficient and converges faster than the DA algorithm. The extra step R, which is the sole difference between the DA and sandwich algorithms, is just a single draw from a univariate gamma density and since the computation of the parameters of this gamma density does not involve any *extra* (compared to the DA algorithm) computationally demanding calculation like matrix inversion, the two algorithms are essentially equivalent in terms of computer time per iteration. Following the proof of Corollary 1 in [32], we can show that the sandwich algorithm inherits the geometric ergodicity of the DA algorithm.

Corollary 2. *Under conditions A1 and A2, the Markov chain underlying the sandwich algorithm is geometrically ergodic if* $\Sigma_a = cX^T X$, $\nu > 2$ *and* $n < c\nu/(\nu + 1)$.

Note that, unlike most examples of sandwich algorithm available in literature, in the above we do not let the group G act on $\mathbb{R}_+^n \times \mathcal{Z}$ through component-wise multiplication; i.e., we do *not* define $g(\lambda, z) = (g\lambda, gz)$. A similar group action is used in [20] to improve a DA algorithm for a Bayesian logistic model. Now we discuss what happens when G is allowed to act on the entire augmented space $\mathbb{R}_+^n \times \mathcal{Z}$ through component-wise multiplication, that is $g(\lambda, z) = (g\lambda, gz)$. Note that, under the above group action, (3.1) holds if we replace $\chi(g)$ with g^{2n}. To construct the sandwich algorithm in this case, we must first demonstrate that there is a probability density (with respect to the Haar measure $\varrho(dg)$) that is proportional to $\pi(g\lambda, gz|y)g^{2n}$. In other words, we must show that $d_1(\lambda, z) := \int_G \pi(g\lambda, gz|y)g^{2n}\varrho(dg) < \infty$ for all $(\lambda, z) \in \mathbb{R}_+^n \times \mathcal{Z}$. It can be shown that, as a function of g,

$$\pi(g\lambda, gz|y)g^{2n} = ag^{n(\frac{3+\nu}{2})}e^{-\frac{g\nu\sum\lambda_i}{2}}e^{-\frac{1}{2}[g^3 z^T\Lambda z - g^4 z^T\Lambda X(gX^T\Lambda X+\Sigma_a)^{-1}X^T\Lambda z]}|gX^T\Lambda X + \Sigma_a|^{-\frac{1}{2}},$$

where the constant a does not depend on g. Since Σ_a is positive semidefinite, it follows that [28, page 70] $(gX^T\Lambda X)^{-1} - (gX^T\Lambda X + \Sigma_a)^{-1}$ is positive semidefinite. Thus,

$$z^T\Lambda X(gX^T\Lambda X + \Sigma_a)^{-1}X^T\Lambda z \le z^T\Lambda X(gX^T\Lambda X)^{-1}X^T\Lambda z,$$

and hence

$$\begin{aligned} &-\frac{1}{2}[g^3 z^T\Lambda z - g^4 z^T\Lambda X(gX^T\Lambda X + \Sigma_a)^{-1}X^T\Lambda z] \\ &\quad\le -\frac{g^3}{2}z^T\Lambda^{1/2}[I - \Lambda^{1/2}X(X^T\Lambda X)^{-1}X^T\Lambda^{1/2}]\Lambda^{1/2}z \le 0, \end{aligned}$$

where the last inequality follows since $I - \Lambda^{1/2}X(X^T\Lambda X)^{-1}X^T\Lambda^{1/2}$ is an idempotent matrix. Since $|gX^T\Lambda X + \Sigma_a| \ge |\Sigma_a|$ [28, page 70] it follows that

$$\pi(gz, g\lambda|y)g^{2n} \le a|\Sigma_a|^{-\frac{1}{2}}g^{n(\frac{3+\nu}{2})}e^{-\frac{g\nu\sum\lambda_i}{2}},$$

so

$$d_1(\lambda, z) \le a|\Sigma_a|^{-\frac{1}{2}}\int_G g^{n(\frac{3+\nu}{2})-1}e^{-\frac{\nu\sum\lambda_i}{2}g}dg < \infty.$$

Now in order to construct an effective sandwich algorithm, we need a fast and efficient way of sampling from the following density on $\mathbb{R}_+$ (with respect to Lebesgue measure on $\mathbb{R}_+$)

$$t_{\lambda,z}(g) := \frac{g^{n(\frac{3+\nu}{2})-1}e^{-\frac{g\nu\sum\lambda_i}{2}}e^{-\frac{1}{2}[g^3 z^T\Lambda z - g^4 z^T\Lambda X(gX^T\Lambda X+\Sigma_a)^{-1}X^T\Lambda z]}}{d_1(\lambda, z)|gX^T\Lambda X + \Sigma_a|^{\frac{1}{2}}}.$$

Of course, in this case the Step 2 of the sandwich algorithm will be to draw $g \sim t_{\lambda,z}(g)$ and set $z' = gz$ as well as $\lambda' = g\lambda$. Note that, even though the above is a univariate density, it may be difficult to efficiently sample from $t_{\lambda,z}(g)$. Also in this case, since λ is updated in step 2, an extra (compared to the DA algorithm) matrix inversion is required in step 3 of the sandwich algorithm. This is why we prefer the sandwich algorithm with Step. 2 as defined in (3.2).

