INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES
Série des Documents de Travail du CREST
(Centre de Recherche en Economie et Statistique)

n° 2010-29

Improving the Convergence Properties of the Data Augmentation Algorithm with an Application to Bayesian Mixture Modelling

J. P HOBERT[1] - C. P. ROBERT[2]
V. ROY[3]

1 Department of Statistics, University of Florida.
2 Université Paris-Dauphine, CEREMADE and CREST-INSEE, Paris.
3 Department of Statistics, Iowa State University.

Improving the Convergence Properties of the Data Augmentation Algorithm with an Application to Bayesian Mixture Modelling

James P. Hobert
Department of Statistics
University of Florida

Vivekanada Roy
Department of Statistics
Iowa State University

Christian P. Robert
Université Paris Dauphine
& CREST, INSEE

November 2009

Abstract

Every reversible Markov chain defines an operator whose spectrum encodes the convergence properties of the chain. When the state space is finite, the spectrum is just the set of eigenvalues of the corresponding Markov transition matrix. However, when the state space is infinite, the spectrum may be uncountable, and is nearly always impossible to calculate. In most applications of the data augmentation (DA) algorithm, the state space of the DA Markov chain is infinite. However, we show that, under regularity conditions that include the finiteness of the augmented space, the operators defined by the DA chain and Hobert and Marchev's (2008) alternative chain are both compact, and the corresponding spectra are both finite subsets of $[0, 1)$. Moreover, we prove that the spectrum of Hobert and Marchev's (2008) chain dominates the spectrum of the DA chain in the sense that the ordered elements of the former are all less than or equal to the corresponding elements of the latter. As a concrete example, we study a widely used DA algorithm for the exploration of posterior densities associated with Bayesian mixture models (Diebolt and Robert, 1994). In particular, we compare this mixture DA algorithm with an alternative algorithm proposed by Frühwirth-Schnatter (2001) that is based on random label switching.

1 Introduction

The data augmentation algorithm (Tanner and Wong, 1987) is a Markov chain Monte Carlo method that can be used to approximate intractable expectations. Let $f_X : \mathbb{R}^p \to [0, \infty)$ denote a probability density function. Assume that expectations with respect to f_X cannot be computed analytically, and that direct simulation from f_X is impossible. To build a data augmentation (DA) algorithm, one must identify a joint density, say $f : \mathbb{R}^p \times \mathbb{R}^q \to [0, \infty)$, that satisfies two conditions: (i) the x-marginal is f_X, and (ii) sampling from the associated conditional densities, $f_{X|Y}(\cdot|y)$ and $f_{Y|X}(\cdot|x)$, is straightforward. The first of the two conditions allows us to construct a Markov chain having f_X as an invariant density, and the second ensures that we are able to simulate this chain. Indeed, let $\{X_n\}_{n=0}^{\infty}$ be a Markov chain whose dynamics are defined (implicitly) through the following two-step procedure for moving from the current state, $X_n = x$, to X_{n+1}.

Iteration n of the DA Algorithm:

1. Draw $Y \sim f_{Y|X}(\cdot|x)$, and call the observed value y
2. Draw $X_{n+1} \sim f_{X|Y}(\cdot|y)$

It is well known and easy to establish that this Markov chain, which we call the DA chain, is reversible with respect to f_X, and this of course implies that f_X is an invariant density. Consequently, if the chain is well-behaved (e.g. Harris ergodic), then we can use empirical ergodic averages based on simulation to estimate intractable expectations with respect to f_X (Tierney, 1994). The resulting Markov chain Monte Carlo (MCMC) algorithm is known as a DA algorithm for f_X.

When designing a DA algorithm, one is free to choose any joint density that satisfies conditions (i) and (ii). Obviously, different joint densities will yield different DA chains, and the goal is to find a joint whose DA chain has good convergence properties. Unfortunately, the "ideal" joint density, which yields the DA chain with the fastest possible rate of convergence, does not satisfy the simulation requirement. Consider $f_\perp(x, y) = f_X(x)\, g_Y(y)$, where $g_Y(y)$ is any density (or mass) function on $\mathbb{R}^q$. Since $f_\perp(x, y)$ factors, $f_{X|Y}(x|y) = f_X(x)$ and the DA chain is just an iid sequence from the target density. Of course, this ideal DA algorithm is useless from a practical standpoint because, in order to simulate the chain, we must draw from f_X, which is impossible. We return to this example later in this section.

It is important to keep in mind that there is no inherent interest in the joint density $f(x, y)$. It is merely a construct that facilitates exploration of the target density, $f_X(x)$. This is the reason why the DA chain does not possess a y-coordinate. In contrast, the two-variable Gibbs sampler based on $f_{X|Y}(\cdot|y)$ and $f_{Y|X}(\cdot|x)$, which is used to explore $f(x, y)$, has both x and y-coordinates. So, while the two-step procedure described above can be used to simulate both the DA and Gibbs chains, there is one key difference. When simulating the DA chain, we do not keep track of the y-coordinate.

Every reversible Markov chain defines an operator whose spectrum encodes the convergence properties of the chain (Diaconis, Khare and Saloff-Coste, 2008; Mira and Geyer, 1999; Rosenthal, 2003). Let $X \sim f_X$ and consider the space of functions g such that the random variable $g(X)$ has finite variance and mean zero. More precisely, define

$$L_0^2(f_X) = \left\{ g : \mathbb{R}^p \to \mathbb{R} : \int_{\mathbb{R}^p} g^2(x)\, f_X(x)\, dx < \infty \text{ and } \int_{\mathbb{R}^p} g(x)\, f_X(x)\, dx = 0 \right\} .$$

Let $k(x'|x)$ be the Markov transition density (Mtd) of the DA chain. This Mtd defines an operator, $K : L_0^2 \to L_0^2$, that maps $g(x)$ to

$$(Kg)(x) := \int_{\mathbb{R}^p} g(x')\, k(x'|x)\, dx' .$$

Of course, $(Kg)(x)$ is just the expected value of $g(X_1)$ given that $X_0 = x$. Let $I : L_0^2 \to L_0^2$ denote the identity operator, which leaves functions unaltered, and consider the operator $K - \lambda I$, where $\lambda \in \mathbb{R}$. By definition, $K - \lambda I$ is invertible if, for each $h \in L_0^2$, there exists a unique $g \in L_0^2$ such that $((K - \lambda I)g)(x) = (Kg)(x) - \lambda g(x) = h(x)$. The spectrum of K, which we denote by $\mathrm{Sp}(K)$, is simply the set of λ such that $K - \lambda I$ is *not* invertible. The fact that K is defined through a DA chain implies that $\mathrm{Sp}(K) \subset [0, 1]$ (see Section 2), and the number of elements in $\mathrm{Sp}(K)$ may be finite, countably infinite or uncountable.

In order to understand what "good" spectra look like, consider the ideal DA algorithm introduced earlier. Let $k_\perp$ and $K_\perp$ denote the Mtd and the corresponding operator, respectively. In the ideal case, X_{n+1} is independent of X_n and has density f_X. Therefore, the Mtd is just $k_\perp(x'|x) = f_X(x')$ and

$$(K_\perp g)(x) = \int_{\mathbb{R}^p} g(x')\, f_X(x')\, dx' = 0 ,$$

which implies that

$$((K_\perp - \lambda I)g)(x) = -\lambda g(x) .$$

It follows that $K_\perp - \lambda I$ is invertible as long as $\lambda \neq 0$. Hence, the "ideal spectrum" is $\mathrm{Sp}(K_\perp) = \{0\}$. Loosely speaking, the closer $\mathrm{Sp}(K)$ is to $\{0\}$, the faster the DA algorithm converges (Diaconis et al., 2008).

In general, the spectrum of a Markov chain that lives on a finite space with d points consists exactly of the $d-1$ smallest eigenvalues of the corresponding $d \times d$ Markov transition matrix (Mtm). Hence, as long as d is not too large, and the Mtm is available in closed form, the spectrum can actually be computed. Unfortunately, if the chain lives on a state space that is not finite, then the spectrum can be quite complex, and may have an uncountable number of elements. Moreover, even if the Mtd that defines the chain is available in closed form, there is no simple method for calculating the spectrum. (Of course, the ideal DA algorithm is an exception.)

The state space of the DA chain is given by $\mathsf{X} = \left\{x \in \mathbb{R}^p : f_X(x) > 0\right\}$, which is uncountable. It follows that $\mathrm{Sp}(K)$ is potentially complex, and cannot be calculated directly. One of the main results in this paper states that, no matter what the space X looks like, if the augmented space, $\mathsf{Y} = \left\{y \in \mathbb{R}^q : f_Y(y) > 0\right\}$, is finite, then $\mathrm{Sp}(K)$ has a finite number of elements. Moreover, these elements are directly related to the Mtm of the so-called conjugate Markov chain that lives on Y and makes the transition $y \to y'$ with probability $\int_{\mathsf{X}} f_{Y|X}(y'|x)\, f_{X|Y}(x|y)\, dx$. In particular, we show that, if $|\mathsf{X}| = \infty$ and $|\mathsf{Y}| = d < \infty$, then $\mathrm{Sp}(K)$ consists of the point $\{0\}$ together with the $d-1$ smallest eigenvalues of the Mtm associated with the conjugate chain. We use this result to prove that the spectrum associated with a particular alternative to the DA chain is closer than $\mathrm{Sp}(K)$ to the ideal spectrum, $\{0\}$.

