Estimating the spectral gap of a trace-class Markov operator

Qian Qin, James P. Hobert and Kshitij Khare

Department of Statistics

University of Florida

April 2017

Abstract

The utility of a Markov chain Monte Carlo algorithm is, in large part, determined by the size of the spectral gap of the corresponding Markov operator. However, calculating (and even approximating) the spectral gaps of practical Monte Carlo Markov chains in statistics has proven to be an extremely difficult and often insurmountable task, especially when these chains move on continuous state spaces. In this paper, a method for accurate estimation of the spectral gap is developed for general state space Markov chains whose operators are non-negative and trace-class. The method is based on the fact that the second largest eigenvalue (and hence the spectral gap) of such operators can be bounded above and below by simple functions of the power sums of the eigenvalues. These power sums often have nice integral representations. A classical Monte Carlo method is proposed to estimate these integrals, and a simple sufficient condition for finite variance is provided. This leads to asymptotically valid confidence intervals for the second largest eigenvalue (and the spectral gap) of the Markov operator. For illustration, the method is applied to Albert and Chib's (1993) data augmentation (DA) algorithm for Bayesian probit regression, and also to a DA algorithm for Bayesian linear regression with non-Gaussian errors (Liu, 1996).

1 Introduction

Markov chain Monte Carlo (MCMC) is widely used to estimate intractable integrals that represent expectations with respect to complicated probability distributions. Let $\pi : S \to [0, \infty)$ be a probability density function (pdf) with respect to a σ-finite measure μ, where $(S, \mathcal{U}, \mu)$ is some measure space. Suppose we

Key words and phrases. Data augmentation algorithm, Eigenvalues, Hilbert-Schmidt operator, Markov chain, Monte Carlo

want to approximate the integral

$$J := \int_S f(u)\pi(u)\mu(du)$$

for some function $f : S \to \mathbb{R}$. Then J can be estimated by $\hat{J}_m := m^{-1}\sum_{k=0}^{m-1} f(\Phi_k)$, where $\{\Phi_k\}_{k=0}^{m-1}$ are the first m elements of a well-behaved Markov chain with stationary density $\pi(\cdot)$. Unlike classical Monte Carlo estimators, $\hat{J}_m$ is not based on iid random elements. Indeed, the elements of the chain are typically neither identically distributed nor independent. Given $\mathrm{var}_\pi f$, the variance of $f(\cdot)$ under the stationary distribution, the accuracy of $\hat{J}_m$ is primarily determined by two factors: (i) the convergence rate of the Markov chain, and (ii) the correlation between the $f(\Phi_k)$s when the chain is stationary. These two factors are related, and can be analyzed jointly under an operator theoretic framework.

The starting point of the operator theoretic approach is the Hilbert space of functions that are square integrable with respect to the target pdf, $\pi(\cdot)$. The Markov transition function that gives rise to $\Phi = \{\Phi_k\}_{k=0}^{\infty}$ defines a linear (Markov) operator on this Hilbert space. (Formal definitions are given in Section 2.) If Φ is reversible, then it is geometrically ergodic if and only if the corresponding Markov operator admits a positive *spectral gap* (Roberts and Rosenthal, 1997; Kontoyiannis and Meyn, 2012). The gap, which is real number in $(0, 1]$, plays a fundamental role in determining the mixing properties of the Markov chain, with larger values corresponding to better performance. For instance, suppose Φ_0 has pdf $\pi_0(\cdot)$ such that $d\pi_0/d\pi$ is in the Hilbert space, and let $d(\Phi_k; \pi)$ denote the total variation distance between the distribution of Φ_k and the chain's stationary distribution. Then, if δ denotes the spectral gap, we have

$$d(\Phi_k; \pi) \leq C(1 - \delta)^k$$

for all positive integers k, where C depends on π_0 but not on k (Roberts and Rosenthal, 1997). Furthermore, $(1 - \delta)^k$ gives the maximal absolute correlation between Φ_j and Φ_{j+k} as $j \to \infty$. It follows (see e.g. Mira and Geyer, 1999) that the asymptotic variance of $\sqrt{m}(\hat{J}_m - J)$ as $m \to \infty$ is bounded above by

$$\frac{2 - \delta}{\delta}\mathrm{var}_\pi f\ .$$

Unfortunately, it is impossible to calculate the spectral gaps of the Markov operators associated with practically relevant MCMC algorithms, and even accurately approximating these quantities has proven extremely difficult. In this paper, we develop a method of estimating the spectral gaps of Markov operators corresponding to a certain class of data augmentation (DA) algorithms (Tanner and Wong, 1987), and then show that the method can be extended to handle a much larger class of reversible MCMC algorithms.

DA Markov operators are necessarily non-negative. Moreover, any non-negative Markov operator that is compact has a pure eigenvalue spectrum that is contained in the set $[0, 1]$, and $1 - \delta$ is precisely the second largest eigenvalue. We propose a classical Monte Carlo estimator of $1-\delta$ for DA Markov operators that are *trace-class*, i.e. compact with summable eigenvalues. While compact operators were once thought to be rare in MCMC problems with uncountable state spaces (Chan and Geyer, 1994), a string of recent results suggests that trace-class DA Markov operators are not at all rare (see e.g. Qin and Hobert, 2016; Chakraborty and Khare, 2017; Choi and Román, 2017; Pal et al., 2017). Furthermore, by exploiting a simple trick, we are able to broaden the applicability of our method well beyond DA algorithms. Indeed, if a reversible Monte Carlo Markov chain has a Markov transition density (Mtd), and the corresponding Markov operator is Hilbert-Schmidt, then our method can be utilized to estimate its spectral gap. This is because the square of such a Markov operator can be represented as a trace-class DA Markov operator. A detailed explanation is provided in Section 4.

Of course, there is a large literature devoted to developing theoretical bounds on the second largest eigenvalue of a Markov operator (see e.g. Lawler and Sokal, 1988; Sinclair and Jerrum, 1989; Diaconis and Stroock, 1991). However, these results are typically not useful in situations where the state space, S, is uncountable and multi-dimensional, which is our main focus. There also exist computational methods for approximating the eigenvalues of a Hilbert-Schmidt operator (see e.g. Koltchinskii and Giné, 2000; Ahues et al., 2001, §4.2). Unfortunately, these methods require a closed form kernel function, which is typically not available in the MCMC context. There are still other methods based on simulation. Most notably, Garren and Smith (2000) used simulations of a reversible chain to estimate the second largest eigenvalue of its operator (assuming it's Hilbert-Schmidt). Their approach is reminiscent of the so-called power method from computer science, and we use these ideas as well. The main difference between their method and ours is that we exploit the specific structure of the Mtd associated with the DA algorithm. This makes our method much simpler to implement computationally, and our results easier to interpret. The power of our method comes at the price of being computationally intensive, especially when the target posterior is based on a large sample.

The rest of the paper is organized as follows. The notion of Markov operator is formalized in Section 2. In Section 3, it is shown that the second largest eigenvalue of a non-negative trace-class operator can be bounded above and below by functions of the power sums of the operator's eigenvalues. In Section 4, DA Markov operators are formally defined, and the sum of the kth ($k \in \mathbb{N}$) power of the eigenvalues of a trace-class DA Markov operator is related to a functional of its Mtd. This functional is usually a multi-dimensional integral, and a classical Monte Carlo estimator of it is developed in Section 5. Finally, in Section 6 we apply our methods to a few well-known MCMC algorithms. Our examples include Albert and Chib's (1993) DA algorithm for Bayesian probit regression, and a DA algorithm for Bayesian linear regression with non-Gaussian errors (Liu, 1996).

2 Markov operators

Assume that the Markov chain Φ has a Markov transition density, $p(u, \cdot)$, $u \in S$, such that, for any measurable $A \subset S$ and $u \in S$,

$$\mathbb{P}(\Phi_k \in A | \Phi_0 = u) = \int_A p^{(k)}(u, u') \, \mu(du') \,,$$

where

$$p^{(k)}(u, \cdot) := \begin{cases} p(u, \cdot) & k = 1 \\ \int_S p^{(k-1)}(u, u') p(u', \cdot) \, \mu(du') & k > 1 \end{cases}$$

is the k-step Mtd corresponding to $p(u, \cdot)$. We will assume throughout that Φ is Harris ergodic, i.e. irreducible, aperiodic and Harris recurrent. Define a Hilbert space consisting of complex valued functions on S that are square integrable with respect to $\pi(\cdot)$, namely

$$L^2(\pi) := \left\{ f : S \to \mathbb{C} \;\middle|\; \int_S |f(u)|^2 \pi(u) \, \mu(du) < \infty \right\} .$$

For $f, g \in L^2(\pi)$, their inner product is given by

$$\langle f, g \rangle_\pi = \int_S f(u) \overline{g(u)} \pi(u) \, \mu(du) \,.$$

We assume that $\mathcal{U}$ is countably generated, which implies that $L^2(\pi)$ is separable and admits a countable orthonormal basis (see e.g. Billingsley, 1995, Theorem 19.2). The transition density $p(u, \cdot)$, $u \in S$ defines the following linear operator P. For any $f \in L^2(\pi)$,

$$Pf(u) = \int_S p(u, u') f(u') \, \mu(du') \,.$$

The spectrum of a linear operator L is defined to be

$$\sigma(L) = \left\{ \lambda \in \mathbb{C} \,\middle|\, (L - \lambda I)^{-1} \text{ doesn't exist or is unbounded} \right\} ,$$

where I is the identity operator. It is well-known that $\sigma(P)$ is a closed subset of the unit disk in $\mathbb{C}$. Let $f_0 \in L^2(\pi)$ be the normalized constant function, i.e. $f_0(u) \equiv 1$, then $Pf_0 = f_0$. (This is just a fancy way of saying that 1 is an eigenvalue with constant eigenfunction, which is true of *all* Markov operators defined by ergodic chains.) Denote by P_0 the operator such that $P_0 f = Pf - \langle f, f_0 \rangle_\pi f_0$ for all $f \in L^2(\pi)$. Then the spectral gap of P is defined as

$$\delta = 1 - \sup \left\{ |\lambda| \;\middle|\; \lambda \in \sigma(P_0) \right\} .$$

For the remainder of this section, we assume that P is non-negative (and thus self-adjoint) and compact. This implies that $\sigma(P) \subset [0, 1]$, and that any non-vanishing element of $\sigma(P)$ is necessarily an eigenvalue of P. Furthermore, there are at most countably many eigenvalues, and they can accumulate only at the origin. Let $\lambda_0, \lambda_1, \dots, \lambda_\kappa$ be the decreasingly ordered strictly positive eigenvalues of P taking into account multiplicity, where $0 \le \kappa \le \infty$. Then $\lambda_0 = 1$ and λ_1 is what we previously referred to as the "second largest eigenvalue" of the Markov operator. If $\kappa = 0$, we set $\lambda_1 = 0$ (which corresponds to the trivial case where $\{\Phi_k\}_{k=0}^\infty$ are iid). Since Φ is Harris ergodic, λ_1 must be strictly less than 1. Also, the compactness of P implies that of P_0, and it's easy to show that $\sigma(P_0) = \sigma(P)\backslash\{1\}$. Hence, Φ is geometrically ergodic and the spectral gap is

$$\delta = 1 - \lambda_1 > 0\,.$$

For further background on the spectrum of a linear operator, see e.g. Helmberg (2014) or Ahues et al. (2001).

