Bernoulli **25**(4B), 2019, 3832–3863
https://doi.org/10.3150/19-BEJ1112

Consistent estimation of the spectrum of trace class Data Augmentation algorithms

SAPTARSHI CHAKRABORTY[*] and KSHITIJ KHARE[**]

Department of Statistics, 102 Griffin Floyd Hall, University of Florida, Gainesville, FL 32611, USA.
E-mail: [*] c7rishi@ufl.edu; [**] kdkhare@stat.ufl.edu

Markov chain Monte Carlo is widely used in a variety of scientific applications to generate approximate samples from intractable distributions. A thorough understanding of the convergence and mixing properties of these Markov chains can be obtained by studying the spectrum of the associated Markov operator. While several methods to bound/estimate the second largest eigenvalue are available in the literature, very few general techniques for consistent estimation of the entire spectrum have been proposed. Existing methods for this purpose require the Markov transition density to be available in closed form, which is often not true in practice, especially in modern statistical applications. In this paper, we propose a novel method to consistently estimate the entire spectrum of a general class of Markov chains arising from a popular and widely used statistical approach known as Data Augmentation. The transition densities of these Markov chains can often only be expressed as intractable integrals. We illustrate the applicability of our method using real and simulated data.

Keywords: Data Augmentation algorithms; eigenvalues of Markov operators; MCMC convergence; trace class Markov operators

1. Introduction

Markov chain Monte Carlo (MCMC) techniques have become an indispensable tool in modern computations. With major applications in high dimensional settings, MCMC methods are routinely applied in various scientific disciplines. A major application of MCMC is to evaluate intractable integrals. To elaborate, let $(\mathcal{X}, \mathcal{B}, \nu)$ be an arbitrary measure space and let Π be a probability measure on $\mathcal{X}$, with associated probability density $\pi(\cdot)$ (with respect to the measure ν). The quantity of interest is the integral

$$\pi g := \int_{\mathcal{X}} g(x)\, d\Pi(x)$$

where g is a well-behaved function. In many modern applications, the above integral is highly intractable. In particular, it is not available in closed form, a (deterministic) numerical integration is extremely inefficient (often due to the high dimensionality of $\mathcal{X}$), and it can not be estimated by classical Monte Carlo techniques, as random (IID) generation from π is not feasible. In such cases, one typically resorts to Markov Chain Monte Carlo (MCMC) methods. Here, a Markov chain $\widetilde{X} = (X_n)_{n \geq 0}$ with equilibrium probability distribution Π is generated (using any standard MCMC strategies such as Metropolis Hastings, Gibbs sampler etc.) and then a Monte Carlo average based on those Markov chain realizations is used to estimate πg.

If the Markov chain $\widetilde{X} = (X_n)_{n\geq 0}$ is Harris ergodic (which is the case if the corresponding Markov transition density is strictly positive everywhere), then the cumulative averages based on the Markov chain realizations consistently estimate the integral of interest (see Asmussen and Glynn [3]). The accuracy of the estimate depends on two factors: (a) the convergence behavior of the Markov chain to its stationary distribution, and (b) the dependence between the successive realizations of the chain at stationarity. An operator theoretic framework provides a unified way of analyzing these two related factors. Let us consider the Hilbert space $L^2(\pi)$ of real valued functions f with finite second moment with respect to π. This is a Hilbert space where the inner product of $f, h \in L^2(\pi)$ is defined as

$$\langle f, h\rangle = \int_{\mathcal{X}} f(x)h(x)\pi(x)\,d\nu(x) = \int_{\mathcal{X}} f(x)h(x)\,d\Pi(x)$$

and the corresponding norm is defined by $\|f\|_{L^2(\pi)} = \sqrt{\langle f, f\rangle}$. Then the Markov transition density $k(\cdot,\cdot)$ corresponding to the Markov chain $\widetilde{X}$ defines an operator $K : L^2(\pi) \to L^2(\pi)$ that maps f to

$$(Kf)(x) = \int_{\mathcal{X}} k\big(x, x'\big) f\big(x'\big)\,d\nu\big(x'\big) = \int_{\mathcal{X}} \frac{k(x, x')}{\pi(x')} f\big(x'\big)\,d\Pi\big(x'\big). \tag{1}$$

We will assume that the Markov chain $\widetilde{X}$ is reversible. In terms of the associated operator K, this means that K is self-adjoint. The spectrum of the self-adjoint operator K, denoted by $\lambda(K)$, is the set of λ for which $K - \lambda I$ is non-invertible (here I denotes the identity operator that leaves a function unchanged). It is known that if K is positive, i.e., if $\langle Kf, f\rangle \geq 0$ for all $f \in L^2(\pi)$, (which is the case when K is the operator corresponding to a Data Augmentation (DA) Markov chain, see Section 3), then $\lambda(K) \subseteq [0, 1]$ (see, e.g., Retherford [39]).

In this paper, we will focus on situations when the (positive, self-adjoint) operator K is trace class, i.e., $\lambda(K)$ is countable and its elements are summable (Conway [11], p. 214). All finite state space Markov chains trivially correspond to trace class operators. Also, in recent years, an increasingly large class of continuous state space Markov chains from statistical applications have been shown to correspond to trace class operators (see, e.g., Choi and Román [10], Chakraborty and Khare [7], Pal, Khare and Hobert [32], Qin and Hobert [34], Hobert et al. [21], Rajaratnam et al. [38]). Let $\lambda(K) = \{\lambda_i\}_{i=0}^{\infty}$, where $(\lambda_i)_{i=0}^{\infty}$ are the decreasingly ordered eigenvalues of K. Then $\lambda_0 = 1$ and the difference $\gamma = \lambda_0 - \lambda_1 = 1 - \lambda_1$ is called the spectral gap for the compact Markov operator K. The spectral gap plays a major role in determining the convergence behavior of the Markov chain. In particular, any $g \in L^2(\pi)$ can be expressed as $g = \sum_{i=0}^{\infty} \eta_i \phi_i$ where $(\phi_i)_{i=0}^{\infty}$ is the sequence of eigenfunctions corresponding to K, and

$$\big\|K^m g - \pi g\big\|_{L^2(\pi)} = \left(\sum_{i=1}^{\infty} \eta_i^2 \lambda_i^{2m}\right)^{1/2} \leq \|g\|\lambda_1^m = \|g\|(1-\gamma)^m \tag{2}$$

for any positive interger m. Hence, γ determines the asymptotic rate of convergence of $\widetilde{X}$ to the stationary distribution. Furthermore, $(1-\gamma)^m$ provides maximal absolute correlation between X_j and X_{j+m} when j is large (i.e., X_j is sufficiently close to the target), and enables us to compute upper bounds of the asymptotic variance of MCMC estimators based on ergodic averages.

There is a substantial literature devoted to finding a theoretical bound for the second largest eigenvalue $\lambda_1 = 1 - \gamma$ of a Markov operator. For finite state space Markov chains, see Lawler and Sokal [29], Sinclair and Jerrum [43], Diaconis and Stroock [16], Saloff-Coste [42], Yuen [46], Diaconis and Saloff-Coste [14,15], François [18] to name just a few. In many statistical applications, the Markov chains move on large continuous state spaces, and techniques based on drift and minorization (see Rosenthal [40], Jones and Hobert [24]) have been used to get bounds on λ_1 for some of these Markov chains. However, these bounds can in many cases be way off. Techniques to estimate the spectral gap based on simulation have been developed in Garren and Smith [20], Raftery and Lewis [37], and more recently in Qin, Hobert and Khare [35] for trace class data augmentation Markov chains.

While bounding or estimating the spectral gap is clearly useful, a much more detailed and accurate picture of the convergence can be obtained by analyzing the entire spectrum of the Markov operator, as explained below.

(i) If we have two competing Markov chains to sample from the same stationary density, having knowledge of their respective spectra allows for a detailed and careful comparison (see Section 4.3 for an illustration).

(ii) For positive integer m, let $k^m(\cdot,\cdot)$ denote the m-step transition density of the associated Markov chain $\widetilde{X}$. The chi-square distance to stationarity after m steps, starting at state x is defined as:

$$\chi_x^2(m) := \int_{\mathcal{X}} \frac{|k^m(x,x') - \pi(x')|^2}{\pi(x')}\, d\nu\big(x'\big).$$

Since K is assumed to be trace class (and hence Hilbert Schmidt), it follows that Diaconis, Khare and Saloff-Coste [13] $\chi_x^2(m) = \sum_{i=1}^{\infty} \lambda_i^{2m} \phi_i(x)^2$. The average or expected chi-square distance to stationarity after m steps is therefore $\pi\chi^2(m) := \int_x \chi_m^2\, d\Pi = \sum_{i=1}^{\infty} \lambda_i^{2m}$ (since $\pi\phi_i^2 = 1$ for all i). Thus, having knowledge of the entire spectrum enables one to compute these average or expected chi-square distances.

(iii) From (2), it is apparent that if η_i's are known, then the knowledge of the entire spectrum enables us to compute the exact L^2 distance to stationarity. While finding the exact η_i's in general will be difficult, specific examples can be found in Diaconis, Khare and Saloff-Coste [13], Hobert, Roy and Robert [22], Khare and Zhou [27].

The literature for general methods to evaluate/estimate the entire spectrum (all the eigenvalues) of a Markov operator is, however, rather sparse. Adamczak and Bednorz [1] provide an elegant and simple way of consistently estimating the spectrum of a general Hilbert–Schmidt integral operator with symmetric kernel using approximations based on random matrices simulated from a Markov chain. The approach in Adamczak and Bednorz [1] can in particular be adapted for estimating the spectra of Markov operators. In fact, as we show in Section 2, in this context, the regularity condition needed for their method is exactly equivalent to the underlying Markov operator being trace class.

However, in order for the approach (and the technical consistency results) in Adamczak and Bednorz [1] to be applicable, the Markov transition density $k(\cdot,\cdot)$ and the stationary density $\pi(\cdot)$ are required to be available in closed form. These assumptions are not satisfied by an overwhelming majority of Markov chains arising in modern statistical applications. This is particularly true

for the so-called Data Augmentation (DA) algorithm, which is a widely used technique for constructing Markov chains by introducing unobserved/latent random variables. In this context, often, (a) the transition density can only be expressed as an intractable high-dimensional integral, and/or (b) the stationary density is only available up to an unknown normalizing constant,[1] see Albert and Chib [2], Hobert, Roy and Robert [22], Roy [41], Polson, Scott and Windle [33], Choi and Hobert [9], Hobert et al. [21], Qin and Hobert [34], Pal, Khare and Hobert [32] to name just a few.

The main objective of this paper is to develop a random matrix approximation method to consistently estimate the spectrum of DA Markov operators for situations where (a) and/or (b) holds. In particular, we show that if the transition densities in the method of Adamczak and Bednorz [1] are replaced by appropriate Monte Carlo based approximations, the spectrum of the resulting random matrix consistently estimates the spectrum of the underlying Markov operator (Theorem 3.1). More generally, we show that the method and the result can be easily adapted to situations where the stationary density is known only up to a normalizing constant (Theorem 3.2).

No regularity conditions are needed for our results if the state space $\mathcal{X}$, or the latent variable space $\mathcal{Z}$ is finite. We would like to mention that in many statistical applications with finite state spaces, the state space can be extremely large, with millions/billions of states. The intractability of the transition density and the size of the state space often make numerical techniques for eigenvalue estimation completely infeasible. However, as we show in the context of the example in Section 4.3, our method can provide reasonable answers in less than 5 minutes using modern parallel processing machinery. If both the state space $\mathcal{X}$ and the latent variable space $\mathcal{Z}$ are infinite, two regularity conditions need to be verified in order to use our results. One of them requires the Markov operator to be trace class, and the other one is a variance condition; each require checking that an appropriate integral is finite. An illustration is provided in Section 4.2 for the Gibbs sampler of Polson, Scott and Windle [33].

The remainder of the article is organized as follows. In Section 2, we first review the approach developed by Adamczak and Bednorz [1], which is applicable when the Markov transition densities have closed form expressions. Then we show that in the context of Markov operators, their regularity condition for consistency is equivalent to assuming that the operator is trace class. In Section 3, we introduce our approach for estimating the spectrum of DA Markov operators with intractable Markov transition densities and establish weak and strong consistency of the resulting estimates under a mild regularity assumption. In Section 4.1, we consider a toy normal-normal DA Markov chain Diaconis, Khare and Saloff-Coste [13], where all the eigenvalues are known, and examine the accuracy of the eigenvalue estimates provided by our algorithm. We then compare the convergence rates of the estimated spectrum to those of an estimated functional of interest (mean second Hermite polynomial), and also make a comparative analysis of the performances of our method to the method of Qin, Hobert and Khare [35] in estimating the second largest eigenvalue. We then move on to real applications. In Section 4.2, we illustrate our method on the Polya Gamma Markov chain of Polson, Scott and Windle [33]. We verify that this Markov chain satisfies the regularity condition needed for consistency and work out the first few eigenvalue estimates for the `nodal` dataset provided in the `boot` (Canty and Ripley [6]) R

[1] One would typically need to evaluate a complicated high-dimensional integral to obtain this constant, which is often infeasible.

package. In Section 4.3, we consider a Bayesian analysis of the two component normal mixture model and examine two competing DA Markov chains proposed in Hobert, Roy and Robert [22] to sample from the resulting posterior distribution. We illustrate the usefulness and applicability of our method by estimating and comparing the first few eigenvalues of the two DA chains for simulated data. We end with a discussion in Section 5. Proofs of all theorems and lemmas introduced in this paper are provided in a supplemental document [8].

2. Random matrix approximation method of Adamczak and Bednorz [1]

The objective of this section is to describe the method of operator spectra estimation via random matrices, first proposed in Koltchinskii and Giné [28] and then in Adamczak and Bednorz [1] in the context of Markov operators. We begin this section with a brief description of the general method, and then discuss how one can potentially use it to estimate spectra of trace class Markov operators. This discussion is followed by a short lemma that establishes an equivalence between the regularity condition used in Adamczak and Bednorz [1], and the condition of the Markov chain being trace class.

