Geometric convergence of the Haar PX-DA algorithm for the Bayesian multivariate regression model with Student t errors

Vivekananda Roy
Department of Statistics
Iowa State University

James P. Hobert
Department of Statistics
University of Florida

February 2009

Abstract

We consider Bayesian analysis of data from multivariate linear regression models whose errors have a distribution that is a scale mixture of normals. Such models are used to analyze data on financial returns, which are notoriously heavy-tailed. Let π denote the intractable posterior density that results when this regression model is combined with the standard non-informative prior on the unknown regression coefficients and scale matrix of the errors. Roughly speaking, the posterior is proper if and only if $n \geq d + k$, where n is the sample size, d is the dimension of the response, and k is number of covariates. We provide a method of making exact draws from π in the special case where $n = d + k$, and we study Markov chain Monte Carlo (MCMC) algorithms that can be used to explore π when $n > d + k$. In particular, we show how the Haar PX-DA technology of Hobert and Marchev (2008) can be used to improve upon Liu's (1996) data augmentation (DA) algorithm. Indeed, the new algorithm that we introduce is theoretically superior to the DA algorithm, yet equivalent to DA in terms of computational complexity. Moreover, we analyze the convergence rates of these MCMC algorithms in the important special case where the regression errors have a Student's t distribution. We prove that, under conditions on n, d, k, and the degrees of freedom of the t distribution, both 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.

Key words and phrases. Data augmentation algorithm, Drift condition, Markov chain, Minorization condition, Monte Carlo, Robust multivariate regression

1 Introduction

Let $y_1, y_2, \dots, y_n$ be d-dimensional random vectors satisfying the linear regression model

$$y_i = \beta^T x_i + \varepsilon_i \ , \tag{1}$$

where β is a $k \times d$ matrix of unknown regression coefficients, the x_i's are known $k \times 1$ covariate vectors, and the d-dimensional error vectors, $\varepsilon_1, \dots, \varepsilon_n$, are iid with common density

$$f_H(\varepsilon) = \int_0^\infty \frac{\delta^{\frac{d}{2}}}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\Big\{ -\frac{\delta}{2} \varepsilon^T \Sigma^{-1} \varepsilon \Big\} \, dH(\delta) \ ,$$

where $H(\cdot)$ is the distribution function of some non-negative random variable. In practice, H will be fixed, but, for the time being, we leave it unspecified because many of our results hold true for any H. The density f_H is a multivariate scale mixture of normals and it belongs to the class of elliptically symmetric distributions. It can be made heavy-tailed by choosing H appropriately. In particular, when H is the distribution function corresponding to a Gamma$(\frac{\nu}{2}, \frac{\nu}{2})$ random variable, then f_H becomes the multivariate Student's t density with $\nu > 0$ degrees of freedom, which, aside from a normalizing constant, is given by $\big[1 + \nu^{-1} \varepsilon^T \Sigma^{-1} \varepsilon\big]^{-\frac{d+\nu}{2}}$. These heavy-tailed error distributions are often used when modelling financial data (see, e.g., Affleck-Grave and McDonald, 1989; Bradley and Taqqu, 2003; Cademartori, Romo, Campos and Galea, 2003; Fama, 1965; Zhou, 1993).

Let y denote the $n \times d$ matrix whose ith row is y_i^T, and let X stand for the $n \times k$ matrix whose ith row is x_i^T, and, finally, let ε represent the $n \times d$ matrix whose ith row is ε_i^T. Using this notation, we can state the n equations in (1) more succinctly as follows

$$y = X\beta + \varepsilon \ .$$

We assume throughout the paper that X has full column rank. The density of y is given by

$$f(y|\beta, \Sigma) = \prod_{i=1}^n \left[\int_0^\infty \frac{\delta^{\frac{d}{2}}}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\Big\{ -\frac{\delta}{2} \Big(y_i - \beta^T x_i\Big)^T \Sigma^{-1} \Big(y_i - \beta^T x_i\Big) \Big\} \, dH(\delta) \right] . \tag{2}$$

We consider a Bayesian analysis using the standard non-informative prior for multivariate location scale problems given by $\pi(\beta, \Sigma) \propto |\Sigma|^{-\frac{d+1}{2}}$. Of course, whenever an improper prior is used, one must check that the resulting posterior is proper. Fernandez and Steel (1999) provide necessary and sufficient conditions for propriety (see Section 2). Roughly speaking, these authors show that the posterior is proper if and only if $n \geq d + k$. Assuming it is proper, the posterior density is characterized by

$$\pi(\beta, \Sigma | y) \propto f(y|\beta, \Sigma) \pi(\beta, \Sigma) \ .$$

Unfortunately, this posterior density is intractable in the sense that posterior expectations, which are required for Bayesian inference, cannot be computed in closed form. In the special case where $n = d+k$, we provide a simple algorithm for making exact draws from $\pi(\beta, \Sigma|y)$, which allows one to use classical Monte Carlo methods (based on iid simulations) to approximate posterior expectations. Unfortunately, this method breaks down when $n > d + k$, and in these cases we must resort to Markov chain Monte Carlo (MCMC) methods.

The first MCMC method developed for this problem was the data augmentation (DA) algorithm of Liu (1996). Our main result in this paper is a proof that, in the important special case where the ε_is have a multivariate Student's t distribution, the Markov chain that drives Liu's (1996) algorithm converges at a geometric rate. This result is important from a practical standpoint because it guarantees the existence of central limit theorems (CLTs) for ergodic averages, which in turn allows for the calculation of valid asymptotic standard errors for the MCMC estimates of posterior expectations (Flegal, Haran and Jones, 2008; Jones, Haran, Caffo and Neath, 2006; Jones and Hobert, 2001). We also present an alternative MCMC algorithm, called the Haar PX-DA algorithm, that is computationally equivalent to Liu's (1996) algorithm, but is theoretically superior.

The remainder of this paper is organized as follows. In Section 2, we present an alternative proof of the sufficiency half of Fernandez and Steel's (1999) propriety result, which yields a simple method of making exact draws from the posterior, $\pi(\beta, \Sigma|y)$, in the special case where $n = d + k$. In Section 3, we consider MCMC methods for simulating from $\pi(\beta, \Sigma|y)$ in cases where $n > d + k$. In particular, we describe the DA algorithm of Liu (1996) and we show how the Haar PX-DA technology of Hobert and Marchev (2008) can be used to improve upon it. Finally, in Section 4 we investigate the rate of convergence of these two MCMC algorithms in the important special case where the regression errors have a Student's t distribution. Our main result is that, for certain configurations of (n, d, k, ν), both of the algorithms converge at a geometric rate.

2 An Alternative Proof of Propriety and an Exact Sampling Algorithm

As mentioned in the introduction, Fernandez and Steel (1999) studied the propriety of the posterior density that results when the multivariate regression likelihood, $f(y|\beta, \Sigma)$, is combined with the standard non-informative prior, $\pi(\beta, \Sigma)$. In order to state their result precisely, define

$$c(y) = \int_W \int_{\mathbb{R}^{dk}} f(y|\beta, \Sigma)\pi(\beta, \Sigma) \, d\beta \, d\Sigma \ ,$$

where $W \subset \mathbb{R}^{\frac{d(d+1)}{2}}$ denotes the space of $d \times d$ positive definite matrices. (Note that dependence on X is being suppressed in the notation.) For a fixed data set, $y \in \mathbb{R}^{dn}$, the posterior distribution is proper if and only if $c(y) < \infty$. Of course, when $c(y)$ is finite, the posterior density is given by

$$\pi(\beta, \Sigma | y) = f(y|\beta, \Sigma)\pi(\beta, \Sigma)/c(y) \ .$$

Fernandez and Steel (1999) sketched a proof of the following result.