3 Power sums of eigenvalues

We now develop some results relating λ_1 to the power sum of P's eigenvalues. We assume throughout this section that P is non-negative and trace-class (compact with summable eigenvalues). For any positive integer k, let

$$s_k = \sum_{i=0}^{\kappa} \lambda_i^k\,,$$

and define s_0 to be infinity. The first power sum, s_1, is the trace norm of P (see e.g. Conway, 1990, 2000), while $\sqrt{s_2}$ is the Hilbert-Schmidt norm of P. That P is trace-class implies $s_1 < \infty$, and it's clear that s_k is decreasing in k.

Observe that,

$$\lambda_1 \le u_k := (s_k - 1)^{1/k}, \quad \forall\, k \in \mathbb{N}\,.$$

Moreover, if $\kappa \ge 1$, then it's easy to show that

$$\lambda_1 \ge l_k := \frac{s_k - 1}{s_{k-1} - 1}, \quad \forall\, k \in \mathbb{N}\,.$$

We now show that, in fact, these bounds are monotone in k and converge to λ_1.

Proposition 1. *As $k \to \infty$,*

$$u_k \downarrow \lambda_1\,, \tag{1}$$

and if furthermore $\kappa \ge 1$,

$$l_k \uparrow \lambda_1\,. \tag{2}$$

Proof. We begin with (1). When $\kappa = 0$, $s_k \equiv 1$ and the conclusion follows. Suppose $\kappa \geq 1$, and that the second largest eigenvalue is of multiplicity m, i.e.

$$1 = \lambda_0 > \lambda_1 = \lambda_2 = \cdots = \lambda_m > \lambda_{m+1} \geq \cdots \geq \lambda_\kappa > 0.$$

If $\kappa = m$, then $s_k - 1 = m\lambda_1^k$ for all $k \geq 1$ and the proof is trivial. Suppose for the rest of the proof that $\kappa \geq m + 1$. For positive integer k, let $r_k = \sum_{i=m+1}^{\kappa} \lambda_i^k < \infty$. Then $r_k > 0$, and

$$\frac{r_{k+1}}{r_k} \leq \lambda_{m+1} < \lambda_1 \,.$$

Hence,

$$\lim_{k\to\infty} \frac{r_k}{s_k - 1 - r_k} = \lim_{k\to\infty} \frac{r_k}{m\lambda_1^k} \leq \lim_{k\to\infty} \frac{r_1 \lambda_{m+1}^{k-1}}{m\lambda_1^k} = 0\,.$$

It follows that

$$\log u_k = \log \lambda_1 + \frac{1}{k} \log m + \frac{1}{k} \log(1 + o(1)) \to \log \lambda_1\,.$$

Finally,

$$u_{k+1} < \lambda_1^{1/(k+1)} \Big(\sum_{i=1}^{\kappa} \lambda_i^k \Big)^{1/(k+1)} \leq \Big(\sum_{i=1}^{\kappa} \lambda_i^k \Big)^{1/[k(k+1)]} \Big(\sum_{i=1}^{\kappa} \lambda_i^k \Big)^{1/(k+1)} = u_k\,,$$

and (1) follows.

Now onto (2). We have already shown that

$$s_k - 1 = m\lambda_1^k (1 + o(1))\,.$$

Thus,

$$l_k = \frac{m\lambda_1^k(1 + o(1))}{m\lambda_1^{k-1}(1 + o(1))} \to \lambda_1\,.$$

To show that l_k is increasing in k, which would complete the proof, we only need note that

$$\begin{aligned}
(s_{k+1} - 1)(s_{k-1} - 1) &= \sum_{i=1}^{\kappa} \lambda_i^{k+1} \sum_{j=1}^{\kappa} \lambda_j^{k-1} \\
&= \frac{1}{2} \sum_{i=1}^{\kappa} \sum_{j=1}^{\kappa} \lambda_i^{k-1} \lambda_j^{k-1} (\lambda_i^2 + \lambda_j^2) \\
&\geq \sum_{i=1}^{\kappa} \sum_{j=1}^{\kappa} \lambda_i^k \lambda_j^k \\
&= (s_k - 1)^2\,.
\end{aligned}$$

$\square$

Suppose now that we are interested in the convergence behavior of a particular Markov operator that is known to be non-negative and trace-class. If it is possible to estimate s_k, then Proposition 1 provides a method of getting approximate bounds on λ_1. When a DA Markov operator is trace-class, there is a nice integral representation of s_k that leads to a simple, classical Monte Carlo estimator of s_k. In the following section, we describe some theory for DA Markov operators, and in Section 5, we develop a classical Monte Carlo estimator of s_k.

4 Data augmentation operators and an integral representation of s_k

In order to formally define DA, we require a second measure space. Let $(S_V, \mathcal{V}, \nu)$ be a σ-finite measure space such that $\mathcal{V}$ is countably generated. Also, rename S and π, S_U and π_U, respectively. Consider the random element (U, V) taking values in $S_U \times S_V$ with joint pdf $\pi_{U,V}(\cdot, \cdot)$. Suppose the marginal pdf of U is the target, $\pi_U(\cdot)$, and denote the marginal pdf of V by $\pi_V(\cdot)$. We further assume that the conditional densities $\pi_{U|V}(u|v) := \pi_{U,V}(u, v)/\pi_V(v)$ and $\pi_{V|U}(v|u) := \pi_{U,V}(u, v)/\pi_U(u)$ are well defined almost everywhere in $S_U \times S_V$. Recall that Φ is a Markov chain on the state space S_U with Mtd $p(u, \cdot)$, $u \in S_U$. We call Φ a DA chain, and accordingly, P a DA operator, if $p(u, \cdot)$ can be expressed as

$$p(u, \cdot) = \int_{S_V} \pi_{U|V}(\cdot|v)\pi_{V|U}(v|u)\, \nu(dv)\,. \tag{3}$$

Such a chain is necessarily reversible with respect to $\pi_U(\cdot)$. To simulate it, in each iteration, one first draws the latent element V using $\pi_{V|U}(\cdot|u)$, where $u \in S_U$ is the current state, and then given $V = v$, one updates the current state according to $\pi_{U|V}(\cdot|v)$. A DA operator is non-negative, and thus possesses a positive spectrum (Liu et al., 1994).

Assume that (3) holds. Given $k \in \mathbb{N}$, the power sum of P's eigenvalues, s_k, if well defined, is closely related to the diagonal components of $p^{(k)}(\cdot, \cdot)$. Just as we can calculate the sum of the eigenvalues of a matrix by summing its diagonals, we can obtain s_k by evaluating $\int_{S_U} p^{(k)}(u, u)\, \mu(du)$. Here is a formal statement.

Theorem 1. *The DA operator P is trace-class if and only if*

$$\int_{S_U} p(u, u)\, \mu(du) < \infty\,. \tag{4}$$

If (4) *holds, then for any positive integer k,*

$$s_k := \sum_{i=0}^{\kappa} \lambda_i^k = \int_{S_U} p^{(k)}(u, u)\, \mu(du)\,. \tag{5}$$

Theorem 1 is a combination of a few standard results in classical functional analysis. It is fairly well-known, but we were unable to find a complete proof in the literature. An elementary proof is given in the appendix for completeness. For a more modern version of the theorem, see Brislawn (1988).

It is often possible to exploit Theorem 1 even when Φ is *not* a DA Markov chain. Indeed, suppose that Φ is reversible, but is not a DA chain. Then P is not a DA operator, but P^2 is, in fact, a DA operator. (Just take $\pi_{U,V}(u,v) = \pi_U(u)p(u,v)$.) If, in addition, P is Hilbert-Schmidt, which is equivalent to

$$\int_{S_U}\int_{S_U} \frac{(p(u,u'))^2\pi_U(u)}{\pi_U(u')}\,\mu(du)\,\mu(du') < \infty\,,$$

then by a simple spectral decomposition (see e.g. Helmberg, 2014, §28 Corollary 2.1) one can show that P^2 is trace-class, and its eigenvalues are precisely the squares of the eigenvalues of P. In this case, the spectral gap of P can be expressed as 1 minus the square root of P^2's second largest eigenvalue. Moreover, by Theorem 1, for $k \in \mathbb{N}$, the sum of the kth power of P^2's eigenvalues is equal to $\int_{S_U} p^{(2k)}(u,u)\,\mu(du) < \infty$.

We now briefly describe the so-called sandwich algorithm, which is a variant of DA that involves an extra step sandwiched between the two conditional draws of DA (Liu and Wu, 1999; Hobert and Marchev, 2008). Let $s(v,\cdot)$, $v \in S_V$ be a Markov transition function (Mtf) with invariant density $\pi_V(\cdot)$. Then

$$\tilde{p}(u,\cdot) = \int_{S_V}\int_{S_V} \pi_{U|V}(\cdot|v')s(v,dv')\pi_{V|U}(v|u)\nu(dv) \tag{6}$$

is an Mtd with invariant density $\pi_U(\cdot)$. This Mtd defines a new Markov chain, call it $\tilde{\Phi}$, which we refer to as a sandwich version of the original DA chain, Φ. To simulate $\tilde{\Phi}$, in each iteration, the latent element is first drawn from $\pi_{V|U}(\cdot|u)$, and then updated using $s(v,\cdot)$ before the current state is updated according to $\pi_{U|V}(\cdot|v')$. Sandwich chains often converge much faster than their parent DA chains (see e.g. Khare and Hobert, 2011).

