A note on the convergence rate of MCMC for robust Bayesian multivariate linear regression with proper priors

Grant Backlund and James P. Hobert
Department of Statistics
University of Florida

January 2020

Abstract

The multivariate linear regression model with errors from a scale mixture of Gaussian densities yields a complex likelihood function. Combining this likelihood with any nontrivial prior distribution leads to a highly intractable posterior density. If a *conditionally conjugate* prior is used, then there is a well-known and easy-to-implement data augmentation (DA) algorithm available for exploring the posterior. Hobert et al. (2018) recently showed that, under an *improper* conditionally conjugate prior (and weak regularity conditions), the Markov chain that drives the DA algorithm converges at a geometric rate. Unfortunately, the model studied by Hobert et al. (2018) can only be used in situations where the X matrix has full column rank. In this note, analogous convergence rate results are established for a *proper* conditionally conjugate prior. An important advantage of using a proper prior is that, not only is the X matrix allowed to be column rank deficient, but it can also have more columns than rows, i.e., our model is applicable in cases where $p > n$. This is an important extension in the era of big data.

1 Introduction

Let $Y_1, Y_2, \ldots, Y_n$ be independent d-dimensional random vectors from the multivariate linear regression model

$$Y_i = \beta^T x_i + \Sigma^{\frac{1}{2}} \varepsilon_i \; , \tag{1}$$

where x_i is a $p \times 1$ vector of known covariates associated with Y_i, β is a $p \times d$ matrix of unknown regression coefficients, and Σ is an unknown $d \times d$ positive definite scale matrix. The random vectors $\varepsilon_1, \ldots, \varepsilon_n$ are iid d-dimensional errors from a density taking the form

$$f_h(\varepsilon) = \int_{\mathbb{R}_+} \frac{u^{\frac{d}{2}}}{(2\pi)^{\frac{d}{2}}} \exp\Big\{ -\frac{u}{2} \varepsilon^T \varepsilon \Big\} h(u) \, du \; ,$$

Key words and phrases. Data augmentation algorithm, Drift condition, Geometric ergodicity, Heavy-tailed distribution, Scale mixture

where $\mathbb{R}_+ := (0,\infty)$, and $h : \mathbb{R}_+ \to [0,\infty)$ is the density function of some positive random variable. We shall refer to h as a *mixing density*. Heavy-tailed error densities can be produced by choosing h with appropriate behavior near the origin (see, e.g., Andrews and Mallows, 1974; Fernández and Steel, 2000; West, 1984). Some typical choices for h are the gamma, inverse gamma, generalized inverse Gaussian, and log-normal densities. However, in principle, h can be taken to be any density on the positive half-line. We assume throughout that

$$\int_0^\infty u^{\frac{d}{2}} h(u)\, du < \infty \ ,$$

and we refer to this as "condition $\mathcal{M}$." As we shall see, this condition is required for the existence of the data augmentation (DA) algorithm.

Denote the n equations in (1) collectively as

$$Y = X\beta + \varepsilon\,\Sigma^{\frac{1}{2}} \ , \tag{2}$$

where Y is the $n \times d$ matrix whose ith row is Y_i^T, X is the $n \times p$ matrix whose ith row is x_i^T, and ε represents the $n \times d$ matrix whose ith row is ε_i^T. Also, let y and y_i denote the observed values of Y and Y_i, respectively. The joint density of the data from model (2) is given by