Proposition 1. *Let ϕ denote Lebesgue measure on $\mathbb{R}^{dn}$. If $n \geq d + k$, then there exists a set $S \subset \mathbb{R}^{dn}$ (that depends on X) with $\phi(S) = 0$ such that, if $y \notin S$, then $c(y) < \infty$. Conversely, if $n < d + k$, then $c(y) = \infty$.*

In words, Proposition 1 says that, if there are too few data vectors ($n < d + k$), then the posterior is improper, and, if there is "enough" data ($n \geq d + k$), then, with probability one, the observed data vector will result in a proper posterior. In this section, we provide the details of an alternative proof of the sufficiency part of Proposition 1 that leads to a method of making exact draws from the posterior density in the special case where $n = d + k$. (For a careful proof of the necessity part, see Roy (2008).) Our proof is based on the standard missing data model that includes both the regression data, y, and the missing observations from the mixing distribution H.

Suppose that, conditional on (β, Σ), $\{(y_i, q_i)\}_{i=1}^n$ are iid pairs such that

$$y_i | q_i, \beta, \Sigma \sim \mathrm{N}_d\big(\beta^T x_i, \Sigma/q_i\big)$$

$$q_i | \beta, \Sigma \sim H(\cdot) \ .$$

Now write the vector of missing data as $q = (q_1, \dots, q_n)$ and assume (for convenience) that H has a derivative, call it h. Then we can write the joint density of these iid pairs as

$$f(y, q \big| \beta, \Sigma) = \prod_{i=1}^n f(y_i | q_i, \beta, \Sigma) h(q_i) \ , \tag{3}$$

and the corresponding y-marginal is given by

$$\begin{aligned} f(y \big| \beta, \Sigma) &= \int_{\mathbb{R}_+^n} f(y, q | \beta, \Sigma) \, dq \\ &= \prod_{i=1}^n \left[\int_0^\infty \frac{\delta^{\frac{d}{2}}}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left\{ -\frac{\delta}{2} \Big(y_i - \beta^T x_i\Big)^T \Sigma^{-1} \Big(y_i - \beta^T x_i\Big) \right\} h(\delta) \, d\delta \right] , \end{aligned} \tag{4}$$

which is the same as (2). This shows that the marginal density of y has not been altered by the introduction of q into the model. Using (4) and Fubini's theorem, we have

$$c(y) = \int_W \int_{\mathbb{R}^{dk}} \left[\int_{\mathbb{R}_+^n} f(y, q | \beta, \Sigma) \, dq \right] \pi(\beta, \Sigma) \, d\beta \, d\Sigma = \int_{\mathbb{R}_+^n} \int_W \int_{\mathbb{R}^{dk}} f(y, q | \beta, \Sigma) \pi(\beta, \Sigma) \, d\beta \, d\Sigma \, dq \ , \tag{5}$$

where $\mathbb{R}_+ := (0, \infty)$. As we will see, this representation simplifies things since two of the three integrals on the right-hand side of (5) can be evaluated in closed.

Alternative proof of the sufficiency part of Proposition 1. First, note that

$$
\begin{aligned}
& f(y,q|\beta,\Sigma)\pi(\beta,\Sigma) \\
&\quad = \frac{1}{(2\pi)^{\frac{nd}{2}}|\Sigma|^{\frac{n}{2}}} \left[\prod_{i=1}^n q_i^{\frac{d}{2}} \exp\left\{ -\frac{q_i}{2}(\beta^T x_i - y_i)^T \Sigma^{-1} (\beta^T x_i - y_i) \right\} \right] \left[\prod_{i=1}^n h(q_i) \right] |\Sigma|^{-\frac{d+1}{2}} \\
&\quad = \frac{|Q|^{-\frac{d}{2}}}{(2\pi)^{\frac{nd}{2}}|\Sigma|^{\frac{n+d+1}{2}}} \left[\exp\left\{ -\frac{1}{2}\sum_{i=1}^n q_i(\beta^T x_i - y_i)^T \Sigma^{-1} (\beta^T x_i - y_i) \right\} \right] \left[\prod_{i=1}^n h(q_i) \right],
\end{aligned}
$$

where Q denotes an $n \times n$ diagonal matrix whose ith diagonal element is q_i^{-1}. Now

$$
\begin{aligned}
& \sum_{j=1}^n q_j(\beta^T x_j - y_j)^T \Sigma^{-1} (\beta^T x_j - y_j) \\
&\qquad = \sum_{j=1}^n q_j \mathrm{tr}\Big((\beta^T x_j - y_j)^T \Sigma^{-1} (\beta^T x_j - y_j) \Big) \\
&\qquad = \sum_{j=1}^n q_j \mathrm{tr}\Big(\Sigma^{-1} (\beta^T x_j - y_j)(\beta^T x_j - y_j)^T \Big) \\
&\qquad = \mathrm{tr}\Big(\Sigma^{-1} \sum_{j=1}^n q_j (\beta^T x_j - y_j)(\beta^T x_j - y_j)^T \Big) \\
&\qquad = \mathrm{tr}\Big[\Sigma^{-1} \Big(\beta^T X^T - y^T\Big) Q^{-1} \Big(\beta^T X^T - y^T\Big)^T \Big] \\
&\qquad = \mathrm{tr}\Big[\Sigma^{-1} \Big(\beta^T X^T Q^{-1} X \beta - \beta^T X^T Q^{-1} y - y^T Q^{-1} X \beta + y^T Q^{-1} y\Big) \Big] \\
&\qquad = \mathrm{tr}\Big[\Sigma^{-1} \Big\{ \Big(\beta^T - \mu^T\Big) \Omega^{-1} \Big(\beta^T - \mu^T\Big)^T - \mu^T \Omega^{-1} \mu + y^T Q^{-1} y \Big\} \Big] \\
&\qquad = \mathrm{tr}\Big[\Sigma^{-1} \Big(\beta^T - \mu^T\Big) \Omega^{-1} \Big(\beta^T - \mu^T\Big)^T \Big] + \mathrm{tr}\Big[\Sigma^{-1} \Big(y^T Q^{-1} y - \mu^T \Omega^{-1} \mu \Big) \Big],
\end{aligned}
$$

where $\Omega = (X^T Q^{-1} X)^{-1}$ and $\mu = (X^T Q^{-1} X)^{-1} X^T Q^{-1} y$. Thus, as a function of β,

$$
f(y,q|\beta,\Sigma)\pi(\beta,\Sigma) \propto \exp\left\{ -\frac{1}{2}\mathrm{tr}\Big[\Sigma^{-1} \Big(\beta^T - \mu^T\Big) \Omega^{-1} \Big(\beta^T - \mu^T\Big)^T \Big] \right\}. \tag{6}
$$

Using results from Appendix A concerning the matrix normal density, we have

$$
\int_{\mathbb{R}^{dk}} \exp\left\{ -\frac{1}{2}\mathrm{tr}\Big[\Sigma^{-1} \Big(\beta^T - \mu^T\Big) \Omega^{-1} \Big(\beta^T - \mu^T\Big)^T \Big] \right\} d\beta^T = (2\pi)^{\frac{kd}{2}} |\Sigma|^{\frac{k}{2}} |\Omega|^{\frac{d}{2}}.
$$

Thus, we have shown that

$$
\begin{aligned}
& \int_{\mathbb{R}^{dk}} f(y,q|\beta,\Sigma)\pi(\beta,\Sigma)\, d\beta^T \\
&\quad = \frac{|\Omega|^{\frac{d}{2}} |Q|^{-\frac{d}{2}}}{(2\pi)^{\frac{d(n-k)}{2}} |\Sigma|^{\frac{n-k+d+1}{2}}} \exp\left\{ -\frac{1}{2}\mathrm{tr}\Big[\Sigma^{-1} \Big(y^T Q^{-1} y - \mu^T \Omega^{-1} \mu \Big) \Big] \right\} \left[\prod_{i=1}^n h(q_i) \right].
\end{aligned} \tag{7}
$$