Of course, $\tilde{p}(u,\cdot)$ defines a Markov operator on $L^2(\pi_U)$, which we refer to as $\tilde{P}$. It is easy to see that, if the Markov chain corresponding to $s(v,\cdot)$ is reversible with respect to $\pi_V(\cdot)$, then $\tilde{p}(u,\cdot)$ is reversible with respect to $\pi_U(\cdot)$. Thus, when $s(v,\cdot)$ is reversible, $\tilde{P}^2$ is a DA operator. Interestingly, it turns out that $\tilde{p}(u,\cdot)$ can often be re-expressed as the Mtd of a DA chain, in which case $\tilde{P}$ itself is a DA operator. Indeed, a sandwich Mtd $\tilde{p}(u,\cdot)$ is said to be "representable" if there exists a random element $\tilde{V}$ in S_V such that

$$\tilde{p}(u,u') = \int_{S_V} \pi_{U|\tilde{V}}(u'|v)\pi_{\tilde{V}|U}(v|u)\,\nu(dv)\,, \tag{7}$$

where $\pi_{U|\tilde{V}}(u'|v)$ and $\pi_{\tilde{V}|U}(v|u)$ have the apparent meanings (see, e.g. Hobert, 2011). It is shown in Proposition 2 in Section 5 that when P is trace-class and $\tilde{p}(u,\cdot)$ is representable, $\tilde{P}$ is also trace-class. In this case, let $\{\tilde{\lambda}_i\}_{i=0}^{\tilde{\kappa}}$ be the decreasingly ordered positive eigenvalues of $\tilde{P}$ taking into account multiplicity, where

$0 \le \tilde{\kappa} \le \infty$. Then $\tilde{\lambda}_0 = 1$, and $\tilde{\lambda}_1 \le \lambda_1 < 1$ (Hobert and Marchev, 2008). For a positive integer k, we will denote $\sum_{i=0}^{\tilde{\kappa}} \tilde{\lambda}_i^k$ by $\tilde{s}_k$. Henceforth, we assume that $\tilde{p}(u, u')$ is representable and we treat $\tilde{P}$ as a DA operator.

It follows from Theorem 1 that in order to find s_k or $\tilde{s}_k$, all we need to do is evaluate $\int_{S_U} p^{(k)}(u, u)\, \mu(du)$ or $\int_{S_U} \tilde{p}^{(k)}(u, u)\, \mu(du)$, where $\tilde{p}^{(k)}(u, \cdot)$ is the k-step Mtd of the sandwich chain. Of course, calculating these integrals (in non-toy problems) is nearly always impossible, even for $k = 1$. In the next section, we introduce a method of estimating these two integrals using classical Monte Carlo.

Throughout the remainder of the paper, we assume that P is a DA operator with Mtd given by (3), and that (4) holds.

5 Classical Monte Carlo

Consider the Mtd given by

$$a(u, \cdot) = \int_{S_V} \int_{S_V} \pi_{U|V}(\cdot|v') r(v, dv') \pi_{V|U}(v|u)\, \nu(dv)\,, \tag{8}$$

where $r(v, \cdot)$, $v \in S_V$ is an Mtf on S_V with invariant pdf $\pi_V(\cdot)$. We will show in this section that this form has utility beyond constructing sandwich algorithms. Indeed, the k-step Mtd of a DA algorithm (or a sandwich algorithm) can be re-expressed in the form (8). This motivates the development of a general method for estimating the integral $\int_{S_U} a(u, u)\, \mu(du)$, which is the main topic of this section.

We begin by showing how $p^{(k)}(u, \cdot)$, $u \in S_U$ can be written in the form (8). The case $k = 1$ is trivial. Indeed, if $r(v, \cdot)$ is taken to be the kernel of the identity operator, then $a(u, \cdot) = p(u, \cdot)$. Define an Mtd $q(v, \cdot)$, $v \in S_V$ by

$$q(v, \cdot) = \int_{S_U} \pi_{V|U}(\cdot|u) \pi_{U|V}(u|v)\, \mu(du),$$

and let $q^{(k)}(v, \cdot)$, $k \ge 1$ denote the corresponding k-step Mtd. If we let

$$r(v, dv') = q^{(k-1)}(v, v')\, \nu(dv'),\ v \in S_V$$

for $k \ge 2$, then $a(u, \cdot) = p^{(k)}(u, \cdot)$. Next, consider the sandwich Mtd $\tilde{p}^{(k)}(u, \cdot)$, $u \in S_U$. Again, the $k = 1$ case is easy. Taking

$$r(v, \cdot) = s(v, \cdot)$$

yields $a(u, \cdot) = \tilde{p}(u, \cdot)$. Now let

$$\tilde{q}(v, \cdot) = \int_{S_U} \int_{S_V} s(v', \cdot) \pi_{V|U}(v'|u) \pi_{U|V}(u|v)\, \nu(dv')\, \mu(du)\,,$$

and denote the corresponding k-step transition function by $\tilde{q}^{(k)}(v, \cdot)$. Then taking

$$r(v, \cdot) = \int_{S_V} \tilde{q}^{(k-1)}(v', \cdot) s(v, dv')$$

when $k \geq 2$ yields $a(u, \cdot) = \tilde{p}^{(k)}(u, \cdot)$.

The following proposition shows that, when P is trace-class, $\int_{S_U} a(u,u)\,\mu(du)$ is finite.

Proposition 2. $\int_{S_U} a(u,u)\,\mu(du) < \infty$.

Proof. That $\int_{S_U} a(u,u)\,\mu(du) < \infty$ is equivalent to

$$\int_{S_U} \int_{S_V} \left(\int_{S_V} \frac{\pi_{U,V}(u,v')}{\pi_U(u)\pi_V(v')} r(v, dv') \right) \left(\frac{\pi_{U,V}(u,v)}{\pi_U(u)\pi_V(v)} \right) \pi_U(u)\pi_V(v)\,\nu(dv)\,\mu(du) < \infty. \tag{9}$$

Note that

$$\int_{S_U} \left(\frac{\pi_{U,V}(u,v)}{\pi_U(u)\pi_V(v)} \right)^2 \pi_U(u)\pi_V(v)\,\mu(du)\nu(dv) = \int_{S_U} p(u,u)\,\mu(du) < \infty. \tag{10}$$

and by Jensen's inequality,

$$\begin{aligned}
&\int_{S_U} \int_{S_V} \left(\int_{S_V} \frac{\pi_{U,V}(u,v')}{\pi_U(u)\pi_V(v')} r(v, dv') \right)^2 \pi_U(u)\pi_V(v)\,\nu(dv)\,\mu(du) \\
&\leq \int_{S_U} \int_{S_V} \int_{S_V} \left(\frac{\pi_{U,V}(u,v')}{\pi_U(u)\pi_V(v')} \right)^2 r(v, dv')\pi_U(u)\pi_V(v)\,\nu(dv)\,\mu(du) \\
&= \int_{S_U} \int_{S_V} \left(\frac{\pi_{U,V}(u,v')}{\pi_U(u)\pi_V(v')} \right)^2 \pi_U(u)\pi_V(v')\,\nu(dv')\,\mu(du) \\
&= \int_{S_U} p(u,u)\,\mu(du) \\
&< \infty.
\end{aligned} \tag{11}$$

The inequality (9) follows from (10), (11), and the Cauchy-Schwarz inequality. $\square$

Combining Proposition 2 and Theorem 1 leads to the following result: If P is trace-class and $\tilde{p}(u,\cdot)$ is representable, then $\tilde{P}$ is also trace-class. This is a generalization of Khare and Hobert's (2011) Theorem 1, which states that, under a condition on $s(v, dv')$ that implies representability, the trace-class-ness of P implies that of $\tilde{P}$.

We now develop a classical Monte Carlo estimator of $\int_{S_U} a(u,u)\,\mu(du)$. Let $\omega : S_V \to [0, \infty)$ be a pdf that is almost everywhere positive. We will exploit the following representation of the integral of interest:

$$\int_{S_U} a(u,u)\,\mu(du) = \int_{S_V} \int_{S_U} \left(\frac{\pi_{V|U}(v|u)}{\omega(v)} \right) \Big(\int_{S_V} \pi_{U|V}(u|v') r(v, dv') \Big) \omega(v)\,\mu(du)\,\nu(dv)\,. \tag{12}$$

Clearly,

$$\eta(u,v) := \Big(\int_{S_V} \pi_{U|V}(u|v') r(v, dv') \Big) \omega(v)$$

defines a pdf on $S_U \times S_V$, and if (U^*, V^*) has joint pdf $\eta(\cdot,\cdot)$, then

$$\int_{S_U} a(u,u)\,\mu(du) = \mathbb{E}\Big(\frac{\pi_{V|U}(V^*|U^*)}{\omega(V^*)}\Big)\,.$$

Therefore, if $\{(U_i^*, V_i^*)\}_{i=1}^N$ are iid random elements from $\eta(\cdot,\cdot)$, then

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\pi_{V|U}(V_i^*|U_i^*)}{\omega(V_i^*)} \tag{13}$$

is a strongly consistent and unbiased estimator of $\int_{S_U} a(u,u)\,\mu(du)$. This is the Monte Carlo formula that is central to our discussion.

Of course, we are mainly interested in the cases $a(u,\cdot) = p^{(k)}(u,\cdot)$ or $a(u,\cdot) = \tilde{p}^{(k)}(u,\cdot)$. We now develop algorithms for drawing from $\eta(\cdot,\cdot)$ in these two situations. First, assume $a(u,\cdot) = p^{(k)}(u,\cdot)$. If $k = 1$, then $r(u,\cdot)$ is the kernel of the identity operator, and

$$\eta(u,v) = \pi_{U|V}(u|v)\omega(v)\,.$$

If $k \geq 2$, then $r(v,dv') = q^{(k-1)}(v,v')\,dv'$, and

$$\eta(u,v) = \Big(\int_{S_V} \pi_{U|V}(u|v')q^{(k-1)}(v,v')\,\nu(dv')\Big)\omega(v) = \Big(\int_{S_U} p^{(k-1)}(u',u)\pi_{U|V}(u'|v)\,\mu(du')\Big)\omega(v)\,.$$

Thus, when $k \geq 2$, we can draw from $\eta(u,v)$ as follows: Draw $V^* \sim \omega(\cdot)$, then draw $U' \sim \pi_{U|V}(\cdot|v^*)$, then draw $U^* \sim p^{(k-1)}(u',\cdot)$, and return (u^*, v^*). Of course, we can draw from $p^{(k-1)}(u',\cdot)$ by simply running $k-1$ iterations of the original DA algorithm from starting value u'. We formalize all of this in Algorithm 1.

Algorithm 1: Drawing $(U^*, V^*) \sim \eta(u,v)\,\mu(du)\,\nu(dv)$ when $a(u,\cdot) = p^{(k)}(u,\cdot)$.

1. Draw V^* from $\omega(\cdot)$.

2. Given $V^* = v^*$, draw U' from $\pi_{U|V}(\cdot|v^*)$.

3. If $k = 1$, set $U^* = U'$. If $k \geq 2$, given $U' = u'$, draw U^* from $p^{(k-1)}(u',\cdot)$ by running $k-1$ iterations of the DA algorithm.