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

Consider a Bayesian analysis with a proper conditionally conjugate prior that is defined sequentially as follows: $\omega(\beta,\Sigma) = \omega(\beta|\Sigma)\omega(\Sigma)$, where $\beta|\Sigma \sim \mathrm{N}_{p,d}(\theta, A, \Sigma)$ and $\Sigma \sim \mathrm{IW}_d(\nu, \Theta)$. Here, $\mathrm{N}_{p,d}$ and IW_d denote the matrix normal and inverse Wishart distributions, respectively, and the associated densities are defined in the Appendix. The hyperparameters are θ (a $p \times d$ matrix), A (a $p \times p$ positive definite matrix), $\nu > d-1$, and Θ (a $d \times d$ positive definite matrix). This prior is standard in multivariate regression settings, and is often used in conjunction with model (2) (see, e.g. Broemeling, 1985, Chapter 8). The highly intractable posterior density that results from this model is, of course, given by

$$\pi(\beta,\Sigma|y) = \frac{f(y|\beta,\Sigma)\,\omega(\beta,\Sigma)}{m(y)} \ ,$$

where

$$m(y) = \int_{\mathcal{S}_d} \int_{\mathbb{R}^{p\times d}} f(y|\beta,\Sigma)\,\omega(\beta,\Sigma)\, d\beta\, d\Sigma \ ,$$

and $\mathcal{S}_d \subset \mathbb{R}^{\frac{d(d+1)}{2}}$ denotes the space of $d \times d$ positive definite matrices. Note that there are no restrictions on X in our model. In particular, X is allowed to have more columns than rows.

We now introduce the latent data that is used to construct the DA algorithm. Conditional on (β,Σ), let $\{(Y_i, Z_i)\}_{i=1}^n$ be independent pairs such that $Y_i|Z_i,\beta,\Sigma \sim \mathrm{N}_d(\beta^T x_i, \Sigma/Z_i)$, and

$Z_i|\beta, \Sigma \sim h$. Let $f(y, z|\beta, \Sigma)$ denote the joint density of Y and $Z := (Z_1, \dots, Z_n)$ given (β, Σ). It's easy to see that

$$\int_{\mathbb{R}_+^n} f(y, z|\beta, \Sigma)\, dz = f(y|\beta, \Sigma) \; .$$

If we now define the so-called complete data posterior density $\pi : \mathbb{R}^{p \times d} \times \mathcal{S}_d \times \mathbb{R}_+^n \to [0, \infty)$ as

$$\pi(\beta, \Sigma, z|y) = \frac{f(y, z|\beta, \Sigma)\omega(\beta, \Sigma)}{m(y)} \; ,$$

then it's clear that

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

which is our target posterior. The DA algorithm simulates a Markov chain, $\Phi = \{(\beta_m, \Sigma_m)\}_{m=0}^{\infty}$, with state space $\mathsf{X} := \mathbb{R}^{p \times d} \times \mathcal{S}_d$, by alternating between draws from $\pi(z|\beta, \Sigma, y)$ and $\pi(\beta, \Sigma|z, y)$. Two important facts about the conditional density $\pi(z|\beta, \Sigma, y)$: (1) It does not depend on the prior, and (2) it is a product of n univariate densities. We now describe its form in more detail. As in Hobert et al. (2018) (HJK&Q), we define a parametric family of univariate density functions indexed by $s \geq 0$ as follows:

$$\psi(u; s) = b(s) u^{\frac{d}{2}} e^{-\frac{su}{2}} h(u) \; ,$$

where $b^{-1}(s) = \int_0^\infty v^{\frac{d}{2}} e^{-\frac{sv}{2}} h(v)\, dv$. Using this notation, we can see that

$$\pi(z|\beta, \Sigma, y) = \prod_{i=1}^{n} \psi(z_i; r_i) \; ,$$

where $r_i = (\beta^T x_i - y_i)^T \Sigma^{-1} (\beta^T x_i - y_i)$. We must be able to draw from $\psi(\cdot; s)$ in order to run the DA algorithm. When h is a density from a standard parametric family, ψ is often standard as well. For example, when h is gamma, ψ is also gamma, and when h is inverse gamma, ψ is generalized inverse Gaussian. If ψ is not a standard density, it can often be efficiently sampled using a rejection sampler with h as the candidate. Because the prior on (β, Σ) is conditionally conjugate, the density $\pi(\beta, \Sigma|z, y)$ takes the same form as the prior, i.e., $\pi(\beta|\Sigma, z, y)$ is matrix normal, and $\pi(\Sigma|z, y)$ is inverse Wishart.