Let $H : L^2(\pi) \to L^2(\pi)$ be a Hilbert–Schmidt integral operator (an integral operator whose eigenvalues are square summable) defined through a symmetric (in argmuents) kernel $h(\cdot, \cdot)$ as:

$$(Hg)(x) = \int_{\mathcal{X}} h\big(x, x'\big) g\big(x'\big)\, d\Pi\big(x'\big), \tag{3}$$

and interest lies in obtaining $\lambda(H)$, the spectrum of H. In general, there does not exist any method of evaluating $\lambda(H)$ for arbitrary H. However, Koltchinskii and Giné [28] suggest a novel, elegant and simple approach of *estimating* $\lambda(H)$ via random matrices. Let $X_0, \ldots, X_{m-1}$ denote an IID sample of size m (≥ 1) from the distribution Π. Then the authors show that a (strong) consistent estimator of $\lambda(H)$ is given by the set of eigenvalues of the random matrix

$$H_m = \frac{1}{m}\big((1 - \delta_{jj'})h(X_j, X_{j'})\big)_{0 \leq j, j' \leq m-1}$$

for large m, where $\delta_{jj'}$ denotes the Dirac delta function. The strength of the result lies in the fact that it works for *any* Hilbert Schmidt operator, irrespective of the dimension and structure of $\mathcal{X}$, as long as an IID sample from Π can be drawn. Unfortunately, IID simulations are not always feasible, especially in high dimensional settings (otherwise there would be no need for MCMC!), thus limiting the applicability of the method. Adamczak and Bednorz [1] generalize Koltchinskii and Giné's [28] result by allowing $X_0, \ldots, X_{m-1}$ to be an MCMC sample (i.e., realizations of a Markov chain with equilibrium distribution Π), and prove consistency of the resulting estimates.

Let K denote a positive self-adjoint trace class Markov operator as defined in (1). Of course K is Hilbert Schmidt (eigenvalues are summable *implies* they are square summable), and $h(x, x') = k(x, x')/\pi(x')$ is symmetric in its argument due to reversibility of the associated Markov chain. Thus by expressing K in the form (3) $\lambda(K)$ can potentially be estimated by Adamczak and Bednorz's [1] method, which only requires an MCMC sample from Π. The resulting method,

Algorithm 1 Random Matrix Approximation (RMA) method of estimating $\lambda(K)$ for a Markov operator K with Markov transition density $k(\cdot,\cdot)$ and stationary density $\pi(\cdot)$ available in closed form

Step 0: Given a starting point X_0, draw realizations $X_1, X_2, \ldots, X_{m-1}$ from the Markov chain $\widetilde{X}$ associated with K.

Step 1: Given $X_0, \ldots, X_{m-1}$, for each pair (j, j') with $0 \le j,\ j' \le m-1$, compute the Markov transition densitys $k(X_j, X_{j'})$ and the kernels $h(X_j, X_{j'}) = k(X_j, X_{j'})/\pi(X_{j'})$, and construct the matrix

$$H_m = \frac{1}{m}\big((1-\delta_{jj'})h(X_j, X_{j'})\big)_{0\le j,j'\le m-1} \tag{4}$$

where $\delta_{jj'} = \mathbb{1}(j = j')$ is the Dirac delta function.

Step 2: Calculate the eigenvalues $\widehat{\lambda}_0 \ge \widehat{\lambda}_1 \ge \cdots \ge \widehat{\lambda}_{m-1}$ of H_m and estimate $\lambda(K)$ by $\widehat{\lambda}(K) = \lambda(H_m) := \{\widehat{\lambda}_0, \widehat{\lambda}_1, \ldots, \widehat{\lambda}_{m-1}\}$.

which uses the same random data generated during the original run of the Markov chain in the recipe proposed in Adamczak and Bednorz [1] to estimate the spectrum, will be called the *Random Matrix Approximation (RMA)* method henceforth, and is described in Algorithm 1.

Sacrificing independence and identicalness of the random sample in Adamczak and Bednorz's [1] RMA method, however, comes at a price (as compared to Koltchinskii and Giné's [28] method, which uses IID samples). In particular, to ensure strong consistency in Adamczak and Bednorz's [1] method, an additional regularity condition is required to be satisfied by the Markov operator K, namely, a $L^2(\pi)$ function $F : \mathcal{X} \to \mathbb{R}$ needs to exist for which $|h(x,x')| \le F(x)F(x')$ for all $x, x' \in X$. Interestingly, as we show in the following lemma (Lemma 2.1), this condition for K is equivalent to that of K being trace-class in the current setting. The proof of Lemma 2.1 is provided in Section A of the supplemental document. At the core of the proof, the following two alternative characterizations of trace class and Hilbert Schmidt operators (see, e.g., Jörgens [25]) are used. The operator K as defined in (1) is trace class if and only if

$$\int_{\mathcal{X}} k(x,x)\,d\nu(x) = \int_{\mathcal{X}} \frac{k(x,x)}{\pi(x)}\,d\Pi(x) < \infty \tag{5}$$

whereas it is Hilbert Schmidt if and only if

$$\int_{\mathcal{X}}\int_{\mathcal{X}} k\big(x,x'\big)^2 \frac{\pi(x)}{\pi(x')}\,d\nu(x)\,d\nu\big(x'\big) = \int_{\mathcal{X}}\int_{\mathcal{X}}\left[\frac{k(x,x')}{\pi(x')}\right]^2 d\Pi(x)\,d\Pi\big(x'\big) < \infty. \tag{6}$$

Lemma 2.1. *Consider a reversible Markov operator K as defined in (1). Define $h(x,x') = k(x,x')/\pi(x')$ for $x, x' \in \mathcal{X}$. Then the following two conditions are equivalent*:

(i) *there exists $F : \mathcal{X} \to \mathbb{R}$ such that $\pi F^2 < \infty$ and $|h(x,x')| \le F(x)F(x')$ for all $x, x' \in X$.*
(ii) *K is trace class.*

As a consequence of Lemma 2.1, we are now in a position to adapt the consistency result from Adamczak and Bednorz [1] for the RMA method described in Algorithm 1. Before stating the result, we introduce required notations from Koltchinskii and Giné [28] and Adamczak and Bednorz [1]. Recall that for any operator A (finite or infinite) we use the notation $\lambda(A)$ to denote its spectrum. Thus, for a finite matrix A, $\lambda(A)$ will denote the set of its eigenvalues. Since the Markov operators we consider are trace class (and therefore, Hilbert Schmidt), their spectra can be identified with the sequences $(\lambda_m)_{m=0}^{\infty} \in \ell_2$ of eigenvalues, where ℓ_2 is the Hilbert space of all square summable real sequences. Because our goal is to approximate the (possibly infinite) spectrum of an integral operator by the finite spectrum of a matrix, we will identify the latter with an element of ℓ_2, by appending an infinite sequence of zeros to it. As in Koltchinskii and Giné [28], the metric we use for comparing spectra is the δ_2 metric, which is defined for $x, y \in \ell_2$ as

$$\delta_2(x, y) = \inf_{\zeta \in \mathcal{P}} \left[\sum_{m=0}^{\infty} (x_m - y_{\zeta(m)})^2 \right]^{1/2} \tag{7}$$

where $\mathcal{P}$ is the set of all permutations of natural numbers. Note that for any two points on ℓ_2, the above metric can be expressed as an ℓ_2 distance of the sorted versions of the two points, as explained below. Following Koltchinskii and Giné [28], for any $x = (x_m)_{m=0}^{\infty} \in \ell_2$, we set $x = x_+ + x_-$, where $x_+ = (\max\{x_m, 0\})_{m=0}^{\infty}$ and $x_- = x - x_+$. We denote by $x_+^{\downarrow}$ ($x_-^{\uparrow}$) the point in ℓ_2 with the same coordinates of x_+ (x_-), but arranged in non-increasing (non-decreasing) order, and set $x^{\uparrow\downarrow} = x_-^{\uparrow} \oplus x_+^{\downarrow}$, where $u \oplus v = (u_0, \ldots, u_m, \ldots, v_0, \ldots, v_m, \ldots) \in \ell_2$ for $u = (u_m)_{m=0}^{\infty}$ and $v = (v_m)_{m=0}^{\infty} \in \ell_2$. Then

$$\delta_2(x, y) = \left\| x^{\uparrow\downarrow} - y^{\uparrow\downarrow} \right\|_{\ell_2}. \tag{8}$$

From the Hoffman–Wielandt inequality (Hoffman and Wielandt [23], Koltchinskii and Giné [28], Theorem 2.2), it follows that for normal operators A and B,

$$\delta_2\big(\lambda(A), \lambda(B)\big) \leq \|A - B\|_{\mathrm{HS}}, \tag{9}$$

where $\|A\|_{\mathrm{HS}}$ denotes the Hilbert Schmidt norm of an operator $A \in L^2(\pi)$ defined by

$$\|A\|_{\mathrm{HS}} = \left(\sum_{m=0}^{\infty} \|A\varphi_m\|^2 \right)^{1/2},$$

for any orthonormal basis $(\varphi_m)_{m=0}^{\infty}$ of $L^2(\pi)$. Note that if A is finite (i.e., a matrix), say $A = (a_{ij})$, then $\|A\|_{\mathrm{HS}} = \|A\|_{\mathrm{F}}$ where $\|A\|_{\mathrm{F}}$ denotes the Frobenious norm of A defined as

$$\|A\|_{\mathrm{F}} = \left(\sum_i \sum_j a_{ij}^2 \right)^{1/2}.$$

The following theorem, a rephrasing of Theorem 2.1 from Adamczak and Bednorz [1] adapted to the current setting using Lemma 2.1, establishes (strong) consistency of the spectrum estimator obtained by RMA method for a positive self-adjoint trace class Markov operator.

Theorem 2.1. *Let $\widetilde{X} = (X_n)_{n\geq 0}$ be a reversible Markov chain with Markov transition density $k(\cdot,\cdot)$, invariant measure Π, and suppose the associated Markov operator K as given in (1) is positive and trace class. Let $\Phi_m = \{X_0, \ldots, X_{m-1}\}$ denote the first m realizations of the Markov chain, and given Φ_m, construct the matrix H_m as given in (4). Then, for every initial measure ν_0 for the chain $\widetilde{X}$, with probability one, as $m \to \infty$,*

$$\delta_2\big(\lambda(H_m), \lambda(K)\big) \to 0.$$

3. A novel Monte Carlo based random matrix approximation method for DA Markov chains

As we see in Section 2, the RMA method of Adamczak and Bednorz [1] requires evaluation of the ratio $k(X_j, X_{j'})/\pi(X_{j'})$ for every pair (j, j'). Unfortunately, as mentioned in the Introduction, one or both of $k(\cdot,\cdot)$ and $\pi(\cdot)$ are often intractable and do not have closed form expressions in many statistical applications. This is particularly true in the context of the Data Augmentation (DA) algorithm, where along with the variable of interest X, one introduces a latent variable Z such that generations from the conditional distributions of $X|Z$ (X given Z) and $Z|X$ are possible. Then, given a starting point X_0, at each iteration $m \geq 1$, one first simulates z from the distribution of $Z|X = X_{m-1}$ and then generates X_m, from the distribution of $X|Z = z$. The X_m's generated in this method are retained and used as the required MCMC sample. Hence, the Markov transition density can be written as

$$k\big(x, x'\big) = \int_{\mathcal{Z}} f_{X|Z}\big(x'|z\big) f_{Z|X}(z|x)\, d\zeta(z). \tag{10}$$

Here $f_{Z|X}$ and $f_{X|Z}$ are conditional densities with respect to the measures ξ and ν respectively, and are simple and easy to sample from in a typical DA algorithm. A DA Markov chain is necessarily reversible, which means the associated Markov operator is self-adjoint. The operator is also positive with a positive spectrum, as shown in Liu, Wong and Kong [30].

However, the integral in (10) providing the Markov transition density of a DA Markov chain often does not have a closed form expression, and cannot be efficiently approximated via deterministic numerical integration (usually due to high dimensionality). Intractability of the integral precludes applicability of Adamczak and Bednorz's [1] RMA method of estimating spectrum (Algorithm 1) in such cases. In this section, we propose a Monte Carlo based random matrix approximation (MCRMA) algorithm to estimate the spectrum of DA algorithms with intractable transition densities (Algorithm 2). To contrast with MCRMA, we shall call the RMA method of Adamczak and Bednorz [1] (Algorithm 1) the exact RMA. Consistency of MCRMA spectrum estimates is established in Theorem 3.1.

Often, in addition to the intractability of the transition density, the stationary density is also available only up to an unknown normalizing constant (which is again hard to estimate in many modern applications as the stationary density is supported on a high-dimensional space). We adapt our algorithm to this situation (Algorithm 3), and establish consistency of the resulting estimates as well (Theorem 3.2).

3.1. Monte Carlo Random Matrix Approximation (MCRMA) method

In this section, we will present a method to estimate the spectrum of a DA Markov operator where the transition density in (10) is intractable, but the stationary density π is available in closed form. Given m realizations $\Phi_m = \{X_0, X_1, \ldots, X_{m-1}\}$ of the positive trace class reversible Markov chain $\widetilde{X}$ with transition density in the form (10), the key idea is to approximate $k(X_j, X_{j'})$ for each pair (j, j') using classical Monte Carlo technique, and then construct an analogue of the RMA estimator that uses the approximate kernels instead of the original. The details of the method are provided in Algorithm 2.

It is to be noted that for Step 1 in the MCRMA algorithm to be feasible, the density $f_{Z|X}$ should be easy to sample from. This is typically true for DA algorithms that are used in practice. In fact, the major motivation for using a DA algorithm is that the conditional densities $f_{X|Z}$ and $f_{Z|X}$ are easy to sample from, whereas it is hard to directly generate samples from π. For Step 2

Algorithm 2 Monte Carlo Random Matrix Approximation (MCRMA) method of estimating $\lambda(K)$ for a positive reversible Markov operator K with associated Markov transition density $k(x, x') = \int_{\mathcal{Z}} f_{X|Z}(x'|z) f_{Z|X}(z|x)\, d\zeta(z)$

Step 0: Given a starting point X_0, draw realizations $X_1, X_2, \ldots, X_{m-1}$ from the associated Markov chain $\widetilde{X}$. Call $\Phi_m = \{X_0, \ldots, X_{m-1}\}$.