Similar arguments lead to the following algorithm for the sandwich algorithm

---

Algorithm 1S: Drawing $(U^*, V^*) \sim \eta(u,v)\,\mu(du)\,\nu(dv)$ when $a(u,\cdot) = \tilde{p}^{(k)}(u,\cdot)$

1. Draw V^* from $\omega(\cdot)$.

2. Given $V^* = v^*$, draw V' from $s(v^*, \cdot)$.

3. Given $V' = v'$ draw U' from $\pi_{U|V}(\cdot|v')$.

4. If $k = 1$, set $U^* = U'$. If $k \geq 2$, given $U' = u'$, draw U^* from $\tilde{p}^{(k-1)}(u', \cdot)$ by running $k-1$ iterations of the sandwich algorithm.

It is important to note that we do not need to know the representing conditionals $\pi_{U|\tilde{V}}(\cdot|v)$ and $\pi_{\tilde{V}|U}(\cdot|u)$ from (7) in order to run Algorithm 1S.

As with all classical Monte Carlo techniques, the key to successful implementation is a finite variance. Define

$$D^2 = \text{var}\bigg(\frac{\pi_{V|U}(V^*|U^*)}{\omega(V^*)}\bigg).$$

Of course, $D^2 < \infty$ if and only if

$$\int_{S_V} \int_{S_U} \bigg(\frac{\pi_{V|U}(v|u)}{\omega(v)}\bigg)^2 \eta(u, v)\, \mu(du)\, \nu(dv) < \infty\,. \tag{14}$$

The following theorem provides a sufficient condition for finite variance.

Theorem 2. *The variance, D^2, is finite if*

$$\int_{S_V} \int_{S_U} \frac{\pi^3_{V|U}(v|u)\pi_{U|V}(u|v)}{\omega^2(v)}\, \mu(du)\, \nu(dv) < \infty. \tag{15}$$

Proof. First, note that (14) is equivalent to

$$\int_{S_V} \int_{S_U} \bigg(\frac{\pi^2_{V|U}(v|u)}{\pi_V(v)\omega(v)}\bigg)\bigg(\frac{\int_{S_V} \pi_{U|V}(u|v')r(v, dv')}{\pi_U(u)}\bigg)\pi_U(u)\pi_V(v)\, \mu(du)\, \nu(dv) < \infty.$$

Now, it follows from (15) that

$$\int_{S_V} \int_{S_U} \bigg(\frac{\pi^2_{V|U}(v|u)}{\pi_V(v)\omega(v)}\bigg)^2 \pi_U(u)\pi_V(v)\, \mu(du)\, \nu(dv) < \infty. \tag{16}$$

Moreover, by Jensen's inequality,

$$\begin{aligned}
&\int_{S_V} \int_{S_U} \bigg(\frac{\int_{S_V} \pi_{U|V}(u|v')r(v, dv')}{\pi_U(u)}\bigg)^2 \pi_U(u)\pi_V(v)\, \mu(du)\, \nu(dv) \\
&\leq \int_{S_V} \int_{S_U} \int_{S_V} \bigg(\frac{\pi_{U|V}(u|v')}{\pi_U(u)}\bigg)^2 r(v, dv')\pi_U(u)\pi_V(v)\, \mu(du)\, \nu(dv) \\
&= \int_{S_V} \int_{S_U} \bigg(\frac{\pi_{U|V}(u|v')}{\pi_U(u)}\bigg)^2 \pi_U(u)\pi_V(v')\, \mu(du)\, \nu(dv') \\
&= \int_{S_U} p(u, u)\, \mu(du) \\
&< \infty.
\end{aligned} \tag{17}$$

The conclusion now follows from (16), (17), and Cauchy-Schwarz. □

Theorem 2 implies that an $\omega(\cdot)$ with heavy tails is more likely to result in finite variance (which is not surprising). It might seem natural to take $\omega(\cdot) = \pi_V(\cdot)$. However, in practice, we are never able to draw from $\pi_V(\cdot)$. (If we could do that, we would not need MCMC). Moreover, in Section 6 we provide an example where taking $\omega(\cdot)$ to be $\pi_V(\cdot)$ leads to an infinite variance, whereas a heavier-tailed alternative leads to a finite variance.

Let $\psi : S_U \to [0,\infty)$ be a pdf that is positive almost everywhere. The following dual of (12) may also be used to represent $\int_{S_U} a(u,u)\,\mu(du)$:

$$\int_{S_U} a(u,u)\,\mu(du) = \int_{S_U}\int_{S_V}\int_{S_V} \frac{\pi_{U|V}(u|v)}{\psi(u)} r(v',dv)\pi_{V|U}(v'|u)\psi(u)\,\nu(dv')\,\mu(du)\,.$$

Now suppose that $\{(U_i^*, V_i^*)\}_{i=1}^N$ are iid from

$$\zeta(u,v)\,\mu(du)\,\nu(dv) = \left(\int_{S_V} r(v',dv)\pi_{V|U}(v'|u)\,\nu(dv')\right)\psi(u)\,\mu(du)\,.$$

The analogue of (13) is the following classical Monte Carlo estimator of $\int_{S_U} a(u,u)\,\mu(du)$:

$$\frac{1}{N}\sum_{i=1}^{N} \frac{\pi_{U|V}(U_i^*|V_i^*)}{\psi(U_i^*)}\,. \tag{18}$$

We now state the obvious analogues of Algorithms 1 and 1S.

Algorithm 2: Drawing $(U^*, V^*) \sim \zeta(u,v)\,\mu(du)\,\nu(dv)$ when $a(u,\cdot) = p^{(k)}(u,\cdot)$.

1. Draw U^* from $\psi(\cdot)$.
2. If $k = 1$, set $U' = U^*$. If $k \geq 2$, given $U^* = u^*$, draw U' from $p^{(k-1)}(u^*,\cdot)$.
3. Given $U' = u'$, draw V^* from $\pi_{V|U}(\cdot|u')$.

Algorithm 2S: Drawing $(U^*, V^*) \sim \zeta(u,v)\,\mu(du)\,\nu(dv)$ when $a(u,\cdot) = \tilde{p}^{(k)}(u,\cdot)$.

1. Draw U^* from $\psi(\cdot)$.
2. If $k = 1$, set $U' = U^*$. If $k \geq 2$, given $U^* = u^*$, draw U' from $\tilde{p}^{(k-1)}(u^*,\cdot)$.
3. Given $U' = u'$, draw V' from $\pi_{V|U}(\cdot|u')$.
4. Given $V' = v'$, draw V^* from $s(v',\cdot)$.

To ensure that the variance of (18) is finite, we only need

$$\int_{S_U}\int_{S_V}\int_{S_V}\left(\frac{\pi_{U|V}(u|v)}{\psi(u)}\right)^2 r(v',dv)\pi_{V|U}(v'|u)\psi(u)\,\nu(dv')\,\mu(du)<\infty\,. \tag{19}$$

The following result is the analogue of Theorem 2.

Corollary 1. *The variance of* (18) *is finite if*

$$\int_{S_U}\int_{S_V}\frac{\pi^3_{U|V}(u|v)\pi_{V|U}(v|u)}{\psi^2(u)}\,\nu(dv)\,\mu(du)<\infty\,. \tag{20}$$

Proof. Note that the left hand side of (19) is equal to

$$\int_{S_U}\int_{S_V}\left(\int_{S_V}\frac{\pi^2_{U|V}(u|v)}{\psi(u)\pi_U(u)}r(v',dv)\right)\left(\frac{\pi_{V|U}(v'|u)}{\pi_V(v')}\right)\pi_U(u)\pi_V(v')\,\nu(dv')\,\mu(du)\,.$$

Apply the Cauchy-Schwarz inequality, then utilize Jensen's inequality to get rid of $r(v',dv)$, and finally make use of (20) and the fact that P is trace-class. $\square$

6 Examples

In this section, we apply our Monte Carlo technique to several common Markov operators. In particular, we examine one toy Markov chain, and two practically relevant Monte Carlo Markov chains. In the two real examples, we are able to take advantage of existing trace-class proofs to establish that (15) (or (20)) hold for suitable $\omega(\cdot)$ (or $\psi(\cdot)$).

6.1 Gaussian chain

We begin with a toy example. Let $S_U = S_V = \mathbb{R}$, $\pi_U(u)\propto \exp(-u^2)$, and

$$\pi_{V|U}(v|u)\propto \exp\left\{-4\left(v-\frac{u}{2}\right)^2\right\}.$$

Then

$$\pi_{U|V}(u|v)\propto \exp\{-2(u-v)^2\}\,.$$

This leads to one of the simplest DA chains known. Indeed, the Mtd,

$$p(u,\cdot)=\int_{\mathbb{R}}\pi_{U|V}(\cdot|v)\pi_{V|U}(v|u)\,dv\,,$$

can be evaluated in closed form, and turns out to be a normal pdf. The spectrum of the corresponding Markov operator, P, has been studied thoroughly (see e.g. Diaconis et al., 2008). It is easy to verify that (4) holds, so P is trace-class. In fact, $\kappa = \infty$, and for any non-negative integer i, $\lambda_i = 1/2^i$. Thus, the second largest eigenvalue, λ_1, and the spectral gap, δ, are both equal to $1/2$. Moreover, for any positive integer k,

$$s_k = \sum_{i=0}^{\infty} \frac{1}{2^{ik}} = \frac{1}{1 - 2^{-k}} \,.$$

We now pretend to be unaware of this spectral information, and we use (13) to estimate $\{s_k, l_k, u_k\}_{k=1}^4$. Recall that l_k and u_k are lower and upper bounds for λ_1, respectively. Note that

$$\int_{\mathbb{R}} \pi_{V|U}^3(v|u) \pi_{U|V}(u|v) \, du \propto \exp\Big(- \frac{6}{5} v^2 \Big) \,.$$

It follows that, if we take $\omega(v) \propto \exp(-v^2/2)$, then (15) holds, and our estimator of s_k has finite variance. We used a Monte Carlo sample size of $N = 1 \times 10^5$ to form our estimates, and the results are shown in Table 1.

Table 1: Estimated power sums of eigenvalues for the Gaussian chain

k	Est. s_k	Est. $D/\sqrt{N}$	Est. l_k	Est. u_k
1	1.996	0.004	0.000	0.996
2	1.331	0.004	0.333	0.575
3	1.142	0.004	0.429	0.522
4	1.068	0.004	0.482	0.511

Note that the estimates of the s_ks are quite good. We constructed 95% confidence intervals (CIs) for l_4 and u_4 via the delta method, and the results were $(0.442, 0.522)$ and $(0.498, 0.524)$, respectively.