In order to formally state the DA algorithm, we need to introduce a bit more notation. For $z = (z_1, \dots, z_n) \in \mathbb{R}_+^n$, let Q be an $n \times n$ diagonal matrix whose ith diagonal element is z_i^{-1}. Also, define $\Omega = (X^T Q^{-1} X + A^{-1})^{-1}$ and $\mu = (X^T Q^{-1} X + A^{-1})^{-1} (X^T Q^{-1} y + A^{-1} \theta)$. We can now formally state the DA algorithm. If the current state of the DA Markov chain is $(\beta_m, \Sigma_m) = (\beta, \Sigma)$, then we simulate the new state, $(\beta_{m+1}, \Sigma_{m+1})$, using the following three-step procedure.

Iteration $m + 1$ of the DA algorithm:

1. Draw $\{Z_i\}_{i=1}^n$ independently with $Z_i \sim \psi\Big(\cdot\;; \big(\beta^T x_i - y_i\big)^T \Sigma^{-1} \big(\beta^T x_i - y_i\big)\Big)$, and call the result $z = (z_1, \dots, z_n)$.

2. Draw
$$\Sigma_{m+1} \sim \mathrm{IW}_d\Big(n+\nu, \Big(\Theta^{-1} + \theta^T A^{-1}\theta + y^T Q^{-1} y - \mu^T \Omega^{-1}\mu\Big)^{-1}\Big) \, .$$

3. Draw $\beta_{m+1} \sim \mathrm{N}_{p,d}\big(\mu, \Omega, \Sigma_{m+1}\big)$

HJK&Q considered the same likelihood function, but a different prior on (β, Σ). In particular, they used an improper conditionally conjugate prior that takes the form $\omega^*(\beta,\Sigma) \propto |\Sigma|^{-a}\, I_{\mathcal{S}_d}(\Sigma)$, where a is a hyperparameter. Taking $a = (d+1)/2$ yields the independence Jeffreys prior, which is a standard default prior for multivariate location scale problems. Let Λ denote the $n \times (p+d)$ matrix $(X : y)$. That is, Λ is the matrix that results when the $n \times d$ matrix y is appended to the right of the $n \times p$ matrix X. Under the prior ω^*, the following conditions are *necessary* for propriety:

$(N1)$ rank$(\Lambda) = p + d$;

$(N2)$ $n > p + 2d - 2a$.

Clearly, $(N1)$ cannot hold unless X has full column rank. This obviously rules out cases in which $p > n$.

Not surprisingly, HJK&Q's DA algorithm is quite similar to ours. Let $\Omega_* = (X^T Q^{-1} X)^{-1}$ and $\mu_* = (X^T Q^{-1} X)^{-1} X^T Q^{-1} y$. If the current state of HJK&Q's DA Markov chain is $(\beta_m^*, \Sigma_m^*) = (\beta, \Sigma)$, then we simulate the new state, $(\beta_{m+1}^*, \Sigma_{m+1}^*)$, using the following three-step procedure.

Iteration $m+1$ of HJK&Q's DA algorithm:

1. Draw $\{Z_i\}_{i=1}^n$ independently with $Z_i \sim \psi\Big(\cdot\, ; \big(\beta^T x_i - y_i\big)^T \Sigma^{-1} \big(\beta^T x_i - y_i\big)\Big)$, and call the result $z = (z_1, \dots, z_n)$.

2. Draw
$$\Sigma_{m+1}^* \sim \mathrm{IW}_d\Big(n - p + 2a - d - 1, \Big(y^T Q^{-1} y - \mu_*^T \Omega_*^{-1} \mu_*\Big)^{-1}\Big) \, .$$

3. Draw $\beta_{m+1}^* \sim \mathrm{N}_{p,d}\big(\mu_*, \Omega_*, \Sigma_{m+1}^*\big)$

Under conditions $(N1)$ and $(N2)$, this algorithm is well-defined.