DA algorithms often suffer from slow convergence, which is not surprising given the close connection between DA and the notoriously slow to converge EM algorithm (Dempster, Laird and Rubin, 1977). Over the last decade, a great deal of effort has gone into modifying the DA algorithm to speed convergence. See, for example, Meng and van Dyk (1999), Liu and Wu (1999), Liu and Sabatti (2000), van Dyk and Meng (2001), Papaspiliopoulos, Roberts and Sköld (2007), Hobert and Marchev (2008) and Yu and Meng (2009). Here we focus on a simple, yet powerful technique for improving the DA algorithm that was introduced and studied by Hobert and Marchev (2008). Let $r(y'|y)$ be an auxiliary Markov transition density (or mass function) that is reversible with respect to f_Y. The method is based on the improved DA (IDA) chain, $\{\tilde{X}_n\}_{n=0}^{\infty}$, which moves from $\tilde{X}_n = x$ to $\tilde{X}_{n+1}$ via the following *three-step* procedure.

Iteration n of the IDA Algorithm:

1. Draw $Y \sim f_{Y|X}(\cdot|x)$, and call the observed value y
2. Draw $Y' \sim r(\cdot|y)$, and call the observed value y'
3. Draw $\tilde{X}_{n+1} \sim f_{X|Y}(\cdot|y')$

A routine calculation shows that the IDA chain remains reversible with respect to f_X, so it is a viable alternative to the DA chain. Note that steps 1 and 3 are the same as the two steps in the DA algorithm. Hence, on a per iteration basis, it is more expensive to simulate the IDA chain. On the other hand, the extra step should speed convergence by reducing the correlation between X_n and X_{n+1}. In fact, experience has shown that it is often possible to find an $r(\cdot|\cdot)$ such that (empirical convergence measures suggest that) the IDA chain converges *much* faster than the DA chain, and, at the same time, the extra effort required to draw from r is insignificant. Concrete examples can be found in Meng and van Dyk (1999), Liu and Wu (1999), van Dyk and Meng (2001) and Roy and Hobert (2007). Let $\text{Sp}(\tilde{K})$ denote the spectrum of the operator defined by the IDA chain. We prove a theoretical result showing that (under regularity conditions) $\text{Sp}(\tilde{K})$ dominates $\text{Sp}(K)$.

We know from the result described above that when $|\mathsf{X}| = \infty$ and $|\mathsf{Y}| = d < \infty$, $\text{Sp}(K)$ contains the point $\{0\}$ along with the $d-1$ smallest eigenvalues of the Mtm of the conjugate chain. We prove that if $|\mathsf{X}| = \infty$, $|\mathsf{Y}| = d < \infty$, and r is idempotent (see Section 3 for the definition), then $\text{Sp}(\tilde{K})$ contains the point $\{0\}$ along with the $d-1$ smallest eigenvalues of a different $d \times d$ Mtm, and $0 \leq \tilde{\lambda}_i \leq \lambda_i$ for all $i \in \{1, 2, \ldots, d-1\}$ where $\tilde{\lambda}_i$ and λ_i are the ith largest elements of $\text{Sp}(\tilde{K})$ and $\text{Sp}(K)$, respectively. So the ordered elements of $\text{Sp}(\tilde{K})$ are uniformly less than or equal to the corresponding elements of $\text{Sp}(K)$. One might hope for a stronger result that quantifies the extent to which the IDA chain is better than the DA chain, but such a result is impossible without further assumptions. Indeed, note that if we take the auxiliary Markov chain on Y to be the degenerate chain that is absorbed at its starting point, then the IDA chain is the same as the DA chain.

To illustrate the huge gains that are possible by using IDA instead of DA, we introduce a new example involving a Bayesian mixture model. Let $Z_1, \ldots, Z_m$ be a random sample from a k-

component mixture density taking the form

$$\sum_{j=1}^{k} p_j h_{\theta_j}(z) \ , \tag{1}$$

where $\theta_1, \ldots, \theta_k \in \Theta \subset \mathbb{R}^l$, $\{h_\theta(\cdot) : \theta \in \Theta\}$ is a parametric family of densities, and the p_js are nonnegative weights that sum to one. Of course, a Bayesian analysis requires priors for the unknown parameters, which are $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T$ and $\boldsymbol{p} = (p_1, \ldots, p_k)^T$. In typical applications we have no prior information on $\boldsymbol{p}$, and the same (lack of) prior information about each of the components in the mixture. Thus, it makes sense to put a symmetric Dirichlet prior on the weights, and to take a prior on $\boldsymbol{\theta}$ that has the form $\prod_{j=1}^{k} \pi(\theta_j)$, where $\pi : \Theta \to [0, \infty)$ is a proper prior density on Θ. Let $\boldsymbol{z} = (z_1, \ldots, z_m)$ denote the observed data. It is well known that the resulting posterior, $\pi(\boldsymbol{\theta}, \boldsymbol{p}|\boldsymbol{z})$, is intractable and highly multi-modal (see, for example, Jasra, Holmes and Stephens, 2005). Indeed, let E denote any one of the $k!$ permutation matrices of dimension k and note that $\pi(\boldsymbol{\theta}, \boldsymbol{p}|\boldsymbol{z}) = \pi(E\boldsymbol{\theta}, E\boldsymbol{p}|\boldsymbol{z})$. Thus, every local maximum of the posterior has $k! - 1$ exact replicas somewhere else in the parameter space.

The standard DA algorithm for this mixture problem was introduced by Diebolt and Robert (1994) and is based on the following augmented model. Assume that $\{(Y_i, Z_i)\}_{i=1}^m$ are iid pairs such that $Y_i = j$ with probability p_j, and, conditional on $Y_i = j$, $Z_i \sim h_{\theta_j}(\cdot)$. Note that the marginal density of Z_i under this two level hierarchy is just (1). Let $\boldsymbol{y} = (y_1, \ldots, y_m)$ denote a realization of the Y_is. The so-called complete posterior density, $\pi((\boldsymbol{\theta}, \boldsymbol{p}), \boldsymbol{y}|\boldsymbol{z})$, is just the posterior that results when we combine our model for $\{(Y_i, Z_i)\}_{i=1}^m$ with the priors on $\boldsymbol{p}$ and $\boldsymbol{\theta}$ defined above. It is easy to see that

$$\sum_{\boldsymbol{y} \in \mathsf{Y}} \pi((\boldsymbol{\theta}, \boldsymbol{p}), \boldsymbol{y}|\boldsymbol{z}) = \pi(\boldsymbol{\theta}, \boldsymbol{p}|\boldsymbol{z}) \ ,$$

where Y is the set of all sequences of length m consisting of integers from the set $\{1, \ldots, k\}$. Hence, $\pi((\boldsymbol{\theta}, \boldsymbol{p}), \boldsymbol{y}|\boldsymbol{z})$ can be used to build a DA algorithm as long as it is possible to sample from the conditionals, $\pi((\boldsymbol{\theta}, \boldsymbol{p})|\boldsymbol{y}, \boldsymbol{z})$ and $\pi(\boldsymbol{y}|(\boldsymbol{\theta}, \boldsymbol{p}), \boldsymbol{z})$. We call it the mixture DA algorithm. Note that, in this example, $|\mathsf{Y}| = k^m < \infty$, while X is the Cartesian product of $\mathbb{R}^{kl}$ and the k-dimensional simplex.

The mixture DA algorithm often converges very slowly because it moves between the symmetric modes of $\pi(\boldsymbol{\theta}, \boldsymbol{p}|\boldsymbol{z})$ too infrequently (Celeux, Hurn and Robert, 2000; Lee, Marin, Mengersen and Robert, 2008). Frühwirth-Schnatter (2001) suggested adding a random label switching step to each

iteration of the DA algorithm in order to force movement between the modes. The first of the two steps that define one iteration of the DA algorithm entails drawing $\boldsymbol{y} \sim \pi(\boldsymbol{y}|(\boldsymbol{\theta}, \boldsymbol{p}), \boldsymbol{z})$. This $\boldsymbol{y}$ represents a *clustering* of the m observations into groups in the sense that all the observations that have the same value of y are from the same component of the mixture. Frühwirth-Schnatter's (2001) idea was to randomly permute the numbers $\{1, \ldots, k\}$ within $\boldsymbol{y}$, which leaves the clustering intact, but changes the mode that is being explored (unless the identity permutation is chosen). We show that the resulting Markov chain, which we call the FS chain, is a special case of the IDA chain. Hence, our theoretical results imply that the spectrum of the operator defined the FS chain dominates the spectrum of the DA operator. Moreover, from a computational standpoint, the two algorithms are essentially equivalent. To illustrate just how much improvement is possible, we study two specific mixture models and compare the spectra associated with the FS and mixture DA chains. The first example is a toy problem in which we are able to get exact formulas for the eigenvalues. The second example is a normal mixture model that is frequently used in practice, and we approximate the eigenvalues via classical Monte Carlo methods. The conclusions from the two examples are quite similar. Firstly, the mixture DA chain converges slowly and the rate deteriorates very rapidly as the sample size, m, increases. Secondly, the FS chain converges much faster and does not seem to be as adversely affected by increasing sample size.

The remainder of this paper is organized as follows. Section 2 contains a review of the operator theory used for analyzing reversible Markov chains, as well as a string of results about $\mathrm{Sp}(K)$ in the case where the augmented space is finite. Section 3 contains a formal description of the IDA chain and a theoretical result comparing the DA and IDA chains in the case where $|\mathsf{Y}| < \infty$. Section 4 contains a review of the standard DA algorithm for exploring Bayesian mixture posteriors and Frühwirth-Schnatter's (2001) alternative. In this same section, the FS chain is shown to be a special case of the IDA chain. Finally, in Section 5, we compare the mixture DA and FS chains in the context of two specific examples. The Appendix contains an eigen-analysis of a 4×4 Mtm that has a special form.