Remark 1. *It is interesting that, if $\omega(\cdot)$ is set to be $\pi_V(\cdot)$, which seems natural, then* (15) *fails to hold. In fact, s_1 actually has infinite variance in this case. Indeed, recall that the estimator of s_1 given in* (13) *has the form*

$$\frac{1}{N} \sum_{i=1}^{N} \frac{\pi_{V|U}(V_i^*|U_i^*)}{\omega(V_i^*)},$$

where (U_i^*, V_i^*) are iid, and (U_1^*, V_1^*) has pdf given by

$$\eta(u,v) = \pi_{U|V}(u|v)\omega(v).$$

Hence, the variance of the estimator, D^2, is finite if and only if

$$\int_{S_U} \int_{S_V} \frac{\pi_{V|U}^2(v|u) \pi_{U|V}(u|v)}{\omega(v)} \, dv \, du < \infty. \tag{21}$$

If $\omega(\cdot) = \pi_V(\cdot) \propto \exp(-2v^2)$, *then the left hand side of* (21) *becomes proportional to*

$$\int_{S_U}\int_{S_V} \exp\Big\{-8\Big(v-\frac{u}{2}\Big)^2 - 2(u-v)^2 + 2v^2\Big\}\, dv\, du = \int_{S_U}\int_{S_V} \exp(-8v^2 + 12uv - 4u^2)\, dv\, du,$$

which is infinite.

6.2 Bayesian linear regression model with non-Gaussian errors

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

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

where $x_i \in \mathbb{R}^p$ is known, while $\beta \in \mathbb{R}^{p\times d}$ and the $d \times d$ positive definite matrix Σ are to be estimated. The iid errors, $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$, are assumed to have a pdf that is a scaled mixture of Gaussian densities:

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

where $h(\cdot)$ is a pdf with positive support, and $\mathbb{R}_+ := (0,\infty)$. For instance, if $d = 1$ and $h(u) \propto u^{-2}e^{-1/(8u)}$, then ε_1 has pdf proportional to $e^{-|\varepsilon|/2}$.

To perform a Bayesian analysis, we require a prior on the unknown parameter, (β, Σ). We adopt the (improper) Jeffreys prior, given by $1/|\Sigma|^{(d+1)/2}$. Let y represent the $n \times d$ matrix whose ith row is the observed value of Y_i. The following four conditions, which are sufficient for the resulting posterior to be proper (Qin and Hobert, 2016; Fernandez and Steel, 1999), will be assumed to hold:

1. $n \geq p + d$,
2. $(X : y)$ is full rank, where X is the $n \times p$ matrix whose ith row is x_i^T,
3. $\int_{\mathbb{R}_+} u^{d/2} h(u)\, du < \infty$, and
4. $\int_{\mathbb{R}_+} u^{-(n-p-d)/2} h(u)\, du < \infty$.

The posterior density is highly intractable, but there is a well-known DA algorithm to sample from it (Liu, 1996). Under our framework, the DA chain Φ is characterized by the Mtd

$$p\big((\beta,\Sigma),\cdot\big) = \int_{\mathbb{R}_+^n} \pi_{U|V}(\cdot|z)\pi_{V|U}(z|\beta,\Sigma)\, dz,$$

where $z = (z_1, z_2, \dots, z_n)^T$,

$$\pi_{U|V}(\beta,\Sigma|z) \propto |\Sigma|^{-(n+d+1)/2} \prod_{i=1}^n \exp\Big\{-\frac{z_i}{2}\big(y_i - \beta^T x_i\big)^T \Sigma^{-1}\big(y_i - \beta^T x_i\big)\Big\}, \text{ and}$$

$$\pi_{V|U}(z|\beta,\Sigma) \propto \prod_{i=1}^n z_i^{d/2} \exp\Big\{-\frac{z_i}{2}\big(y_i - \beta^T x_i\big)^T \Sigma^{-1}\big(y_i - \beta^T x_i\big)\Big\} h(z_i)\,.$$

The first conditional density, $\pi_{U|V}(\cdot,\cdot|z)$, characterizes a multivariate normal distribution on top of an inverse Wishart distribution, i.e. $\beta|\Sigma, z$ is multivariate normal, and $\Sigma|z$ is inverse Wishart. The second conditional density, $\pi_{V|U}(\cdot|\beta,\Sigma)$, is a product of n univariate densities. Moreover, when $h(\cdot)$ is a standard pdf on $\mathbb{R}_+$, these univariate densities are often members of a standard parametric family. The following proposition about the resulting DA operator is proved in Qin and Hobert (2016).

Proposition 3. *Suppose $h(\cdot)$ is strictly positive in a neighborhood of the origin. If there exists $\xi \in (1,2)$ and $\delta > 0$ such that*

$$\int_0^\delta \frac{u^{d/2}h(u)}{\int_0^{\xi u} v^{d/2}h(v)\,dv}\,du < \infty,$$

then P is trace-class.

When P is trace-class, we can pick an $\omega(\cdot)$ and try to make use of (13). A sufficient condition for the estimator's variance to be finite is stated in the following proposition, whose proof is given in the appendix.

Proposition 4. *Suppose that $h(\cdot)$ is strictly positive in a neighborhood of the origin. If $\omega(z)$ can be written as $\prod_{i=1}^n \omega_i(z_i)$, and there exists $\xi \in (1,4/3)$ such that for all $i \in \{1,2,\ldots,n\}$,*

$$\int_{\mathbb{R}_+} \frac{u^{3d/2}h^3(u)}{(\int_0^{\xi u} v^{d/2}h(v)\,dv)^3 \omega_i^2(u)}\,du < \infty, \tag{22}$$

then (15) *holds, and thus by Theorem 2, the estimator* (13) *has finite variance.*

For illustration, take $d = 1$ and $h(u) \propto u^{-2}e^{-1/(8u)}$. Then ε_1 follows a scaled Laplace distribution, and the model can be viewed as a median regression model with variance Σ unknown. It's easy to show that $h(\cdot)$ satisfies the assumptions in Proposition 3, so the resultant DA operator is trace-class. (This result was actually first proven by Choi and Hobert (2013).) Now let

$$\omega(z) = \prod_{i=1}^n \omega_i(z_i) \propto \prod_{i=1}^n z_i^{-3/2} e^{-1/(32 z_i)}\,.$$

The following result shows that this will lead to an estimator with finite variance.

Corollary 2. *Suppose $d = 1$, $h(u) \propto u^{-2}e^{-1/(8u)}$, and*

$$\omega(z) = \prod_{i=1}^n \omega_i(z_i) \propto \prod_{i=1}^n z_i^{-\alpha-1} e^{-\gamma/z_i}\,,$$

where $0 < \alpha < 3/4$ and $0 < \gamma < 3/64$. Then (15) *holds.*

Proof. In light of Proposition 4, we only need to show that (22) holds for some $\xi \in (1, 4/3)$. Making use of the fact that (by monotone convergence theorem)

$$\lim_{u\to\infty}\int_0^{\xi u} v^{1/2}h(v)\,dv = \int_{\mathbb{R}_+} u^{1/2}h(u)\,du > 0\,,$$

one can easily show for any $\delta > 0$,

$$\int_\delta^\infty \frac{u^{3/2}h^3(u)}{(\int_0^{\xi u} v^{1/2}h(v)\,dv)^3\omega_i^2(u)}\,du = \int_\delta^\infty \frac{u^{2\alpha-5/2}\exp\{2\gamma/u - 3/(8u)\}}{(\int_0^{\xi u} v^{1/2}h(v)\,dv)^3}\,du < \infty. \tag{23}$$

On the other hand, using L'Hôpital's rule, we can see for $(1-16\gamma/3)^{-1} < \xi < 4/3$,

$$\begin{aligned}\lim_{u\to 0}\left(\frac{u^{3/2}h^3(u)}{(\int_0^{\xi u} v^{1/2}h(v)\,dv)^3\omega_i^2(u)}\right)^{1/3} &= \lim_{u\to 0}\frac{u^{2\alpha/3-5/6}\exp\{2\gamma/(3u)-1/(8u)\}}{\int_0^{\xi u} v^{-3/2}e^{-1/(8v)}\,dv}\\ &= \lim_{u\to 0} R(u)\exp\Big\{-\Big(-\frac{2\gamma}{3}-\frac{1}{8\xi}+\frac{1}{8}\Big)\frac{1}{u}\Big\}\\ &= 0,\end{aligned}$$

where $R(u)$ is a function that is either bounded near the origin or goes to ∞ at the rate of some power function as $u \to 0$. It follows that for some small enough δ,

$$\int_0^\delta \frac{u^{3/2}h^3(u)}{(\int_0^{\xi u} v^{1/2}h(v)\,dv)^3\omega_i^2(u)}\,du < \infty. \tag{24}$$

Combining (23) and (24) yields (22). □

We now test the efficiency of the Monte Carlo estimator (13) using some simulated data with $d = 1$. Here are our simulated X and y:

$$X = \begin{pmatrix} 1 & 2.32 & 1 & 0 & 2.32 & 0\\ 1 & 5.65 & 1 & 0 & 5.65 & 0\\ 1 & -7.69 & 1 & 0 & -7.69 & 0\\ 1 & 3.59 & 0 & 1 & 0 & 3.59\\ 1 & 5.57 & 0 & 1 & 0 & 5.57\\ 1 & -9.99 & 0 & 1 & 0 & -9.99\\ 1 & -18.88 & 0 & 1 & 0 & -18.88\\ 1 & 5.95 & 0 & 0 & 0 & 0\\ 1 & -16.39 & 0 & 0 & 0 & 0\\ 1 & 4.75 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad y = \begin{pmatrix} 0.14\\ 2.99\\ 1.37\\ -2.55\\ -3.60\\ 14.86\\ 21.24\\ 2.78\\ 9.14\\ 2.06 \end{pmatrix}.$$

The simulation was was based on a linear model containing an intercept, one continuous covariate, a single factor with three levels, and an interaction between the two. The elements in the second column of X were independently generated from $N(0,100)$. Once X was simulated, we generated the data according to

$$Y_i = \beta_*^T x_i + \sqrt{\Sigma^*}\epsilon_i \,,$$

where $\beta_* = (1,0,-0.5,0.5,0,-1)^T$, $\Sigma^* = 1$, $\{\epsilon_i\}_{i=1}^{10}$ are iid, and ϵ_1 has pdf given by $f_h(\cdot)$ when $h(u) \propto u^{-2}e^{-1/(8u)}$. That is, the errors have a Laplace distribution. Then the DA chain Φ lives in $S_U = \mathbb{R}^6 \times \mathbb{R}_+$, and S_V is $\mathbb{R}_+^{10}$. We used a Monte Carlo sample size of $N = 2 \times 10^6$ to form estimates of $\{s_k, l_k, u_k\}_{k=1}^4$, and the results are shown in Table 2.