Now assume that $n \geq d+k$. We first establish that, for ϕ-almost all y, the matrix $y^T Q^{-1} y - \mu^T \Omega^{-1} \mu$ is positive definite. We use the following representation:

$$\begin{aligned} y^T Q^{-1} y - \mu^T \Omega^{-1} \mu &= y^T Q^{-1} y - y^T Q^{-1} X (X^T Q^{-1} X)^{-1} X^T Q^{-1} y \\ &= y^T Q^{-\frac{1}{2}} \left[I - Q^{-\frac{1}{2}} X (X^T Q^{-1} X)^{-1} X^T Q^{-\frac{1}{2}} \right] Q^{-\frac{1}{2}} y \ . \end{aligned}$$

Since $I - Q^{-\frac{1}{2}} X (X^T Q^{-1} X)^{-1} X^T Q^{-\frac{1}{2}}$ is idempotent, the matrix above is positive semi-definite. We can now establish positive definiteness by showing that its determinant is nonzero. To this end, let Λ be the $n \times (d+k)$ augmented matrix $(X : y)$. Then,

$$\begin{aligned} \Lambda^T Q^{-1} \Lambda &= \begin{bmatrix} X^T \\ y^T \end{bmatrix} Q^{-1} \begin{bmatrix} X & y \end{bmatrix} \\ &= \begin{bmatrix} X^T Q^{-1} X & X^T Q^{-1} y \\ y^T Q^{-1} X & y^T Q^{-1} y \end{bmatrix} \ . \end{aligned}$$

Therefore,

$$\begin{aligned} |\Lambda^T Q^{-1} \Lambda| &= |X^T Q^{-1} X| |y^T Q^{-1} y - y^T Q^{-1} X (X^T Q^{-1} X)^{-1} X^T Q^{-1} y| \\ &= |X^T Q^{-1} X| |y^T Q^{-1} y - \mu^T \Omega^{-1} \mu| \\ &= |\Omega|^{-1} |y^T Q^{-1} y - \mu^T \Omega^{-1} \mu| \ . \end{aligned} \tag{8}$$

Let S denote the set of ys that lead to linear dependencies among the columns of Λ, and note that S has ϕ-measure zero. Throughout the remainder of this proof, we assume that the observed data, y, are not from this set of measure zero. Under this assumption, Λ has full column rank, which implies that $|\Lambda^T Q^{-1} \Lambda| > 0$. It now follows from (8) that $|y^T Q^{-1} y - \mu^T \Omega^{-1} \mu| > 0$ and hence $y^T Q^{-1} y - \mu^T \Omega^{-1} \mu$ is positive definite. We may now apply the results from Appendix A concerning the inverse Wishart distribution to conclude that

$$\int_W \int_{\mathbb{R}^{dk}} f(y, q|\beta, \Sigma) \pi(\beta, \Sigma) \, d\beta^T \, d\Sigma = \frac{|\Omega|^{\frac{d}{2}} \left[\prod_{i=1}^n h(q_i) \right] \left[\prod_{j=1}^d \Gamma\left(\frac{1}{2}(n-k+1-j) \right) \right]}{\pi^{\frac{d(2(n-k)-d+1)}{4}} |Q|^{\frac{d}{2}} \left| y^T Q^{-1} y - \mu^T \Omega^{-1} \mu \right|^{\frac{n-k}{2}}} \ . \tag{9}$$

Now suppose that $n = d+k$. Then, using (8) we have

$$\begin{aligned} \int_W \int_{\mathbb{R}^{dk}} f(y, q|\beta, \Sigma) \pi(\beta, \Sigma) \, d\beta^T \, d\Sigma &= \frac{\left[\prod_{i=1}^n h(q_i) \right] \left[\prod_{j=1}^d \Gamma\left(\frac{1}{2}(d+1-j) \right) \right]}{\pi^{d(d+1)/4} |\Lambda^T Q^{-1} \Lambda|^{\frac{d}{2}} |Q|^{\frac{d}{2}}} \\ &= \frac{\left[\prod_{i=1}^n h(q_i) \right] \left[\prod_{j=1}^d \Gamma\left(\frac{1}{2}(d+1-j) \right) \right]}{\pi^{d(d+1)/4} |\Lambda|^d} \ . \end{aligned} \tag{10}$$

Note that, as a function of q, the expression above is proportional to $\prod_{i=1}^n h(q_i)$. Hence,

$$\int_{\mathbb{R}_+^n}\int_W\int_{\mathbb{R}^{dk}} f(y,q|\beta,\Sigma)\pi(\beta,\Sigma)\,d\beta^T\,d\Sigma\,dq = \frac{\prod_{j=1}^d \Gamma\big(\frac{1}{2}(d+1-j)\big)}{\pi^{d(d+1)/4}|\Lambda|^d} < \infty\ .$$

Therefore, when $n = d+k$, the posterior density is finite for all y outside a set of ϕ-measure zero. The sufficiency part of the theorem now follows directly from Marchev and Hobert's (2004) Lemma 2, which states that, if a posterior density is proper at a fixed sample size, say n', then it is also proper for all sample sizes larger than n'. □

We now show how the proof above leads directly to a method of making exact draws from $\pi(\beta,\Sigma|y)$ when $n = d+k$. The missing data model allows us to construct a so-called complete data posterior density whose marginal is the target posterior density. For the time being, assume that $n \geq d+k$ and that $c(y) < \infty$. Now define $\pi(q,\beta,\Sigma|y) = f(y,q|\beta,\Sigma)\pi(\beta,\Sigma)\big/c(y)$ and note that

$$\int_{\mathbb{R}_+^n} \pi(q,\beta,\Sigma|y)\,dq = \pi(\beta,\Sigma|y)\ .$$

It follows that a draw from $\pi(q,\beta,\Sigma|y)$ provides us with a draw from $\pi(\beta,\Sigma|y)$ (just throw away the q component). We now explain how to draw from $\pi(q,\beta,\Sigma|y)$ using the representation

$$\pi(q,\beta,\Sigma|y) = \pi(\beta|\Sigma,q,y)\pi(\Sigma|q,y)\pi(q|y)\ . \tag{11}$$

If we could sample from each of the three densities on the right-hand side of (11), then we could make exact draws from $\pi(q,\beta,\Sigma|y)$. Two of these three are easy to sample. First,

$$\pi(\beta|\Sigma,q,y) = \frac{\pi(q,\beta,\Sigma|y)}{\int_{R^{dk}}\pi(q,\beta,\Sigma|y)\,d\beta} \propto f(y,q|\beta,\Sigma)\pi(\beta,\Sigma)\ ,$$

so it follows from (6) that

$$\beta^T|\Sigma,q,y \sim \mathrm{N}_{d,k}\big(\mu^T,\Sigma,\Omega\big)\ ,$$

where, again, $\mu = (X^TQ^{-1}X)^{-1}X^TQ^{-1}y$ and $\Omega = (X^TQ^{-1}X)^{-1}$. Similarly, it follows from (7) that

$$\Sigma|q,y \sim \mathrm{IW}_d\bigg(n-k, \Big(y^TQ^{-1}y - \mu^T\Omega^{-1}\mu\Big)^{-1}\bigg)\ .$$

However, (9) implies that

$$\pi(q|y) \propto \frac{|\Omega|^{\frac{d}{2}}\Big[\prod_{i=1}^n h(q_i)\Big]}{|Q|^{\frac{d}{2}}\Big|y^TQ^{-1}y - \mu^T\Omega^{-1}\mu\Big|^{\frac{n-k}{2}}}\ ,$$

which is a highly intractable n-dimensional density. Thus, in general, it is impossible to sample from $\pi(q|y)$. However, in the special case where $n = d + k$, this density is *extremely* simple. Indeed, (10) shows that, in this special case,

$$\pi(q|y) = \prod_{i=1}^{n} h(q_i) \ ,$$

so $q_1, \ldots, q_n$ are iid with density h. We conclude that, when $n = d+k$, an exact draw from $\pi(q, \beta, \Sigma|y)$ can be made by performing the following three steps:

1. Draw $q_1, q_2, \ldots q_n$ independently with $q_i \sim h$
2. Draw $\Sigma \sim \text{IW}_d\Big(n-k, \big(y^T Q^{-1} y - \mu^T \Omega^{-1} \mu\big)^{-1}\Big)$
3. Draw $\beta^T \sim \text{N}_{d,k}\big(\mu^T, \Sigma, \Omega\big)$

Standard statistical packages like R (R Development Core Team, 2006) have functions for generating random matrices from the inverse Wishart distribution. Appendix A describes a simple method of simulating from the matrix normal distribution.