4. A numerical example

Suppose that $n = 7, \nu = 3, p = 2, \beta = (\beta_0, \beta_1)$. Let the data be $y = (0, 0, 0, 1, 1, 0, 1)$, and the design matrix $X = (\mathbf{1}, \mathbf{x})$, where $\mathbf{1}$ is a vector of 1's, and $\mathbf{x}^T = (0.010, 0.020, 0.030, 0.050, 0.060, 0.075, 0.100)$. In this section we illustrate our theoretical results with the above data set.

Consider the prior on β to be $\phi(\beta, 0, \Sigma_a^{-1})$, where $\Sigma_a = cX^TX$ for some c to be chosen later. We want to approximate the posterior expectation of β_1, that is, $E_{\tilde{\pi}}h$, where $h(\beta) = \beta_1$. Since $\text{rank}(X) = 2$, the condition A1 holds. As mentioned in [32], using the "simplex" function from the "boot" library in R [27] we check that the condition A2 is satisfied for the above data. From Corollary 1 we know that if $c > n(\nu+1)/\nu$ then the DA algorithm is geometrically ergodic. We take $c = n(\nu+1)/\nu + 0.005 \approx 9.338$. We ran the DA algorithms for 2,000 iterations starting at $\beta = (0, 0)$ and discard first 1,000 iteration as burn-in. The estimate of $E_{\tilde{\pi}}\beta_1$ based on the remaining 1,000 iterations is $\bar{h}_m$ =1.51. Since the posterior density $\tilde{\pi}(\beta|y)$ has finite moments of all orders, the results in [12] imply that consistent estimator of asymptotic variance in the CLT can be obtained by the method of batch means with batch size $b_m = \lfloor m^{0.5} \rfloor$. We use the "mcmcse" package in R [4] to compute the standard errors by batch means with the default batch size $b_m = \lfloor m^{0.5} \rfloor$. Note that we could also use the regenerative simulation method to calculate the standard errors by constructing a *minorization condition* as in [32]. The standard error for $\bar{h}_m$ based on 1,000 iterations is 0.144. (The mean and standard deviation for this standard error estimate based on 1,000 independent repetitions of the above simulation are 0.133 and 0.017 respectively.) Next, we find out how large m needs to be for the half-width of the 95% interval to be below 0.10. Based on the standard error calculated above, we need to run the DA chain for $4 \times 1000 \times (.144)^2/(0.1)^2 \approx 8,294$ iterations to achieve this level of accuracy. In fact after 9,000 iterations of the chain, the 95% CI for $E_{\tilde{\pi}}\beta_1$ is given by $1.540 \pm (2 \times 0.049)$. (The mean and standard deviation for the corresponding standard error estimate based on 1,000 independent repetitions are 0.044 and 0.003 respectively.)

Next, we want to compare the performance of the DA and the sandwich algorithm in the context of this example. From Corollary 2 we know that the sandwich algorithm is also geometrically ergodic. The standard error estimate corresponding to the sandwich chain based on 1,000 iterations is 0.124, which is not significantly less than the estimate (0.144) corresponding to the DA chain. In fact, based on 1,000 independent repetitions of the sandwich chain simulation the mean and standard deviation for the standard error estimate are 0.132 and 0.017 respectively. The reason there is not much increase in efficiency by running the sandwich chain instead of the DA chain is that here we are using a very informative prior. In fact, the prior variances of β_0 and β_1 are 0.057 and 17.493 respectively. On the other hand, when we use a small value for c, and hence the prior is more spread-out, we observe huge gains in efficiency by using the sandwich algorithm instead of the DA algorithm. For example, if we use $c = 0.005$, the prior variances of β_0 and β_1 become 107.935 and 32672.112. In this case after 20,000 iterations of the chains, the estimates of

the standard errors corresponding to the DA and sandwich algorithms are 2.684 and 1.169 respectively. (Note that, Corollary 1 or Corollary 2 are not applicable here anymore as $c \not> n(\nu+1)/\nu$; we are simply assuming that the CLT still holds in this case.) These estimates suggest that, even in this simple example, the DA algorithm requires about $2.684^2/1.169^2 \approx 5.3$ times as many iterations as the sandwich algorithm to achieve the same level of precision. We repeated the simulation 1,000 times independently and the above ratio estimates ranged between 2.34 and 24.41.

5. Discussion

We present two MCMC algorithms for exploring the posterior density associated with a Bayesian robit model with a normal prior on the regression coefficients. The first one is a DA algorithm which is obtained using the fact that t distribution can be expressed as a scale mixture of normal distributions. We then use the group action recipe given in [10] to construct an efficient sandwich algorithm. Unlike most of the sandwich algorithms available in the literature, we do not consider a group action on the entire augmented space defined through componentwise multiplication. The sandwich algorithm converges faster than the DA algorithm in the Markov operator norm sense. Also, the sandwich algorithm is more efficient than the DA algorithm in the sense that the asymptotic variance in the CLT under the sandwich algorithm is no larger than that under the DA algorithm. Since the only difference between a single iteration of the sandwich algorithm and that of the DA algorithm is a univariate draw from a gamma distribution, the two algorithms are essentially equivalent in terms of computational effort. Thus, we prefer the sandwich algorithm to the DA algorithm.

We prove that, under certain conditions, both DA and sandwich algorithms converge at a geometric rate. These convergence rate results are important from a practical standpoint because geometric ergodicity guarantees the existence of central limit theorems which are essential for the calculation of valid asymptotic standard errors for MCMC based estimates. Our results are illustrated through a numerical example.