2 The Spectrum of the DA Chain

Consider a generalized version of the problem described in Section 1. Let X be a general space (equipped with a countably generated σ-algebra) and let $f_X : \mathsf{X} \to [0, \infty)$ be an intractable prob-

ability density with respect to the measure μ. Suppose that Y is a second general space and that ν is a measure on Y. Let $f : \mathsf{X} \times \mathsf{Y} \to [0, \infty)$ be a joint probability density with respect to $\mu \times \nu$. Assume that $\int_{\mathsf{Y}} f(x, y)\, \nu(dy) = f_X(x)$ and that simulating from the associated conditional densities, $f_{X|Y}(\cdot|y)$ and $f_{Y|X}(\cdot|x)$, is straightforward. (For convenience, we assume that f_X and f_Y are strictly positive on X and Y, respectively.) The DA chain, $\{X_n\}_{n=0}^{\infty}$, has Mtd (with respect to μ) given by

$$k(x'|x) = \int_{\mathsf{Y}} f_{X|Y}(x'|y)\, f_{Y|X}(y|x)\, \nu(dy)\ . \tag{2}$$

It is easy to see that $k(x'|x)\, f_X(x)$ is symmetric in (x, x'), so the DA chain is reversible with respect to f_X. We assume throughout that it is also Harris ergodic. (See Hobert (2009) for a simple sufficient condition for Harris ergodicity of the DA chain.) If the integral in (2) is intractable, as is nearly always the case in practice, then direct simulation from $k(\cdot|x)$ will be problematic. This is why the indirect two-step procedure is used.

We now give a brief review of the use of linear operator theory for analyzing the convergence of reversible Markov chains, and, in particular, the DA chain. Consider the Hilbert space

$$L_0^2(f_X) = \left\{ g : \mathsf{X} \to \mathbb{R} : \int_{\mathsf{X}} g^2(x) f_X(x) \mu(dx) < \infty \ \text{ and } \ \int_{\mathsf{X}} g(x) f_X(x) \mu(dx) = 0 \right\},$$

where inner product is defined as

$$\langle g, h \rangle = \int_{\mathsf{X}} g(x)\, h(x)\, f_X(x)\, \mu(dx)\ .$$

The corresponding norm is given by $\|g\| = \sqrt{\langle g, g \rangle}$. Let $p(x'|x)$ be a Mtd (with respect to μ) such that the corresponding Markov chain, which we call the generic chain, is Harris ergodic and reversible with respect to $f_X(x)$. (Of course, $k(x'|x)$ is a special case of $p(x'|x)$). This Mtd defines an operator, $P : L_0^2(f_X) \to L_0^2(f_X)$, that acts on $g \in L_0^2(f_X)$ as follows:

$$(Pg)(x) = \int_{\mathsf{X}} g(x')\, p(x'|x)\, \mu(dx')\ .$$

It is easy to show, using reversibility, that for $g, h \in L_0^2(f_X)$, $\langle Pg, h \rangle = \langle g, Ph \rangle$; that is, P is a self-adjoint operator. Let $L_{0,1}^2(f_X)$ denote the subset of functions in $L_0^2(f_X)$ that satisfy $\int_{\mathsf{X}} g^2(x)\, f_X(x)\, \mu(dx) = 1$. The (operator) norm of P is defined as

$$\|P\| = \sup_{g \in L_{0,1}^2(f_X)} \|Pg\|\ .$$

A simple application of Jensen's inequality shows that the nonnegative quantity $\|P\|$ is bounded above by 1. It is well known that $\|P\|$ is closely related to the (asymptotic) rate of convergence of the generic chain to its stationary distribution (Liu, Wong and Kong, 1995; Roberts and Rosenthal, 1997; Rosenthal, 2003). However, as shown in Diaconis et al. (2008), a more complete version of the convergence picture can be gleaned from the spectrum of P, which is defined as

$$\text{Sp}(P) = \Big\{ \lambda \in \mathbb{R} : P - \lambda I \text{ is not invertible} \Big\} .$$

As described in Rudin (1991, Chapter 4) and Mira and Geyer (1999), there are two ways in which $P - \lambda I$ can fail to be invertible. Firstly, $P - \lambda I$ may not be onto; that is, if there exists $h \in L_0^2(f_X)$ such that there is no $g \in L_0^2(f_X)$ for which $((P - \lambda I)g) = h$, then the range of $P - \lambda I$ is not all of $L_0^2(f_X)$, so $P - \lambda I$ is not invertible and $\lambda \in \text{Sp}(P)$. Secondly, $P - \lambda I$ may not be one-to-one; that is, if there exist two different functions $g, h \in L_0^2(f_X)$ such that $((P - \lambda I)g) = ((P - \lambda I)h)$, then $P - \lambda I$ is not one-to-one, so $P - \lambda I$ is not invertible and $\lambda \in \text{Sp}(P)$. Note that, if $((P - \lambda I)g) = ((P - \lambda I)h)$, then $Pg^* = \lambda g^*$ with $g^* = g - h$, and λ is called an eigenvalue with eigen-function g^*. We call the pair (λ, g^*) an eigen-solution.

The quantity $\|P\|$ is a good univariate summary of $\text{Sp}(P)$. Indeed, define

$$u_P = \sup_{g \in L_{0,1}^2(f_X)} \langle Pg, g \rangle \qquad \text{and} \qquad l_P = \inf_{g \in L_{0,1}^2(f_X)} \langle Pg, g \rangle .$$

Standard linear operator theory implies that $\sup \text{Sp}(P) = u_P$, $\inf \text{Sp}(P) = l_P$, and that $\|P\| = \max\big\{ -l_P, u_P \big\}$. Consequently, $\text{Sp}(P) \subset \big[-\|P\|, \|P\| \big] \subset [-1, 1]$. Another name for $\|P\|$ in this context is the *spectral radius*, which makes sense since $\|P\|$ represents the maximum distance that $\text{Sp}(P)$ extends away from the origin. The quantity $1 - \|P\|$ is called the *spectral gap*.

Liu, Wong and Kong (1994) showed that the DA chain satisfies an important additional property that results in a positive spectrum. Let K denote the operator defined by the DA chain. For $g \in L_0^2(f_X)$, we have

$$\begin{aligned}
\langle Kg, g \rangle &= \int_{\mathsf{X}} (Kg)(x)\, g(x)\, f_X(x)\, \mu(dx) \\
&= \int_{\mathsf{X}} \left[\int_{\mathsf{X}} g(x')\, k(x'|x)\, \mu(dx') \right] g(x)\, f_X(x)\, \mu(dx) \\
&= \int_{\mathsf{X}} \left[\int_{\mathsf{X}} g(x') \left[\int_{\mathsf{Y}} f_{X|Y}(x'|y)\, f_{Y|X}(y|x)\, \nu(dy) \right] \mu(dx') \right] g(x)\, f_X(x)\, \mu(dx) \\
&= \int_{\mathsf{Y}} \left[\int_{\mathsf{X}} g(x)\, f_{X|Y}(x|y)\, \mu(dx) \right]^2 f_Y(y)\, \nu(dy) \geq 0 ,
\end{aligned}$$

which shows that K is a positive operator. It follows that $l_K \geq 0$, so $\mathrm{Sp}(K) \subset [0, \|K\|] \subset [0, 1]$ and $\|K\| = \sup \mathrm{Sp}(K)$.

In most applications of the DA algorithm, the state space, X, is uncountable. In these cases, K is potentially difficult to analyze and $\mathrm{Sp}(K)$ may contain an uncountable number of points. One exception is when K is a compact operator (see Retherford (1993) for a definition of compactness). Indeed, if $|\mathsf{X}| = \infty$ and K is compact, then the following all hold: (i) the number of points in $\mathrm{Sp}(K)$ is at most countably infinite, (ii) $\{0\} \in \mathrm{Sp}(K)$, (iii) $\{0\}$ is the only possible accumulation point, and (iv) any point in $\mathrm{Sp}(K)$ other than $\{0\}$ is an eigenvalue. In the remainder of this section we prove that, if the augmented space has d elements, then K is a compact operator and $\mathrm{Sp}(K)$ contains $d - 1$ eigenvalues that are determined by the conjugate chain. Hence, K has a finite spectral decomposition which provides very precise information about the convergence of the DA chain (Diaconis et al., 2008). Indeed, assume that $|\mathsf{X}| = \infty$, $|\mathsf{Y}| = d < \infty$, and let (λ_i, g_i), $i = 1, \ldots, d-1$, denote a set of (orthonormal) eigen-solutions of K. If the chain is started at $X_0 = x$, then the χ^2-distance between the distribution of X_n and the stationary distribution can be expressed as

$$\sum_{i=1}^{d-1} \lambda_i^{2n} g_i^2(x) \; . \tag{3}$$

Of course, the χ^2-distance is an upper bound on the total variation distance (see, for example, Liu et al., 1995). As we demonstrate below, the λ_is are the eigenvalues of the Mtm of the conjugate chain, so there is some hope of calculating, or at least bounding them.