2 The Main Result

The Markov transition density (Mtd) of Φ is given by
$$k(\beta, \Sigma | \tilde{\beta}, \tilde{\Sigma}) = \int_{\mathbb{R}_+^n} \pi(\beta, \Sigma | z, y) \pi(z | \tilde{\beta}, \tilde{\Sigma}, y)\, dz \, .$$

It is easy to show that Φ is Harris ergodic (irreducible, aperiodic and positive Harris recurrent), and that $\pi(\beta, \Sigma|y)$ is the stationary density. The chain Φ is geometrically ergodic if there exist $M : \mathsf{X} \to [0,\infty)$ and $\rho \in [0,1)$ such that, for all $m \in \mathbb{N}$,

$$\int_{\mathcal{S}_d} \int_{\mathbb{R}^{p\times d}} \left| k^m(\beta, \Sigma|\tilde{\beta}, \tilde{\Sigma}) - \pi(\beta, \Sigma|y) \right| \, d\beta \, d\Sigma \leq M(\tilde{\beta}, \tilde{\Sigma})\rho^m \ ,$$

where k^m is the m-step Mtd. The importance of using geometrically ergodic Markov chains in MCMC has been well-documented (see, e.g. Flegal et al., 2008).

In order to state our main result concerning the convergence rate of Φ, we must introduce several classes of mixing densities that were defined in HJK&Q. Let $\mathcal{Z}$ denote the set of mixing densities that are zero near the origin, that is, $h \in \mathcal{Z}$ if there exists $\delta > 0$ such that $h(u) = 0$ for all $u \in (0, \delta)$. Similarly, let $\mathcal{P}$ denote the set of mixing densities that are strictly positive near the origin, that is, $h \in \mathcal{P}$ if there exists $\delta > 0$ such that $h(u) > 0$ for all $u \in (0, \delta)$. If $h \in \mathcal{P}$ and there exists a $c > -1$ such that

$$\lim_{u\to 0} \frac{h(u)}{u^c} \in \mathbb{R}_+ \ ,$$

then we say that h is *polynomial near the origin with power c*. Finally, if $h \in \mathcal{P}$ and, for every $c > 0$, there exists an $\eta_c > 0$ such that the ratio $\frac{h(u)}{u^c}$ is strictly increasing in $(0, \eta_c)$, then we say that h is *faster than polynomial near the origin*. HJK&Q demonstrated that every mixing density that is a member of a standard parametric family is either polynomial near the origin, or faster than polynomial near the origin. Here is our main result.

Proposition 1. *Let $h : \mathbb{R}_+ \to [0,\infty)$ be a mixing density that satisfies condition $\mathcal{M}$. If any one of the following conditions holds, then the DA Markov chain Φ is geometrically ergodic.*

1. *$h \in \mathcal{Z}$.*
2. *h is faster than polynomial near the origin.*
3. *h is polynomial near the origin with power $c > \frac{n+\nu}{2}$.*

Proposition 1 is an extension of the following result.

Theorem 1 (HJK&Q)**.** *Assume that $(N1)$ and $(N2)$ hold. Let $h : \mathbb{R}_+ \to [0,\infty)$ be a mixing density that satisfies condition $\mathcal{M}$. If any one of the following conditions holds, then the posterior distribution is proper and the DA Markov chain $\Phi^* = \{(\beta^*_m, \Sigma^*_m)\}_{m=0}^\infty$ is geometrically ergodic.*

1. *$h \in \mathcal{Z}$.*
2. *h is faster than polynomial near the origin.*
3. *h is polynomial near the origin with power $c > \frac{n-p+2a-d-1}{2}$.*

We reiterate that HJK&Q's model cannot be used unless X has full column rank.