In this paper we have considered a multivariate normal prior distribution for the regression coefficients β. As a possible avenue for future work, it would be interesting to see if the methods presented here can be used to establish geometric convergence of MCMC algorithms for Bayesian robit models with other priors on β, e.g., a multivariate Student's t prior or an improper uniform prior. Recently [20] [also see 11] showed that a DA algorithm can be obtained for Bayesian logistic models using the mixture normal representation of logistic distribution. It would also be interesting to see if the DA algorithms for Bayesian logistic regression converge at a geometric rate.

Appendix A: A Mill's ratio type result for Student's t distribution

The following result can be gleaned from [25], but here we give a proof for completeness.

Lemma 1. *For $u > 0$ we have*

$$\frac{1}{(1 - F_\nu(u))(\nu + u^2)^{\frac{\nu-1}{2}}} \geq \frac{u}{\kappa},$$

where $\kappa = \Gamma((\nu-1)/2)\,\nu^{\nu/2}/(2\sqrt{\pi}\Gamma(\nu/2))$ and as before $F_\nu(\cdot)$ is the cdf of $t_\nu(0,1)$.

Proof. Let $U \sim t_\nu(0,1)$, that is, U follows the t-distribution with mean 0, variance 1 and ν d.f. The pdf of U is

$$f_\nu(u) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}}\,\frac{1}{(1+\frac{u^2}{\nu})^{\frac{\nu+1}{2}}}, -\infty < u < \infty\ .$$

Then,

$$\begin{aligned}
1 - F_\nu(u) &= \int_u^\infty f_\nu(t)dt \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}} \int_u^\infty \frac{1}{(1+\frac{t^2}{\nu})^{\frac{\nu+1}{2}}}\,dt \\
&= \frac{1}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}} \int_u^\infty \int_0^\infty w^{\frac{\nu+1}{2}-1} \exp\Big[-\Big(1+\frac{t^2}{\nu}\Big)w\Big]dw\,dt \\
&= \frac{1}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}} \int_0^\infty w^{\frac{\nu-1}{2}} e^{-w} \int_u^\infty e^{-\frac{t^2 w}{\nu}}dt\,dw\ .
\end{aligned}$$

Since $u > 0$, using a well known bound for Mill's ratio [3, p. 175] we have

$$\begin{aligned}
\int_u^\infty e^{-\frac{t^2 w}{\nu}}dt &= \sqrt{\frac{\nu}{2w}} \int_{\sqrt{\frac{2u}{\nu}w}}^\infty e^{-\frac{v^2}{2}}dv \\
&\leq \sqrt{\frac{\nu}{2w}}\,\frac{1}{\sqrt{\frac{2w}{\nu}}u}\,e^{-\frac{u^2 w}{\nu}} \\
&= \frac{\nu}{2uw}\,e^{-\frac{u^2 w}{\nu}}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
1 - F_\nu(u) &\leq \frac{\nu}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}}\,\frac{1}{2u} \int_0^\infty w^{\frac{\nu-1}{2}-1} e^{-w(1+\frac{u^2}{\nu})}dw \\
&= \frac{1}{2u}\sqrt{\frac{\nu}{\pi}}\,\frac{1}{\Gamma(\frac{\nu}{2})}\,\frac{\Gamma(\frac{\nu-1}{2})}{(1+\frac{u^2}{\nu})^{\frac{\nu-1}{2}}}\,,
\end{aligned}$$

or, equivalently

$$\frac{1}{(1 - F_\nu(u))(\nu + u^2)^{\frac{\nu-1}{2}}} \geq \frac{u}{\kappa}.$$

□

Appendix B: Proof of Theorem 1

Proof. We use $V(\beta) = \beta^T X^T X \beta$ to establish a drift condition. Since X is assumed to have full rank (condition $A1$), the level sets $\{\beta : V(\beta) \le \alpha\}$ are compact. By Fubini's theorem, we have

$$\begin{aligned}
(KV)(\beta) &= \int_{\mathbb{R}^p} V(\beta')\, k(\beta' \mid \beta)\, d\beta' \\
&= \int_{\mathcal{Z}} \int_{\mathbb{R}^n_+} \int_{\mathbb{R}^p} V(\beta')\, \pi(\beta' \mid \lambda, z, y)\, \pi(\lambda, z \mid \beta, y)\, d\beta'\, d\lambda\, dz \\
&= \int_{\mathcal{Z}} \int_{\mathbb{R}^n_+} \int_{\mathbb{R}^p} V(\beta')\, \pi(\beta' \mid \lambda, z, y)\, \pi(\lambda \mid z, \beta, y) \pi(z \mid \beta, y)\, d\beta'\, d\lambda\, dz \\
&= \int_{\mathcal{Z}} \left\{ \int_{\mathbb{R}^n_+} \left(\int_{\mathbb{R}^p} V(\beta')\, \pi(\beta' \mid \lambda, z, y)\, d\beta' \right) \pi(\lambda \mid z, \beta, y)\, d\lambda \right\} \pi(z \mid \beta, y)\, dz \\
&= E\Big[E\Big\{ E\big(V(\beta') \,\big|\, \lambda, z, y\big) \,\Big|\, z, \beta, y \Big\} \,\Big|\, \beta, y \Big]\,, \qquad \text{(B.1)}
\end{aligned}$$

where, as the notation suggests, the (conditional) expectations are with respect to the densities $\pi(\beta \mid \lambda, z, y)$, $\pi(\lambda \mid z, \beta, y)$ and $\pi(z \mid \beta, y)$ in the given order. The inner-most expectation in (B.1) is with respect to $\pi(\beta \mid \lambda, z, y)$, which is a multivariate normal density. The next level expectation is with respect to $\pi(\lambda \mid z, \beta, y)$, which is a product of univariate gamma densities. And, lastly, the outer-most expectation is with respect to $\pi(z \mid \beta, y)$, which is a product of truncated Student's t densities.