Let $L_0^2(f_Y)$ be the set of mean-zero, square integrable functions with respect to f_Y. In a slight abuse of notation, we will let $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ do double duty as inner product and norm on both $L_0^2(f_X)$ and on $L_0^2(f_Y)$. We now describe a representation of the operator K that was developed and exploited by Diaconis et al. (2008). Define $Q : L_0^2(f_X) \to L_0^2(f_Y)$ and $Q^* : L_0^2(f_Y) \to L_0^2(f_X)$ as follows:

$$(Qg)(y) = \int_{\mathsf{X}} g(x)\, f_{X|Y}(x|y)\, \mu(dx) \qquad \text{and} \qquad (Q^*h)(x) = \int_{\mathsf{Y}} h(y)\, f_{Y|X}(y|x)\, \nu(dy) \; .$$

Note that

$$\begin{aligned}
\langle Qg, h\rangle &= \int_{\mathsf{Y}} (Qg)(y)\, h(y)\, f_Y(y)\, \nu(dy) \\
&= \int_{\mathsf{Y}} \left[\int_{\mathsf{X}} g(x)\, f_{X|Y}(x|y)\, \mu(dx) \right] h(y)\, f_Y(y)\, \nu(dy) \\
&= \int_{\mathsf{X}} g(x) \left[\int_{\mathsf{Y}} h(y)\, f_{Y|X}(y|x)\, \nu(dy) \right] f_X(x) \mu(dx) \\
&= \langle g, Q^* h\rangle \ ,
\end{aligned}$$

which shows that Q^* is the *adjoint* of Q. (Note that we are using the term adjoint in a somewhat non-standard way since $\langle Qg, h\rangle$ is an inner product on $L_0^2(f_Y)$, while $\langle g, Q^* h\rangle$ is an inner product on $L_0^2(f_X)$.) Moreover,

$$\begin{aligned}
(Kg)(x) &= \int_{\mathsf{X}} g(x')\, k(x'|x)\, \mu(dx') \\
&= \int_{\mathsf{X}} g(x') \left[\int_{\mathsf{Y}} f_{X|Y}(x'|y)\, f_{Y|X}(y|x)\, \nu(dy) \right] \mu(dx') \\
&= \int_{\mathsf{Y}} \left[\int_{\mathsf{X}} g(x')\, f_{X|Y}(x'|y)\, \mu(dx') \right] f_{Y|X}(y|x)\, \nu(dy) \\
&= \int_{\mathsf{Y}} (Qg)(y)\, f_{Y|X}(y|x)\, \nu(dy) \\
&= ((Q^*Q)g)(x) \ ,
\end{aligned}$$

which shows that $K = Q^*Q$. As in Section 1, consider the conjugate Markov chain whose Mtd (with respect to ν) is given by

$$\hat{k}(y'|y) = \int_{\mathsf{X}} f_{Y|X}(y'|x)\, f_{X|Y}(x|y)\, \mu(dx) \ . \tag{4}$$

Obviously, $\hat{k}(y'|y)$ is reversible with respect to f_Y. Furthermore, it is easy to see that $\hat{K} = QQ^*$, where $\hat{K} : L_0^2(f_Y) \to L_0^2(f_Y)$ is the operator associated with $\hat{k}$.

Now suppose that (λ, g) is an eigen-solution for K; that is, $(Kg)(x) = \lambda g(x)$, or, equivalently, $((Q^*Q)g)(x) = \lambda g(x)$. Applying the operator Q to both sides yields, $(Q((Q^*Q)g))(y) = \lambda(Qg)(y)$, but we can rewrite this as $(\hat{K}(Qg))(y) = \lambda(Qg)(y)$, which shows that (λ, Qg) is an eigen-solution for $\hat{K}$. Of course, the same argument can be used to convert an eigen-solution for $\hat{K}$ into an eigen-solution for K. We conclude that $\hat{K}$ and K share the same eigenvalues. Here is a precise statement.

Proposition 1. *If (λ, g) is an eigen-solution for K, then $(\lambda, (Qg))$ is an eigen-solution for $\hat{K}$. Conversely, if (λ, h) is an eigen-solution for $\hat{K}$, then $(\lambda, (Q^*h))$ is an eigen-solution for K.*

Remark 1. *Diaconis et al. (2008) describe several examples where the eigen-solutions of K and $\hat{K}$ can be calculated explicitly. These authors studied the case where $f_{X|Y}(x|y)$ is a univariate exponential family (with y playing the role of the parameter), and $f_Y(y)$ is the conjugate prior.*

The next result, which is easily established using minor extensions of results in Retherford's (1993) Chapter VII, shows that compactness is a solidarity property for K and $\hat{K}$.

Proposition 2. *K is compact if and only if $\hat{K}$ is compact.*

Recall that, if $|\mathsf{Y}| = d < \infty$, then $\mathrm{Sp}(\hat{K})$ consists of the (smallest) $d-1$ eigenvalues of the Mtm $\hat{k}(y'|y)$, which are all in $[0, 1)$. Here is the main result of this section, which relates the spectrum of the DA chain to the spectrum of the conjugate chain.

Proposition 3. *Assume that $|\mathsf{X}| = \infty$ and $|\mathsf{Y}| = d < \infty$. Then K is a compact operator and $Sp(K) = \{0\} \cup Sp(\hat{K})$.*

Proof. Since $|\mathsf{Y}| < \infty$, $\hat{K}$ is a compact operator. It follows from Proposition 2 that K is also compact. Hence, $\{0\} \in \mathrm{Sp}(K)$, and aside from $\{0\}$, all the elements of $\mathrm{Sp}(K)$ are eigenvalues of K. But we know from Proposition 1 that K and $\hat{K}$ share the same eigenvalues. □

Remark 2. *Liu et al.'s (1994) Theorem 3.2 states that $\|K\| = \|\hat{K}\|$ (regardless of the cardinalities of X and Y). Proposition 3 can be viewed as a refinement of this result in the case where $|\mathsf{Y}| < \infty$. See also Roberts and Rosenthal (2001).*

In the next section, we use Proposition 3 to prove that the spectrum of the IDA chain dominates the spectrum of the DA chain.

3 Improving the DA algorithm

Suppose that $R(y, dy')$ is a Markov transition kernel on Y that is reversible with respect to $f_Y(y)$. Let $\{\tilde{X}_n\}_{n=0}^{\infty}$ be a Markov chain on X whose Mtd is given by

$$\tilde{k}(x'|x) = \int_{\mathsf{Y}} \int_{\mathsf{Y}} f_{X|Y}(x'|y')\, R(y, dy')\, f_{Y|X}(y|x)\, \nu(dy) \; . \tag{5}$$

As in Section 1, we call it the IDA chain. Again, routine calculations show that the IDA chain remains reversible with respect to the target density f_X. Moreover, if we can draw from $R(y, \cdot)$, then we can draw from $\tilde{k}(\cdot|x)$ in three steps. First, draw $Y \sim f_{Y|X}(\cdot|x)$, call the result y, then draw $Y' \sim$

$R(y,\cdot)$, call the result y', and finally draw $X' \sim f_{X|Y}(\cdot|y')$. Hobert and Marchev (2008) provide general conditions under which the IDA chain outperforms the DA chain in both convergence rate and asymptotic variance. Here, we refine their results in the case where the augmented space is finite.

At first glance, $\tilde{k}$ does not appear to be the Mtd of a DA chain. Indeed, it is not defined as the integral of the product of two conditional densities, as in (2). However, as we now explain, if R satisfies a certain property, called idempotence, then $\{\tilde{X}_n\}_{n=0}^{\infty}$ is, in fact, a DA chain. The transition kernel $R(y,dy')$ is called *idempotent* if $R^2(y,dy') = R(y,dy')$ where $R^2(y,dy') = \int_{\mathsf{Y}} R(y,dw)\,R(w,dy')$. This property implies that, if we start the Markov chain at a fixed point y, then the distribution of the chain after one step *is the same* as the distribution after two steps. For example, if $R(y,dy')$ does not depend on y, which implies that the Markov chain is just an iid sequence, then R is idempotent. Here is a more interesting example. Take $\mathsf{Y} = \mathbb{R}$ and $R(y,dy') = r(y'|y)\,dy'$ with

$$r(y'|y) = e^{-|y'|}\Big[I_{[0,\infty)}(y)I_{[0,\infty)}(y') + I_{(-\infty,0)}(y)I_{(-\infty,0)}(y')\Big]\,.$$

It is easy to show that $\int_{\mathbb{R}} r(y'|w)\,r(w|y)\,dw = r(y'|y)$, so R is indeed idempotent. Note that the chain is reducible since, for example, if it is started on the positive half-line, it can never get to the negative half-line. In fact, reducibility is a common feature of idempotent chains. Fortunately, the IDA chain usually does not inherit this property.

Hobert and Marchev (2008) proved that if R is idempotent, then

$$\tilde{k}(x'|x) = \int_{\mathsf{Y}} f^*_{X|Y}(x'|y)\,f^*_{Y|X}(y|x)\,\nu(dy)\,, \tag{6}$$

where

$$f^*(x,y) = f_Y(y)\int_{\mathsf{Y}} f_{X|Y}(x|y')\,R(y,dy')\,.$$

Note that f^* is a probability density (with respect to $\mu\times\nu$) whose x and y-marginals are f_X and f_Y. What is important here is not the particular form of f^*, but the fact that such a density exists, because this shows that the IDA chain is actually a DA chain based on the joint density $f^*(x,y)$. Therefore, we can use the theory developed in Section 2 to analyze the IDA chain. Let $\tilde{K} : L^2_0(f_X) \to L^2_0(f_X)$ denote the operator defined by the Mtd $\tilde{k}$. Hobert and Marchev's (2008) Corollary 1 states that $\|\tilde{K}\| \le \|K\|$. Here is a refinement of that result.