Table 2: Estimated power sums of eigenvalues for the DA chain for Bayesian linear regression with non-Gaussian errors

k	Est. s_k	Est. $D/\sqrt{N}$	Est. l_k	Est. u_k
1	35.587	0.121	0.000	34.587
2	2.465	0.020	0.042	1.210
3	1.325	0.014	0.222	0.687
4	1.102	0.012	0.313	0.564

The asymptotic 95% CIs for l_4 and u_4 are $(0.241, 0.383)$ and $(0.532, 0.597)$, respectively. Also, using a Bonferroni argument, we may conclude that asymptotically, with at least 95% confidence, $\lambda_1 \in (0.241, 0.597)$. The Monte Carlo sample size required to secure a reasonable estimate does increase with n and p. For example, in another example we considered where $n = 20$ and $p = 4$, we needed a Monte Carlo sample size of 1×10^7 to get decent results.

6.3 Bayesian probit regression

Let $Y_1, Y_2, \ldots, Y_n$ be independent Bernoulli random variables with $\mathbb{P}(Y_1 = 1|\beta) = \Phi(x_i^T\beta)$, where $x_i, \beta \in \mathbb{R}^p$, and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Take the prior on β to be $N_p(Q^{-1}v, Q^{-1})$, where $v \in \mathbb{R}^p$ and Q is positive definite. The resulting posterior distribution is intractable, but Albert and Chib (1993) devised a DA algorithm to sample from it. Let $z = (z_1, z_2, \ldots, z_n)^T$ be a vector of latent variables, and let X be the design matrix whose ith row is x_i^T. The Mtd of the Albert and Chib (AC) chain, $p(\beta, \cdot), \beta \in \mathbb{R}^p$, is characterized by

$$\pi_{U|V}(\beta|z) \propto \exp\left[-\frac{1}{2}\left\{\beta - \left(X^TX+Q\right)^{-1}\left(v+X^Tz\right)\right\}^T\left(X^TX+Q\right)\left\{\beta - \left(X^TX+Q\right)^{-1}\left(v+X^Tz\right)\right\}\right],$$

and

$$\pi_{V|U}(z|\beta) \propto \prod_{i=1}^n \exp\left\{-\frac{1}{2}\left(z_i - x_i^T\beta\right)^2\right\} I_{\mathbb{R}_+}\left((y_i - 0.5)z_i\right).$$

The first conditional density, $\pi_{U|V}(\cdot|z)$, is a multivariate normal density, and the second conditional density, $\pi_{V|U}(\cdot|\beta)$, is a product of univariate truncated normal pdfs.

A sandwich step can be added to facilitate the convergence of the AC chain. Chakraborty and Khare (2017) constructed a Haar PX-DA variant of the chain, which is a sandwich chain with transition density of the form (6) (see also Roy and Hobert (2007)). The sandwich step $s(v, dv')$ is equivalent to the following update: $z \mapsto z' = gz$, where the scalar g is drawn from the following density:

$$\pi_G(g|z) \propto g^{n-1} \exp\left[-\frac{1}{2}z^T\left\{I_n - X(X^TX+Q)^{-1}X^T\right\}zg^2 + z^TX(X^TX+Q)^{-1}vg\right].$$

Note that this pdf is particularly easy to sample from when $v = 0$.

Chakraborty and Khare showed that P is trace-class when one uses a concentrated prior (corresponding to Q having large eigenvalues). In fact, the following is shown to hold in their proof.

Proposition 5. *Suppose that X is full rank. If all the eigenvalues of $Q^{-1/2}X^TXQ^{-1/2}$ are less than $7/2$, then for any polynomial function $t : \mathbb{R}^p \to \mathbb{R}$,*

$$\int_{\mathbb{R}^p} |t(\beta)| p(\beta, \beta)\, d\beta < \infty.$$

We will use the estimator (18). The following proposition provides a class of $\psi(\cdot)$s that lead to estimators with finite variance.

Proposition 6. *Suppose the hypothesis in Proposition 5 holds. If $\psi(\cdot)$ is the pdf of a p-variate t-distribution, i.e.*

$$\psi(\beta) \propto \left\{1 + \frac{1}{a}(\beta - b)^T\Sigma^{-1}(\beta - b)\right\}^{-(a+p)/2}$$

for some $b \in \mathbb{R}^p$, positive definite matrix $\Sigma \in \mathbb{R}^{p\times p}$, and positive integer a, then the estimator (18) *has finite variance.*

Proof. Note that for every β and z

$$\pi_{U|V}^3(\beta|z) \leq C\pi_{U|V}(\beta|z),$$

where C is a constant. Hence, for any polynomial function $t : \mathbb{R}^p \to \mathbb{R}$,

$$\int_{\mathbb{R}^p} \int_{\mathbb{R}^n} |t(\beta)| \pi_{U|V}^3(\beta|z) \pi_{V|U}(z|\beta)\, dz\, d\beta \le C \int_{\mathbb{R}^p} |t(\beta)| p(\beta, \beta)\, d\beta < \infty.$$

Since $\psi^{-2}(\cdot)$ is a polynomial function on $\mathbb{R}^p$, the moment condition (20) holds. The result follows from Corollary 1. $\square$

As a numerical illustration, we apply our method to the Markov operator associated with the AC chain corresponding to the famous "lupus data" of van Dyk and Meng (2001). In this data set, $n = 55$ and $p = 3$. As in Chakraborty and Khare (2017), we will let $v = 0$ and $Q = X^T X / g$, where $g = 3.499999$. It can be easily shown that the assumptions in Proposition 5 are met. Chakraborty and Khare compared the AC chain, Φ, and its Haar PX-DA variant, $\tilde{\Phi}$, defined a few paragraphs ago. This comparison was done using estimated autocorrelations. Their results suggest that $\tilde{\Phi}$ outperforms Φ significantly when estimating a certain test function. We go further and estimate the second largest eigenvalue of each operator.

Let $\hat{\beta}$ be the MLE of β, and $\hat{\Sigma}$ its estimated (asymptotic) variance. We pick $\psi(\cdot)$ to be the pdf of $t_1(\hat{\beta}, \hat{\Sigma})$. This is to say, for any $\beta \in \mathbb{R}^p$,

$$\psi(\beta) \propto \left\{ 1 + (\beta - \hat{\beta})^T \hat{\Sigma}^{-1} (\beta - \hat{\beta}) \right\}^{-(p+1)/2}.$$

By Proposition 6, this choice of $\psi(\cdot)$ is justified.

We used a Monte Carlo sample size of $N = 4 \times 10^6$ to form our estimates for the DA operator, and the results are shown in Table 3. Asymptotic 95% CIs for l_4 and u_4 are $(0.389, 0.553)$ and $(0.622, 0.691)$, respectively. Again, using Bonferroni, we can state that asymptotically, with at least 95% confidence, $\lambda_1 \in (0.389, 0.691)$.

Table 3: Estimated power sums of eigenvalues for the AC chain

k	Est. s_k	Est. $D'/\sqrt{N}$	Est. l_k	Est. u_k
1	6.805	0.049	0.000	5.805
2	2.065	0.026	0.183	1.032
3	1.393	0.022	0.369	0.733
4	1.185	0.020	0.471	0.656

We now consider the sandwich chain, $\tilde{\Phi}$. One can show that the Mtd of any Haar PX-DA chain is representable (Hobert and Marchev, 2008). Hence, $\tilde{P}$ is indeed a DA operator. Recall that $\{\tilde{\lambda}_i\}_{i=0}^{\tilde{\kappa}}$, $0 \le \tilde{\kappa} \le$

∞, denote the decreasingly ordered positive eigenvalues of $\tilde{P}$. It was shown in Khare and Hobert (2011) that $\tilde{\lambda}_i \leq \lambda_i$ for $i \in \mathbb{N}$ with at least one strict inequality. For a positive integer k, $\sum_{i=0}^{\tilde{\kappa}} \tilde{\lambda}_i^k$ is denoted by $\tilde{s}_k$. Let $\tilde{u}_k$, $\tilde{l}_k$, and $\tilde{D}'$ be the respective counterparts of u_k, l_k, and D'. Estimates of $\tilde{s}_k,\ k = 1, 2, 3, 4$ using 4×10^6 Monte Carlo samples are given in Table 4. Our estimate of $\tilde{s}_1 - 1$ is less than half of $s_1 - 1$, implying that, in an average sense, the sandwich version of the AC chain reduces the nontrivial eigenvalues of P by more than half. Asymptotic 95% CIs for l_4 and u_4 are $(0.175, 0.604)$ and $(0.454, 0.587)$. Thus, asymptotically, with at least 95% confidence, $\lambda_1 \in (0.175, 0.587)$. The rather wide confidence intervals are due to the large coefficient of variation of $\tilde{s}_k - 1$ when $k = 4$. If we instead use $\tilde{u}_3$ and $\tilde{l}_3$ to construct a CI for $\tilde{\lambda}_1$, we obtain $(0.273, 0.610)$, which is actually narrower than what we got from $\tilde{u}_4$ and $\tilde{l}_4$. To decrease the coefficient of variation of $\tilde{s}_4 - 1$, and thus improve the CI of $\tilde{\lambda}_1$ constructed based on $\tilde{u}_4$ and $\tilde{l}_4$, one would have to increase the Monte Carlo sample size N.

Table 4: Estimated power sums of eigenvalues for the Haar PX-DA version of the AC chain

k	Est. $\tilde{s}_k$	Est. $\tilde{D}'/\sqrt{N}$	Est. $\tilde{l}_k$	Est. $\tilde{u}_k$
1	3.774	0.038	0.000	2.774
2	1.557	0.023	0.201	0.746
3	1.189	0.020	0.339	0.573
4	1.073	0.019	0.390	0.521

Acknowledgment. The second and third authors were supported by NSF Grant DMS-15-11945.