Starting with the innermost expectation, we have

$$\begin{aligned}
E\Big(V(\beta') | \lambda, z, y\Big) &= E\Big(\beta'^T X^T X \beta' | \lambda, z, y\Big) \\
&= \operatorname{tr}\Big((X^T X)(X^T \Lambda X + \Sigma_a)^{-1}\Big) \\
&\quad + (z^T \Lambda X + \beta_a^T \Sigma_a)(X^T \Lambda X + \Sigma_a)^{-1} \\
&\quad \times X^T X (X^T \Lambda X + \Sigma_a)^{-1} (X^T \Lambda z + \Sigma_a \beta_a), \qquad \text{(B.2)}
\end{aligned}$$

where $\operatorname{tr}(\cdot)$ denote the trace of a matrix. Note that,

$$\begin{aligned}
\operatorname{tr}\Big((X^T X)(X^T \Lambda X + \Sigma_a)^{-1}\Big) &= \operatorname{tr}\Big(\Big(\sum_{i=1}^n x_i x_i^T\Big)(X^T \Lambda X + \Sigma_a)^{-1}\Big) \\
&= \sum_{i=1}^n x_i^T (X^T \Lambda X + \Sigma_a)^{-1} x_i \\
&= \sum_{i=1}^n x_i^T \Big(\sum_{j=1}^n (\lambda_j + c) x_j x_j^T\Big)^{-1} x_i,
\end{aligned}$$

where the last equality follows from our assumption that $\Sigma_a = c X^T X$. Now

$$x_i^T \Big(\sum_{j=1}^n (\lambda_j + c) x_j x_j^T\Big)^{-1} x_i = \frac{1}{\lambda_i + c} x_i^T \Big(\sum_{j=1}^n \frac{\lambda_j + c}{\lambda_i + c} x_j x_j^T\Big)^{-1} x_i$$

Since

$$\sum_{j=1}^{n} \frac{\lambda_j + c}{\lambda_i + c} x_j x_j^T = \frac{1}{\lambda_i + c}\Big(X^T \Lambda X + \Sigma_a\Big)$$

is a positive definite matrix and

$$\sum_{j=1}^{n} \frac{\lambda_j + c}{\lambda_i + c} x_j x_j^T - x_i x_i^T = \sum_{j \neq i} \frac{\lambda_j + c}{\lambda_i + c} x_j x_j^T$$

is positive semidefinite, from Roy and Hobert's [2010] Lemma 3, it follows that

$$x_i^T \Big(\sum_{j=1}^{n} (\lambda_j + c) x_j x_j^T \Big)^{-1} x_i \leq \frac{1}{\lambda_i + c}.$$

Therefore,

$$\mathrm{tr}\Big((X^T X)(X^T \Lambda X + \Sigma_a)^{-1}\Big) \leq \sum_{i=1}^{n} \frac{1}{\lambda_i + c} \leq \frac{n}{c}.$$

Now we consider the second term in (B.2). Note that

$$\begin{aligned}
&(z^T \Lambda X + \beta_a^T \Sigma_a)(X^T \Lambda X + \Sigma_a)^{-1} X^T X (X^T \Lambda X + \Sigma_a)^{-1} (X^T \Lambda z + \Sigma_a \beta_a) \\
&\quad = z^T \Lambda X (X^T \Lambda X + \Sigma_a)^{-1} X^T X (X^T \Lambda X + \Sigma_a)^{-1} X^T \Lambda z \\
&\qquad + \beta_a^T \Sigma_a (X^T \Lambda X + \Sigma_a)^{-1} X^T X (X^T \Lambda X + \Sigma_a)^{-1} \Sigma_a \beta_a \\
&\qquad + 2 z^T \Lambda X (X^T \Lambda X + \Sigma_a)^{-1} X^T X (X^T \Lambda X + \Sigma_a)^{-1} \Sigma_a \beta_a \\
&\quad = A + B + C, \qquad \text{(B.3)}
\end{aligned}$$

where A, B and C denote the first, second and the third terms in the above expression. Note that, all of these three terms are functions of random vectors z and λ as well as the prior covariance matrix Σ_a^{-1} and the data (y, X). We will analyze each of these three terms separately. We begin with the second term. Note that

$$\begin{aligned}
B &= \beta_a^T \Sigma_a (X^T \Lambda X + \Sigma_a)^{-1} X^T X (X^T \Lambda X + \Sigma_a)^{-1} \Sigma_a \beta_a \\
&\leq \beta_a^T \Sigma_a (X^T \Lambda X + \Sigma_a)^{-1} \Big[\frac{1}{c}\Big(X^T \Lambda X + c X^T X\Big)\Big] (X^T \Lambda X + \Sigma_a)^{-1} \Sigma_a \beta_a \\
&= \frac{1}{c} \beta_a^T \Sigma_a (X^T \Lambda X + \Sigma_a)^{-1} \Big(X^T \Lambda X + \Sigma_a\Big) (X^T \Lambda X + \Sigma_a)^{-1} \Sigma_a \beta_a \\
&= \frac{1}{c} \beta_a^T \Sigma_a (X^T \Lambda X + \Sigma_a)^{-1} \Sigma_a \beta_a,
\end{aligned}$$