Theorem 1. *Assume that* $|\mathsf{X}| = \infty$ *and* $|\mathsf{Y}| = d < \infty$. *Assume further that* R *is idempotent and that the Markov chains defined by* k *and* $\tilde{k}$ *are both Harris ergodic. Then* K *and* $\tilde{K}$ *are both compact operators and each has a spectrum that consists exactly of the point* $\{0\}$ *and* $d-1$ *eigenvalues in* $[0,1)$. *Furthermore, if we denote the eigenvalues of* K *by*

$$0 \leq \lambda_{d-1} \leq \lambda_{d-2} \leq \cdots \leq \lambda_1 < 1 \ ,$$

and those of $\tilde{K}$ *by*

$$0 \leq \tilde{\lambda}_{d-1} \leq \tilde{\lambda}_{d-2} \leq \cdots \leq \tilde{\lambda}_1 < 1 \ ,$$

then $\tilde{\lambda}_i \leq \lambda_i$ *for each* $i \in \{1, 2, \ldots, d-1\}$.

Proof. Since R is idempotent, both k and $\tilde{k}$ are DA Markov chains. Moreover, in both cases, the conjugate chain lives on the finite space Y, which has d elements. Therefore, Proposition 3 implies that K and $\tilde{K}$ are both compact and each has a spectrum consisting of the point $\{0\}$ and $d-1$ eigenvalues in $[0,1)$. Now, Corollary 1 of Hobert and Marchev (2008) implies that $K - \tilde{K}$ is a positive operator. Thus, for any $g \in L_0^2(f_X)$,

$$\frac{\langle \tilde{K}g, g\rangle}{\langle g, g\rangle} \leq \frac{\langle Kg, g\rangle}{\langle g, g\rangle} \ .$$

The eigenvalue ordering now follows from an extension of the argument used to prove Mira and Geyer's (1999) Theorem 3.3. Indeed, the Courant-Fischer-Weyl minmax characterization of eigenvalues of compact, self-adjoint operators (see, e.g., Voss, 2003) yields

$$\tilde{\lambda}_i = \min_{\dim(V)=i-1} \max_{g \in V^\perp, g \neq 0} \frac{\langle \tilde{K}g, g\rangle}{\langle g, g\rangle} \leq \min_{\dim(V)=i-1} \max_{g \in V^\perp, g \neq 0} \frac{\langle Kg, g\rangle}{\langle g, g\rangle} = \lambda_i \ .$$

□

Theorem 1 shows that, unless the two spectra are exactly the same, $\mathrm{Sp}(\tilde{K})$ is closer than $\mathrm{Sp}(K)$ to the ideal spectrum, $\{0\}$. In fact, in all of the numerical comparisons of DA and IDA that we have performed, it has always turned out that there is strict inequality between the eigenvalues (except, of course, when they are both zero). When the domination is strict, there exists a positive integer N such that, for all $n \geq N$, the χ^2-distance between $\tilde{X}_n$ and the stationary distribution is smaller than the χ^2-distance between X_n and the stationary distribution. Indeed, let $(\tilde{\lambda}_i, \tilde{g}_i)$, $i = 1, \ldots, d-1$, denote a set of (orthonormal) eigen-solutions of $\tilde{K}$. Then, according to (3), the χ^2-distance between

$\tilde{X}_n$ and the stationary distribution is given by

$$\sum_{i=1}^{d-1} \tilde{\lambda}_i^{2n} \tilde{g}_i^2(x) \ . \tag{7}$$

Now, fix $i \in \{1, \ldots, d-1\}$. If $\tilde{\lambda}_i = \lambda_i = 0$, then the ith term in the sum is irrelevant. On the other hand, if $0 \leq \tilde{\lambda}_i < \lambda_i$, then, no matter what the values of $g_i(x)$ and $\tilde{g}_i(x)$ are, $\tilde{\lambda}_i^{2n} \tilde{g}_i^2(x)$ will be less than $\lambda_i^{2n} g_i^2(x)$ for all n eventually.

In the next section, we provide examples where the IDA chain converges *much* faster than the DA chain, and the two are essentially equivalent in terms of computer time per iteration.

4 Improving the DA Algorithm for Bayesian Mixtures

4.1 The model and the DA algorithm

Let $\Theta \subset \mathbb{R}^l$ and consider a parametric family of densities (with respect to Lebesgue or counting measure on $\mathbb{R}^s$) given by $\big\{h_\theta(\cdot) : \theta \in \Theta\big\}$. We work with a k-component mixture of these densities that takes the form

$$f(z|\boldsymbol{\theta}, \boldsymbol{p}) = \sum_{j=1}^{k} p_j h_{\theta_j}(z) \ , \tag{8}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T \in \Theta^k$ and $\boldsymbol{p} = (p_1, \ldots, p_k)^T \in S_k$, where

$$S_k := \Big\{ \boldsymbol{p} \in \mathbb{R}^k : p_i \in [0,1] \ \text{ and } \ p_1 + \cdots + p_k = 1 \Big\} \ .$$

Let $Z_1, \ldots, Z_m$ be a random sample from f and consider a Bayesian analysis of these data. We take the prior for $\boldsymbol{\theta}$ to be $\prod_{j=1}^k \pi(\theta_j)$, where $\pi : \Theta \to [0, \infty)$ is a proper prior density on Θ. The prior on $\boldsymbol{p}$ is taken to be the uniform distribution on S_k, which is, of course, a special case of the Dirichlet distribution. (The results that we describe in this section all go through with obvious minor changes if $\boldsymbol{p}$ is known and all of its components are equal to $1/k$.) Letting $\boldsymbol{z} = (z_1, \ldots, z_m)$ denote the observed data, the posterior density is given by

$$\pi(\boldsymbol{\theta}, \boldsymbol{p}|\boldsymbol{z}) = \frac{(k-1)!\, I_{S_k}(\boldsymbol{p}) \Big[\prod_{j=1}^k \pi(\theta_j) \Big] f(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{p})}{m(\boldsymbol{z})} \ , \tag{9}$$

where

$$f(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{p}) = \prod_{i=1}^{m} \left[\sum_{j=1}^{k} p_j h_{\theta_j}(z_i) \right] \ ,$$

and $m(\boldsymbol{z})$ denotes the marginal density. The complexity of this posterior density obviously depends on many factors, including the choices of h_θ and π and the actual data that is observed. However, the versions of $\pi(\boldsymbol{\theta}, \boldsymbol{p}|\boldsymbol{z})$ that arise in practice are nearly always highly intractable. Moreover, as we now explain, every version of this posterior density satisfies an interesting symmetry property, which can render MCMC algorithms ineffectual.

The prior distribution on $(\boldsymbol{\theta}, \boldsymbol{p})$ is exchangeable in the sense that, if E is any permutation matrix of dimension k, then the prior density of the point $(\boldsymbol{\theta}, \boldsymbol{p})$ is equal to that of $(E\boldsymbol{\theta}, E\boldsymbol{p})$. Furthermore, the likelihood function satisfies a similar invariance. Indeed, $f(\boldsymbol{z}|E\boldsymbol{\theta}, E\boldsymbol{p})$ does not vary with E. Consequently, $\pi(E\boldsymbol{\theta}, E\boldsymbol{p}|\boldsymbol{z})$ is invariant to E, which means that any posterior mode has $k! - 1$ exact replicas somewhere else in the space. Now, if a set of symmetric modes are separated by areas of very low (posterior) probability, then it may take a very long time for a Markov chain (with invariant density $\pi(\boldsymbol{\theta}, \boldsymbol{p}|\boldsymbol{z})$) to move from one to the other.

We now describe the standard DA algorithm for exploring the mixture posterior. Despite the fact that this mixture DA algorithm has been around for many years (Diebolt and Robert, 1994), we provide a careful description here as this will facilitate our development and analysis of an IDA algorithm based on a random label switching step. Consider a new (joint) density given by

$$f(z, y|\boldsymbol{\theta}, \boldsymbol{p}) = \sum_{j=1}^{k} p_j I_{\{j\}}(y) h_{\theta_j}(z) \ . \tag{10}$$

Integrating z out yields the marginal mass function of Y, which is $\sum_{j=1}^{k} p_j I_{\{j\}}(y)$. Hence, Y is a multinomial random variable that takes the values $1, \ldots, k$ with probabilities $p_1, \ldots, p_k$. Summing out the y component leads to

$$\sum_{y=1}^{k} f(z, y|\boldsymbol{\theta}, \boldsymbol{p}) = \sum_{j=1}^{k} p_j h_{\theta_j}(z) \ , \tag{11}$$

which is just (8). Equation (11) establishes Y as a latent variable. Now suppose that $\{(Y_i, Z_i)\}_{i=1}^{m}$ are iid pairs from (10). Their joint density is given by

$$f(\boldsymbol{z}, \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{p}) = \prod_{i=1}^{m} \left[\sum_{j=1}^{k} p_j I_{\{j\}}(y_i) h_{\theta_j}(z_i) \right] ,$$

where $\boldsymbol{y} = (y_1, \ldots, y_m)$ takes values in Y, the set of sequences of length m consisting of positive integers between 1 and k. Combining $f(\boldsymbol{z}, \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{p})$ with our prior on $(\boldsymbol{\theta}, \boldsymbol{p})$ yields the so-called

complete data posterior density given by

$$\pi(\boldsymbol{\theta}, \boldsymbol{p}, \boldsymbol{y}|\boldsymbol{z}) = \frac{(k-1)! I_{S_k}(\boldsymbol{p}) \Big[\prod_{j=1}^{k} \pi(\theta_j) \Big] f(\boldsymbol{z}, \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{p})}{m(\boldsymbol{z})} \ . \tag{12}$$

This is a valid density since, by (11),

$$\sum_{\boldsymbol{y} \in \mathsf{Y}} f(\boldsymbol{z}, \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{p}) = f(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{p}) \ ,$$

which in turn implies that

$$\sum_{\boldsymbol{y} \in \mathsf{Y}} \pi(\boldsymbol{\theta}, \boldsymbol{p}, \boldsymbol{y}|\boldsymbol{z}) = \pi(\boldsymbol{\theta}, \boldsymbol{p}|\boldsymbol{z}) \ . \tag{13}$$

In fact, (13) is the key property of the complete data posterior density. In words, when the $\boldsymbol{y}$ coordinate is summed out of $\pi(\boldsymbol{\theta}, \boldsymbol{p}, \boldsymbol{y}|\boldsymbol{z})$, we are left with the target density. Hence, we will have a viable (mixture) DA algorithm as long as straightforward sampling from $\pi(\boldsymbol{\theta}, \boldsymbol{p}|\boldsymbol{y}, \boldsymbol{z})$ and $\pi(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{p}, \boldsymbol{z})$ is possible. Note that the roles of x and y from Sections 1, 2 and 3 are being played here by $(\boldsymbol{\theta}, \boldsymbol{p})$ and $\boldsymbol{y}$, respectively.