Step 1: Given Φ_m, for each $j = 0, 1, \ldots, m-2$, generate generate $N = N(m)$ IID observations $Z_1^{(j)}, \ldots, Z_N^{(j)}$ from the density $f_{Z|X}(\cdot|X_j)$.

Step 2: For each pair (j, j') with $0 \le j < j' \le m-1$, construct the Monte Carlo estimate

$$\widehat{k}_N(X_j, X_{j'}) = \frac{1}{N} \sum_{l=1}^{N} f_{X|Z}\big(X_{j'}|Z_l^{(j)}\big),$$

define the estimated kernel

$$\widehat{h}_N(X_j, X_{j'}) = \begin{cases} \dfrac{\widehat{k}_N(X_j, X_{j'})}{\pi(X_{j'})} & \text{if } j < j' \\ 0 & \text{if } j = j' \\ \widehat{h}_N(X_{j'}, X_j) & \text{if } j > j' \end{cases},$$

and construct the matrix

$$\widehat{H}_m^{(N)} = \frac{1}{m}\big((1 - \delta_{jj'})\widehat{h}_N(X_j, X_{j'})\big)_{0 \le j, j' \le m-1}, \tag{11}$$

where $\delta_{jj'} = \mathbb{1}(j = j')$ is the Dirac delta function. Observe that $\widehat{H}_m^{(N)}$ is symmetric by construction, with zero diagonal entries.

Step 3: Calculate the eigenvalues $\widehat{\lambda}_0 \ge \widehat{\lambda}_1 \ge \cdots \ge \widehat{\lambda}_{m-1}$ of $\widehat{H}_m^{(N)}$ and estimate $\lambda(K)$ by $\widehat{\lambda}(K) = \lambda(\widehat{H}_m^{(N)}) := \{\widehat{\lambda}_0, \widehat{\lambda}_1, \ldots, \widehat{\lambda}_{m-1}\}$.

---

to be feasible, we need $f_{X|Z}$ to be available in closed form. Again, this is true in most statistical applications, where $f_{X|Z}$ is typically a standard density such as multivariate normal, gamma etc. Another crucial thing to note, from a computational point of view, is that the rows of the matrix $\widehat{H}_m^{(N)}$ can be constructed in an embarrassingly parallel fashion (since no relationship is assumed among the elements of $\widehat{H}_m^{(N)}$), thereby reducing the running time of the algorithm significantly.

Note that the MCRMA algorithm (Algorithm 2) provides a coarser approximation to the spectrum of K as compared to the the exact RMA algorithm (Algorithm 1). This is because we use the additional Monte Carlo based approximation $\widehat{k}_N$ (for k) in the constructed random matrices. An obvious and important question is: does an analog of the consistency result for the RMA algorithm (Theorem 2.1) hold in this more complex setting of the MCRMA algorithm? We state a consistency result below, and the proof is provided in Section B of the supplemental document.

Theorem 3.1. *Let $\widetilde{X} = (X_n)_{m\geq 0}$ be a positive, reversible Markov chain with transition density $k(\cdot,\cdot)$ in the form (10), invariant measure Π and associated Markov operator K as given in (1). Let $\Phi_m = \{X_0, \ldots, X_{m-1}\}$ denote the first m realizations of the Markov chain, and given Φ_m, construct the matrix $\widehat{H}_m^{(N)}$ as given in (11). Then the following hold*:

(I) *If $\mathcal{X}$ is finite, then (strong consistency) for every initial measure ν_0 of the chain $\widetilde{X}$, as $m \to \infty$ and $N \to \infty$,*

$$\delta_2\big(\lambda\big(\widehat{H}_m^{(N)}\big), \lambda(K)\big) \to 0 \text{ almost surely.}$$

(II) *If $\mathcal{X}$ is infinite (countable or uncountable), and*
(A) *K is trace class, and*
(B) *(variance condition)*

$$\sup_{m\geq 1} \max_{0\leq j<j'\leq m-1} \int_{\mathcal{X}}\int_{\mathcal{X}}\int_{\mathcal{Z}} \left(\frac{f_{X|Z}(x_{j'}|z)}{\pi(x_{j'})}\right)^2 f_{Z|X}(z|x_j) \times q_{jj'}(x_j, x_{j'})\, d\zeta(z)\, d\nu(x_j)\, d\nu(x_{j'}) < \infty,$$

where $q_{j_1 j_2 \cdots j_k}$ denotes the joint density of $X_{j_1}, \ldots, X_{j_k}$, $0 \leq j_1 < \cdots < j_k \leq m-1$, $1 \leq k \leq m$,

then
(i) *(weak consistency) if $\frac{1}{N(m)} \to 0$ as $m \to \infty$, then $\delta_2(\lambda(\widehat{H}_m^{(N)}), \lambda(K)) \xrightarrow{P} 0$,*
(ii) *(strong consistency) if $\sum_{m=0}^{\infty} \frac{1}{N(m)} < \infty$, then $\delta_2(\lambda(\widehat{H}_m^{(N)}), \lambda(K)) \to 0$ almost surely.*

Remark 3.1. Let K^* denote the Markov operator associated with the Z chain (the Markov chain of the generated latent data), defined for $f \in L^2(\pi^*)$ as

$$\big(K^* f\big)(z) = \int_{\mathcal{Z}} k^*\big(z, z'\big) f\big(z'\big)\, d\zeta\big(z'\big) = \int_{\mathcal{X}} \frac{k^*(z, z')}{\pi^*(z')} f\big(z'\big)\, d\Pi^*\big(z'\big) \tag{12}$$

where $k^*(\cdot,\cdot)$ denotes the Markov transition density of the Z chain, π^* denotes the stationary density for Z (associated with k^*) and Π^* is the probability measure associated with π^*. Then Khare and Hobert [26] show that $\lambda(K) = \lambda(K^*)$, which implies that instead of estimating $\lambda(K)$,

one can equivalently estimate $\lambda(K^*)$. Note that,

$$k^*(z, z') = \int_{\mathcal{X}} f_{Z|X}(z'|x) f_{X|Z}(x|z)\, d\nu(x)$$

with $f_{X|Z}$ and $f_{Z|X}$ being the same conditional densities as before. Therefore, an analogous MCRMA algorithm for estimating $\lambda(K^*)$ can be similarly formulated. Here, given the realizations $Z_0, Z_1, \ldots, Z_{m-1}$, one first finds the Monte Carlo approximates of $k^*(z, z')$ (via IID samples generated from $f_{X|Z}(\cdot|z)$) at every paired realization $(Z_j, Z_{j'})$, then defines a random matrix with the ratios $k^*(z, z')/\pi^*(z')$ (times an adjustment factor $1/m$), similar to the MCRMA with X observations, and finally evaluates eigenvalues of the resulting random matrix. Consequently, an analogous consistency theorem will also hold for the resulting algorithm.

Because the Z chain is automatically generated as a by-product of the DA algorithm, from a practitioner's point of view, using Z instead of X makes little difference in MCRMA. However, substantial simplifications on the regularity conditions may be achieved by using Z. This is particularly true in cases where the latent variable space $\mathcal{Z}$ is finite (however large). In such cases, no regularity condition is required to be satisfied (case (I) in Theorem 3.1) to achieve strong consistency. See Section 4.3 for examples.

Remark 3.2. The variance condition (B) is more restrictive than the trace class condition (A) because of the square term $\{f_{X|Z}(x_{j'}|z)/\pi(x_{j'})\}^2$. These types of second moment conditions are often necessary to guarantee good behavior of eigenvalue estimators; a somewhat similar second moment condition appears in Qin, Hobert and Khare [35], equation 14 and Theorem 2, to ensure finite variance of their second largest eigenvalue estimator. Proofs of Theorems 4.1 and 4.2 in Section C of the supplemental document provide illustrations on how the integrals in conditions (A) and (B) can be handled.

Remark 3.3. When $\mathcal{X}$ is finite, strong consistency is guaranteed as long as $m \to \infty$ and $N \to \infty$ (no relationship between the rate of growth of m and N is necessary). When $\mathcal{X}$ is infinite and the conditions (A) and (B) hold, the conditions (II)(i) and (ii) on in Theorem 3.1 are required to justify weak and strong consistency, respectively. These conditions on N and m, are however, not very demanding. For example, when $N(m) = O(m)$ or even, $N(m) = O(\log m)$, (II)(i) is satisfied, and weak convergence holds. On the other hand when $N(m) = O(m^{1+\delta})$, for some $\delta > 0$, condition (ii) is satisfied, ensuring strong convergence. In practice, as long as both N and m are sufficiently large, reasonable results can be expected.

3.2. MCRMA with π specified only up to a constant

Note that Step 2 of MCRMA method requires construction of a symmetric matrix whose (j, j')th entry has $\pi(X_{j'})$ in the denominator. This is clearly not feasible in cases where π is known up to a constant, that is, π is of the form $\pi(\cdot) = \eta(\cdot)/c$, where $c \in (0, \infty)$ is an unknown constant, and the functional form of $\eta(\cdot)$ is completely known. In this section, we propose a simple strategy that adapts Algorithm 2 for such cases. The basic idea, displayed formally in Algorithm 3, is to follow the steps of Algorithm 2 but now with $\eta(\cdot)$ in the denominator of the random matrix instead of $\pi(\cdot)$, and then simply rescale the eigenvalues so that the largest eigenvalue is 1.

Algorithm 3 MCRMA estimation of $\lambda(K)$ for a trace class Markov operator K when $\pi(\cdot) \propto \eta(\cdot)$, and the functional form for $\eta(\cdot)$ is known

Step 0: Given a starting point X_0, draw realizations $X_1, X_2, \ldots, X_{m-1}$ from the associated Markov chain $\widetilde{X}$. Call $\Phi_m = \{X_0, \ldots, X_{m-1}\}$.

Step 1: Given Φ_m, for each $j = 0, 1, \ldots, m-1$, generate generate $N = N(m)$ IID observations $Z_1^{(j)}, \ldots, Z_N^{(j)}$ from the density $f_2(\cdot|X_j)$.

Step 2: For each pair (j, j') with $0 \le j < j' \le m$, construct the Monte Carlo estimate

$$\widehat{k}_N(X_j, X_{j'}) = \frac{1}{N}\sum_{l=1}^{N} f_1\big(X_{j'}|Z_l^{(j)}\big),$$

define the estimated kernel

$$\widehat{s}_N(X_j, X_{j'}) = \begin{cases} \dfrac{\widehat{k}_N(X_j, X_{j'})}{\eta(X_{j'})} & \text{if } j < j' \\ 0 & \text{if } j = j' \\ \widehat{s}_N(X_{j'}, X_j) & \text{if } j > j' \end{cases},$$

and construct the matrix

$$\widehat{S}_m^{(N)} = \frac{1}{m+1}\big((1-\delta_{jj'})\widehat{s}_N(X_j, X_{j'})\big)_{0\le j,j'\le m-1} \tag{13}$$

Step 3: Calculate the eigenvalues $\widehat{\kappa}_0 \ge \widehat{\kappa}_1 \ge \cdots \ge \widehat{\kappa}_{m-1}$ of $\widehat{S}_m^{(N)}$, and estimate $\lambda(K)$ by $\lambda(\widehat{S}_m^{(N)})/\lambda_{\max}(\widehat{S}_m^{(N)}) := \{1, \widehat{\kappa}_1/\widehat{\kappa}_0, \ldots, \widehat{\kappa}_{m-1}/\widehat{\kappa}_0\}$, where $\lambda_{\max}(\widehat{S}_m^{(N)}) = \widehat{\kappa}_0$ is the largest eigenvalue of $\widehat{S}_m^{(N)}$.

Clearly, this nullifies any estimation/evaluation of the normalizing constant. Theorem 3.2 establishes consistency for the resulting estimator by exploiting the fact that the largest eigenvalue of any Markov operator is 1. A detailed proof is given in the supplemental document (Section B).

Theorem 3.2. *Let $\widetilde{X} = (X_n)_{m\ge 0}$ be a positive, reversible Markov chain with transition density $k(\cdot,\cdot)$ in the form (10), invariant measure Π and associated Markov operator K as given in (1). Further, suppose that $\pi(\cdot) = \eta(\cdot)/c$, where $c \in (0,\infty)$ is a possibly unknown normalizing constant, and the functional form for $\eta(\cdot)$ is known. Let $\Phi_m = \{X_0, \ldots, X_{m-1}\}$ denote the first m realizations of the chain. Given Φ_m, construct the matrix $\widehat{S}_m^{(N)}$ as given in (13). Then*

(I) *if $\mathcal{X}$ is finite, then (strong consistency) for every initial measure ν_0 for the chain $\widetilde{X}$, as $m \to \infty$ and $N \to \infty$,*

$$\delta_2\left(\frac{\lambda(\widehat{S}_m^{(N)})}{\lambda_{\max}(\widehat{S}_m^{(N)})}, \lambda(K)\right) \to 0 \quad \textit{almost surely.}$$

(II) *if* $\mathcal{X}$ *is infinite* (*countable or uncountable*), *and condition* (*A*) *and* (*B*) *in Theorem* 3.1 *hold*, *then*

(i) (*weak consistency*) *if* $\frac{1}{N(m)} \to 0$ *as* $m \to \infty$, *then* $\delta_2(\frac{\lambda(\widehat{S}_m^{(N)})}{\lambda_{\max}(\widehat{S}_m^{(N)})}, \lambda(K)) \xrightarrow{P} 0$,

(ii) (*strong consistency*) *if* $\sum_{m=0}^{\infty} \frac{1}{N(m)} < \infty$, *then* $\delta_2(\frac{\lambda(\widehat{S}_m^{(N)})}{\lambda_{\max}(\widehat{S}_m^{(N)})}, \lambda(K)) \to 0$ *almost surely.*

Remark 3.4. The quantity $\widehat{\kappa}_0 = \max \lambda(\widehat{S}_m^{(N)})$, obtained as a by-product of Algorithm 3, is in fact a consistent estimator of the normalizing constant $1/c$ (see the proof of Theorem 3.2 in Section B of the supplemental document). This estimator is implicitly used during spectrum estimation in Algorithm 3, and nullifies the need of any separate estimation of the normalizing constant. It is to be noted that estimation of the constant $1/c$ is an interesting problem on its own, and appears in many statistical and machine learning problems; one notable example being marginal likelihood estimation in Bayesian model selection. Clearly, $\widehat{\kappa}_0$ can be used for estimating $1/c$ outside the context of eigenvalue estimation, where the only goal is to estimate the normalizing constant; consistency of the estimator is however guaranteed only when the assumptions in Theorem 3.2 are met. Comparative assessment of the estimator $\widehat{\kappa}_0$ with other estimators of the normalizing constant, such as the bridge sampling estimator Bennett [5], Meng and Wong [31], is a topic of further research.