where the inequality follows from the fact that $\frac{1}{c} X^T \Lambda X$ is a positive semidefinite matrix and the second equality is due to our assumption that $\Sigma_a = c X^T X$. Since $X^T \Lambda X$ is positive semidefinite, it follows that [see e.g. 28, page 70] $\Sigma_a^{-1} - (X^T \Lambda X + \Sigma_a)^{-1}$ is positive semidefinite. So we have

$$B \leq \frac{1}{c} \beta_a^T \Sigma_a \Sigma_a^{-1} \Sigma_a \beta_a = \frac{1}{c} \beta_a^T \Sigma_a \beta_a = \beta_a^T X^T X \beta_a.$$

Let us denote the constant $\beta_a^T X^T X \beta_a$ by L_1. Next, we consider the last term in (B.3). By Cauchy-Schwartz inequality, we have

$$z^T \Lambda X (X^T \Lambda X + \Sigma_a)^{-1} X^T X (X^T \Lambda X + \Sigma_a)^{-1} \Sigma_a \beta_a \leq \sqrt{A.B} \leq \sqrt{L_1}\sqrt{A}.$$

Therefore, we have $C \leq 2\sqrt{L_1}\sqrt{A}$. We will analyze the first term A in (B.3) later. Putting all of this together, from (B.2) we have

$$E(V(\beta')|z, \lambda, y) \leq \frac{n}{c} + A + B + C \leq \frac{n}{c} + L_1 + A + 2\sqrt{L_1}\sqrt{A}. \tag{B.4}$$

Now we consider the second level expectation in (B.1), that is $E\{E(V(\beta') \mid \lambda, z, y) \mid z, \beta, y\}$. Note that

$$\begin{aligned} E\Big\{E\big(V(\beta') \,\big|\, \lambda, z, y\big) \,\big|\, z, \beta, y\Big\} &\leq \frac{n}{c} + L_1 + E(A|z, \beta, y) + 2\sqrt{L_1}E(\sqrt{A}|z, \beta, y) \\ &\leq \frac{n}{c} + L_1 + E(A|z, \beta, y) + 2\sqrt{L_1}\sqrt{E(A|z, \beta, y)} \\ &\leq \frac{n}{c} + L_1 + E(A|z, \beta, y) + 2\sqrt{L_1}(1 + E(A|z, \beta, y)) \\ &= \frac{n}{c} + L_1 + 2\sqrt{L_1} + (1 + 2\sqrt{L_1})E(A|z, \beta, y), \end{aligned} \tag{B.5}$$

where the first inequality follows from (B.4), second inequality is an application of the Jensen's inequality and the last inequality is due to the fact that $\sqrt{x} \leq 1 + x$ for $x \geq 0$. Now we will construct an upper bound of $E(A|z, \beta, y)$. By similar arguments that was used before to bound the term B, we have

$$\begin{aligned} A &= z^T \Lambda X (X^T \Lambda X + \Sigma_a)^{-1} X^T X (X^T \Lambda X + \Sigma_a)^{-1} X^T \Lambda z \\ &\leq z^T \Lambda X (X^T \Lambda X + \Sigma_a)^{-1} \Big[\frac{1}{c}\Big(X^T \Lambda X + cX^T X\Big)\Big] (X^T \Lambda X + \Sigma_a)^{-1} X^T \Lambda z \\ &= \frac{1}{c} z^T \Lambda X (X^T \Lambda X + \Sigma_a)^{-1} (X^T \Lambda X + \Sigma_a)(X^T \Lambda X + \Sigma_a)^{-1} X^T \Lambda z \\ &= \frac{1}{c} z^T \Lambda X (X^T \Lambda X + \Sigma_a)^{-1} X^T \Lambda z. \end{aligned}$$

Since Σ_a is positive semidefinite, we know that $(X^T \Lambda X)^{-1} - (X^T \Lambda X + \Sigma_a)^{-1}$ is positive semidefinite. So we have

$$A \leq \frac{1}{c} z^T \Lambda X (X^T \Lambda X)^{-1} X^T \Lambda z.$$

Now,

$$z^T \Lambda z - z^T \Lambda X (X^T \Lambda X)^{-1} X^T \Lambda z = z^T \Lambda^{\frac{1}{2}} \Big(I - \Lambda^{\frac{1}{2}} X (X^T \Lambda X)^{-1} X^T \Lambda^{\frac{1}{2}}\Big) \Lambda^{\frac{1}{2}} z \geq 0.$$

Therefore, we have

$$A \leq \frac{1}{c} z^T \Lambda z = \frac{1}{c} \sum_{i=1}^n \lambda_i z_i^2 \leq \frac{1}{c} \sum_{i=1}^n \lambda_i \sum_{i=1}^n z_i^2.$$

Thus we have

$$E(A|z,\beta,y) \le \frac{1}{c}\Big(\sum_{i=1}^n z_i^2\Big) E\Big(\sum_{i=1}^n \lambda_i | z,\beta,y\Big).$$