Unfortunately, it appears that there are no efficient methods of making exact draws from the posterior density when $n > d + k$. In this case, we must resort to MCMC algorithms and this is the topic of the next section.

3 The DA and Haar PX-DA Algorithms

Liu (1996) developed a DA algorithm for exploring $\pi(\beta, \Sigma|y)$ that is based on the missing data model from the previous section. The algorithm simulates a Markov chain, $\{(\beta_m, \Sigma_m)\}_{m=0}^{\infty}$, with Markov transition density

$$k(\beta', \Sigma'|\beta, \Sigma) = \int_{\mathbb{R}_+^n} \pi(\beta', \Sigma'|q, y)\pi(q|\beta, \Sigma, y)\, dq \ .$$

Note that, while the Markov transition density does depend on the data, y, and the matrix X, these quantities are fixed, so this dependence is suppressed in the notation. Simulating this Markov chain is straightforward. Indeed, given the value of the current state, (β_m, Σ_m), we move to the next state, $(\beta_{m+1}, \Sigma_{m+1})$, by first drawing q from $\pi(q|\beta_m, \Sigma_m, y)$ and then using this new value of q to draw $(\beta_{m+1}, \Sigma_{m+1})$ from $\pi(\beta, \Sigma|q, y)$. Of course, we already know from the previous section how to draw from $\pi(\beta, \Sigma|q, y)$ by exploiting the factorization $\pi(\beta, \Sigma|q, y) = \pi(\beta|\Sigma, q, y)\, \pi(\Sigma|q, y)$. As for drawing from $\pi(q|\beta, \Sigma, y)$, note that

$$\pi(q|\beta, \Sigma, y) = \frac{\pi(q, \beta, \Sigma|y)}{\int_{\mathbb{R}_+^n} \pi(q, \beta, \Sigma|y)\, dq} \propto f(y, q|\beta, \Sigma) \ .$$

Hence, it follows from (3) that, conditional on (β, Σ, y) the elements of q are independent and q_i has density proportional to

$$\pi(q_i|\beta, \Sigma, y) \propto h(q_i)\, q_i^{\frac{d}{2}} \exp\left\{ -\frac{q_i}{2}(\beta^T x_i - y_i)^T \Sigma^{-1} (\beta^T x_i - y_i) \right\} . \tag{12}$$

Putting all of this together, a single iteration of the DA algorithm uses the current state (β, Σ) to produce the new state (β', Σ') through the following three steps:

1. Draw $q_1, q_2, \ldots q_n$ independently with q_i drawn from the univariate density proportional to the right-hand side of (12).
2. Draw $\Sigma' \sim \text{IW}_d\Big(n-k, \big(y^T Q^{-1} y - \mu^T \Omega^{-1} \mu\big)^{-1}\Big)$
3. Draw $\beta'^T \sim \text{N}_{d,k}\big(\mu^T, \Sigma', \Omega\big)$

No matter what the form of h, (12) is just a univariate density which can be sampled using rejection sampling. Moreover, in the important special case where $h(q_i)$ is a $\text{Gamma}\big(\frac{\nu}{2}, \frac{\nu}{2}\big)$ density, sampling from $\pi(q_i|\beta, \Sigma, y)$ is especially simple since

$$q_i|\beta, \Sigma, y \sim \text{Gamma}\left(\frac{\nu+d}{2}, \frac{\nu + (\beta^T x_i - y_i)^T \Sigma^{-1} (\beta^T x_i - y_i)}{2}\right) .$$

Recall from the previous section that an exact draw from the posterior density $\pi(\beta, \Sigma|y)$ can (at least theoretically) be made by drawing sequentially from $\pi(q|y)$, $\pi(\Sigma|q, y)$ and $\pi(\beta|\Sigma, q, y)$. Note that one iteration of the DA algorithm also involves simulating from $\pi(\Sigma|q, y)$ and $\pi(\beta|\Sigma, q, y)$, but, instead of a draw from $\pi(q|y)$, the DA algorithm only requires n independent univariate draws. It is interesting that, if we are willing to settle for a Markov chain that converges to $\pi(\beta, \Sigma|y)$ instead of an exact draw from $\pi(\beta, \Sigma|y)$, then the complicated draw from $\pi(q|y)$ can be replaced by the much simpler task of making n univariate draws from (12).

The basic theory of DA algorithms implies that the Markov chain $\{(\beta_m, \Sigma_m)\}_{m=0}^{\infty}$ is reversible with respect to the posterior density $\pi(\beta, \Sigma|y)$; i.e., for all $(\beta, \Sigma), (\beta', \Sigma') \in R^{dk} \times W$, we have

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

It follows immediately that the posterior density is invariant for the chain; i.e.,

$$\int_W \int_{\mathbb{R}^{dk}} k(\beta, \Sigma|\beta', \Sigma')\pi(\beta', \Sigma'|y)\, d\beta'\, d\Sigma' = \pi(\beta, \Sigma|y) .$$

Moreover, Roy (2008) shows that this Markov chain is Harris ergodic; that is, irreducible, aperiodic and Harris recurrent. Harris ergodicity implies that the chain converges to its stationary distribution and that ergodic averages based on the DA algorithm converge almost surely to their population counterparts, which are, of course, posterior expectations with respect to $\pi(\beta, \Sigma|y)$.

Over last decade, several authors have shown that it is possible to drastically improve the convergence behavior of DA algorithms by adding a computationally simple "extra step" (see, e.g., Hobert and Marchev, 2008; Liu and Wu, 1999; Meng and van Dyk, 1999; van Dyk and Meng, 2001). Indeed, suppose that $R(q, dq')$ is a Markov transition function on $\mathbb{R}_+^n$ that is reversible with respect to $\pi(q \,|\, y)$. Consider adding an extra step to the DA algorithm where, after q is drawn in the first step, we move to a new value, $q' \sim R(q, \cdot)$, before drawing new values of β and Σ. To be more specific, let $\{(\tilde{\beta}_m, \tilde{\Sigma}_m)\}_{m=0}^{\infty}$ be a new Markov chain that proceeds from the current state $(\tilde{\beta}, \tilde{\Sigma})$ to the next state $(\tilde{\beta}', \tilde{\Sigma}')$ via the following *four* steps

1. Draw $q_1, q_2, \ldots, q_n$ independently with q_i drawn from the density proportional to the right-hand side of (12).
2. Draw $q' \sim R(q, \cdot)$
3. Draw $\Sigma' \sim \mathrm{IW}_d\left(n - k, \left(y^T Q'^{-1} y - \mu'^T \Omega'^{-1} \mu'\right)^{-1}\right)$
4. Draw $\beta'^T \sim \mathrm{N}_{d,k}\left(\mu'^T, \Sigma', \Omega'\right)$

where Q' is a diagonal matrix with the reciprocals of the q_i' on the diagonal, and μ' and Ω' are just μ and Ω with Q' in place of Q.