3 Proof of Proposition 1

Recall that the Mtd of Φ is given by

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

The Mtd of the DA Markov chain studied in HJK&Q takes the form

$$k^*(\beta,\Sigma|\tilde{\beta},\tilde{\Sigma}) = \int_{\mathbb{R}_+^n} \pi^*(\beta,\Sigma|z,y)\pi(z|\tilde{\beta},\tilde{\Sigma},y)\,dz\ .$$

Note that $\pi(z|\beta,\Sigma,y)$ appears in both integrands. (Recall that this density does not depend on the prior.) Because the improper prior is conditionally conjugate, $\pi^*(\beta,\Sigma|z,y)$ has the same form as $\pi(\beta,\Sigma|z,y)$, i.e., it is the product of a matrix normal and an inverse Wishart. Due to the similarities between k and k^*, we are able to reuse many of the calculations in HJK&Q's proof of Theorem 1.

As in HJK&Q, we prove our result by establishing drift and minorization conditions with the following drift function:

$$V(\beta,\Sigma) = \sum_{i=1}^n (y_i - \beta^T x_i)^T \Sigma^{-1} (y_i - \beta^T x_i)\ .$$

For an introduction to this method, see Jones and Hobert (2001). The minorization condition follows immediately from a calculation in HJK&Q. Indeed, fix $l > 0$ and define

$$B_l = \big\{(\beta,\Sigma) : V(\beta,\Sigma) \le l\big\}\ .$$

HJK&Q construct $\epsilon \in (0,1)$ and a density function $\hat{f} : \mathbb{R}_+^n \to [0,\infty)$ such that, for all $(\tilde{\beta},\tilde{\Sigma}) \in B_l$,

$$\pi(z|\tilde{\beta},\tilde{\Sigma},y) \ge \epsilon \hat{f}(z)\ .$$

Thus, for all $(\tilde{\beta},\tilde{\Sigma}) \in B_l$, we have

$$k\big(\beta,\Sigma\big|\tilde{\beta},\tilde{\Sigma}\big) = \int_{\mathbb{R}_+^n} \pi(\beta,\Sigma|z,y)\,\pi(z|\tilde{\beta},\tilde{\Sigma},y)\,dz \ge \epsilon \int_{\mathbb{R}_+^n} \pi(\beta,\Sigma|z,y)\,\hat{f}(z)\,dz = \epsilon f^*(\beta,\Sigma)\ .$$

This is the required minorization condition. Now we move on to the drift condition. We must show that there exists $\lambda \in [0,1)$ and $L < \infty$ such that

$$\int_{\mathsf{X}} V(\beta,\Sigma)k(\beta,\Sigma|\tilde{\beta},\tilde{\Sigma})\,d\beta\,d\Sigma \le \lambda V(\tilde{\beta},\tilde{\Sigma}) + L\ ,$$

for all $(\tilde{\beta},\tilde{\Sigma}) \in \mathsf{X}$. Note that

$$\begin{aligned}
&\int_{\mathsf{X}} V(\beta,\Sigma)k(\beta,\Sigma|\tilde{\beta},\tilde{\Sigma})\,d\beta\,d\Sigma \\
&\qquad = \int_{\mathbb{R}_+^n} \bigg[\int_{\mathsf{X}} V(\beta,\Sigma)\pi(\beta,\Sigma|z,y)\,d\beta\,d\Sigma\bigg]\pi(z|\tilde{\beta},\tilde{\Sigma},y)\,dz \\
&\qquad = \int_{\mathbb{R}_+^n} \bigg[\int_{\mathcal{S}_d} \bigg[\int_{\mathbb{R}^{p\times d}} V(\beta,\Sigma)\pi(\beta|\Sigma,z,y)\,d\beta\bigg]\pi(\Sigma|z,y)\,d\Sigma\bigg]\pi(z|\tilde{\beta},\tilde{\Sigma},y)\,dz\ . \qquad (3)
\end{aligned}$$