Now recall that $\pi(\lambda|z,\beta,y)$ is product of n univariate Gamma densities with

$$\lambda_i|z,\beta,y \sim \text{Gamma}\Big(\frac{\nu+1}{2}, \frac{\nu+(z_i-x_i^T\beta)^2}{2}\Big).$$

So we have

$$E\Big(\sum_{i=1}^n \lambda_i|z,\beta,y\Big) \le \sum_{i=1}^n \frac{\nu+1}{\nu+(z_i-x_i^T\beta)^2} \le \frac{n(\nu+1)}{\nu}.$$

Thus

$$E\Big(A|z,\beta,y\Big) \le \frac{n(\nu+1)}{c\nu}\sum_{i=1}^n z_i^2,$$

and so from (B.5) we have

$$E\Big\{E\big(V(\beta')\,\big|\,\lambda,z,y\big)\,\big|\,z,\beta,y\Big\} \le \frac{n}{c} + L_1 + 2\sqrt{L_1} + (1+2\sqrt{L_1})\frac{n(\nu+1)}{c\nu}\sum_{i=1}^n z_i^2.$$

Finally, we consider the outer-most expectation in (B.1). Note that

$$\begin{aligned}(KV)(\beta) &= E\Big[E\Big\{E\big(V(\beta')\,\big|\,\lambda,z,y\big)\,\big|\,z,\beta,y\Big\}\,\big|\,\beta,y\Big] \\ &\le \frac{n}{c} + L_1 + 2\sqrt{L_1} + (1+2\sqrt{L_1})\frac{n(\nu+1)}{c\nu}\sum_{i=1}^n E(z_i^2|\beta,y). \qquad \text{(B.6)}\end{aligned}$$

In order to calculate $\sum_{i=1}^n E(z_i^2|\beta,y)$, recall that conditional on (β,y), $z_1,\ldots,z_n$ are independent with $z_i|\beta,y \sim Tt_\nu(x_i^T\beta,y_i)$. Since $\nu>2$, simple calculations using the results in [15] show that

$$E(z_i^2|\beta,y) = \begin{cases} \frac{\nu}{\nu-2} + (x_i^T\beta)^2 - \frac{\kappa(x_i^T\beta)}{1-F_\nu(x_i^T\beta)}\cdot\frac{1}{(\nu+(x_i^T\beta)^2)^{\frac{\nu-1}{2}}} & \text{if} \quad y_i=0 \\ \frac{\nu}{\nu-2} + (x_i^T\beta)^2 + \frac{\kappa(x_i^T\beta)}{F_\nu(x_i^T\beta)}\cdot\frac{1}{(\nu+(x_i^T\beta)^2)^{\frac{\nu-1}{2}}} & \text{if} \quad y_i=1 \end{cases}$$

where κ is as defined in Appendix A or we can simply write

$$E(z_i^2|\beta,y) = \frac{\nu}{\nu-2} + (w_i^T\beta)^2 - \kappa\frac{w_i^T\beta}{1-F_\nu(w_i^T\beta)}\cdot\frac{1}{(\nu+(w_i^T\beta)^2)^{\frac{\nu-1}{2}}}.$$

So

$$\sum_{i=1}^n E(z_i^2|\beta,y) = \frac{n\nu}{\nu-2} + \sum_{i=1}^n (w_i^T\beta)^2 - \kappa\sum_{i=1}^n \frac{w_i^T\beta}{1-F_\nu(w_i^T\beta)}\cdot\frac{1}{(\nu+(w_i^T\beta)^2)^{\frac{\nu-1}{2}}} \qquad \text{(B.7)}$$

Note that $\sum_{i=1}^n E(z_i^2|\beta=0,y) = \frac{n\nu}{\nu-2}$. In order to bound the above expression when $\beta \in \mathbb{R}^p \setminus \{0\}$, as in [32] we construct a partition of the set $\mathbb{R}^p \setminus \{0\}$ using the n hyperplanes defined by $w_i^T\beta = 0$. For a positive integer m, define $\mathbb{N}_m = \{1,2,\ldots,m\}$. Let $D_1, D_2, \ldots, D_{2^n}$ denote all the subsets of $\mathbb{N}_n$, and, for each $j \in \mathbb{N}_{2^n}$, define a subset of the p-dimensional Euclidean space as follows:

$$S_j = \left\{\beta \in \mathbb{R}^p \setminus \{0\} : w_i^T\beta \le 0 \text{ for all } i \in D_j \text{ and } w_i^T\beta > 0 \text{ for all } i \in \bar{D}_j\right\}$$

where $\bar{D}_j$ denotes the complement of D_j; that is, $\bar{D}_j = \mathbb{N}_n \setminus D_j$. Note that the sets S_j's are disjoint, $\cup_{j=1}^{2^n} S_j = \mathbb{R}^p \setminus \{0\}$, and some of the S_j's may be empty.