Hobert and Marchev (2008) proved that, if R is constructed according to a specific recipe involving group actions and left Haar measure, then the MCMC algorithm based on the new Markov chain is at least as good as the original DA algorithm in that the new chain converges faster and has smaller asymptotic variances in the CLT. They called the resulting MCMC algorithm the *Haar PX-DA algorithm*. There are currently no theoretical results that *quantify* the improvement of Haar PX-DA over DA. However, there is a great deal of *empirical* evidence suggesting that Haar PX-DA algorithms can be far superior to the DA algorithms upon which they are based (Roy and Hobert, 2007; van Dyk and Meng, 2001). We now build a Haar PX-DA algorithm for our problem.

Let G be the multiplicative group $\mathbb{R}_+$ where group composition is defined as multiplication. This is a unimodular group, so the left and right Haar measures are the same and are given by $\nu(dg) = dg/g$, where dg denotes Lebesgue measure on $\mathbb{R}_+$. Allow G to act on the space $\mathbb{R}_+^n$ through component-wise multiplication; i.e., $gq = (gq_1, gq_2, \cdots, gq_n)$. Note that, if $g \in G$ and $h : \mathbb{R}_+^n \to \mathbb{R}$ is an integrable function (with respect to Lebesgue measure), then

$$\int_{\mathbb{R}_+^n} h(z)dz = g^n \int_{\mathbb{R}_+^n} h(gz)dz \,.$$

This shows that Lebesgue measure on $\mathbb{R}_+^n$ is relatively invariant with *multiplier* $\chi(g) = g^n$. (See Eaton (1989) for details on the group theory we are using here.) To construct the Haar PX-DA algorithm using the group structure introduced above, we must first demonstrate that there is a probability density (with respect to Haar measure) that is proportional to $\pi(gq \,|\, y)\,\chi(g)$. In other words, we must show that $m(q) = \int_G \pi(gq \,|\, y)\,\chi(g)\,\nu(dg) < \infty$ for all $q \in \mathbb{R}_+^n$. Note that we can re-express $\pi(q \,|\, y)$ as follows:

$$\pi(q \,|\, y) \propto \left| \frac{X^T Q^{-1} X}{q.} \right|^{-\frac{d}{2}} \left| \frac{y^T Q^{-1} y - y^T Q^{-1} X (X^T Q^{-1} X)^{-1} X^T Q^{-1} y}{q.} \right|^{-\frac{n-k}{2}} \prod_{i=1}^n \left(\frac{q_i}{q.} \right)^{\frac{d}{2}} h(q_i) \; ,$$

where $q. = \sum_{i=1}^n q_i$. Therefore, as a function of g, $\pi(gq \,|\, y) = c \prod_{i=1}^n h(gq_i)$ where c is a constant. Hence, using the transformation $g \to 1/g$, we have

$$m(q) = c \int_0^\infty \left[\prod_{i=1}^n h(gq_i) \right] g^n \frac{dg}{g} = c \int_0^\infty \left[\prod_{i=1}^n \frac{1}{\sigma} h\Big(\frac{q_i}{\sigma} \Big) \right] \frac{d\sigma}{\sigma} \; . \tag{13}$$

We now use a simple Bayesian statistical argument to show that $m(q)$ is finite. Indeed, suppose that W is a random variable from the scale family $\frac{1}{\sigma} h\big(\frac{w}{\sigma}\big)$ and, as a prior density on the scale parameter σ we use $1/\sigma$. Then the posterior density of σ given w is proper because

$$\int_0^\infty \frac{1}{\sigma} h\Big(\frac{w}{\sigma}\Big) \frac{d\sigma}{\sigma} = \frac{1}{w} \int_0^\infty h(u)\, du = \frac{1}{w} < \infty \; .$$

Now, the final integral in (13) is just the marginal density associated with a random sample of size n from $\frac{1}{\sigma} h\big(\frac{w}{\sigma}\big)$. Since the posterior is proper for a sample of size one, it is proper for any larger sample size (Marchev and Hobert, 2004, Lemma 2), and it follows that $m(q) < \infty$ for all $q \in \mathbb{R}_+^n$.

We now describe the Haar PX-DA algorithm by providing a detailed description of Step 2. Define a density on G (with respect to Lebesgue measure on $\mathbb{R}_+$) as

$$e_q(g) = \frac{g^{n-1}}{m(q)} \prod_{i=1}^n h(gq_i) \; .$$

In Step 2, we make the transition $q \to q'$ by drawing $g \sim e_q(g)$ and setting $q' = gq$. In the special case where the regression errors have a multivariate student's t distribution, it's easy to show that the density e_q is Gamma$(n\nu/2, \nu q./2)$. Hence, the extra step, which is the sole difference between Haar PX-DA and DA, is just a single draw from a univariate gamma density. Therefore, in terms of computer time per iteration, the Haar PX-DA algorithm is essentially equivalent to the DA algorithm. Roy (2008) also establishes that Harris ergodicity of the Haar PX-DA algorithm.

In the next section, we prove that when the regression errors have a Student's t distribution, the Markov chain underlying the DA algorithm converges to its stationary distribution at a geometric rate.

This implies the same of the Haar PX-DA algorithm. These results have important practical implications. For example, geometric ergodicity implies that ergodic averages obey central limit theorems, which allow for the calculation of valid asymptotic standard errors (see, e.g., Flegal et al., 2008; Jones et al., 2006; Jones and Hobert, 2001).

4 Convergence Rates of the DA and Haar PX-DA Algorithms

We now study the convergence rates of the MCMC algorithms introduced in the previous section in the important special case where the regression errors have a Student's t distribution. In particular, we prove that Liu's (1996) DA algorithm converges at a geometric rate. This result is then used in conjunction with results in Hobert and Marchev (2008) to show that the Haar PX-DA algorithm is also geometric.

We begin by defining what it means for the DA algorithm to converge at a geometric rate. For $m \in \{1, 2, 3, \dots\}$, let $P^m(\cdot, \cdot)$ denote the m-step Markov transition function associated with the Markov chain $\{(\beta_m, \Sigma_m)\}_{m=0}^{\infty}$. That is, for $(\beta, \Sigma) \in R^{dk} \times W$ and a measurable set $A \subset R^{dk} \times W$,

$$P^m\big((\beta, \Sigma), A\big) = \Pr\big((\beta_m, \Sigma_m) \in A | (\beta_0, \Sigma_0) = (\beta, \Sigma)\big) \; .$$

As usual, we write P instead of P^1. Note that if $A_1 \in R^{dk}$ and $A_2 \in W$, then

$$P\big((\beta, \Sigma), A_1 \times A_2\big) = \int_{A_2} \int_{A_1} k(\beta', \Sigma' | \beta, \Sigma) \, d\beta' \, d\Sigma' \; .$$

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

$$\big\| P^m\big((\beta, \Sigma), \cdot\big) - \Pi(\cdot|y) \big\| \downarrow 0 \quad \text{as } m \to \infty \; .$$

Note that there is no information about the *rate* of convergence in the above statement. The Markov chain is called geometrically ergodic if there exists a function $M : R^{dk} \times W \to [0, \infty)$ and a constant $\rho \in (0, 1)$ such that, for all m,

$$\big\| P^m\big((\beta, \Sigma), \cdot\big) - \Pi(\cdot|y) \big\| \leq M(\beta, \Sigma) \rho^m \; .$$

We will establish geometric ergodicity of the DA algorithm using the so-called drift and minorization technique, which requires the construction of a *drift condition* and an associated *minorization condition* (Meyn and Tweedie, 1993; Rosenthal, 1995). (See Jones and Hobert (2001) for a gentle introduction to these ideas.) For the drift condition, we must find a function $V : R^{dk} \times W \to \mathbb{R}_+$, $\lambda \in [0,1)$, and $L \in \mathbb{R}$ such that, for all $(\beta,\Sigma) \in R^{dk} \times W$,

$$(PV)(\beta,\Sigma) \leq \lambda V(\beta,\Sigma) + L$$

where $(PV)(\beta,\Sigma)$ is the expected value of $V(\beta_1,\Sigma_1)$ given that the starting value is $(\beta_0,\Sigma_0) = (\beta,\Sigma)$. More formally,

$$(PV)(\beta,\Sigma) = \int_W \int_{R^{dk}} V(\beta',\Sigma')\, k(\beta',\Sigma'|\beta,\Sigma)\, d\beta'\, d\Sigma' \;.$$

We now describe the minorization condition. First, let

$$C = \Big\{(\beta,\Sigma) \in R^{dk} \times W : V(\beta,\Sigma) \leq l\Big\} \;,$$

for some $l > 2L/(1-\lambda)$. For the associated minorization condition, we must find a density function $u : R^{dk} \times W \to [0,\infty)$ and an $\varepsilon \in (0,1]$ such that, for all $(\beta',\Sigma') \in C$ and all $(\beta,\Sigma) \in R^{dk} \times W$,

$$k(\beta,\Sigma|\beta',\Sigma') \geq \varepsilon\, u(\beta,\Sigma) \;.$$

If we can establish these two conditions, then the Markov chain is geometrically ergodic by Theorem 12 of Rosenthal (1995). We begin with the drift condition.