Now consider sampling from the two conditionals. First, it follows from (12) that

$$\pi(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{p}, \boldsymbol{z}) = \prod_{i=1}^{m} \left[\frac{\sum_{j=1}^{k} p_j I_{\{j\}}(y_i) h_{\theta_j}(z_i)}{\sum_{l=1}^{k} p_l h_{\theta_l}(z_i)} \right] \ . \tag{14}$$

Therefore, conditional on $(\boldsymbol{\theta}, \boldsymbol{p}, \boldsymbol{z})$, the Y_is are independent multinomial random variables and Y_i takes the value j with probability $p_j h_{\theta_j}(z_i) / \big(\sum_{l=1}^{k} p_l h_{\theta_l}(z_i) \big)$ for $j \in \{1, \ldots, k\}$. Consequently, simulating from $\pi(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{p}, \boldsymbol{z})$ is simple.

A two-step method is used to sample from $\pi(\boldsymbol{\theta}, \boldsymbol{p}|\boldsymbol{y}, \boldsymbol{z})$. Indeed, we draw from $\pi(\boldsymbol{p}|\boldsymbol{y}, \boldsymbol{z})$ and then from $\pi(\boldsymbol{\theta}|\boldsymbol{p}, \boldsymbol{y}, \boldsymbol{z})$. It follows from (12) that

$$\pi(\boldsymbol{p}|\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{z}) \propto I_{S_k}(\boldsymbol{p}) \prod_{j=1}^{k} p_j^{c_j} \ ,$$

where $c_j = \sum_{i=1}^{m} I_{\{j\}}(y_i)$. This formula reveals two facts: (i) given $(\boldsymbol{z}, \boldsymbol{y})$, $\boldsymbol{p}$ is conditionally independent of $\boldsymbol{\theta}$, and (ii) the conditional distribution of $\boldsymbol{p}$ given $(\boldsymbol{z}, \boldsymbol{y})$ is Dirichlet. Thus, it is easy to draw from $\pi(\boldsymbol{p}|\boldsymbol{y}, \boldsymbol{z})$, and our sequential strategy will be viable as long as we can draw from $\pi(\boldsymbol{\theta}|\boldsymbol{p}, \boldsymbol{y}, \boldsymbol{z})$. Our ability to sample from $\pi(\boldsymbol{\theta}|\boldsymbol{p}, \boldsymbol{y}, \boldsymbol{z})$ will depend on the particular forms of h_θ and the prior π. In cases where π is a conjugate prior for the family h_θ, it is usually straightforward to draw from $\pi(\boldsymbol{\theta}|\boldsymbol{p}, \boldsymbol{y}, \boldsymbol{z})$. For several detailed examples, see Chapter 9 of Robert and Casella (2004).

The state space of the mixture DA chain is $\mathsf{X} = \Theta^k \times S_k$ and its Mtd is given by

$$k(\boldsymbol{\theta}', \boldsymbol{p}'|\boldsymbol{\theta}, \boldsymbol{p}) = \sum_{\boldsymbol{y} \in \mathsf{Y}} \pi(\boldsymbol{\theta}', \boldsymbol{p}'|\boldsymbol{y}, \boldsymbol{z})\, \pi(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{p}, \boldsymbol{z}) \ .$$

The augmented space, Y, contains $d = k^m$ elements. Hence, Proposition 3 implies that the operator $K : L_0^2\big(\pi(\boldsymbol{\theta}, \boldsymbol{p}|\boldsymbol{z})\big) \to L_0^2\big(\pi(\boldsymbol{\theta}, \boldsymbol{p}|\boldsymbol{z})\big)$ defined by $k(\boldsymbol{\theta}', \boldsymbol{p}'|\boldsymbol{\theta}, \boldsymbol{p})$ is compact and

$$\mathrm{Sp}(K) = \{0, \lambda_{d-1}, \lambda_{d-2}, \ldots, \lambda_1\} \ ,$$

where $0 \leq \lambda_{d-1} \leq \lambda_{d-2} \leq \cdots \leq \lambda_1 < 1$, and the λ_is are the eigenvalues of the $k^m \times k^m$ Mtm $\hat{k} : \mathsf{Y} \times \mathsf{Y} \to [0, 1]$ whose elements are given by

$$\hat{k}(\boldsymbol{y}'|\boldsymbol{y}) = \int_{\Theta^k} \int_{S_k} \pi(\boldsymbol{y}'|\boldsymbol{\theta}, \boldsymbol{p}, \boldsymbol{z})\, \pi(\boldsymbol{\theta}, \boldsymbol{p}|\boldsymbol{y}, \boldsymbol{z})\, d\boldsymbol{p}\, d\boldsymbol{\theta} \ .$$

As far as we know, there are no theoretical results available concerning the magnitude of the λ_is. On the other hand, as mentioned in Section 1, there is a great deal of empirical evidence suggesting that the mixture DA chain convergences very slowly. This is because the mixture DA chain moves between the symmetric modes of the posterior too infrequently. In the next section, we describe an alternative chain that moves easily among the modes.

4.2 An IDA chain based on random label switching

One iteration of the mixture DA chain can be represented graphically as $(\boldsymbol{\theta}, \boldsymbol{p}) \to \boldsymbol{y} \to (\boldsymbol{\theta}', \boldsymbol{p}')$. To encourage transitions between the symmetric modes of the posterior, Frühwirth-Schnatter (2001) suggested adding an extra step to get $(\boldsymbol{\theta}, \boldsymbol{p}) \to \boldsymbol{y} \to \boldsymbol{y}' \to (\boldsymbol{\theta}', \boldsymbol{p}')$, where the transition $\boldsymbol{y} \to \boldsymbol{y}'$ is a random label switching move that proceeds as follows. Randomly choose one of the $k!$ permutations of the integers $1, \ldots, k$, and then switch the labels in $\boldsymbol{y}$ according to the chosen permutation to get $\boldsymbol{y}'$. For example, suppose that $m = 8$, $k = 4$, $\boldsymbol{y} = (3, 3, 4, 1, 3, 3, 4, 3)$, and that the chosen permutation is (1324). Then we move from $\boldsymbol{y}$ to $\boldsymbol{y}' = (2, 2, 1, 3, 2, 2, 1, 2)$. Using both theory and examples, we will demonstrate that Frühwirth-Schnatter's (2001) Markov chain, which we call the FS chain, explores $\pi(\boldsymbol{\theta}, \boldsymbol{p}|\boldsymbol{z})$ much more effectively than the mixture DA chain.

To establish that the results in Section 3 can be used to compare the FS and mixture DA chains, we must show that the FS chain is an IDA chain with an idempotent r. That is, we must demonstrate that the Mtd of the FS chain can be expressed in the form

$$\tilde{k}(\boldsymbol{\theta}', \boldsymbol{p}'|\boldsymbol{\theta}, \boldsymbol{p}) = \sum_{\boldsymbol{y} \in \mathsf{Y}} \sum_{\boldsymbol{y}' \in \mathsf{Y}} \pi(\boldsymbol{\theta}', \boldsymbol{p}'|\boldsymbol{y}', \boldsymbol{z})\, r(\boldsymbol{y}'|\boldsymbol{y})\, \pi(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{p}, \boldsymbol{z}) \ , \tag{15}$$

where $r(\boldsymbol{y}'|\boldsymbol{y})$ is a Mtm (on Y) that is both reversible with respect to

$$\pi(\boldsymbol{y}|\boldsymbol{z}) = \int_{S_k} \int_{\Theta^k} \pi(\boldsymbol{\theta}, \boldsymbol{p}, \boldsymbol{y}|\boldsymbol{z})\, d\boldsymbol{\theta}\, d\boldsymbol{p} \ ,$$

and idempotent. We begin by developing a formula for $r(\boldsymbol{y}'|\boldsymbol{y})$. Let $\mathfrak{S}_k$ denote the set (group) of permutations of the integers $1, \ldots, k$. For $\sigma \in \mathfrak{S}_k$, let $\sigma\boldsymbol{y}$ represent the permuted version of $\boldsymbol{y}$. For example, if $\boldsymbol{y} = (3, 3, 4, 1, 3, 3, 4, 3)$ and $\sigma = (1324)$, then $\sigma\boldsymbol{y} = (2, 2, 1, 3, 2, 2, 1, 2)$. The label switching move, $\boldsymbol{y} \to \boldsymbol{y}'$, in the FS algorithm can now be represented as follows. Choose σ uniformly at random from $\mathfrak{S}_k$ and move from $\boldsymbol{y}$ to $\boldsymbol{y}' = \sigma\boldsymbol{y}$. Define the *orbit* of $\boldsymbol{y} \in \mathsf{Y}$ as

$$O_{\boldsymbol{y}} = \left\{ \boldsymbol{y}' \in \mathsf{Y} : \boldsymbol{y}' = \sigma\boldsymbol{y} \ \text{ for some } \sigma \in \mathfrak{S}_k \right\} \ .$$

The set $O_{\boldsymbol{y}}$ simply contains all the points in Y that represent a particular clustering (or partitioning) of the m observations. For example, the point $\boldsymbol{y} = (3, 3, 4, 1, 3, 3, 4, 3)$ represents the clustering of the $m = 8$ observations into the three sets: $\{1, 2, 5, 6, 8\}$, $\{3, 7\}$, $\{4\}$. And, for any $\sigma \in \mathfrak{S}_k$, $\sigma\boldsymbol{y}$ represents that same clustering because all we're doing is changing the labels.