Appendix

A Proof of Theorem 1

Theorem 1. *The DA operator P is trace-class if and only if*

$$\int_{S_U} p(u, u)\, \mu(du) < \infty. \tag{4}$$

If (4) *holds, then for any positive integer k,*

$$s_k := \sum_{i=0}^{\kappa} \lambda_i^k = \int_{S_U} p^{(k)}(u, u)\, \mu(du) < \infty. \tag{5}$$

Proof. Note that P is self-adjoint and non-negative. Let $\{g_i\}_{i=0}^{\infty}$ be an orthonormal basis of $L^2(\pi_U)$. The operator P is defined to be trace-class if (see e.g. Conway, 2000)

$$\sum_{i=0}^{\infty} \langle P g_i, g_i \rangle_{\pi_U} < \infty. \tag{25}$$

This condition is equivalent to P being compact with summable eigenvalues. To show that P being trace-class is equivalent to (4), we will prove a stronger result, namely

$$\sum_{i=0}^{\infty} \langle P g_i, g_i \rangle_{\pi_U} = \int_{S_U} p(u,u) \mu(du). \tag{26}$$

We begin by defining two new Hilbert spaces. Let $L^2(\pi_V)$ be the Hilbert space consisting of functions that are square integrable with respect to the weight function $\pi_V(\cdot)$. For $f, g \in L^2(\pi_V)$, their inner product is defined, as usual, by

$$\langle f, g \rangle_{\pi_V} = \int_{S_V} f(v) \overline{g(v)} \pi_V(v) \, \nu(dv).$$

Let $L^2(\pi_U \times \pi_V)$ be the Hilbert space of functions on $S_U \times S_V$ that are square integrable with respect to the weight function $\pi_U(\cdot)\pi_V(\cdot)$. For $f, g \in L^2(\pi_U \times \pi_V)$, their inner product is

$$\langle f, g \rangle_{\pi_U \times \pi_V} = \int_{S_U \times S_V} f(u,v) \overline{g(u,v)} \pi_U(u) \pi_V(v) \, \mu(du) \, \nu(dv).$$

Note that $L^2(\pi_V)$ is separable. Let $\{h_j\}_{j=0}^{\infty}$ be an orthonormal basis of $L^2(\pi_V)$. It can be shown that $\{g_i h_j\}_{(i,j) \in \mathbb{Z}_+^2}$ is an orthonormal basis of $L^2(\pi_U \times \pi_V)$. Of course, $g_i h_j$ denotes the function given by $(g_i h_j)(u,v) = g_i(u) h_j(v)$.

The inequality (4) is equivalent to

$$\int_{S_U \times S_V} \left(\frac{\pi_{U,V}(u,v)}{\pi_U(u)\pi_V(v)} \right)^2 \pi_U(u) \pi_V(v) \, \mu(du) \, \nu(dv) < \infty,$$

which holds if and only if the function $\varphi : S_U \times S_V \to \mathbb{R}$ given by

$$\varphi(u,v) = \frac{\pi_{U,V}(u,v)}{\pi_U(u)\pi_V(v)}$$

is in $L^2(\pi_U \times \pi_V)$. Suppose (4) holds. Then by Parseval's identity,

$$\begin{aligned}
\int_{S_U} p(u,u)\,\mu(du) &= \langle \varphi, \varphi \rangle_{\pi_U \times \pi_V} \\
&= \sum_{(i,j)\in\mathbb{Z}_+^2} |\langle \varphi, g_i h_j \rangle_{\pi_U \times \pi_V}|^2 \\
&= \sum_{(i,j)\in\mathbb{Z}_+^2} \Big| \int_{S_U \times S_V} \overline{g_i(u) h_j(v)} \pi_{U,V}(u,v)\,\mu(du)\,\nu(dv) \Big|^2 \\
&= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \Big| \int_{S_V} \Big(\int_{S_U} \overline{g_i(u)} \pi_{U|V}(u|v)\,\mu(du) \Big) \overline{h_j(v)} \pi_V(v)\,\nu(dv) \Big|^2 .
\end{aligned}$$

Again by Parseval's identity, this time applied to the function on S_V (and in fact, in $L^2(\pi_V)$ by Jensen's inequality) given by

$$\varphi_i(v) = \int_{S_U} \overline{g_i(u)} \pi_{U|V}(u|v)\,\mu(du),$$

we have

$$\begin{aligned}
\int_{S_U} p(u,u)\,\mu(du) &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} |\langle \varphi_i, h_j \rangle_{\pi_V}|^2 \\
&= \sum_{i=0}^{\infty} \langle \varphi_i, \varphi_i \rangle_{\pi_V} \\
&= \sum_{i=0}^{\infty} \int_{S_V} \Big| \int_{S_U} \overline{g_i(u)} \pi_{U|V}(u|v)\,\mu(du) \Big|^2 \pi_V(v)\,\nu(dv) \\
&= \sum_{i=0}^{\infty} \int_{S_V} \int_{S_U} \Big(\int_{S_U} g_i(u') \pi_{U|V}(u'|v) \pi_{V|U}(v|u)\,\mu(du') \Big) \overline{g_i(u)} \pi_U(u)\,\mu(du)\,\nu(dv) \\
&= \sum_{i=0}^{\infty} \int_{S_U} \Big(\int_{S_U} p(u,u') g_i(u')\,\mu(du') \Big) \overline{g_i(u)} \pi_U(u)\,\mu(du).
\end{aligned} \tag{27}$$

Note that the use of Fubini's theorem in the last equality can be easily justified by noting that $g_i \in L^2(\pi_U)$, and making use of Jensen's inequality. But the right hand side of (27) is precisely $\sum_{i=0}^{\infty} \langle P g_i, g_i \rangle_{\pi_U}$. Hence, (26) holds when $\int_{S_U} p(u,u)\,\mu(du)$ is finite.

To finish our proof of (26), we'll show (25) implies (4). Assume that (25) holds. Tracing backwards along (27) yields

$$\sum_{(i,j)\in\mathbb{Z}_+^2} |\langle \varphi_i, h_j \rangle_{\pi_V}|^2 < \infty.$$

This implies that the function

$$\tilde{\varphi} := \sum_{(i,j)\in\mathbb{Z}_+^2} \langle \varphi_i, h_j \rangle_{\pi_V} g_i h_j$$

is in $L^2(\pi_U \times \pi_V)$. Recall that (4) is equivalent to φ being in $L^2(\pi_U \times \pi_V)$. Hence, it suffices to show that $\tilde{\varphi}(u,v) = \varphi(u,v)$ almost everywhere. Define a linear transformation $T : L^2(\pi_U) \to L^2(\pi_V)$ by

$$Tf(v) = \int_{S_U} f(u)\pi_{U|V}(u|v)\,\mu(du), \quad \forall f \in L^2(\pi_U).$$

By Jensen's inequality, T is bounded, and thus, continuous. For any $g = \sum_{i=0}^{\infty} \alpha_i g_i \in L^2(\pi_U)$ and $h = \sum_{j=0}^{\infty} \beta_j h_j \in L^2(\pi_V)$,

$$\begin{aligned}
&\int_{S_V}\int_{S_U} \varphi(u,v)\overline{g(u)h(v)}\,\pi_U(u)\pi_V(v)\,\mu(du)\,\nu(dv) \\
&= \langle T\overline{g}, h\rangle_{\pi_V} \\
&= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \overline{\alpha_i\beta_j}\langle T\overline{g_i}, h_j\rangle_{\pi_V} \\
&= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \overline{\alpha_i\beta_j}\langle \varphi_i, h_j\rangle_{\pi_V} \\
&= \langle \tilde{\varphi}, gh\rangle_{\pi_U\times\pi_V} \\
&= \int_{S_V}\int_{S_U} \tilde{\varphi}(u,v)\overline{g(u)h(v)}\,\pi_U(u)\pi_V(v)\,\mu(du)\,\nu(dv),
\end{aligned}$$

where $\overline{g} \in L^2(\pi_V)$ is given by $\overline{g}(u) := \overline{g(u)}$, and $\overline{g_i}$ is defined similarly for $i \in \mathbb{Z}_+$. This implies that for any $C_1 \in \mathcal{U}$ and $C_2 \in \mathcal{V}$,

$$\int_{C_1\times C_2} \varphi(u,v)\,\pi_U(u)\pi_V(v)\,\mu(du)\,\nu(dv) = \int_{C_1\times C_2} \tilde{\varphi}(u,v)\,\pi_U(u)\pi_V(v)\,\mu(du)\,\nu(dv).$$

Note that

$$\int_{S_U\times S_V} |\tilde{\varphi}(u,v)|\,\pi_U(u)\pi_V(v)\,\mu(du)\,\nu(dv) \leq \langle\tilde{\varphi},\tilde{\varphi}\rangle_{\pi_U\times\pi_V}^{1/2} < \infty. \tag{28}$$

By (28) and the dominated convergence theorem, one can show that

$$\mathcal{A} := \left\{ C \in \mathcal{U}\times\mathcal{V} \,\middle|\, \int_C \varphi(u,v)\,\pi_U(u)\pi_V(v)\,\mu(du)\,\nu(dv) = \int_C \tilde{\varphi}(u,v)\,\pi_U(u)\pi_V(v)\,\mu(du)\,\nu(dv) \right\}$$

is a λ system. An application of Dynkin's π-λ theorem reveals that $\mathcal{U}\times\mathcal{V} \subset \mathcal{A}$. Therefore, $\tilde{\varphi}(u,v) = \varphi(u,v)$ almost everywhere, and (4) follows.

For the rest of the proof, assume that P is trace-class. This implies that P is compact, and thus admits the spectral decomposition (see e.g. Helmberg, 2014, §28 Corollary 2.1) given by

$$Pf = \sum_{i=0}^{\kappa} \lambda_i \langle f, f_i\rangle_{\pi_U} f_i, \quad f \in L^2(\pi_U) \tag{29}$$

where f_i, $i = 0, 1, \ldots, \kappa$, is the normalized eigenfunction corresponding to λ_i. By Parseval's identity,

$$\begin{aligned}\sum_{i=0}^{\infty} \langle P g_i, g_i \rangle_{\pi_U} &= \sum_{i=0}^{\infty} \sum_{j=0}^{\kappa} \lambda_j |\langle g_i, f_j \rangle_{\pi_U}|^2 \\ &= \sum_{j=0}^{\kappa} \lambda_j \langle f_j, f_j \rangle_{\pi_U} \\ &= \sum_{j=0}^{\kappa} \lambda_j .\end{aligned}$$

This equality is in fact a trivial case of Lidskii's theorem (see e.g. Erdös, 1974; Gohberg et al., 2012). It follows that (5) holds for $k = 1$.