We attack (3) using an argument similar to one found in Roy and Hobert (2010). The innermost integral can be expressed as

$$E\left[\sum_{i=1}^n y_i^T \Sigma^{-1} y_i - 2\sum_{i=1}^n x_i^T \beta \Sigma^{-1} y_i + \sum_{i=1}^n x_i^T \beta \Sigma^{-1} \beta^T x_i \,\Big|\, \Sigma, z, y\right] ,$$

where $\beta | \Sigma, z, y \sim \mathrm{N}_{p,d}\big(\mu, \Omega, \Sigma\big)$. Standard results for the matrix normal distribution (see, e.g., Arnold, 1981, Chapter 17) imply that

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

It follows that

$$\begin{aligned}\int_{\mathbb{R}^{p\times d}} V(\beta,\Sigma)\pi(\beta|\Sigma,z,y)\,d\beta &= \sum_{i=1}^n y_i^T \Sigma^{-1} y_i - 2\sum_{i=1}^n x_i^T \mu \Sigma^{-1} y_i + \sum_{i=1}^n x_i^T \left[d\Omega + \mu\Sigma^{-1}\mu^T\right] x_i \\ &= \sum_{i=1}^n (y_i - \mu^T x_i)^T \Sigma^{-1} (y_i - \mu^T x_i) + d\sum_{i=1}^n x_i^T \Omega x_i \ .\end{aligned}$$

Now recall that $\Sigma | z, y \sim \mathrm{IW}_d\Big(n+\nu, \big(\Theta^{-1} + \theta^T A^{-1}\theta + y^T Q^{-1} y - \mu^T \Omega^{-1} \mu\big)^{-1}\Big)$. Hence, we have

$$E(\Sigma^{-1}|z,y) = (n+\nu)\left(\Theta^{-1} + \theta^T A^{-1}\theta + y^T Q^{-1} y - \mu^T \Omega^{-1}\mu\right)^{-1} \ .$$

Therefore,

$$\begin{aligned}&\int_{\mathcal{S}_d}\left[\int_{\mathbb{R}^{p\times d}} V(\beta,\Sigma)\pi(\beta|\Sigma,z,y)\,d\beta\right]\pi(\Sigma|z,y)\,d\Sigma \\ &= \sum_{i=1}^n (y_i - \mu^T x_i)^T E(\Sigma^{-1}|z,y)(y_i - \mu^T x_i) + d\sum_{i=1}^n x_i^T \Omega x_i \\ &\le (n+\nu)\sum_{i=1}^n (y_i - \mu^T x_i)^T \left(\theta^T A^{-1}\theta + y^T Q^{-1} y - \mu^T \Omega^{-1}\mu\right)^{-1}(y_i - \mu^T x_i) + d\sum_{i=1}^n x_i^T \Omega x_i \ ,\end{aligned}$$

where we have used the fact that Θ, and hence Θ^{-1}, is positive definite. Note that we were able to compute the first two conditional expectations exactly. Unfortunately, we are not able to compute the outer-most expectation in closed form. Instead, we will compute the expectation of a simple upper bound. First, observe that

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

It follows that

$$\begin{aligned}
&(y_i - \mu^T x_i)^T \big(y^T Q^{-1} y - \mu^T \Omega^{-1} \mu + \theta^T A^{-1} \theta \big)^{-1} (y_i - \mu^T x_i) = \\
&\frac{1}{z_i} (y_i - \mu^T x_i)^T \left(\sum_{j=1}^n \frac{z_j}{z_i} (y_j - \mu^T x_j)(y_j - \mu^T x_j)^T + \frac{1}{z_i} (\mu - \theta)^T A^{-1} (\mu - \theta) \right)^{-1} (y_i - \mu^T x_i) \ .
\end{aligned}$$