4. Illustrations

The purpose of this section is to illustrate the applicability and usefulness of the MCRMA algorithm in practical settings. We shall consider two separate examples, one with a finite, and one with an infinite state space. However, before proceeding to these applications, we will start with a toy normal-normal DA Markov chain to understand/illustrate the performance of the MCRMA algorithm in a setting where the entire spectrum is already known. All computations in this section are done in R (R Core Team [36]) with some parts written in C++ to speed up computation. We used the R packages `Rcpp` Eddelbuettel [17] to call C++ functions inside R, and `ggplot2` Wickham [45] and `reshape2` Wickham [44] to create the plots.

4.1. Toy example: The normal-normal DA chain

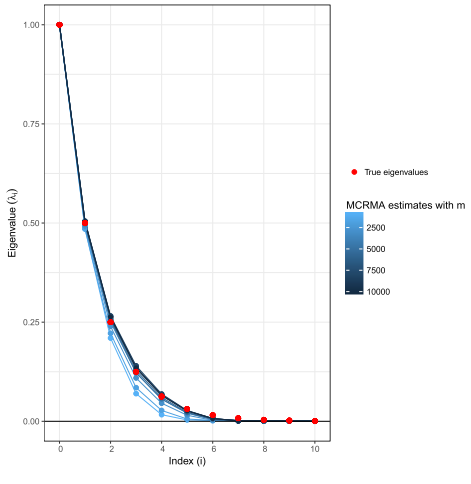
In this section, we consider a toy normal-normal DA Markov chain considered in Diaconis, Khare and Saloff-Coste [13] and then in Qin, Hobert and Khare [35] with known eigenvalues to illustrate the performance of the MCRMA method. Here, given a starting point x_0, one iterates between the following two steps:

(i) generate z from $\mathrm{N}(x/2, 1/8)$,
(ii) generate x from $\mathrm{N}(z, 1/4)$.

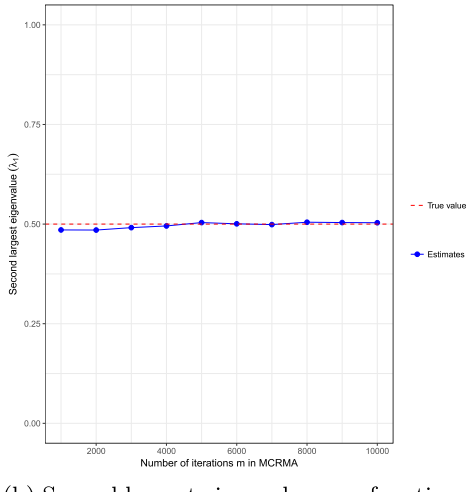
Of course, the stationary density of x is just $\mathrm{N}(0, 1/2)$, and there is no practical need for this MCMC algorithm. However, the spectrum of the corresponding Markov operator K has been

studied thoroughly in Diaconis, Khare and Saloff-Coste [13] and therefore it can be used as a nice toy example to exhibit the performance of MCRMA. It is easy to see that both the trace class condition (A) and the variance condition (B) hold for the operator K (since all the full conditional densities are just normal densities). From Diaconis, Khare and Saloff-Coste [13], it follows that the eigenvalues of K are given by $(\lambda_n)_{n=0}^{\infty}$ with $\lambda_n = 1/2^n$.

Starting from $x = 0$, we first generate 10,000 realizations of the above Markov chain, after discarding a burn-in of size 10,000, and then extract the x chain. Then we run 10 instances of the MCRMA algorithm 3 with $m = 1000, 2000, \ldots, 10{,}000$ (by using the first $1000, 2000, \ldots, 10{,}000$ iterations of the already generated Markov chain), and $N = N(m) = \lceil m^{1+10^{-6}} \rceil$, where for a real number x, $\lceil x \rceil$ denotes the "ceiling" of x, i.e., the smallest integer bigger than x. Then we look at the largest 11 eigenvalues (including $\lambda_0 = 1$) obtained from each instance of MCRMA. Note that the choice of m and N used here ensures strong consistency of the MCRMA estimates; weak consistency only requires $N(m) \to \infty$ as $m \to \infty$ (see Remark 3.3). To understand the accuracy of MCRMA, the estimated eigenvalues are compared to the true eigenvalues, by displaying the estimates and truths on the same plot. The resulting plots are shown in Figure 1. Figure 1(a) displays all 11 eigenvalues obtained from each of the 10 MCRMA instances (shown as 10 curves, one for each MCRMA instance), along with the true eigenvalues (shown as red dots). The second, third, fourth, fifth and sixth largest estimated eigenvalues, viewed as functions of the MCRMA iteration size m, are shown separately in Figures 1(b) through 1(f), with the dotted line indicating the corresponding the true eigenvalue. The noticeable similarity between the truth and the estimates, (especially for the instances with $m \geq 5000$,

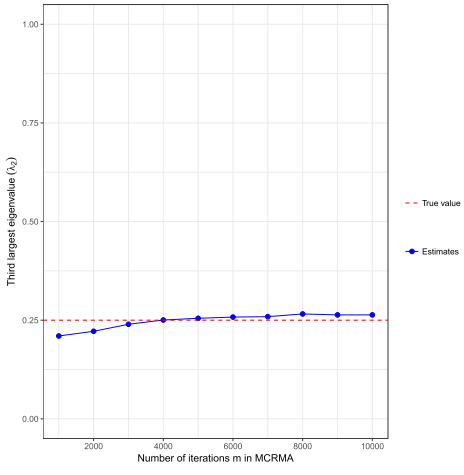


(a) The largest 11 eigenvalues. There are 10 curves, each corresponding to the choices $m = 1000, \cdots, 10,000$ in the MCRMA algorithm. The true eigenvalues are shown as red dots.

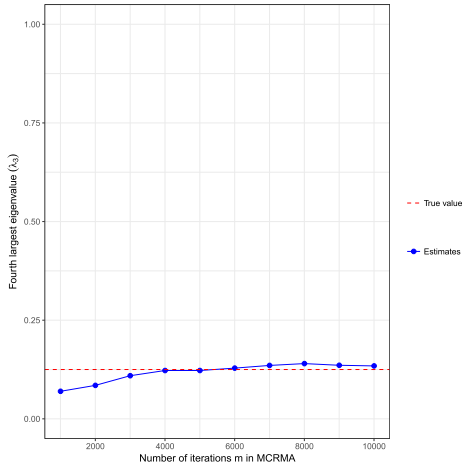


(b) Second largest eigenvalue as a function of iterations m in the MCRMA algorithm. The true eigenvalue is shown as a horizontal line.

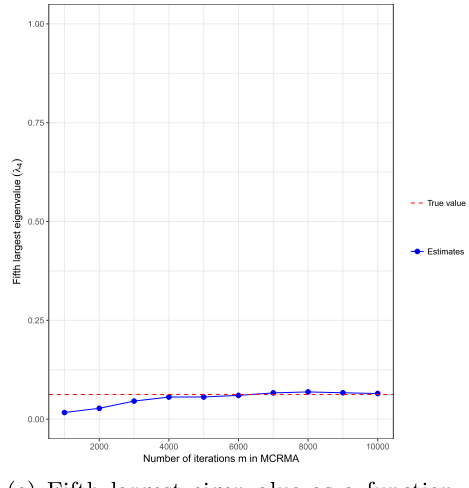
Figure 1. Eigenvalue estimates for the toy normal-normal DA Markov chain using the MCRMA algorithm.



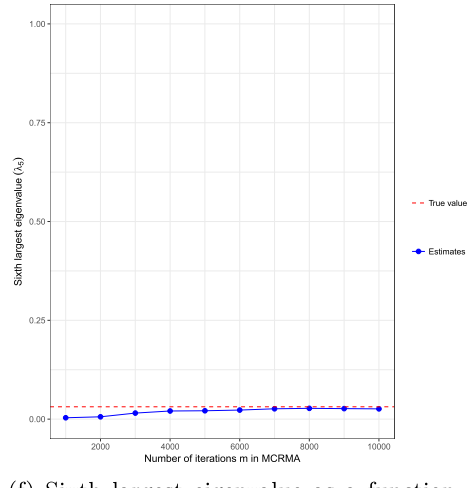
(c) Third largest eigenvalue as a function of iterations m in the MCRMA algorithm. The true eigenvalue is shown as a horizontal line.



(d) Fourth largest eigenvalue as a function of iterations m in the MCRMA algorithm. The true eigenvalue is shown as a horizontal line.



(e) Fifth largest eigenvalue as a function of iterations m in the MCRMA algorithm. The true eigenvalue is shown as a horizontal line.



(f) Sixth largest eigenvalue as a function of iterations m in the MCRMA algorithm. The true eigenvalue is shown as a horizontal line.

Figure 1. (Continued).

where the estimates show satisfactory signs of convergence), illustrates accuracy of the MCRMA method.

In spectral based diagnostics of MCMC algorithms, interest often lies in comparing the convergence rates of the estimated spectra, and that of the estimated functionals of interest (such as the posterior mean in Bayesian statistical analysis). Here we consider the estimated (ergodic

averages) mean of associated second Hermite polynomial for x, $H_2(x) = \frac{1}{\sqrt{2}}(2x^2 - 1)$, and compare its convergence to $\pi H_2 = 0$ with the convergence of MCRMA estimate of the second largest eigenvalue λ_1 to 0.5. In particular, for $m = 1000, 2000, \ldots, 10{,}000$, we compute

(i) $\widehat{\pi H_2}(m) := m^{-1} \sum_{i=0}^{m-1} H_2(x_i)$ using the already generated Markov chain realizations $\{x_i : i = 0, \ldots, 9999\}$
(ii) $\widehat{\lambda}_1(m) - 0.5$, where $\widehat{\lambda}_1(m)$ is the estimated second largest eigenvalue obtained from the MCRMA instance ran with iteration size m

and plot $|\widehat{\pi H_2}(m)|$ and $|\widehat{\lambda}_1(m) - 0.5|$, both as functions of m, in the same diagram. The resulting plots are shown in Figure 2, which shows that the convergence rates of the estimated spectrum

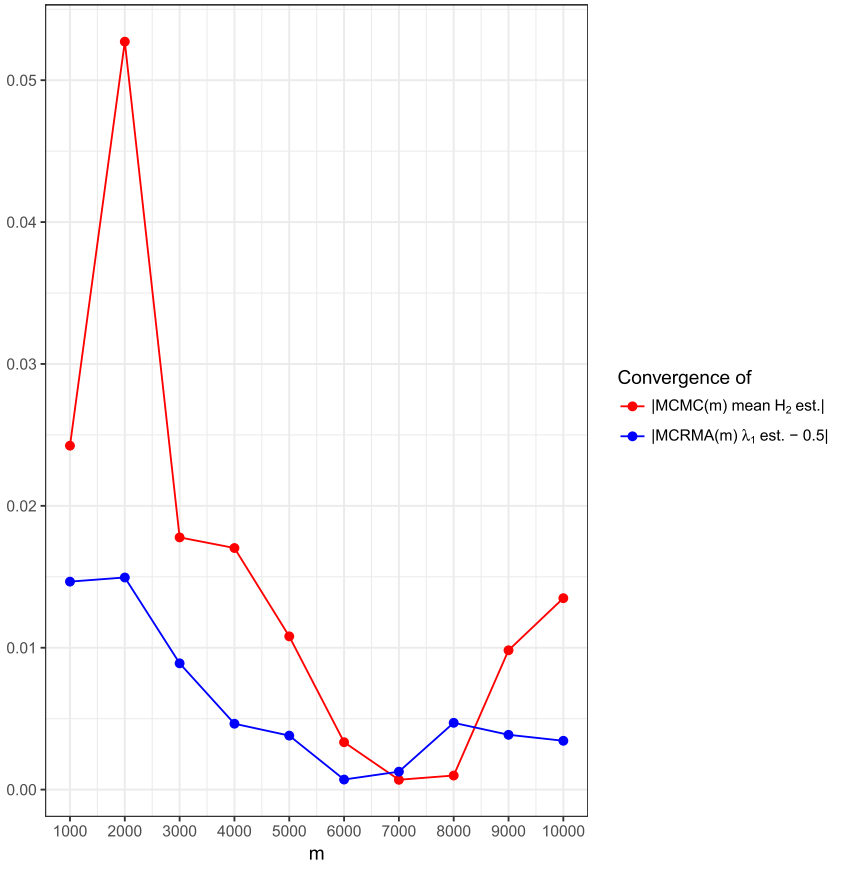


Figure 2. Convergences of the MCMC estimate of mean second Hermite polynomial and MCRMA estimate of second largest eigenvalue, both viewed as functions of iteration size m. The absolute estimated means $|\widehat{\pi H_2}(m)|$ are displayed as red dots, and the absolute differences $|\widehat{\lambda}_1(m) - 0.5|$ are shown as blue dots.

and the estimated mean of second Hermite polynomial are comparable when $m \geq 5000$, and neither convergence is strictly faster than the other.