Lemma 1. *Let* $V(\beta,\Sigma) = \sum_{i=1}^n (y_i - \beta^T x_i)^T \Sigma^{-1} (y_i - \beta^T x_i)$ *and assume that* $\nu + d > 2$. *Then the Markov chain underlying the DA algorithm satisfies the following*

$$(PV)(\beta,\Sigma) \leq \frac{n+d-k}{\nu+d-2}\, V(\beta,\Sigma) + \frac{n\nu(n+d-k)}{\nu+d-2} \;.$$

Proof. First, by Fubini's theorem, we have

$$\begin{aligned}
(PV)(\beta',\Sigma') &= \int_W \int_{R^{dk}} V(\beta,\Sigma)\, k(\beta,\Sigma \,|\, \beta',\Sigma')\, d\beta\, d\Sigma \\
&= \int_{\mathbb{R}_+^n} \int_W \int_{R^{dk}} V(\beta,\Sigma)\, \pi(\beta,\Sigma \,|\, q,y)\, \pi(q \,|\, \beta',\Sigma',y)\, d\beta\, d\Sigma\, dq \\
&= \int_{\mathbb{R}_+^n} \int_W \int_{R^{dk}} V(\beta,\Sigma)\, \pi(\beta \,|\, \Sigma,q,y)\, \pi(\Sigma \,|\, q,y)\, \pi(q \,|\, \beta',\Sigma',y)\, d\beta\, d\Sigma\, dq \\
&= \int_{\mathbb{R}_+^n} \left\{ \int_W \left[\int_{R^{dk}} V(\beta,\Sigma)\, \pi(\beta \,|\, \Sigma,q,y)\, d\beta \right] \pi(\Sigma \,|\, q,y)\, d\Sigma \right\} \pi(q \,|\, \beta',\Sigma',y)\, dq \;. \qquad (14)
\end{aligned}$$

The inner-most integral in (14) can be viewed as an expectation with respect to $\pi(\beta \,|\, \Sigma,q,y)$, which is a matrix normal density. The integral in curly brackets can then be viewed as an expectation with respect to $\pi(\Sigma \,|\, q,y)$, which is an inverse Wishart density. And, finally, the outer-most integral is an expectation with respect to $\pi(q \,|\, \beta',\Sigma',y)$, which is a product of univariate gamma densities. Hence, we write

$$(PV)(\beta',\Sigma') = E\Big[E\Big\{E\big(V(\beta,\Sigma) \,\big|\, \Sigma,q,y\big) \,\Big|\, q,y\Big\} \,\Big|\, \beta',\Sigma',y\Big] \;.$$

Starting with the innermost expectation, we have

$$
\begin{aligned}
&E\big(V(\beta,\Sigma)\,\big|\,\Sigma,q,y\big)\\
&\quad = E\Bigg[\sum_{i=1}^{n} y_i^T\Sigma^{-1}y_i - \sum_{i=1}^{n}\big(y_i^T\Sigma^{-1}\beta^T x_i + x_i^T\beta\Sigma^{-1}y_i\big) + \sum_{i=1}^{n} x_i^T\beta\Sigma^{-1}\beta^T x_i \,\Bigg|\, \Sigma,q,y\Bigg]\\
&\quad = \sum_{i=1}^{n} y_i^T\Sigma^{-1}y_i - \sum_{i=1}^{n} E\Big[\big(y_i^T\Sigma^{-1}\beta^T x_i + x_i^T\beta\Sigma^{-1}y_i\big)\,\Big|\,\Sigma,q,y\Big] + \sum_{i=1}^{n} E\Big[x_i^T\beta\Sigma^{-1}\beta^T x_i\,\big|\,\Sigma,q,y\Big]\ .
\end{aligned}
$$

Recall that $\beta^T\,|\,\Sigma,q,y \sim \mathrm{N}_{d,k}\big(\mu^T,\Sigma,\Omega\big)$. Hence, $E\big(\beta^T\,\big|\,\Sigma,q,y\big) = \mu^T$ and results in Arnold (1981, chap 17) show that

$$E\big(\beta\Sigma^{-1}\beta^T\,\big|\,\Sigma,q,y\big) = d\Omega + \mu\Sigma^{-1}\mu^T\ .$$

It follows that

$$
\begin{aligned}
&E\big(V(\beta,\Sigma)\,\big|\,\Sigma,q,y\big)\\
&\qquad = \sum_{i=1}^{n} y_i^T\Sigma^{-1}y_i - \sum_{i=1}^{n}\big(y_i^T\Sigma^{-1}\mu^T x_i + x_i^T\mu\Sigma^{-1}y_i\big) + \sum_{i=1}^{n} x_i^T\big[d\Omega + \mu\Sigma^{-1}\mu^T\big]x_i\\
&\qquad = \sum_{i=1}^{n}(y_i-\mu^T x_i)^T\Sigma^{-1}(y_i-\mu^T x_i) + d\sum_{i=1}^{n} x_i^T\Omega x_i\ .
\end{aligned}
$$

Now, recall that

$$\Sigma\,|\,q,y \sim \mathrm{IW}_d\Big(n-k,\big(y^TQ^{-1}y-\mu^T\Omega^{-1}\mu\big)^{-1}\Big)\ .$$

Results in Mardia, Kent and Bibby (1979, p. 85) imply that

$$E\big(\Sigma^{-1}\,|\,q,y\big) = (n-k)\big(y^TQ^{-1}y-\mu^T\Omega^{-1}\mu\big)^{-1}\ .$$

Therefore,

$$
\begin{aligned}
&E\Big\{E\big(V(\beta,\Sigma)\,\big|\,\Sigma,q,y\big)\,\big|\,q,y\Big\}\\
&\qquad = \sum_{i=1}^{n}(y_i-\mu^T x_i)^T E\big(\Sigma^{-1}\,|\,q,y\big)(y_i-\mu^T x_i) + d\sum_{i=1}^{n} x_i^T\Omega x_i\\
&\qquad = (n-k)\sum_{i=1}^{n}(y_i-\mu^T x_i)^T\big(y^TQ^{-1}y-\mu^T\Omega^{-1}\mu\big)^{-1}(y_i-\mu^T x_i) + d\sum_{i=1}^{n} x_i^T\Omega x_i\ .
\end{aligned}
\tag{15}
$$