We now show that, if $\boldsymbol{y}$ is fixed and σ is chosen uniformly at random from $\mathfrak{S}_k$, then the random element $\sigma\boldsymbol{y}$ has a uniform distribution on $O_{\boldsymbol{y}}$. Indeed, suppose that $\boldsymbol{y}$ contains u distinct elements, so $u \in \{1, 2, \ldots, k\}$. Then, for any fixed $\boldsymbol{y}' \in O_{\boldsymbol{y}}$, exactly $(k-u)!$ of the $k!$ elements in $\mathfrak{S}_k$ satisfy $\sigma\boldsymbol{y} = \boldsymbol{y}'$. Thus, the probability that $\sigma\boldsymbol{y}$ equals $\boldsymbol{y}'$ is given by $(k-u)!/k!$, which does not depend on $\boldsymbol{y}'$. Hence, the distribution is uniform. (Note that this argument implies that $|O_{\boldsymbol{y}}| = k!/(k-u)!$, which can also be shown directly.) Therefore, we can write the Mtm r as follows:

$$r(\boldsymbol{y}'|\boldsymbol{y}) = \frac{1}{|O_{\boldsymbol{y}}|} I_{\{O_{\boldsymbol{y}}\}}(\boldsymbol{y}') \ .$$

Since the chain driven by r cannot escape from the orbit (clustering) in which it is started, it is reducible. (Recall from Section 3 that reducibility is a common characteristic of idempotent Markov chains.)

A key observation that will allow us to establish the reversibility of r is that $\pi(\boldsymbol{y}|\boldsymbol{z}) = \pi(\sigma\boldsymbol{y}|\boldsymbol{z})$ for all $\boldsymbol{y} \in \mathsf{Y}$ and all $\sigma \in \mathfrak{S}_k$. Indeed,

$$\pi(\boldsymbol{y}|\boldsymbol{z}) = \frac{(k-1)!}{m(\boldsymbol{z})} \int_{\Theta^k} \Big[\pi(\theta_1) \cdots \pi(\theta_k)\Big] \left\{ \int_{S_k} \prod_{i=1}^{m} \left[\sum_{j=1}^{k} p_j I_{\{j\}}(y_i) h_{\theta_j}(z_i) \right] d\boldsymbol{p} \right\} d\boldsymbol{\theta} \ .$$

Let $\sigma\boldsymbol{y} = \boldsymbol{y}' = (y'_1, \ldots, y'_m)$. Now, since $y'_i = \sigma(j) \Leftrightarrow y_i = j$, we have

$$\sum_{j=1}^{k} p_j I_{\{j\}}(y'_i) h_{\theta_j}(z_i) = \sum_{j=1}^{k} p_{\sigma(j)} I_{\{j\}}(y_i) h_{\theta_{\sigma(j)}}(z_i) \ .$$

Hence,

$$\pi(\sigma \boldsymbol{y}|\boldsymbol{z}) = \frac{(k-1)!}{m(\boldsymbol{z})} \int_{\Theta^k} \Big[\pi(\theta_1)\cdots\pi(\theta_k)\Big] \Bigg\{ \int_{S_k} \prod_{i=1}^{m} \Bigg[\sum_{j=1}^{k} p_{\sigma(j)} I_{\{j\}}(y_i) h_{\theta_{\sigma(j)}}(z_i) \Bigg] d\boldsymbol{p} \Bigg\} d\boldsymbol{\theta} \ .$$

The fact that $\pi(\boldsymbol{y}|\boldsymbol{z}) = \pi(\sigma\boldsymbol{y}|\boldsymbol{z})$ can now be established through a couple of simple arguments based on symmetry.

We now demonstrate that the Mtm r satisfies detailed balance with respect to $\pi(\boldsymbol{y}|\boldsymbol{z})$; that is, we will show that, for any $\boldsymbol{y}, \boldsymbol{y}' \in \mathsf{Y}$, $r(\boldsymbol{y}'|\boldsymbol{y})\,\pi(\boldsymbol{y}|\boldsymbol{z}) = r(\boldsymbol{y}|\boldsymbol{y}')\,\pi(\boldsymbol{y}'|\boldsymbol{z})$. First, a little thought reveals that, for any two elements $\boldsymbol{y}$ and $\boldsymbol{y}'$, only one of two things can happen: either $O_{\boldsymbol{y}} = O_{\boldsymbol{y}'}$ or $O_{\boldsymbol{y}} \cap O_{\boldsymbol{y}'} = \emptyset$. If $O_{\boldsymbol{y}} \cap O_{\boldsymbol{y}'} = \emptyset$, then $I_{\{O_{\boldsymbol{y}}\}}(\boldsymbol{y}') = I_{\{O_{\boldsymbol{y}'}\}}(\boldsymbol{y}) = 0$, so $r(\boldsymbol{y}'|\boldsymbol{y}) = r(\boldsymbol{y}|\boldsymbol{y}') = 0$ and detailed balance is satisfied. On the other hand, if $O_{\boldsymbol{y}} = O_{\boldsymbol{y}'}$, then $I_{\{O_{\boldsymbol{y}}\}}(\boldsymbol{y}') = I_{\{O_{\boldsymbol{y}'}\}}(\boldsymbol{y}) = 1$ and $1/|O_{\boldsymbol{y}}| = 1/|O_{\boldsymbol{y}'}|$, so $r(\boldsymbol{y}'|\boldsymbol{y}) = r(\boldsymbol{y}|\boldsymbol{y}')$, and the common value is strictly positive. But $\boldsymbol{y}' \in O_{\boldsymbol{y}}$ implies that $\boldsymbol{y}' = \sigma\boldsymbol{y}$ for some $\sigma \in \mathfrak{S}_k$. Thus, $\pi(\boldsymbol{y}|\boldsymbol{z}) = \pi(\boldsymbol{y}'|\boldsymbol{z})$, and detailed balance holds.

Finally, it is intuitively clear that r is idempotent since, if we start the chain at $\boldsymbol{y}$, then one step results in a uniformly chosen point from $O_{\boldsymbol{y}}$. Obviously, the state after two steps is still uniformly distributed over $O_{\boldsymbol{y}}$. Here's a formal proof that $r^2(\boldsymbol{y}'|\boldsymbol{y}) = r(\boldsymbol{y}'|\boldsymbol{y})$. For $\boldsymbol{y}, \boldsymbol{y}' \in \mathsf{Y}$, we have

$$\begin{aligned}
r^2(\boldsymbol{y}'|\boldsymbol{y}) &= \sum_{\boldsymbol{w}\in\mathsf{Y}} r(\boldsymbol{y}'|\boldsymbol{w})\, r(\boldsymbol{w}|\boldsymbol{y}) \\
&= \sum_{\boldsymbol{w}\in\mathsf{Y}} \frac{1}{|O_{\boldsymbol{w}}|} I_{\{O_{\boldsymbol{w}}\}}(\boldsymbol{y}') \frac{1}{|O_{\boldsymbol{y}}|} I_{\{O_{\boldsymbol{y}}\}}(\boldsymbol{w}) \\
&= \frac{1}{|O_{\boldsymbol{y}}|} \sum_{\boldsymbol{w}\in O_{\boldsymbol{y}}} \frac{1}{|O_{\boldsymbol{w}}|} I_{\{O_{\boldsymbol{w}}\}}(\boldsymbol{y}') \\
&= \frac{1}{|O_{\boldsymbol{y}}|} I_{\{O_{\boldsymbol{y}}\}}(\boldsymbol{y}') \sum_{\boldsymbol{w}\in O_{\boldsymbol{y}}} \frac{1}{|O_{\boldsymbol{y}}|} \\
&= r(\boldsymbol{y}'|\boldsymbol{y}) \ ,
\end{aligned}$$

where the fourth equality follows from the fact that $\boldsymbol{w} \in O_{\boldsymbol{y}} \Rightarrow O_{\boldsymbol{w}} = O_{\boldsymbol{y}}$.

We have now shown that the Mtd of the FS chain can indeed be written in the form (15) with an appropriate r that is reversible and idempotent. Hence, Theorem 1 is applicable and implies that the operators defined by the two chains are both compact and each has a spectrum consisting of the point $\{0\}$ and $k^m - 1$ eigenvalues in $[0, 1)$. Moreover, $\tilde{\lambda}_i \le \lambda_i$ for each $i \in \{1, 2, \ldots, k^m - 1\}$, where $\{\tilde{\lambda}_i\}_{i=1}^{k^m-1}$ and $\{\lambda_i\}_{i=1}^{k^m-1}$ denote the ordered eigenvalues associated with the FS and mixture DA chains, respectively.