We end this example by comparing the MCRMA estimates of λ_1 to the estimates obtained using the power sum estimation technique of Qin, Hobert and Khare [35], which we briefly describe in the following. For a positive integer r, define the power sum $s_r := \sum_{i=0}^{\infty} \lambda_i^r$ of eigenvalues $(\lambda_i)_{i \geq 0}$ of the associated trace class Markov operator K. Then for any $r \geq 1$

$$l_r := \frac{s_r - 1}{s_{r-1} - 1} \leq \lambda_1 \leq (s_r - 1)^{1/r} =: u_r$$

(with $s_0 = \infty$), and in addition, Qin, Hobert and Khare [35], Proposition 1, show that as $1 \leq r \to \infty$, $l_r \uparrow \lambda_1$ and $u_r \downarrow \lambda_1$. For DA Markov operators, the authors provide an IID Monte Carlo based estimation technique for s_r, which in turn, provides estimates of l_r and u_r, thus providing asymptotically consistent interval estimates of λ_1. The authors note that r is not required to be very large in practice (in fact, very large r cause instability in estimation, see Qin, Hobert and Khare [35], Section 6), and they recommend using r large enough so that the difference between estimated s_r and 1 is small.

The key step in the power-sum estimation method of Qin, Hobert and Khare [35] is the step of Monte Carlo estimation of s_r. To aid comparability with MCRMA, we set the associated Monte Carlo sample size to be the same $N(m) = \lceil m^{1+10^{-6}} \rceil$, and run 10 instances of trace sum estimation method with $m = 1000, \ldots, 10{,}000$ and $r = 4$. The estimated (l_r, u_r) are then plotted as functions of m together with the MCRMA estimates of λ_1, and displayed in Figure 3, which shows that MCRMA gives slightly better estimates in the current settings.

4.2. Infinite state space application: Polson and Scott DA Gibbs sampler

We consider the Data Augmentation algorithm for Bayesian logistic regression proposed in Polson, Scott and Windle [33]. Let $Y_1, Y_2, \ldots, Y_n$ be independent Bernoulli random variables with $P(Y_i = 1|\boldsymbol{\beta}) = F(\boldsymbol{u}_i^T \boldsymbol{\beta})$. Here $\boldsymbol{u}_i \in \mathbb{R}^p$ is a vector of known covariates associated with Y_i, $i = 1, \ldots, n$, $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of unknown regression coefficients, and $F : \mathbb{R} \to [0, 1] : t \mapsto e^t/(1 + e^t)$ is the distribution function of a standard logistic distribution. For $y_i \in \{0, 1\}$, the likelihood function under this model is given by:

$$P(Y_1 = y_1, \ldots, Y_n = y_n|\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[F\left(\boldsymbol{u}_i^T \boldsymbol{\beta}\right)\right]^{y_i} \left[1 - F\left(\boldsymbol{u}_i^T \boldsymbol{\beta}\right)\right]^{1-y_i}$$

The objective is to make inferences on the regression parameter $\boldsymbol{\beta}$ and we intend to adopt a Bayesian approach, which requires a prior density for $\boldsymbol{\beta}$ to be specified. To keep parity with the literature, in this section we shall slightly abuse our notation by using $\boldsymbol{\beta}$ (not X) to denote the parameter of interest, U to denote the non-stochastic design matrix with ith row $\boldsymbol{u}_i^T$, and $\pi(\boldsymbol{\beta})$ to denote the prior density for $\boldsymbol{\beta}$. Note that here our target distribution is not the prior density $\pi(\cdot)$,

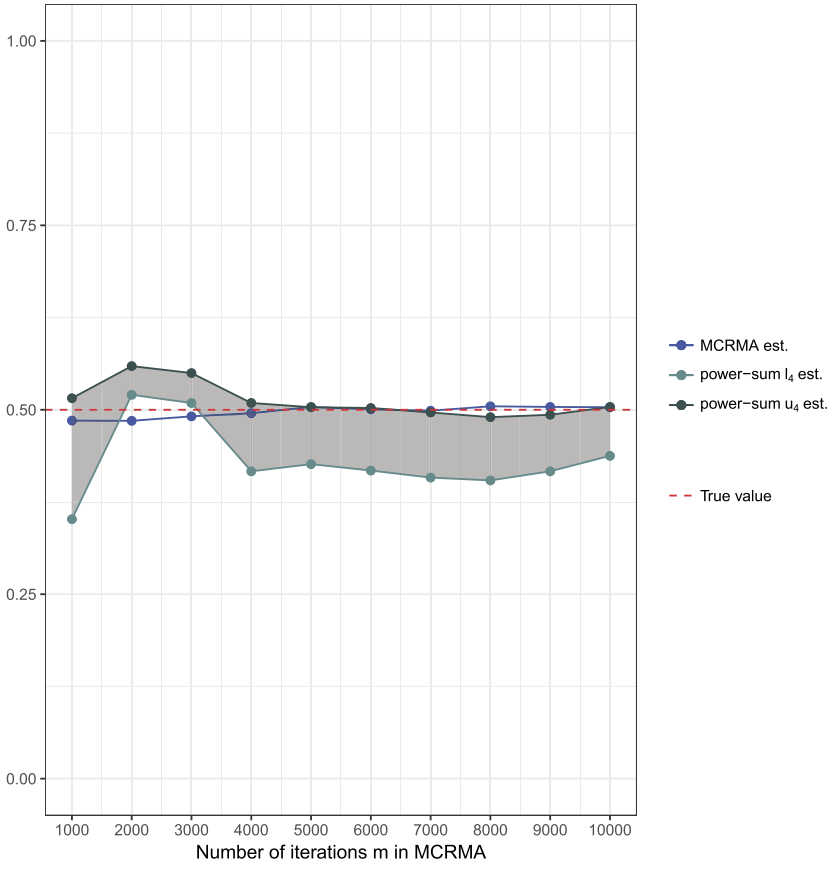


Figure 3. Comparing MCRMA estimates with power-sum estimates of $\lambda_1 = 0.5$. The 10 MCRMA estimates with $m = 1000, \ldots, 10{,}000$ and $N(m) = \lceil m^{1+10^{-6}} \rceil$ are shown as blue dots, the light and dark gray dots are the power-sum estimates of l_r and u_r with $r = 4$ and Monte Carlo sample size $N(m)$, and the shaded gray region provides power-sum interval estimates of λ_1. The true λ_1 is shown as a horizontal red line.

but the posterior density $\pi(\cdot|\boldsymbol{y})$ given the data $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, which is given by

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}) = \frac{1}{c(\boldsymbol{y})}\pi(\boldsymbol{\beta})\left(\prod_{i=1}^{n}\left[F\left(\boldsymbol{u}_i^T\boldsymbol{\beta}\right)\right]^{y_i}\left[1 - F\left(\boldsymbol{u}_i^T\boldsymbol{\beta}\right)\right]^{1-y_i}\right) \tag{14}$$

where

$$c(\boldsymbol{y}) = \int_{\mathbb{R}^p} \pi(\boldsymbol{\beta})\left(\prod_{i=1}^{n}\left[F\left(\boldsymbol{u}_i^T\boldsymbol{\beta}\right)\right]^{y_i}\left[1 - F\left(\boldsymbol{u}_i^T\boldsymbol{\beta}\right)\right]^{1-y_i}\right) d\boldsymbol{\beta} \tag{15}$$

is the normalizing constant dependent of $\boldsymbol{y}$ only. We shall consider a proper $\mathrm{N}_p(\boldsymbol{b}, B)$ prior for $\boldsymbol{\beta}$, as in Choi and Hobert [9]. Note that the posterior density $\pi(\boldsymbol{\beta}|\boldsymbol{y})$ is intractable; it does not have a closed form, and IID sampling is very inefficient even for moderately large p. Polson, Scott and Windle [33] proposed a data augmentation Gibbs sampling algorithm for approximate sampling from $\pi(\boldsymbol{\beta}|\boldsymbol{y})$, which only requires random generation from easy to sample univariate distributions. In the following, we borrow notations from Choi and Hobert [9] where the uniform ergodicity of the Markov chain produced by the Polson, Scott and Windle DA algorithm is proved. Let $\mathbb{R}_+ = (0, \infty)$, and for fixed $\boldsymbol{w} = (w_1, \ldots, w_n)^T \in \mathbb{R}_+^n$

$$\Omega(\boldsymbol{w}) = \mathrm{diag}(w_1, \ldots, w_n),$$

$$\Sigma(\boldsymbol{w}) = \left(U^T \Omega(\boldsymbol{w}) U + B^{-1}\right)^{-1},$$

$$\text{and} \quad \boldsymbol{\mu}(\boldsymbol{y}) = \boldsymbol{\mu} = U^T \left(\boldsymbol{y} - \frac{1}{2}\mathbf{1}_n\right) + B^{-1}\boldsymbol{b}.$$

Then the Polson, Scott and Windle DA Gibbs sampler (Algorithm 4) for generating MCMC samples from the posterior distribution $\pi(\boldsymbol{\beta}|\boldsymbol{y})$ is obtained by iteratively sampling independent w_i from (univariate) Polya-Gamma$(1, |\boldsymbol{u}_i^T \boldsymbol{\beta}|)$ distribution, for $i = 1, \ldots, n$, and then sampling $\boldsymbol{\beta}$ from $\mathrm{N}_p(\Sigma(\boldsymbol{w})\boldsymbol{\mu}, \Sigma(\boldsymbol{w}))$. Here Polya-Gamma$(1, c)$ denotes the Polya-Gamma distribution with parameters 1 and c, which is defined as follows. Let $(E_k)_{k\geq 1}$ be a sequence of IID standard Exponential random variables, and let

$$W = \frac{2}{\pi^2} \sum_{l=1}^{\infty} \frac{E_l}{(2l-1)^2} \tag{16}$$

which has density

$$\widetilde{g}(w) = \sum_{l=1}^{\infty} (-1)^l \frac{(2l+1)}{\sqrt{2\pi w^3}} e^{-\frac{(2l+1)^2}{8w}}; \quad w \geq 0. \tag{17}$$

Then the Polya-Gamma family of densities $\{\widetilde{g}_c : c \geq 0\}$ is obtained through an exponential tilting of the density $\widetilde{g}$:

$$\widetilde{g}_c(x) = \cosh(c/2) e^{-\frac{c^2 x}{2}} \widetilde{g}(x),$$

and a random variable is said to have a Polya-Gamma$(1, c)$ distribution if it has density $\widetilde{g}_c$. (Recall that $\cosh(t) = (e^t + e^{-t})/2$.) An efficient data generating algorithm from Polya-Gamma$(1, c)$ is provided in Polson, Scott and Windle [33].

From the Polson, Scott and Windle Gibbs sampler (Algorithm 4), it follows that

1. For $i = 1, \ldots, n$, the (full) conditional posterior distribution of w_i given $\boldsymbol{\beta}$ is independent Polya-Gamma$(1, |\boldsymbol{u}_i^T \boldsymbol{\beta}|)$, so that the conditional joint density of $\boldsymbol{w} = (w_1, \ldots, w_n)^T$ given $\boldsymbol{\beta}$, $\boldsymbol{y}$ is given by

$$\pi(\boldsymbol{w}|\boldsymbol{\beta}, \boldsymbol{y}) = \prod_{i=1}^{n} \left\{ \cosh\left(\frac{|\boldsymbol{u}_i^T \boldsymbol{\beta}|}{2}\right) \exp\left[-\frac{1}{2}\left(\boldsymbol{u}_i^T \boldsymbol{\beta}\right)^2 w_i\right] \widetilde{g}(w_i) \right\} \tag{18}$$

Algorithm 4 The Polson, Scott and Windle DA Gibbs Sampler

Given a starting value $\boldsymbol{\beta}_0$, iterate between the following two steps:

(i) Draw independent $w_1, \ldots, w_n$ with $w_i \sim \text{Polya-Gamma}(1, |\boldsymbol{u}_i^T \boldsymbol{\beta}|)$, $i = 1, \ldots, n$, and define

$$\Omega(\boldsymbol{w}) = \text{diag}(w_1, \ldots, w_n),$$
$$\Sigma(\boldsymbol{w}) = \left(U^T \Omega(\boldsymbol{w}) U + B^{-1}\right)^{-1},$$
$$\text{and} \quad \boldsymbol{\mu}(\boldsymbol{y}) = \boldsymbol{\mu} = U^T \left(\boldsymbol{y} - \frac{1}{2}\mathbf{1}_n\right) + B^{-1}\boldsymbol{b}.$$

.

(ii) Draw $\boldsymbol{\beta} \sim \text{N}_p(\Sigma(\boldsymbol{w})\boldsymbol{\mu}, \Sigma(\boldsymbol{w}))$.

where $\widetilde{g}$ is as given in (17).

2. The full conditional distribution of $\boldsymbol{\beta}$ given $\boldsymbol{w}$, $\boldsymbol{y}$ is $\text{N}_p(\Sigma(\boldsymbol{w})\boldsymbol{\mu}, \Sigma(\boldsymbol{w}))$ with density

$$\pi(\boldsymbol{\beta}|\boldsymbol{w}, \boldsymbol{y}) = (2\pi)^{-p/2} \left|U^T \Omega(\boldsymbol{w}) U + B^{-1}\right|^{1/2} \times \exp\left[-\frac{1}{2}\left(\boldsymbol{\beta} - \Sigma(\boldsymbol{w})\boldsymbol{\mu}\right)^T \Sigma(\boldsymbol{w})^{-1}\left(\boldsymbol{\beta} - \Sigma(\boldsymbol{w})\boldsymbol{\mu}\right)\right]. \tag{19}$$

Note that the transition density of the associated Markov chain Φ for $\boldsymbol{\beta}$ is given by

$$k(\boldsymbol{\beta}, \boldsymbol{\beta}') = \int_{\mathbb{R}_+^n} \pi(\boldsymbol{\beta}'|\boldsymbol{w}, \boldsymbol{y}) \pi(\boldsymbol{w}|\boldsymbol{\beta}, \boldsymbol{y})\, d\boldsymbol{w}$$

where $\pi(\boldsymbol{\beta}|\boldsymbol{w}, \boldsymbol{y})$ and $\pi(\boldsymbol{w}|\boldsymbol{\beta}, \boldsymbol{y})$ are as given in (19) and (18), respectively. It is clear that this transition density cannot be evaluated in closed form. Moreover, a closed form expression for the normalizing constant $c(\boldsymbol{y})$ in (15) is not available, which means the posterior density $\pi(\boldsymbol{\beta}|\boldsymbol{y})$ in (14) can only be specified up to a constant factor. Thus, exact RMA cannot be applied in this example. However, by letting $\boldsymbol{w}$ play the role of the augmented data z, $f_{Z|X}(\cdot|\cdot)$ the conditional density $\pi(\boldsymbol{w}|\boldsymbol{\beta}, \boldsymbol{y})$ (from which random sampling is easy due to the efficient simulation algorithm from Polya-Gamma$(1, c)$ proposed in Polson, Scott and Windle [33]), and $f_{X|Z}(\cdot|\cdot)$ the simple multivariate normal density $\pi(\boldsymbol{\beta}|\boldsymbol{w}, \boldsymbol{y})$, we can use the extended MCRMA method (Algorithm 3). Since the state space of $\boldsymbol{\beta}$ (and $\boldsymbol{w}$) is infinite, in order to ensure consistency of the MCRMA estimates, we need (A), and (B) in Theorem 3.1 to hold. The following two theorems (Theorem 4.1 and Theorem 4.2) show that the Polson, Scott and Windle Markov chain does indeed satisfy these two conditions, thus guaranteeing consistency of MCRMA estimates in this case. Proofs of Theorem 4.1 and Theorem 4.2 are provided in Section C of the supplemental document.