Note that we were able to compute the first two conditional expectations exactly. Unfortunately, we cannot compute the expectation of (15) with respect to $\pi(q\,|\,\beta',\Sigma',y)$ in closed form. We instead compute the expectation of a simple upper bound on (15). First, note that

$$\begin{aligned} y^T Q^{-1} y - \mu^T \Omega^{-1} \mu &= y^T Q^{-1} y + \mu^T \Omega^{-1} \mu - 2\mu^T \Omega^{-1} \mu \\ &= \sum_{i=1}^n q_i y_i y_i^T + \mu^T (X^T Q^{-1} X)\mu - 2\mu^T X^T Q^{-1} y \\ &= \sum_{i=1}^n q_i y_i y_i^T + \mu^T \left(\sum_{i=1}^n q_i x_i x_i^T \right) \mu - 2\mu^T \sum_{i=1}^n q_i x_i y_i^T \\ &= \sum_{i=1}^n q_i (y_i - \mu^T x_i)(y_i - \mu^T x_i)^T \ . \end{aligned}$$

So,

$$\begin{aligned} &(y_i - \mu^T x_i)^T \big(y^T Q^{-1} y - \mu^T \Omega^{-1} \mu \big)^{-1} (y_i - \mu^T x_i) \\ &\qquad = (y_i - \mu^T x_i)^T \left(\sum_{j=1}^n q_j (y_j - \mu^T x_j)(y_j - \mu^T x_j)^T \right)^{-1} (y_i - \mu^T x_i) \\ &\qquad = \frac{1}{q_i} (y_i - \mu^T x_i)^T \left(\sum_{j=1}^n \frac{q_j}{q_i} (y_j - \mu^T x_j)(y_j - \mu^T x_j)^T \right)^{-1} (y_i - \mu^T x_i) \ . \end{aligned}$$

Now,

$$\frac{1}{q_i} \big(y^T Q^{-1} y - \mu^T \Omega^{-1} \mu \big) = \sum_{j=1}^n \frac{q_j}{q_i} (y_j - \mu^T x_j)(y_j - \mu^T x_j)^T \ ,$$

which is positive definite. Also,

$$\sum_{j=1}^n \frac{q_j}{q_i} (y_j - \mu^T x_j)(y_j - \mu^T x_j)^T - (y_i - \mu^T x_i)(y_i - \mu^T x_i)^T = \sum_{j \neq i} \frac{q_j}{q_i} (y_j - \mu^T x_j)(y_j - \mu^T x_j)^T,$$

which is is positive semidefinite. Hence, we may apply Lemma 3 from Appendix B to conclude that

$$(y_i - \mu^T x_i)^T \big(y^T Q^{-1} y - \mu^T \Omega^{-1} \mu \big)^{-1} (y_i - \mu^T x_i) \leq \frac{1}{q_i} \ .$$

Therefore,

$$(n-k) \sum_{i=1}^n (y_i - \mu^T x_i)^T \big(y^T Q^{-1} y - \mu^T \Omega^{-1} \mu \big)^{-1} (y_i - \mu^T x_i) \leq (n-k) \sum_{i=1}^n \frac{1}{q_i} \ .$$

We now use a similar argument to bound the term $d \sum_{i=1}^n x_i^T \Omega x_i$. Since X has full column rank, $\frac{1}{q_i} X^T Q^{-1} X = \sum_{j=1}^n \frac{q_j}{q_i} x_j x_j^T$ is positive definite. Furthermore, it is clear that $\sum_{j \neq i} \frac{q_j}{q_i} x_j x_j^T$ is positive semidefinite. Another application of Lemma 3 yields

$$q_i x_i^T \Omega x_i = x_i^T \Big(\frac{1}{q_i} X^T Q^{-1} X \Big)^{-1} x_i = x_i^T \left(\sum_{j=1}^n \frac{q_j}{q_i} x_j x_j^T \right)^{-1} x_i \leq 1 \ .$$

It follows that

$$d\sum_{i=1}^{n} x_i^T \Omega x_i \le d\sum_{i=1}^{n} \frac{1}{q_i} \; .$$

Putting all of this together, we have

$$\begin{aligned} E\Big\{ & E\big(V(\beta,\Sigma) \,\big|\, \Sigma, q, y\big) \,\big|\, q, y\Big\} \\ &= (n-k)\sum_{i=1}^{n} (y_i - \mu^T x_i)^T \big(y^T Q^{-1} y - \mu^T \Omega^{-1} \mu\big)^{-1} (y_i - \mu^T x_i) + d\sum_{i=1}^{n} x_i^T \Omega x_i \\ &\le (n+d-k)\sum_{i=1}^{n} \frac{1}{q_i} \; . \end{aligned}$$

Now recall that $\pi(q \,|\, \beta', \Sigma', y)$ is a product of n univariate gamma densities with

$$q_i \,|\, \beta', \Sigma', y \sim \text{Gamma}\bigg(\frac{\nu + d}{2}, \frac{\nu + (\beta'^T x_i - y_i)^T \Sigma'^{-1} (\beta'^T x_i - y_i)}{2} \bigg) \; .$$

So, as long as $\nu + d > 2$, we have

$$E(q_i^{-1} \,|\, \beta', \Sigma', y) = \frac{(y_i - \beta'^T x_i)^T \Sigma'^{-1} (y_i - \beta'^T x_i) + \nu}{\nu + d - 2} \; .$$

Finally,

$$\begin{aligned} (PV)(\beta', \Sigma') &\le \frac{n+d-k}{\nu+d-2} \Bigg[n\nu + \sum_{i=1}^{n} (y_i - \beta'^T x_i)^T \Sigma'^{-1} (y_i - \beta'^T x_i) \Bigg] \\ &= \frac{n+d-k}{\nu+d-2} V(\beta', \Sigma') + \frac{n\nu(n+d-k)}{\nu+d-2} \; , \end{aligned}$$

and this completes the proof. □

Here is a formal statement of the required minorization condition.

Lemma 2. *Fix $l > 0$ and let $C = \big\{(\beta, \Sigma) : V(\beta, \Sigma) \le l\big\}$. For all $(\beta', \Sigma') \in C$ and all $(\beta, \Sigma) \in R^{dk} \times W$, we have*

$$k(\beta, \Sigma \,|\, \beta', \Sigma') \ge \varepsilon\, u(\beta, \Sigma) \; ,$$

where the density $u(\beta, \Sigma)$ is given by

$$u(\beta, \Sigma) = \int_{\mathbb{R}_n^+} \pi(\beta, \Sigma | q, y) \Bigg[\prod_{i=1}^{n} \frac{g(q_i)}{\int_0^\infty g(t)dt} \Bigg] dq$$

and $\varepsilon = \big(\int_0^\infty g(t)dt\big)^n$. The function $g(\cdot)$ is given by

$$g(t) = \Gamma\Big(\frac{\nu+d}{2}, \frac{\nu}{2}; t\Big) I_{(0,q^*)}(t) + \Gamma\Big(\frac{\nu+d}{2}, \frac{\nu+l}{2}; t\Big) I_{(q^*,\infty)}(t)$$

where $q^* = \frac{\nu+d}{l} \log\big(1 + \frac{l}{\nu}\big)$ and $\Gamma(a, b; x)$ denotes the Gamma(a,b) density evaluated at the point x.