Now,

$$\frac{1}{z_i} (y^T Q^{-1} y - \mu^T \Omega^{-1} \mu + \theta^T A^{-1} \theta) = \sum_{j=1}^n \frac{z_j}{z_i} (y_j - \mu^T x_j)(y_j - \mu^T x_j)^T + \frac{1}{z_i} (\mu - \theta)^T A^{-1} (\mu - \theta) \ ,$$

is positive definite, and

$$\begin{aligned}
&\sum_{j=1}^n \frac{z_j}{z_i} (y_j - \mu^T x_j)(y_j - \mu^T x_j)^T + \frac{1}{z_i} (\mu - \theta)^T A^{-1} (\mu - \theta) - (y_i - \mu^T x_i)(y_i - \mu^T x_i)^T \\
&\qquad\qquad = \sum_{j \neq i} \frac{z_j}{z_i} (y_j - \mu^T x_j)(y_j - \mu^T x_j)^T + \frac{1}{z_i} (\mu - \theta)^T A^{-1} (\mu - \theta) \ ,
\end{aligned}$$

is positive semi-definite. It then follows from Lemma 3 in Roy and Hobert (2010) that

$$(y_i - \mu^T x_i)^T (y^T Q^{-1} y - \mu^T \Omega^{-1} \mu + \theta^T A^{-1} \theta)^{-1} (y_i - \mu^T x_i) \leq \frac{1}{z_i} \ .$$

Therefore,

$$(n + \nu) \sum_{i=1}^n (y_i - \mu^T x_i)^T (y^T Q^{-1} y - \mu^T \Omega^{-1} \mu + \theta^T A^{-1} \theta)^{-1} (y_i - \mu^T x_i) \leq (n + \nu) \sum_{i=1}^n \frac{1}{z_i} \ .$$

We now focus on bounding the term $d \sum_{i=1}^n x_i^T \Omega x_i$. Since the matrix A is positive definite,

$$\frac{1}{z_i} X^T Q^{-1} X + \frac{1}{z_i} A^{-1} = \sum_{j=1}^n \frac{z_j}{z_i} x_j x_j^T + \frac{1}{z_i} A^{-1}$$

is positive definite. Similarly,

$$\frac{1}{z_i}X^TQ^{-1}X + \frac{1}{z_i}A^{-1} - x_ix_i^T = \sum_{j\neq i}\frac{z_j}{z_i}x_jx_j^T + \frac{1}{z_i}A^{-1}$$

is also positive definite. Another application of Lemma 3 from Roy and Hobert (2010) yields

$$z_ix_i^T\Omega x_i = x_i^T\left(\frac{1}{z_i}X^TQ^{-1}X + \frac{1}{z_i}A^{-1}\right)^{-1}x_i = x_i^T\left(\sum_{j=1}^n\frac{z_j}{z_i}x_jx_j^T + \frac{1}{z_i}A^{-1}\right)^{-1}x_i \leq 1 \ .$$

Thus,

$$d\sum_{i=1}^n x_i^T\Omega x_i \leq d\sum_{i=1}^n\frac{1}{z_i} \ .$$

Putting all of this together, we have

$$\int_{\mathcal{S}_d}\left[\int_{\mathbb{R}^{p\times d}}V(\beta,\Sigma)\pi(\beta|\Sigma,z,y)\,d\beta\right]\pi(\Sigma|z,y)\,d\Sigma \leq (n+\nu+d)\sum_{i=1}^n\frac{1}{z_i} \ ,$$

and hence

$$\begin{aligned}\int_{\mathsf{X}}V(\beta,\Sigma)k(\beta,\Sigma|\tilde{\beta},\tilde{\Sigma})\,d\beta\,d\Sigma &\leq (n+\nu+d)\int_{\mathbb{R}_+^n}\left(\sum_{i=1}^n z_i^{-1}\right)\pi(z|\tilde{\beta},\tilde{\Sigma},y)\,dz \\ &= (n+\nu+d)\sum_{i=1}^n b(\tilde{r}_i)\int_0^\infty u^{\frac{d-2}{2}}e^{-\frac{\tilde{r}_iu}{2}}h(u)\,du \ ,\end{aligned}$$