Theorem 4.1. *For any choice of the* (*proper multivariate normal*) *prior parameters* $\boldsymbol{b}$ *and* B, *the Markov operator associated with Polson, Scott and Windle Markov chain* Φ *is trace class.*

Theorem 4.2. *Let the initial distribution ν_0 of $\boldsymbol{\beta}$ be such that* $\exp[\frac{1}{2}\sum_{i=1}^{n}|\boldsymbol{u}_i^T\boldsymbol{\beta}|]$ *is ν_0-integrable for all $i = 1, \ldots, n$. Then the operator K associated with the Polson, Scott and Windle algorithm satisfies the variance condition* (B).

Remark 4.1. Note that the integrability condition assumed on the initial measure ν_0 in Theorem 4.2 is not very restrictive, and can be easily ensured in practice for a number of families of distribution. For example, if the initial distribution of $\boldsymbol{\beta}$ is Gaussian, integrability of $\exp[\frac{1}{2}\sum_{i=1}^{n}|\boldsymbol{u}_i^T\boldsymbol{\beta}|]$ is immediate.

4.2.1. *Simulation results*

For simulation, we used the the R package `BayesLogit` Polson, Scott and Windle [33] to efficiently draw random samples from the Polya Gamma density. We generated a Polson, Scott and Windle Markov chain on the `nodal` dataset from the R package `boot` Canty and Ripley [6], Davison and Hinkley [12]. The dataset consists of 53 observations on 5 binary predictors (aged, stage, grade, xray and acid) and one response which indicates whether cancer has spread from prostate to surrounding lymph nodes. Taking the maximum likelihood estimate as the starting value, we first generated 30,000 iterations of the Polson, Scott and Windle chain for the regression coefficient $\boldsymbol{\beta}$ ($\in \mathbb{R}^6$, includes one intercept coefficient). We discarded the first 20,000 iterations as burn-in, and kept the remaining 10,000 as the MCMC sample. Then we ran 10 instances of the MCRMA algorithm with the MCMC sample already generated and with $m = 1000, 2000, \ldots, 10{,}000$ and $N(m) = \lceil m^{1+10^{-6}} \rceil$ (to ensure strong consistency), and recorded the 30 largest eigenvalues. Then we created plots similar to the toy normal-normal DA example, except, the true eigenvalues were of course unknown in this case. The resulting plots are shown in Figure 4.

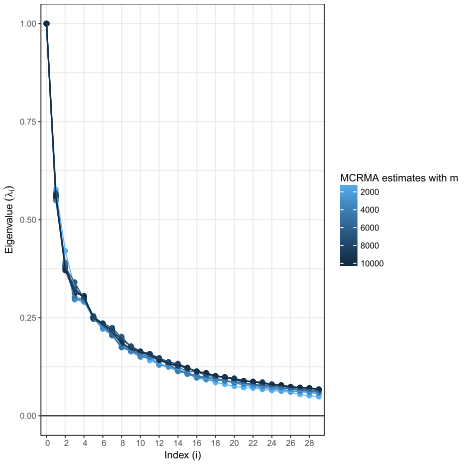
Figure 4(a) shows all 30 eigenvalues obtained from each of the 20 MCRMA instances, plotted as 20 curves, one for each MCRMA instance. The second, third, fourth, fifth and sixth largest estimated eigenvalues, viewed as functions of the MCRMA iteration size m, are shown in Figures 4(b) through 4(f). As is clear from the plots, the MCRMA spectrum estimates for the Polson, Scott and Windle chain show adequate signs of convergence when $m \geq 5000$, thereby providing confidence on the accuracy of estimation.

4.3. Finite state space application: Two component normal mixture

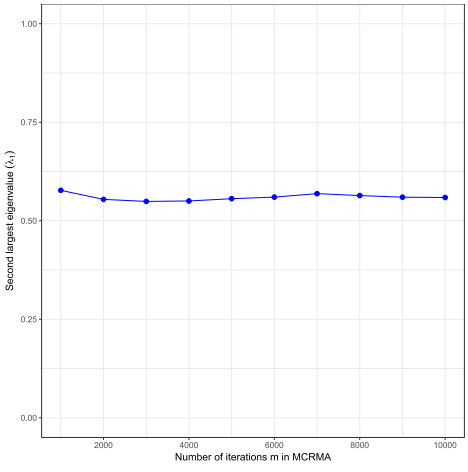
In this section, we consider the problem of Bayesian finite mixture modeling with two components. Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be a random sample from the two component equal variance mixture normal density

$$f(y|\boldsymbol{\mu}, p) = p\frac{1}{\tau}\phi\left(\frac{y-\mu_1}{\tau}\right) + (1-p)\frac{1}{\tau}\phi\left(\frac{y-\mu_2}{\tau}\right) \tag{20}$$

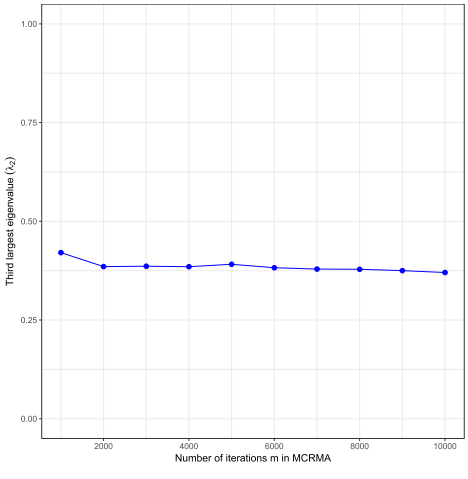
where $p \in [0, 1]$ is the mixing proportion, $\boldsymbol{\mu} = (\mu_1, \mu_2) \in \mathbb{R}^2$ is the vector of component means and $\tau^2 > 0$ is the *known* variance for both components, and $\phi(\cdot)$ is the standard normal density function. The objective is to make inferences on the unknown parameter vector $\boldsymbol{\theta} = (\boldsymbol{\mu}, p)$ through the data $\boldsymbol{y}$, and we adopt a Bayesian approach. The prior density for $\boldsymbol{\theta}$ is taken to be of the form $\pi(\boldsymbol{\theta}) = \pi(p)\pi(\mu_1)\pi(\mu_2)$, with $\pi(p)$ being the Uniform(0, 1) density, and $\pi(\mu_j)$ being



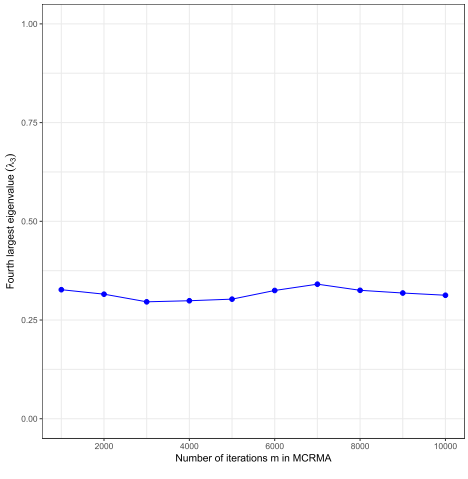
(a) The largest 30 eigenvalues. There are 10 curves, each corresponding to the choices $m = 1000, \cdots, 10,000$ in the MCRMA algorithm.



(b) Second largest eigenvalue as a function of iterations m in the MCRMA algorithm.



(c) Third largest eigenvalue as a function of iterations m in the MCRMA algorithm.

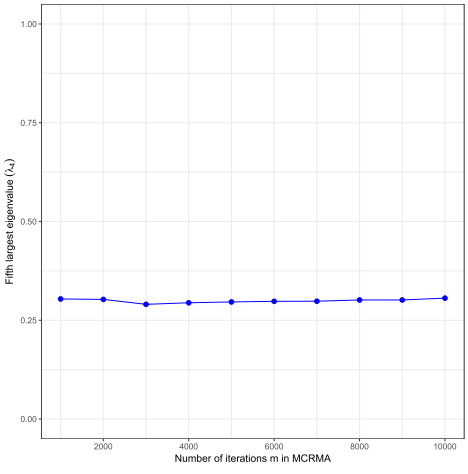


(d) Fourth largest eigenvalue as a function of iterations m in the MCRMA algorithm.

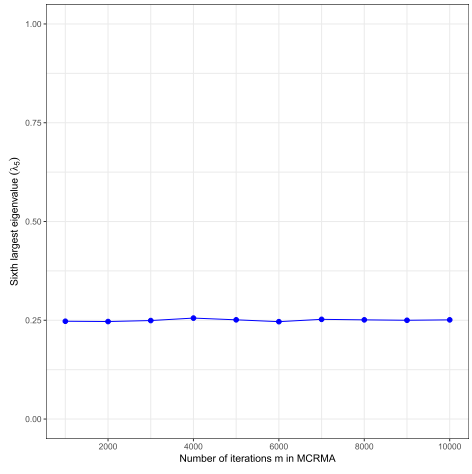
Figure 4. Eigenvalue estimates for the Polson, Scott and Windle DA Markov chain using the MCRMA algorithm.

the $\mathrm{N}(0, \tau^2)$ density. Then the posterior density for $\boldsymbol{\theta}$ is given by

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \pi(\boldsymbol{\mu}, p|\boldsymbol{y}) = \frac{1}{c(\boldsymbol{y})} \prod_{i=1}^{n} \left\{ p \frac{1}{\tau} \phi\left(\frac{y_i - \mu_1}{\tau} \right) + (1-p) \frac{1}{\tau} \phi\left(\frac{y_i - \mu_2}{\tau} \right) \right\} \pi(\boldsymbol{\theta}) \qquad (21)$$



(e) Fifth largest eigenvalue as a function of iterations m in the MCRMA algorithm.



(f) Sixth largest eigenvalue as a function of iterations m in the MCRMA algorithm.

Figure 4. (Continued).

where $c(\boldsymbol{y})$ is the normalizing constant that makes (21) a proper density. It is clear that $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ is intractable, which makes evaluation of posterior mean or IID simulation infeasible. We therefore resort to approximate sampling via MCMC. A slightly general version of this problem (with *unknown* and *different* component variances τ_1^2 and τ_2^2) is considered in Hobert, Roy and Robert [22], Section 6.2, and the authors consider two different Gibbs sampling algorithms, namely, the mixture Data Augmentation algorithm (MDA algorithm, or simply, MDA) and the Frühwirth-Schnatter algorithm (FS algorithm, or simply, FS), to generate MCMC samples from the posterior.

4.3.1. *MDA algorithm*

Let us introduce latent component indicators $z_1, \ldots, z_n$, with $z_i = j$ indicating that the ith observation y_i is coming from the jth component $\mathrm{N}(\mu_j, \tau^2)$ for $j = 1, 2$. Then,

1. the full conditional posterior distribution of the components of $\boldsymbol{\theta}$ given $\boldsymbol{z}$ are independent, with p being $\mathrm{Beta}(c_1 + 1, c_2 + 1)$ and μ_j being $\mathrm{N}(\frac{c_j}{c_j+1}\bar{y}_j, \frac{\tau^2}{c_j+1})$ with $c_j = \sum_{i=1}^n \mathbb{1}_{\{j\}}(z_i)$ and $\bar{y}_j = c_j^{-1} \sum_{i=1}^n y_i \mathbb{1}_{\{j\}}(z_i)$ for $j = 1, 2$. We shall denote the corresponding density of $\boldsymbol{\theta}$ by $\pi(\boldsymbol{\theta}|\boldsymbol{z}, \boldsymbol{y})$.
2. the full conditional posterior density (*mass*, with respect to the counting measure ζ) of $\boldsymbol{z}$ given $\boldsymbol{\theta}$ is given by

$$\pi(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{y}) \propto \prod_{i=1}^{n} \big(\widetilde{p}_i \mathbb{1}_{\{1\}}(z_i) + (1 - \widetilde{p}_i)\mathbb{1}_{\{2\}}(z_i)\big) \tag{22}$$

Algorithm 5 The Mixture DA (MDA) Gibbs Sampler

1 Given a starting value $(\boldsymbol{\mu}_0, p_0)$ for the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\mu}, p)$, iterate between the following two steps:

(i) Draw independent $z_1, \ldots, z_n$ with z_i having a categorical probability distribution with categories 1 and 2, and

$$P(z_i = j) = \begin{cases} \dfrac{p\phi(\frac{y_i-\mu_1}{\tau})}{p\phi(\frac{y_i-\mu_1}{\tau}) + (1-p)\phi(\frac{y_i-\mu_2}{\tau})} & \text{if } j = 1 \\ \dfrac{(1-p)\phi(\frac{y_i-\mu_2}{\tau})}{p\phi(\frac{y_i-\mu_1}{\tau}) + (1-p)\phi(\frac{y_i-\mu_2}{\tau})} & \text{if } j = 2 \end{cases}.$$

(ii) Compute $c_j = \sum_{i=1}^n \mathbb{1}_{\{j\}}(z_i)$ and $\bar{y}_j = c_j^{-1} \sum_{i=1}^n y_i \mathbb{1}_{\{j\}}(z_i)$. Then independently generate:

(a) p from Beta$(c_1 + 1, c_2 + 1)$

(b) μ_j from N$(\frac{c_j}{c_j+1}\bar{y}_j, \frac{\tau^2}{c_j+1})$ for $j = 1, 2$.

where

$$\widetilde{p}_i = \frac{p\phi(\frac{y_i-\mu_1}{\tau})}{p\phi(\frac{y_i-\mu_1}{\tau}) + (1-p)\phi(\frac{y_i-\mu_2}{\tau})}.$$

The MDA algorithm entails iterative generation of $\boldsymbol{z}$ from $\pi(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{y})$, and $\boldsymbol{\theta}$ from $\pi(\boldsymbol{\theta}|\boldsymbol{z}, \boldsymbol{y})$. The resulting Gibbs sampler is formally displayed in Algorithm 5.