We now consider the case where $k \geq 2$. By (29) and a simple induction, we have the following decomposition for P^k.

$$P^k f = \sum_{i=0}^{\kappa} \lambda_i^k \langle f, f_i \rangle_{\pi_U} f_i, \quad f \in L^2(\pi_U) .$$

Hence P^k is trace-class with ordered positive eigenvalues $\{\lambda_i^k\}_{i=0}^{\kappa}$. Note that P^k is a Markov operator whose Mtd is $p^{(k)}(u, \cdot)$, $u \in S_U$. Thus, in order to show that (5) holds for $k \geq 2$, it suffices to verify P^k is a DA operator, for then we can treat P^k as P and repeat our argument for the $k = 1$ case. To be specific, we'll show that there exists a random variable $\tilde{V}$ taking values on $S_{\tilde{V}}$, where $(S_{\tilde{V}}, \tilde{\mathcal{V}}, \tilde{\nu})$ is a σ-finite measure space and $\tilde{\mathcal{V}}$ is countably generated, such that for $u \in S_U$,

$$p^{(k)}(u, \cdot) = \int_{S_{\tilde{V}}} \pi_{U|\tilde{V}}(\cdot|v) \pi_{\tilde{V}|U}(v|u) \, \tilde{\nu}(dv), \tag{30}$$

where $\pi_{\tilde{V}}(\cdot)$, $\pi_{U|\tilde{V}}(\cdot|\cdot)$, and $\pi_{\tilde{V}|U}(\cdot|\cdot)$ have the apparent meanings.

Let $(U_k, V_k)_{k=0}^{\infty}$ be a Markov chain. Suppose that U_0 has pdf $\pi_U(\cdot)$, and for any non-negative integer k, let $V_k | U_k = u$ have pdf $\pi_{V|U}(\cdot|u)$, and let $U_{k+1} | V_k = v$ have pdf $\pi_{U|V}(\cdot|v)$. It's easy to see $\{U_k\}_{k=0}^{\infty}$ is a stationary DA chain with Mtd $p(u, \cdot)$. Suppose k is even. The pdf of $U_k | U_0 = u$ is

$$p^{(k)}(u, \cdot) = \int_{S_U} p^{(k/2)}(u, u') p^{(k/2)}(u', \cdot) \mu(du).$$

Meanwhile, since the chain is reversible and starts from the stationary distribution, $U_0 | U_{k/2} = u$ has the same distribution as $U_{k/2} | U_0 = u$, which is just $p^{(k/2)}(u, \cdot)$. Thus, (30) holds with $\tilde{V} = U_{k/2}$. A similar argument shows that when k is odd, (30) holds with $\tilde{V} = V_{(k-1)/2}$. □

B Proof of Proposition 4

Proposition 4. *Suppose that $h(\cdot)$ is strictly positive in a neighborhood of the origin. If $\omega(z)$ can be written as $\prod_{i=1}^n \omega_i(z_i)$, and there exists $\xi \in (1, 4/3)$ such that for all $i \in \{1, 2, \dots, n\}$,*

$$\int_{\mathbb{R}_+} \frac{u^{3d/2} h^3(u)}{(\int_0^{\xi u} v^{d/2} h(v)\, dv)^3 \omega_i^2(u)}\, du < \infty,$$

then (15) *holds, and thus by Theorem 2, second moment exists for the estimator* (13).

Proof. Let S_d be the set of $d \times d$ positive definite matrices. For any $\beta \in \mathbb{R}^p$, $\Sigma \in S_d$, $z \in \mathbb{R}^n$, and $\xi \in (1, 4/3)$,

$$\begin{aligned}
&\pi_{U|V}(\beta, \Sigma|z) \pi_{V|U}^3(z|\beta, \Sigma) \\
&= \frac{|\Sigma|^{-(n+d+1)/2} \prod_{i=1}^n \exp\{-z_i (y_i - \beta^T x_i)^T \Sigma^{-1} (y_i - \beta^T x_i)/2\}}{\int_{\mathbb{R}^p} \int_{S_d} |\tilde{\Sigma}|^{-(n+d+1)/2} \prod_{i=1}^n \exp\{-z_i (y_i - \tilde{\beta}^T x_i)^T \tilde{\Sigma}^{-1} (y_i - \tilde{\beta}^T x_i)/2\}\, d\tilde{\Sigma}\, d\tilde{\beta}} \times \\
&\quad \prod_{i=1}^n \frac{z_i^{3d/2} \exp\{-3 z_i (y_i - \beta^T x_i)^T \Sigma^{-1} (y_i - \beta^T x_i)/2\} h^3(z_i)}{\{\int_0^\infty v^{d/2} \exp[-v (y_i - \beta^T x_i)^T \Sigma^{-1} (y_i - \beta^T x_i)/2] h(v)\, dv\}^3} \\
&\le \frac{|\Sigma|^{-(n+d+1)/2} \prod_{i=1}^n \exp\{-z_i (y_i - \beta^T x_i)^T [\Sigma/(4 - 3\xi)]^{-1} (y_i - \beta^T x_i)/2\}}{\int_{\mathbb{R}^p} \int_{S_d} |\tilde{\Sigma}|^{-(n+d+1)/2} \prod_{i=1}^n \exp\{-z_i (y_i - \tilde{\beta}^T x_i)^T \tilde{\Sigma}^{-1} (y_i - \tilde{\beta}^T x_i)/2\}\, d\tilde{\Sigma}\, d\tilde{\beta}} \times \\
&\quad \prod_{i=1}^n \frac{z_i^{3d/2} h^3(z_i)}{(\int_0^{\xi z_i} v^{d/2} h(v)\, dv)^3}.
\end{aligned}$$

Note that

$$\begin{aligned}
&\int_{S_d} |\Sigma|^{-(n+d+1)/2} \prod_{i=1}^n \exp\left\{ -\frac{z_i}{2} \big(y_i - \beta^T x_i\big)^T \left(\frac{\Sigma}{4 - 3\xi} \right)^{-1} \big(y_i - \beta^T x_i\big) \right\} d\Sigma \\
&= (4 - 3\xi)^{-nd/2} \int_{S_d} |\Sigma|^{-(n+d+1)/2} \prod_{i=1}^n \exp\left\{ -\frac{z_i}{2} \big(y_i - \beta^T x_i\big)^T \Sigma^{-1} \big(y_i - \beta^T x_i\big) \right\} d\Sigma.
\end{aligned}$$

Thus,

$$\int_{\mathbb{R}^p} \int_{S_d} \pi_{U|V}(\beta, \Sigma|z) \pi_{V|U}^3(z|\beta, \Sigma)\, d\Sigma\, d\beta \le (4 - 3\xi)^{-nd/2} \prod_{i=1}^n \frac{z_i^{3d/2} h^3(z_i)}{(\int_0^{\xi z_i} v^{d/2} h(v)\, dv)^3}.$$

The result follows immediately. □

References

AHUES, M., LARGILLIER, A. and LIMAYE, B. (2001). *Spectral Computations for Bounded Operators.* CRC Press.

ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association* **88** 669–679.

BILLINGSLEY, P. (1995). *Probability and Measure*. 3rd ed. John Wiley & Sons.

BRISLAWN, C. (1988). Kernels of trace class operators. *Proceedings of the American Mathematical Society* **104** 1181–1190.

CHAKRABORTY, S. and KHARE, K. (2017). Convergence properties of Gibbs samplers for Bayesian probit regression with proper priors. *Electronic Journal of Statistics* **11** 177–210.

CHAN, K. S. and GEYER, C. J. (1994). Discussion: Markov chains for exploring posterior distributions. *The Annals of Statistics* **22** 1747–1758.

CHOI, H. M. and HOBERT, J. P. (2013). Analysis of MCMC algorithms for Bayesian linear regression with Laplace errors. *Journal of Multivariate Analysis* **117** 32–40.

CHOI, H. M. and ROMÁN, J. C. (2017). Analysis of Polya-Gamma Gibbs sampler for Bayesian logistic analysis of variance. *Electronic Journal of Statistics* **11** 326–337.

CONWAY, J. B. (1990). *A Course in Functional Analysis*. 2nd ed. Springer-Verlag.

CONWAY, J. B. (2000). *A Course in Operator Theory*. American Mathematical Soc.

DIACONIS, P., KHARE, K. and SALOFF-COSTE, L. (2008). Gibbs sampling, exponential families and orthogonal polynomials. *Statistical Science* **23** 151–178.

DIACONIS, P. and STROOCK, D. (1991). Geometric bounds for eigenvalues of Markov chains. *The Annals of Applied Probability* **1** 36–61.

ERDÖS, J. (1974). On the trace of a trace class operator. *Bulletin of the London Mathematical Society* **6** 47–50.

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

GARREN, S. T. and SMITH, R. L. (2000). Estimating the second largest eigenvalue of a Markov transition matrix. *Bernoulli* **6** 215–242.

GOHBERG, I., GOLDBERG, S. and KRUPNIK, N. (2012). *Traces and Determinants of Linear Operators*, vol. 116. Birkhäuser.

HELMBERG, G. (2014). *Introduction to Spectral Theory in Hilbert Space*. Elsevier.

HOBERT, J. P. (2011). The data augmentation algorithm: Theory and methodology. In *Handbook of Markov Chain Monte Carlo* (S. Brooks, A. Gelman, G. Jones and X.-L. Meng, eds.). Chapman & Hall/CRC Press.

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.

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.

KOLTCHINSKII, V. and GINÉ, E. (2000). Random matrix approximation of spectra of integral operators. *Bernoulli* **6** 113–167.

KONTOYIANNIS, I. and MEYN, S. P. (2012). Geometric ergodicity and the spectral gap of non-reversible Markov chains. *Probability Theory and Related Fields* **154** 327–339.

LAWLER, G. F. and SOKAL, A. D. (1988). Bounds on the l^2 spectrum for Markov chains and Markov processes: A generalization of Cheegers inequality. *Transactions of the American mathematical society* **309** 557–580.

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

LIU, J. S., WONG, W. H. and KONG, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81** 27–40.

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

MIRA, A. and GEYER, C. J. (1999). Ordering Monte Carlo Markov chains. *School of Statistics, University of Minnesota. technical report* .

PAL, S., KHARE, K. and HOBERT, J. P. (2017). Trace class Markov chains for Bayesian inference with generalized double Pareto shrinkage priors. *Scandinavian Journal of Statistics,* to appear .

QIN, Q. and HOBERT, J. P. (2016). Trace-class Monte Carlo Markov chains for Bayesian multivariate linear regression with non-Gaussian errors. *arXiv:1602.00136* .

ROBERTS, G. O. and ROSENTHAL, J. S. (1997). Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability* **2** 13–25.

ROY, V. and HOBERT, J. P. (2007). Convergence rates and asymptotic standard errors for markov chain monte carlo algorithms for bayesian probit regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69** 607–623.

SINCLAIR, A. and JERRUM, M. (1989). Approximate counting, uniform generation and rapidly mixing Markov chains. *Information and Computation* **82** 93–133.

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** 528–540.

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.