Proof. For $i = 1, 2, \ldots, n$, define

$$C_i = \left\{ (\beta, \Sigma) : (y_i - \beta^T x_i)^T \Sigma^{-1} (y_i - \beta^T x_i) \leq l \right\} .$$

Clearly, $C \subset C_i$ for each i. Recall that, $\pi(q \,|\, \beta, \Sigma, y)$ is product of n univariate Gamma densities. Hence,

$$\begin{aligned}
\inf_{(\beta,\Sigma)\in C} \pi(q \,|\, \beta, \Sigma, y) &= \inf_{(\beta,\Sigma)\in C} \left[\prod_{i=1}^{n} \Gamma\left(\frac{\nu + d}{2}, \frac{\nu + (\beta^T x_i - y_i)^T \Sigma^{-1} (\beta^T x_i - y_i)}{2}, q_i \right) \right] \\
&\geq \prod_{i=1}^{n} \left[\inf_{(\beta,\Sigma)\in C} \Gamma\left(\frac{\nu + d}{2}, \frac{\nu + (\beta^T x_i - y_i)^T \Sigma^{-1} (\beta^T x_i - y_i)}{2}, q_i \right) \right] \\
&\geq \prod_{i=1}^{n} \left[\inf_{(\beta,\Sigma)\in C_i} \Gamma\left(\frac{\nu + d}{2}, \frac{\nu + (\beta^T x_i - y_i)^T \Sigma^{-1} (\beta^T x_i - y_i)}{2}, q_i \right) \right] .
\end{aligned}$$

Then, by Hobert's (2001) Lemma 1, it straightforwardly follows that, for any $(\beta', \Sigma') \in C$,

$$\pi(q \,|\, \beta', \Sigma', y) \geq \prod_{i=1}^{n} g(q_i) .$$

Therefore, if $(\beta', \Sigma') \in C$, we have

$$\begin{aligned}
k(\beta, \Sigma \,|\, \beta', \Sigma') &= \int_{\mathbb{R}_+^n} \pi(\beta, \Sigma \,|\, q, y) \, \pi(q \,|\, \beta', \Sigma', y) \, dq \\
&\geq \int_{\mathbb{R}_+^n} \pi(\beta, \Sigma \,|\, q, y) \left[\prod_{i=1}^{n} g(q_i) \right] dq \\
&= \left(\int_0^\infty g(t) \, dt \right)^n \int_{\mathbb{R}_+^n} \pi(\beta, \Sigma \,|\, q, y) \left[\prod_{i=1}^{n} \frac{g(q_i)}{\int_0^\infty g(t) \, dt} \right] dq .
\end{aligned}$$

and the proof is complete. □

In conjunction with Rosenthal's (1995) Theorem 12, Lemmas 1 and 2 immediately yield the following result.

Theorem 1. *The Markov chain underlying the DA algorithm is geometrically ergodic if* $\nu + d > 2$ *and* $n < \nu + k - 2$.

A similar drift and minorization type analysis of the Haar PX-DA algorithm is much more complicated due to the extra step (Roy, 2008). Fortunately, such an analysis is unnecessary because Hobert and Marchev's (2008) Theorem 4 shows that the Haar PX-DA algorithm inherits the geometric ergodicity of the DA algorithm upon which it is based. Hence, we have the following result:

Corollary 1. *The Markov chain underlying the Haar PX-DA algorithm is geometrically ergodic if* $\nu + d > 2$ *and* $n < \nu + k - 2$.

Theorem 1 and Corollary 1 are generalizations of results in Marchev and Hobert (2004), who studied the multivariate location-scale model that is the special case of model (1) in which $k = 1$ and there are no covariates.

Appendices

A Distributional facts

Matrix Normal Distribution Suppose Z is an $r \times c$ random matrix with density

$$f_Z(z) = \frac{1}{(2\pi)^{\frac{rc}{2}} |A|^{\frac{c}{2}} |B|^{\frac{r}{2}}} \exp\left[-\frac{1}{2} \text{tr}\Big\{ A^{-1}(z-\theta)B^{-1}(z-\theta)^T \Big\} \right],$$

where θ is an $r \times c$ matrix, A and B are $r \times r$ and $c \times c$ positive definite matrices. Then Z is said to have a *matrix normal distribution* and we denote this by $Z \sim \mathrm{N}_{r,c}(\theta, A, B)$ (Arnold, 1981, Chapter 17). Simulating a random matrix from the $\mathrm{N}_{r,c}(\theta, A, B)$ distribution is easy. For $i = 1, \ldots, r$, draw $Z_i \sim \mathrm{N}_c(0, B)$ and then set

$$Z = A^{\frac{1}{2}} \begin{bmatrix} Z_1^T \\ Z_2^T \\ \vdots \\ Z_m^T \end{bmatrix} + \theta \, .$$

Inverse Wishart Distribution Suppose W is a $p \times p$ random positive definite matrix with density

$$f_W(w) = \frac{|w|^{-\frac{m+p+1}{2}} \exp\Big\{ -\frac{1}{2}\text{tr}\big(\Theta^{-1} w^{-1}\big) \Big\}}{2^{\frac{mp}{2}} \pi^{\frac{p(p-1)}{4}} |\Theta|^{\frac{m}{2}} \prod_{i=1}^{p} \Gamma\big(\frac{1}{2}(m+1-i)\big)},$$

where $m \geq p$ and Θ is a $p \times p$ positive definite matrix. Then W is said to have an inverse Wishart distribution and this is denoted by $W \sim IW_p(m, \Theta)$.

B A matrix result

Lemma 3. *If* P *is a positive definite matrix and, for some vector* x, *the matrix* $P - xx^T$ *is positive semidefinite, then* $x^T P^{-1} x \leq 1$.

Proof. Calculating the determinant of the matrix

$$\begin{bmatrix} P & x \\ x^T & 1 \end{bmatrix}$$

twice using the Schur complement of P and 1, we get the following identity

$$|P|(1 - x^T P^{-1} x) = |P - xx^T| \ .$$

This implies that

$$x^T P^{-1} x = 1 - \frac{|P - xx^T|}{|P|} \ .$$

Since P is positive definite and $P - xx^T$ is a positive semidefinite, we know that $|P| > 0$ and $|P - xx^T| \geq 0$. Hence, $|P - xx^T|/|P| \geq 0$ and it follows that $x^T P^{-1} x \leq 1$. □

References

AFFLECK-GRAVE, J. and MCDONALD, B. (1989). Nonnormalities and tests of asset pricing theories. *Journal of Finance*, **44** 889–908.

ARNOLD, S. F. (1981). *The Theory of Linear Models and Multivariate Analysis*. Wiley, New York.

BRADLEY, B. O. and TAQQU, M. S. (2003). *Handbook of Heavy-Tailed Distributions in Finance*, chap. Financial Risk and Heavy Tails. Elsevire, Amsterdam, 35–103.

CADEMARTORI, D., ROMO, C., CAMPOS, R. and GALEA, M. (2003). Robust estimation of systematic risk using the t distribution in the Chilean stock markets. *Applied Economics Letters*, **10** 447–453.

EATON, M. L. (1989). *Group Invariance Applications in Statistics*. Institute of Mathematical Statistics and the American Statistical Association, Hayward, California and Alexandria, Virginia.

FAMA, E. (1965). The behavior of stock market prices. *Journal of Business*, **38** 34–105.

FERNANDEZ, C. and STEEL, M. F. J. (1999). Multivariate Student-t regression models: Pitfalls and inference. *Biometrika*, **86** 153–167.

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.

HOBERT, J. P. (2001). Discussion of "The art of data augmentation" by D.A. van Dyk and X.-L. Meng. *Jouranl of Computational and Graphical Statistics*, **10** 59–68.

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.

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.

JONES, G. L. and HOBERT, J. P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science*, **16** 312–34.

LIU, C. (1996). Bayesian robust multivariate linear regression with incomplete data. *Journal of the American Statistical Association*, **91** 1219–1227.

LIU, J. S. and WU, Y. N. (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association*, **94** 1264–1274.

MARCHEV, D. and HOBERT, J. P. (2004). Geometric ergodicity of van Dyk and Meng's algorithm for the multivariate student's model. *Journal of the American Statistical Association*, **99** 228–238.

MARDIA, K., KENT, J. and BIBBY, J. (1979). *Multivariate Analysis*. Academic press, London.

MENG, X.-L. and VAN DYK, D. A. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika*, **86** 301–320.

MEYN, S. P. and TWEEDIE, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer Verlag, London.

R DEVELOPMENT CORE TEAM (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`.

ROSENTHAL, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association*, **90** 558–566.

ROY, V. (2008). *Theoretical and methodological developments for Markov chain Monte Carlo algorithms for Bayesian regression*. Ph.D. thesis, Department of Statistics, University of Florida.

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.

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.

ZHOU, G. (1993). Asset-pricing tests under alternative distributions. *Journal of Finance*, **48** 1927–1942.