Note that, although the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\mu}, p)$ in the MDA algorithm lives on the infinite space $\mathcal{X} = \mathbb{R}^2 \times [0, 1]$, the latent data $\boldsymbol{z} = (z_1, \ldots, z_n)$ lives on the finite state space $\mathcal{Z} = \{1, 2\}^n$. We shall, therefore, study the spectrum of the Markov operator K^* associated with the latent data $\boldsymbol{z}$ (see Remark 3.1). The Markov transition density associated with the operator K^* is given by

$$k^*(\boldsymbol{z}, \boldsymbol{z}') = \int_{\mathcal{X}} \pi(\boldsymbol{z}'|\boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{z}, \boldsymbol{y})d\boldsymbol{\theta} \tag{23}$$

which is, of course, not available in closed form, because of the denominators of the product terms in $\pi(\boldsymbol{z}'|\boldsymbol{\theta}, \boldsymbol{y})$. However, $\pi(\boldsymbol{z}'|\boldsymbol{\theta}, \boldsymbol{y})$ is available in closed form and $\pi(\boldsymbol{\theta}|\boldsymbol{z}, \boldsymbol{y})$ is easy to sample from. Thus, the MCRMA method can be applied here, and the estimates are guaranteed to be strongly consistent (Remark 3.1). Recall that MCRMA requires evaluation of the stationary density $\pi(\boldsymbol{z}|\boldsymbol{y})$ for $\boldsymbol{z}$. Straight-forward calculations show that

$$\pi(\boldsymbol{z}|\boldsymbol{y}) \propto B(c_1 + 1, c_2 + 1) \prod_{j=1}^{2} \left[(1 + c_j)^{-\frac{1}{2}} \exp\left(\frac{c_j^2 \bar{y}_j^2}{2\tau^2(1 + c_j)} \right) \right]. \tag{24}$$

Since the normalizing constant that makes (24) a density is not available in closed form, we shall, therefore use Algorithm 3.

4.3.2. *FS algorithm*

Along with MDA, Hobert, Roy and Robert [22] consider another Gibbs sampling algorithm, called the Frühwirth-Schnatter (FS) algorithm Frühwirth-Schnatter [19], which is obtained by inserting an intermediate random label switching step in between the two steps of MDA. The key idea here is to randomly permute the labels of the latent variable z obtained in the first step of MDA, before moving on to the second step. That is, after generating z from the conditional distribution of $z|\boldsymbol{\theta}$, instead of drawing the next state of $\boldsymbol{\theta}$ directly from $\boldsymbol{\theta}|z$, here one first randomly permutes the labels of components in the mixture model, and switches the labels of z according to that random permutation to get z'. The next state of $\boldsymbol{\theta}$ is then generated from the conditional distribution of $\boldsymbol{\theta}|z'$. In the context of two component mixture models, the intermediate step $z \to z'$ simply entails performing a Bernoulli experiment with probability of success 0.5. One then takes $z' = \bar{z}$ or $z' = z$ according as whether the Bernoulli experiment results in a success or a failure, where $\bar{z}$ denotes z with its 1's and 2's flipped.

The computationally inexpensive label switching step in the FS algorithm is introduced to force movement between the symmetric modes of the posterior density $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. This makes the FS algorithm superior to the MDA algorithm in terms of convergence and mixing. The FS algorithm is in fact a member of a wide class of so-called *sandwich algorithms*, where one inserts an inexpensive intermediate *meat* step inside the two *bread* steps of a DA algorithm to achieve better convergence and mixing. In fact, when the operator associated with a Markov chain is trace class, the spectrum of a sandwich chain is guaranteed to be bounded above by that of the parent DA chain, with at least one strict inequality Khare and Hobert [26]. In the current setting, since the MDA Markov chain is trace class (the latent state space is finite), the FS chain is therefore guaranteed to be better mixing than the DA chain. To visualize or quantify the improvement, however, information on the actual spectra of the two chains is needed. Clearly, the spectrum of the FS chain, similar to the MDA chain, can neither be evaluated analytically nor can be estimated in exact RMA method of Adamczak and Bednorz [1], since the associated Markov transition density is not available in closed form. Instead, we make use of MCRMA estimation, as described in the following.

The usual sandwich representation (with three steps – two *bread* steps similar to MDA and one additional *meat* step) of the FS algorithm does not furnish a Markov transition density in the form (10); however, following Hobert, Roy and Robert [22], Section 5.2, one can represent the algorithm as a DA algorithm with different joint (and hence, different full conditional), but same marginal posterior distributions as MDA. In particular, the DA representation of the FS algorithm entails iterative random generation of z from the conditional density $\widetilde{\pi}(z|\boldsymbol{\theta}, \boldsymbol{y})$, and $\boldsymbol{\theta}$ from $\widetilde{\pi}(\boldsymbol{\theta}|z, \boldsymbol{y})$, where

$$\widetilde{\pi}(\boldsymbol{\theta}|z, \boldsymbol{y}) = \int_{\mathcal{Z}} \pi\left(\boldsymbol{\theta}|z', \boldsymbol{y}\right) r\left(z, z'\right) d\zeta\left(z'\right)$$

$$\text{and} \quad \widetilde{\pi}(z|\boldsymbol{\theta}, \boldsymbol{y}) = \frac{\pi(z|\boldsymbol{y})}{\pi(\boldsymbol{\theta}|\boldsymbol{y})} \int_{\mathcal{Z}} \pi\left(\boldsymbol{\theta}|z'\right) r\left(z, z'\right) d\zeta\left(z'\right)$$

and $r(z, z')$ is the transition density (with respective to the counting measure ζ) associated with the intermediate *meat* step $z \to z'$ in the sandwich representation of FS. Since the intermediate

meat step is that of random label switchings, we have

$$r(z, z') = \frac{1}{2}\mathbb{1}_{\{z\}}(z') + \frac{1}{2}\mathbb{1}_{\{\bar{z}\}}(z'),$$

where $\bar{z}$ is z with its 1's and 2's flipped, and therefore,

$$\widetilde{\pi}(\boldsymbol{\theta}|z, \boldsymbol{y}) = \frac{1}{2}\pi(\boldsymbol{\theta}|z, \boldsymbol{y}) + \frac{1}{2}\pi(\boldsymbol{\theta}|\bar{z}, \boldsymbol{y}) \tag{25}$$

and

$$\widetilde{\pi}(z|\boldsymbol{\theta}, \boldsymbol{y}) = \frac{1}{2}\pi(z|\boldsymbol{\theta}, \boldsymbol{y}) + \frac{1}{2}\pi(\bar{z}|\boldsymbol{\theta}, \boldsymbol{y}) \tag{26}$$

with (26) being a consequence of (24). Note that $\widetilde{\pi}(\boldsymbol{\theta}|z, \boldsymbol{y})$ and $\widetilde{\pi}(z|\boldsymbol{\theta}, \boldsymbol{y})$ are just half-half mixtures of standard densities, and can be easily sampled. The DA form of the FS algorithm is formally displayed in Algorithm 6.

Similar to the MDA case, the spectrum of the Markov operator associated with the $\boldsymbol{\theta}$ sub-chain of an FS Markov chain can be studied through that of the Markov operator $\widetilde{K}^*$ corresponding to the z sub-chain. From the DA representation described in the previous paragraph, it follows that

---

Algorithm 6 The Frühwirth-Schnatter (FS) Gibbs Sampling algorithm (in the DA form)

Given a starting value $(\boldsymbol{\mu}_0, p_0)$ for the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\mu}, p)$, iterate between the following two steps:

(i) Draw independent $z'_1, \ldots, z'_n$ with z'_i having a categorical probability distribution with categories 1 and 2, and

$$P(z'_i = j) = \begin{cases} \dfrac{p\phi(\frac{y_i-\mu_1}{\tau})}{p\phi(\frac{y_i-\mu_1}{\tau}) + (1-p)\phi(\frac{y_i-\mu_2}{\tau})} & \text{if } j = 1 \\ \dfrac{(1-p)\phi(\frac{y_i-\mu_2}{\tau})}{p\phi(\frac{y_i-\mu_1}{\tau}) + (1-p)\phi(\frac{y_i-\mu_2}{\tau})} & \text{if } j = 2 \end{cases},$$

and call $z' = (z'_1, \ldots, z'_n)$. Now perform a Bernoulli experiment with probability of success 0.5. If the experiment results in a success, define $z = z'$, or else define $z = \bar{z'}$, where $\bar{z'}$ is z' with its 1's and 2's flipped.

(ii) Perform another Bernoulli experiment with probability of success 0.5. Define $z^* = z$ if the experiment results in a success, and $z^* = \bar{z}$ otherwise. Compute $c_j = \sum_{i=1}^n \mathbb{1}_{\{j\}}(z_i^*)$ and $\bar{y}_j = c_j^{-1}\sum_{i=1}^n y_i \mathbb{1}_{\{j\}}(z_i^*)$ for $j = 1, 2$. Then independently generate:
   (a) p from Beta$(c_1 + 1, c_2 + 1)$
   (b) μ_j from N$(\frac{c_j}{c_j+1}\bar{y}_j, \frac{\tau^2}{c_j+1})$ for $j = 1, 2$.

---

the Markov transition density associated with $\widetilde{K}^*$ can be written as

$$\widetilde{k}^*(z, z') = \int_{\mathcal{X}} \widetilde{\pi}(z'|\boldsymbol{\theta}, \boldsymbol{y})\widetilde{\pi}(\boldsymbol{\theta}|z, \boldsymbol{y})\, d\boldsymbol{\theta}. \tag{27}$$

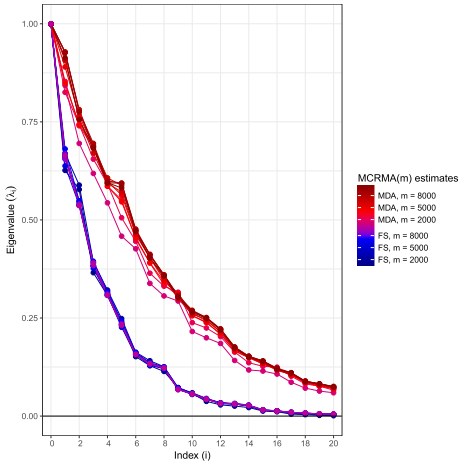
Owing to the above representation and the facts that $\widetilde{\pi}(z|\boldsymbol{\theta}, \boldsymbol{y})$ is available in closed form, and $\widetilde{\pi}(\boldsymbol{\theta}|z, \boldsymbol{y})$ is easy to sample from, one can use the MCRMA method to estimate the spectrum of $\widetilde{K}^*$.

4.3.3. *Simulation study*

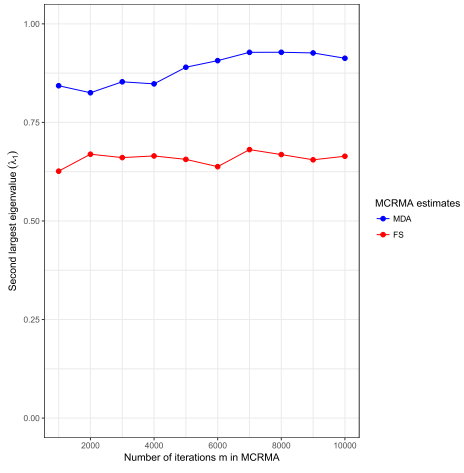
To illustrate the performance of the MCRMA method in estimating the spectra of K^* and $\widetilde{K}^*$ (the MDA and FS Markov operators respectively), we consider a simulated dataset with sample size $n = 20$ from the mixture density (20), with $\mu_1 = 0$, $\mu_2 = 0.1$, $p = 0.5$ and fixed $\tau = 0.1$. Then, with k-means estimates taken as the starting values, we separately generate 10,000 realizations of MDA and FS Markov chains after discarding first 20,000 realizations as burn-in from each chain. Then we extract the z sub-chains from the two MCMC samples and use them in the MCRMA method to estimate their spectra. Note that the latent space $\mathcal{Z}$ in both algorithms consist of $2^{20} = 1{,}048{,}576$ states, which means, each of the associated Markov operators corresponds to a $1{,}048{,}576 \times 1{,}048{,}576$ matrix of transition probabilities. Hence, in order to find the true eigenvalues, one needs to compute the eigenvalues of $1{,}048{,}576 \times 1{,}048{,}576$ matrices, which is practically infeasible even though the state space is finite. However, the MCRMA method can still be applied here to provide estimates, as we discuss in the following.

For each of the two Markov chains, we run 10 separate instances of MCRMA, with number of Markov chain iterations $m = 1000, 2000, \ldots, 10{,}000$, and Monte Carlo sample size $N = 5000$, to estimate the eigenvalues, and then create plots similar to Figure 4. Note that, because the latent state space $\mathcal{Z}$ is finite, strong consitency of the MCRMA estimator is automatically ensured, and no relationship between the rate of growth of N and m is required. For each of the two chains, and for each of the 10 MCRMA instances, we record the first 21 estimated eigenvalues (including the trivial eigenvalue $\lambda_0 = 1$) and plot them in Figure 5. Figure 5(a) shows all 21 eigenvalues obtained from each of the 10 MCRMA instances and for each Markov chain, plotted as 20 curves. The second, third, fourth, fifth and sixth largest estimated eigenvalues, viewed as functions of the MCRMA iteration size m, are shown in Figures 5(b) through 5(f). From these plots, it appears that the MCRMA estimates for the MDA chain show some instability. Most of these estimates eventually stabilize, but it is interesting to note that the behavior of FS spectrum estimates is much more stable than the corresponding MDA spectrum estimates, even for smaller m's. This is due to the fact that the FS chain is better mixing than the MDA chain, which in turn, is a consequence of the theoretically proven fact that the true spectrum of the MDA chain dominates that of the FS chain (see Section 4.3.2). As clearly displayed by the plots, the MCRMA estimates also exhibit this dominance, and provides us a visual idea of the gains achieved in the FS algorithm in terms of convergence and mixing.