where $\tilde{r}_i = (y_i - \tilde{\beta}^Tx_i)^T\tilde{\Sigma}^{-1}(y_i - \tilde{\beta}^Tx_i)$. Now, in conjunction with our assumptions about h, the results in Section 4 of HJK&Q imply the existence of $\lambda \in [0, \frac{1}{n+\nu+d})$ and $L \in \mathbb{R}$ such that

$$\frac{\int_0^\infty u^{\frac{d-2}{2}}e^{-\frac{su}{2}}h(u)du}{\int_0^\infty u^{\frac{d}{2}}e^{-\frac{su}{2}}h(u)du} \leq \lambda s + L$$

for every $s \geq 0$. Therefore, we have

$$\begin{aligned}\int_{\mathsf{X}}V(\beta,\Sigma)k(\beta,\Sigma|\tilde{\beta},\tilde{\Sigma})\,d\beta\,d\Sigma &\leq (n+\nu+d)\left(\lambda\sum_{i=1}^n\tilde{r}_i + nL\right) \\ &= \lambda(n+\nu+d)V(\tilde{\beta},\tilde{\Sigma}) + (n+\nu+d)nL \\ &=: \lambda'V(\tilde{\beta},\tilde{\Sigma}) + L' \ ,\end{aligned}$$

where $\lambda' = \lambda(n+\nu+d) \in [0,1)$ and $L' = (n+\nu+d)nL \in \mathbb{R}$. Hence, the drift condition has been established. Since the minorization condition holds for all $l > 0$, the proof is complete.

Acknowledgments. The authors thank two anonymous referees for helpful comments and suggestions. The second author was supported by NSF Grant DMS-15-11945.

Appendix

A Matrix Normal and Inverse Wishart Densities

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

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

where θ is an $r \times c$ matrix, and 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 $Z \sim N_{r,c}(\theta, A, B)$ (Arnold 1981, Chapter 17).

Inverse Wishart Distribution Suppose W is an $r \times r$ random positive definite matrix with density

$$f_W(w) = \frac{|w|^{-\frac{\nu+r+1}{2}} \exp\left\{ -\frac{1}{2} \text{tr}\left(\Theta^{-1} w^{-1} \right) \right\}}{2^{\frac{\nu r}{2}} \pi^{\frac{r(r-1)}{4}} |\Theta|^{\frac{\nu}{2}} \prod_{i=1}^{r} \Gamma\left(\frac{1}{2}(\nu + 1 - i) \right)} I_{\mathcal{S}_r}(w),$$

where $\nu > r-1$ and Θ is an $r \times r$ positive definite matrix. Then W is said to have an *inverse Wishart distribution* and this is denoted by $W \sim \text{IW}_r(\nu, \Theta)$.

References

ANDREWS, D. and MALLOWS, C. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Series B* **36** 99–102.

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

BROEMELING, L. D. (1985). *Bayesian Analysis of Linear Models*. Marcel Dekker.

FERNÁNDEZ, C. and STEEL, M. F. J. (2000). Bayesian regression analysis with scale mixtures of normals. *Econometric Theory* **16** 80–101.

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., JUNG, Y. J., KHARE, K. and QIN, Q. (2018). Convergence analysis of MCMC algorithms for Bayesian multivariate linear regression with non-Gaussian errors. *Scandinavian Journal of Statistics* **45** 513–533.

JONES, G. L. and HOBERT, J. P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science* **16** 312–334.

ROY, V. and HOBERT, J. P. (2010). On Monte Carlo methods for Bayesian multivariate linear regression models with heavy-tailed errors. *Journal of Multivariate Analysis* **101** 1190–1202.

WEST, M. (1984). Outlier models and prior distributions in Bayesian linear regression. *Journal of the Royal Statistical Society, Series B* **46** 431–439.