Since the condition A2 is in force, following [32] we can show that if S_j is nonempty, then so are D_j and $\bar{D}_j$. Now define $E = \left\{j \in \mathbb{N}_{2^n} : S_j \ne \emptyset\right\}$. For each $j \in E$, define

$$R_j(\beta) = \frac{\sum_{i\in D_j}(w_i^T\beta)^2}{\sum_{i=1}^n (w_i^T\beta)^2} = \frac{\sum_{i\in D_j}(w_i^T\beta)^2}{\sum_{i\in D_j}(w_i^T\beta)^2 + \sum_{i\in \overline{D}_j}(w_i^T\beta)^2}\,.$$

and

$$\rho_j = \sup_{\beta\in S_j} R_j(\beta)\,.$$

From [32] we know that $\rho_j < 1$ for all $j \in E$. Now we consider the following Mill's ratio type expression for Student's t distribution

$$\frac{u}{(1-F_\nu(u))(\nu+u^2)^{\frac{\nu-1}{2}}}.$$

Since $\nu > 2$, it is clear that if we define

$$M = \sup_{u\in(-\infty,0]} \left|\frac{u}{(1-F_\nu(u))(\nu+u^2)^{\frac{\nu-1}{2}}}\right|,$$

then $M \in (0,\infty)$. From Lemma 1 we know that when $u > 0$,

$$\frac{u}{(1-F_\nu(u))(\nu+u^2)^{\frac{\nu-1}{2}}} \ge \frac{u^2}{\kappa}.$$

Fix $j \in E$. It follows from (B.7) and above two results that for all $\beta \in S_j$, we have

$$\begin{aligned}
\sum_{i=1}^n E(z_i^2|\beta,y) &= \frac{n\nu}{\nu-2} + \sum_{i=1}^n (w_i^T\beta)^2 - \kappa \sum_{i\in D_j} \frac{w_i^T\beta}{1-F_\nu(w_i^T\beta)} \cdot \frac{1}{(\nu+(w_i^T\beta)^2)^{\frac{\nu-1}{2}}} \\
&\qquad - \kappa \sum_{i\in \bar{D}_j} \frac{w_i^T\beta}{1-F_\nu(w_i^T\beta)} \cdot \frac{1}{(\nu+(w_i^T\beta)^2)^{\frac{\nu-1}{2}}} \\
&\le \frac{n\nu}{\nu-2} + \sum_{i=1}^n (w_i^T\beta)^2 + \kappa \sum_{i\in D_j} \left|\frac{w_i^T\beta}{1-F_\nu(w_i^T\beta)} \cdot \frac{1}{(\nu+(w_i^T\beta)^2)^{\frac{\nu-1}{2}}}\right| \\
&\qquad - \kappa \sum_{i\in \bar{D}_j} \frac{w_i^T\beta}{1-F_\nu(w_i^T\beta)} \cdot \frac{1}{(\nu+(w_i^T\beta)^2)^{\frac{\nu-1}{2}}}
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{n\nu}{\nu-2} + n\kappa M + \sum_{i=1}^{n}(w_i^T\beta)^2 - \sum_{i\in\bar{D}_j}(w_i^T\beta)^2 \\
&= \frac{n\nu}{\nu-2} + n\kappa M + \sum_{i\in D_j}(w_i^T\beta)^2 \\
&= \frac{n\nu}{\nu-2} + n\kappa M + R_j(\beta)\sum_{i=1}^{n}(w_i^T\beta)^2 \\
&\leq \frac{n\nu}{\nu-2} + n\kappa M + \rho_j V(\beta)
\end{aligned}
$$

Therefore, since $\uplus_{j\in E} S_j = \mathbb{R}^p \setminus \{0\}$, it follows that for all $\beta \in \mathbb{R}^p$, we have

$$
\sum_{i=1}^{n} E(z_i^2|\beta, y) \leq \frac{n\nu}{\nu-2} + n\kappa M + \rho' V(\beta)\ ,
$$

where $\rho' := \max_{j\in E} \rho_j < 1$. Finally from (B.6)

$$
\begin{aligned}
(KV)(\beta) &\leq \frac{n}{c} + L_1 + 2\sqrt{L_1} + (1+2\sqrt{L_1})\frac{n(\nu+1)}{c\nu}\sum_{i=1}^{n} E(z_i^2|\beta, y) \\
&\leq L + \rho V(\beta), \qquad \text{(B.8)}
\end{aligned}
$$

where $L := \frac{n}{c} + L_1 + 2\sqrt{L_1} + \frac{n(\nu+1)}{c\nu}(1+2\sqrt{L_1})(\frac{n\nu}{\nu-2} + n\kappa M)$ and $\rho = \rho'\frac{n(\nu+1)}{c\nu}(1+2\sqrt{\beta_a^T X^T X \beta_a})$. Since $\rho' < 1$ and $n < \frac{c\nu}{(\nu+1)(1+2\sqrt{\beta_a^T X^T X\beta_a})}$, ρ is also less than 1. So the DA chain is geometrically ergodic. □

Acknowledgments

The author thanks two reviewers and an associate editor for helpful comments and valuable suggestions which led to several improvements in the manuscript.

References

[1] Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88** 669–679. MR1224394

[2] Bednorz, W. and Latuszynski, K. (2007). A few remarks on "Fixed-width output analysis for Markov chain Monte Carlo" by Jones et al. *Journal of the American Statistical Association* **102** 1485–1486. MR2412582

[3] Feller, W. (1968). *An Introduction to Probability Theory and its Applications*, vol. **I**, 3rd. ed. John Wiley & Sons, New York. MR0228020

[4] Flegal, J. M. (2012). *mcmcse: Monte Carlo standard errors for MCMC.* R package version 0.1. http://CRAN.R-project.org/package=mcmcse

[5] Flegal, J. M., Haran, M. and Jones, G. L. (2008). Markov Chain Monte Carlo: Can We Trust the Third Significant Figure? *Statistical Science* **23** 250–260. MR2516823

[6] Flegal, J. M. and Jones, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *The Annals of Statistics* **38** 1034–1070. MR2604704

[7] Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, Cambridge University Press.