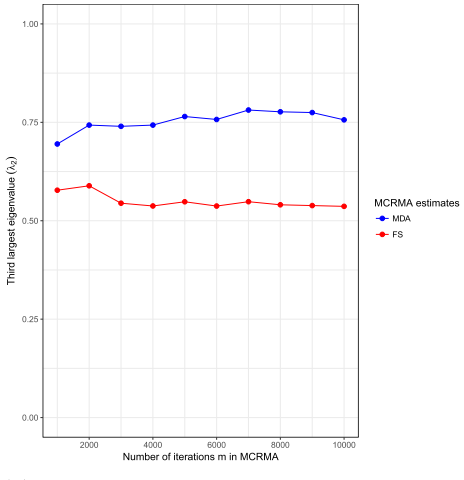
Remark 4.2. It should be noted that the performance of MCRMA can be poor when the state space of the Markov chain is *finite, but extremely large*, especially if the chain is poorly mixing. Although the spectrum estimates are guaranteed to converge to the truth for any Markov chain



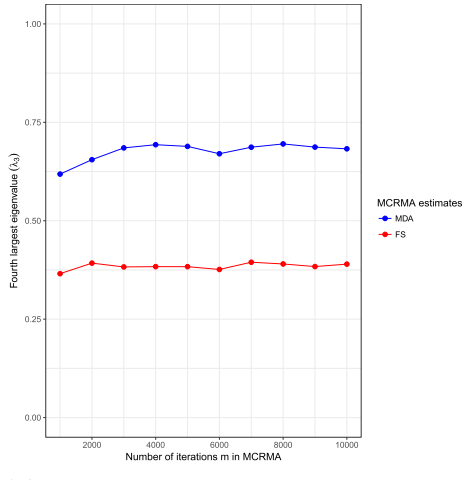
(a) The largest 21 eigenvalues of the MDA and FS chains. There are 10 curves for each Markov chain, each corresponding to the choices $m = 1000, \cdots, 10,000$ in the MCRMA algorithm.



(b) Second largest eigenvalues of the MDA and FS algorithm as functions of iterations m in the MCRMA algorithm.

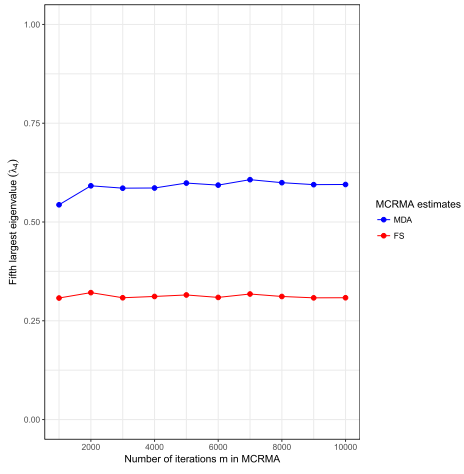


(c) Third largest eigenvalues of the MDA and FS algorithm as functions of iterations m in the MCRMA algorithm.

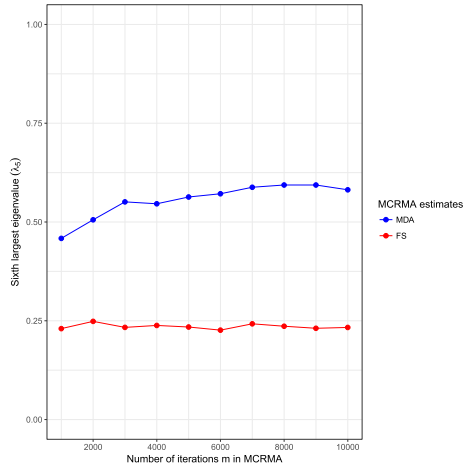


(d) Fourth largest eigenvalues of the MDA and FS algorithm as functions of iterations m in the MCRMA algorithm.

Figure 5. Eigenvalue estimates for the MDA and FS Markov chains using the MCRMA algorithm.



(e) Fifth largest eigenvalues of the MDA and FS algorithm as functions of iterations m in the MCRMA algorithm.



(f) Sixth largest eigenvalues of the MDA and FS algorithm as functions of iterations m in the MCRMA algorithm.

Figure 5. (Continued).

with a finite state space, in practice however, the value of m required for a reasonable approximation can be too large to handle (recall that we need to find the eigenvalues of an $m \times m$ matrix to obtain the eigenvalue estimates). In our case, we tried running the MCRMA algorithm for MDA and FS chains with $n = 30$ (more than a billion states), but the estimates did not show enough signs of convergence with $m \leq 10{,}000$.

5. Discussion

As stated in the Introduction, while bounding or estimating the spectral gap (or equivalently, the second largest eigenvalue) has received a lot of attention over the past three decades, very few methods have been proposed for accurately estimating the entire spectrum of Markov chains arising in modern applications. Building on the work of Koltchinskii and Giné [28], Adamczak and Bednorz [1] develop an elegant method to estimate the spectrum of a trace class Markov operator using random matrix approximations. However, this method requires closed form expressions for the Markov transition density (and the stationary density), which is often unavailable in practice. We consider the general class of Markov chains arising from trace class Data Augmentation algorithms, where the transition density can typically only be expressed as an intractable integral. We develop a Monte Carlo based random matrix approximation method to consistently estimate the entire spectrum of the corresponding DA Markov operators.

The particular integral form of the DA transition density in (10) was critical in the development of our method. This form enables us to provide Monte Carlo based approximations for the

intractable Markov transition density. We are able to show in Theorem 3.1 that the eigenvalues of the subsequently constructed random matrix still consistently estimates the desired spectrum. Methods to approximate general intractable transition densities, which may not necessarily have the integral form in (10), have been proposed in the literature, see, for example, Athreya and Atuncar [4]. The next obvious question in this line of research is: if the intractable transition densities appearing in the random matrix approximation of Adamczak and Bednorz [1] are replaced by approximations based on these methods, does that still lead to consistent estimates of the desired spectrum? This is a challenging question, and will be investigated in future research.

Acknowledgements

Kshitj Khare's work was partially supported by NSF grant DMS-1511945.

Supplementary Material

Supplement to "Consistent estimation of the spectrum of trace class data augmentation algorithms" (DOI: 10.3150/19-BEJ1112SUPP; .pdf). The supplement provides proofs of the theorems and lemmas introduced in this article.

References

[1] Adamczak, R. and Bednorz, W. (2015). Some remarks on MCMC estimation of spectra of integral operators. *Bernoulli* **21** 2073–2092. MR3378459
[2] Albert, J.H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. MR1224394
[3] Asmussen, S. and Glynn, P.W. (2011). A new proof of convergence of MCMC via the ergodic theorem. *Statist. Probab. Lett.* **81** 1482–1485. MR2818658
[4] Athreya, K.B. and Atuncar, G.S. (1998). Kernel estimation for real-valued Markov chains. *Sankhya, Ser. A* **60** 1–17. MR1714774
[5] Bennett, C.H. (1976). Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* **22** 245–268. MR0471852
[6] Canty, A. and Ripley, B.D. (2017). boot: Bootstrap R (S-Plus) Functions. R package version 1.3-19.
[7] Chakraborty, S. and Khare, K. (2017). Convergence properties of Gibbs samplers for Bayesian probit regression with proper priors. *Electron. J. Stat.* **11** 177–210. MR3604022
[8] Chakraborty, S. and Khare, K. (2019). Supplement to "Consistent estimation of the spectrum of trace class data augmentation algorithms." DOI:10.3150/19-BEJ1112SUPP.
[9] Choi, H.M. and Hobert, J.P. (2013). The Polya-gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic. *Electron. J. Stat.* **7** 2054–2064. MR3091616
[10] Choi, H.M. and Román, J.C. (2017). Analysis of Polya-Gamma Gibbs sampler for Bayesian logistic analysis of variance. *Electron. J. Stat.* **11** 326–337. MR3606773
[11] Conway, J.B. (1990). *A Course in Functional Analysis*, 2nd ed. *Graduate Texts in Mathematics* **96**. New York: Springer. MR1070713
[12] Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and Their Application*. *Cambridge Series in Statistical and Probabilistic Mathematics* **1**. Cambridge: Cambridge Univ. Press. MR1478673

[13] Diaconis, P., Khare, K. and Saloff-Coste, L. (2008). Gibbs sampling, exponential families and orthogonal polynomials. *Statist. Sci.* **23** 151–178. MR2446500
[14] Diaconis, P. and Saloff-Coste, L. (1993). Comparison techniques for random walk on finite groups. *Ann. Probab.* **21** 2131–2156. MR1245303
[15] Diaconis, P. and Saloff-Coste, L. (1996). Nash inequalities for finite Markov chains. *J. Theoret. Probab.* **9** 459–510. MR1385408
[16] Diaconis, P. and Stroock, D. (1991). Geometric bounds for eigenvalues of Markov chains. *Ann. Appl. Probab.* **1** 36–61. MR1097463
[17] Eddelbuettel, D. (2013). *Seamless R and C++ Integration with Rcpp*. New York: Springer.
[18] François, O. (2000). Geometric inequalities for the eigenvalues of concentrated Markov chains. *J. Appl. Probab.* **37** 15–28. MR1761658
[19] Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *J. Amer. Statist. Assoc.* **96** 194–209. MR1952732
[20] Garren, S.T. and Smith, R.L. (2000). Estimating the second largest eigenvalue of a Markov transition matrix. *Bernoulli* **6** 215–242. MR1748720
[21] Hobert, J. P., Jung, Y. J., Khare, K. and Qin, Q. (2015). Convergence analysis of the Data Augmentation algorithm for Bayesian linear regression with non-Gaussian errors. ArXiv e-prints.
[22] Hobert, J.P., Roy, V. and Robert, C.P. (2011). Improving the convergence properties of the data augmentation algorithm with an application to Bayesian mixture modeling. *Statist. Sci.* **26** 332–351. MR2918006
[23] Hoffman, A.J. and Wielandt, H.W. (1953). The variation of the spectrum of a normal matrix. *Duke Math. J.* **20** 37–39. MR0052379
[24] Jones, G.L. and Hobert, J.P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statist. Sci.* **16** 312–334. MR1888447
[25] Jörgens, K. (1982). *Linear Integral Operators. Surveys and Reference Works in Mathematics* **7**. Boston, MA–London: Pitman. MR0647629
[26] Khare, K. and Hobert, J.P. (2011). A spectral analytic comparison of trace-class data augmentation algorithms and their sandwich variants. *Ann. Statist.* **39** 2585–2606. MR2906879
[27] Khare, K. and Zhou, H. (2009). Rates of convergence of some multivariate Markov chains with polynomial eigenfunctions. *Ann. Appl. Probab.* **19** 737–777. MR2521887
[28] Koltchinskii, V. and Giné, E. (2000). Random matrix approximation of spectra of integral operators. *Bernoulli* **6** 113–167. MR1781185
[29] Lawler, G.F. and Sokal, A.D. (1988). Bounds on the L^2 spectrum for Markov chains and Markov processes: A generalization of Cheeger's inequality. *Trans. Amer. Math. Soc.* **309** 557–580. MR0930082
[30] 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. MR1279653
[31] Meng, X.-L. and Wong, W.H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statist. Sinica* **6** 831–860. MR1422406
[32] Pal, S., Khare, K. and Hobert, J.P. (2017). Trace class Markov chains for Bayesian inference with generalized double Pareto shrinkage priors. *Scand. J. Stat.* **44** 307–323. MR3658516
[33] Polson, N.G., Scott, J.G. and Windle, J. (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *J. Amer. Statist. Assoc.* **108** 1339–1349. MR3174712
[34] Qin, Q. and Hobert, J.P. (2018). Trace-class Monte Carlo Markov chains for Bayesian multivariate linear regression with non-Gaussian errors. *J. Multivariate Anal.* **166** 335–345. MR3799651
[35] Qin, Q., Hobert, J.P. and Khare, K. (2017). Estimating the spectral gap of a trace-class Markov operator. Preprint. Available at arXiv:1704.00850.
[36] R Core Team (2015). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

[37] Raftery, A.E. and Lewis, S. (1992). How many iterations in the Gibbs sampler. *Bayesian Stat*. **4** 763–773.
[38] Rajaratnam, B., Sparks, D., Khare, K. and Zhang, L. (2017). Scalable Bayesian shrinkage and uncertainty quantification in high-dimensional regression. ArXiv e-prints.
[39] Retherford, J.R. (1993). *Hilbert Space*: *Compact Operators and the Trace Theorem*. *London Mathematical Society Student Texts* **27**. Cambridge: Cambridge Univ. Press. MR1237405
[40] Rosenthal, J.S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *J. Amer. Statist. Assoc*. **90** 558–566. MR1340509
[41] Roy, V. (2012). Convergence rates for MCMC algorithms for a robust Bayesian binary regression model. *Electron. J. Stat*. **6** 2463–2485. MR3020272
[42] Saloff-Coste, L. (2004). Total variation lower bounds for finite Markov chains: Wilson's lemma. In *Random Walks and Geometry* 515–532. Berlin: de Gruyter. MR2087800
[43] Sinclair, A. and Jerrum, M. (1989). Approximate counting, uniform generation and rapidly mixing Markov chains. *Inform. and Comput*. **82** 93–133. MR1003059
[44] Wickham, H. (2007). Reshaping data with the reshape package. *J. Stat. Softw.* **21** 1–20.
[45] Wickham, H. (2016). *ggplot2*: *Elegant Graphics for Data Analysis*. New York: Springer.
[46] Yuen, W.K. (2000). Applications of geometric bounds to the convergence rate of Markov chains on $\mathbf{R}^n$. *Stochastic Process. Appl*. **87** 1–23. MR1751162

Received November 2017 and revised July 2018