[8] Hobert, J. P. (2011). *Handbook of Markov chain Monte Carlo*. The data augmentation algorithm: theory and methodology, 253–293. CRC Press, Boca Raton, FL. MR2858452

[9] Hobert, J. P., Jones, G. L., Presnell, B. and Rosenthal, J. S. (2002). On the applicability of regenerative simulation in Markov chain Monte Carlo. *Biometrika* **89** 731–743. MR1946508

[10] Hobert, J. P. and Marchev, D. (2008). A theoretical comparison of the data augmentation, marginal augmentation and PX-DA algorithms. *The Annals of Statistics* **36** 532-554. MR2396806

[11] Holmes, C. C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis* **1** 145-168. MR2227368

[12] Jones, G. L., Haran, M., Caffo, B. S. and Neath, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association* **101** 1537–1547. MR2279478

[13] Jones, G. L. and Hobert, J. P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science* **16** 312–34. MR1888447

[14] Khare, K. and Hobert, J. P. (2011). A spectral analytic comparison of trace-class data augmentation algorithms and their sandwich variants. *The Annals of Statistics* **39** 2585-2606. MR2906879

[15] Kim, H. J. (2008). Moments of truncated Student-t distribution. *Journal of the Korean Statistical Society* **37** 81-87. MR2409373

[16] Liu, C. (2004). Robit regression: A simple robust alternative to logistic and probit regression. In *Applied Bayesian Modeling and Casual Inference from Incomplete-Data Perspectives* (A. Gelman and X. L. Meng, eds.) 227-238. Wiley, London. MR2138259

[17] Liu, J. S. and Sabatti, C. (2000). Generalised Gibbs sampler and multigrid Monte Carlo for Bayesian computation. *Biometrika* **87** 353–369. MR1782484

[18] Liu, J. S., Wong, W. H. and Kong, A. (1995). Covariance Structure and Convergence Rate of the Gibbs Sampler with Various Scans. *Journal of the Royal Statistical Society,* Series B **57** 157-169. MR1325382

[19] Liu, J. S. and Wu, Y. N. (1999). Parameter Expansion for Data Augmentation. *Journal of the American Statistical Association* **94** 1264–1274. MR1731488

[20] Marchev, D. (2011). Markov chain Monte Carlo algorithms for the Bayesian logistic regression model. In *Proceeding of the Annual Interna-*

tional Conference on Operations Research and Statistics (C. B. Gupta, ed.) 154-159. Global Science and Technology Forum, Penang, Malaysia.
[21] Meng, X.-L. and van Dyk, D. A. (1999). Seeking Efficient Data Augmentation Schemes via Conditional and Marginal Augmentation. *Biometrika* **86** 301–320. MR1705351
[22] Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer Verlag, London. MR1287609
[23] Mira, A. and Geyer, C. J. (1999). Ordering Monte Carlo Markov chains. Technical Report No. 632, School of Statistics, University of Minnesota.
[24] Mudholkar, G. S. and George, E. O. (1978). A remark on the shape of the logistic distribution. *Biometrika* **65** 667-668.
[25] Pinkham, R. S. and Wilk, M. B. (1963). Tail areas of the t-distribution from a Mills'-ratio-like expansion. *Annals of Mathematical Statistics* **34** 335-337. MR0144409
[26] Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrics* **38** 485-498.
[27] R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria ISBN 3-900051-07-0. http://www.R-project.org
[28] Rao, C. R. (1973). *Linear statistical inference and its applications*, 2nd. ed. John Wiley & Sons, New York. MR0346957
[29] Roberts, G. O. and Rosenthal, J. S. (1997). Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability* **2** 13-25. MR1448322
[30] Roberts, G. O. and Tweedie, R. L. (2001). Geometric $L2$ and $L1$ convergence are equivalent for reversible Markov chains. *Journal of Applied Probability* **38A** 37–41. MR1915532
[31] Roy, V. (2012). Spectral analytic comparisons for data augmentation. *Stat. and Prob. Letters* **82** 103-108. MR2863030
[32] Roy, V. and Hobert, J. P. (2007). Convergence rates and asymptotic standard errors for MCMC algorithms for Bayesian probit regression. *Journal of the Royal Statistical Society,* Series B **69** 607-623. MR2370071
[33] Roy, V. and Hobert, J. P. (2010). On Monte Carlo methods for Bayesian regression models with heavy-tailed errors. *Journal of Multivariate Analysis* **101** 1190-1202. MR2595301
[34] Tan, A. and Hobert, J. P. (2009). Block Gibbs sampling for Bayesian random effects models with improper priors: convergence and regeneration. *Journal of Computational and Graphical Statistics* **18** 861-878. MR2598033
[35] Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation(with discussion). *Journal of the American Statistical Association* **82** 528–550. MR0898357
[36] van Dyk, D. A. and Meng, X.-L. (2001). The Art of Data Augmentation (with discussion). *Journal of Computational and Graphical Statistics* **10** 1–50. MR1936358
[37] Yu, Y. and Meng, X. L. (2011). To center or not to center: that is not the question - An ancillarity-sufficiency interweaving strategy (ASIS) for boost-

ing MCMC efficiency. *Journal of Computational and Graphical Statistics* **20** 531-570. MR2878987

[38] Zellner, A. (1983). Applications of Bayesian analysis in econometrics. *The Statistician* **32** 23-34.