Interestingly, in the special case where $m = 1$, the FS algorithm actually produces an iid sequence from the target distribution. Recall that $\pi(\boldsymbol{y}|\boldsymbol{z}) = \pi(\sigma\boldsymbol{y}|\boldsymbol{z})$ for all $\boldsymbol{y} \in \mathsf{Y}$ and all $\sigma \in \mathfrak{S}_k$. Thus, all the points in $O_{\boldsymbol{y}}$ share the same value of $\pi(\cdot|\boldsymbol{z})$. When $m = 1$, Y contains only k points and they all exist in the same orbit. Thus, $\pi(\boldsymbol{y}|\boldsymbol{z}) = 1/k$ for all $\boldsymbol{y} \in \mathsf{Y}$. Moreover, since there is only one orbit, $r(\boldsymbol{y}'|\boldsymbol{y}) = 1/k$ for all $\boldsymbol{y}' \in \mathsf{Y}$; i.e., the Markov chain corresponding to r is just an iid sequence from the uniform distribution on Y. In other words, the label switching move results in an exact draw from $\pi(\boldsymbol{y}'|\boldsymbol{z})$. Now recall the graphical representation of one iteration of the FS algorithm: $(\boldsymbol{\theta}, \boldsymbol{p}) \to \boldsymbol{y} \to \boldsymbol{y}' \to (\boldsymbol{\theta}', \boldsymbol{p}')$. When $m = 1$, the arguments above imply that, given $(\boldsymbol{\theta}, \boldsymbol{p})$, the density of $(\boldsymbol{y}, \boldsymbol{y}', \boldsymbol{\theta}', \boldsymbol{p}')$ is

$$\pi(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{p}, \boldsymbol{z}) r(\boldsymbol{y}'|\boldsymbol{y}) \pi(\boldsymbol{\theta}', \boldsymbol{p}'|\boldsymbol{y}', \boldsymbol{z}) = \pi(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{p}, \boldsymbol{z}) \pi(\boldsymbol{y}'|\boldsymbol{z}) \pi(\boldsymbol{\theta}', \boldsymbol{p}'|\boldsymbol{y}', \boldsymbol{z}) \ .$$

Thus, conditional on $(\boldsymbol{\theta}, \boldsymbol{p})$, $\boldsymbol{y}$ and $(\boldsymbol{y}', \boldsymbol{\theta}', \boldsymbol{p}')$ are independent, and the latter has density

$$\pi(\boldsymbol{y}'|\boldsymbol{z}) \pi(\boldsymbol{\theta}', \boldsymbol{p}'|\boldsymbol{y}', \boldsymbol{z}) = \pi(\boldsymbol{\theta}', \boldsymbol{p}', \boldsymbol{y}'|\boldsymbol{z}) \ .$$

It follows that, marginally, $(\boldsymbol{\theta}', \boldsymbol{p}') \sim \pi(\boldsymbol{\theta}', \boldsymbol{p}'|\boldsymbol{z})$, so the FS algorithm produces an iid sequence from the target posterior density. When $m = 1$, $|\mathsf{Y}| = k^m - 1 = k - 1$. Thus, while the spectrum associated with the DA chain contains $k - 1$ eigenvalues, at least one of which is strictly positive, the spectrum of the IDA chain is the ideal spectrum, $\{0\}$.

In the next section, we consider two specific mixture models and, for each one, we compare the spectra associated with FS and mixture DA chains. The first example is a toy problem where we are able to get exact formulas for the eigenvalues. The second example is a normal mixture model that is frequently used in practice, and we approximate the eigenvalues via classical Monte Carlo methods.

5 Examples

5.1 A toy Bernoulli mixture

Take the parametric family h_θ to be the family of Bernoulli mass functions, and consider a two-component version of the mixture with known weights both equal to $1/2$. This mixture density takes the form

$$f(z|r, s) = \frac{1}{2} r^z (1 - r)^{1-z} + \frac{1}{2} s^z (1 - s)^{1-z} \ ,$$

where $z \in \{0,1\}$ and $\boldsymbol{\theta} = (r,s)$. To simplify things ever further, assume that $r,s \in \{\rho, 1-\rho\}$ where $\rho \in (0,1/2)$ is fixed; that is, the two success probabilities, r and s, can only take the values ρ and $1-\rho$. Hence, $(r,s) \in \mathsf{X} = \big\{(\rho,\rho),(\rho,1-\rho),(1-\rho,\rho),(1-\rho,1-\rho)\big\}$. Our prior for (r,s) puts mass 1/4 on each of these four points. A simple calculation shows that the posterior mass function takes the form

$$\pi(r,s|\boldsymbol{z}) = \frac{I_{\{\rho,1-\rho\}}(r) I_{\{\rho,1-\rho\}}(s) (r+s)^{m_1} (2-r-s)^{m-m_1}}{2^m \rho^{m_1} (1-\rho)^{m-m_1} + 2^m \rho^{m-m_1} (1-\rho)^{m_1} + 2} \ ,$$

where $\boldsymbol{z} = (z_1, \ldots, z_m) \in \{0,1\}^m$ denotes the observed data, and m_1 denotes the number of successes among the m Bernoulli trials; that is, $m_1 = \sum_{i=1}^m z_i$. While we would never actually use MCMC to explore this simple four-point posterior, it is both interesting and useful to compare the FS and mixture DA algorithms in this context.

As described in Section 4.1, the mixture DA algorithm is based on the complete data posterior density, which is denoted here by $\pi(r,s,\boldsymbol{y}|\boldsymbol{z})$. (The fact that $\boldsymbol{p}$ is known in this case doesn't really change anything.) Of course, all we really need are the specific forms of the conditional mass functions, $\pi(\boldsymbol{y}|r,s,\boldsymbol{z})$ and $\pi(r,s|\boldsymbol{y},\boldsymbol{z})$. It follows from the general development in Subsection 4.1 that, given $(r,s,\boldsymbol{z})$, the components of $\boldsymbol{y} = (y_1, y_2, \ldots, y_m)$ are independent multinomials with mass functions given by

$$\pi(y_i|r,s,\boldsymbol{z}) = \frac{I_{\{1\}}(y_i) r^{z_i} (1-r)^{1-z_i} + I_{\{2\}}(y_i) s^{z_i} (1-s)^{1-z_i}}{r^{z_i}(1-r)^{1-z_i} + s^{z_i}(1-s)^{1-z_i}} \ .$$

Furthermore, it is easy to show that, given $(\boldsymbol{y},\boldsymbol{z})$, r and s are independent so $\pi(r,s|\boldsymbol{y},\boldsymbol{z}) = \pi(r|\boldsymbol{y},\boldsymbol{z})\,\pi(s|\boldsymbol{y},\boldsymbol{z})$. Now, for $j \in \{1,2\}$ and $k \in \{0,1\}$, let m_{jk} denotes the number of (y_i, z_i) pairs that take the value (j,k). (Note that $m_{10} + m_{11} = c_1$ and $m_{11} + m_{21} = m_1$.) Then we have

$$\pi(r|\boldsymbol{y},\boldsymbol{z}) = \frac{I_{\{\rho\}}(r) \rho^{m_{11}} (1-\rho)^{m_{10}} + I_{\{1-\rho\}}(r) \rho^{m_{10}} (1-\rho)^{m_{11}}}{\rho^{m_{11}} (1-\rho)^{m_{10}} + \rho^{m_{10}} (1-\rho)^{m_{11}}} \ ,$$

and

$$\pi(s|\boldsymbol{y},\boldsymbol{z}) = \frac{I_{\{\rho\}}(s) \rho^{m_{21}} (1-\rho)^{m_{20}} + I_{\{1-\rho\}}(s) \rho^{m_{20}} (1-\rho)^{m_{21}}}{\rho^{m_{21}} (1-\rho)^{m_{20}} + \rho^{m_{20}} (1-\rho)^{m_{21}}} \ .$$

The state space of the mixture DA chain is $\mathsf{X} = \big\{(\rho,\rho),(\rho,1-\rho),(1-\rho,\rho),(1-\rho,1-\rho)\big\}$, which has only four points. Hence, in this toy Bernoulli example, we can analyze the mixture DA chain directly. Its Mtm is 4×4 and the transition probabilities are given by

$$k(r',s'|r,s) = \sum_{\boldsymbol{y} \in \mathsf{Y}} \pi(r',s'|\boldsymbol{y},\boldsymbol{z})\,\pi(\boldsymbol{y}|r,s,\boldsymbol{z}) \ , \tag{16}$$

where $\mathsf{Y} = \{1,2\}^m$. We now perform an eigen-analysis of this Mtm. Note that $\pi(r',s'|\boldsymbol{y},\boldsymbol{z})$ and $\pi(\boldsymbol{y}|r,s,\boldsymbol{z})$ depend on $\boldsymbol{y}$ only through m_{10}, m_{11}, m_{20} and m_{21}. If we let $m_0 = m - m_1$, then we can express the transition probabilities as follows:

$$k(r',s'|r,s) = \sum_{i=0}^{m_1}\sum_{j=0}^{m_0}\binom{m_1}{i}\binom{m_0}{j}\left[\frac{I_{\{\rho\}}(r')\rho^i(1-\rho)^j + I_{\{1-\rho\}}(r')\rho^j(1-\rho)^i}{\rho^i(1-\rho)^j + \rho^j(1-\rho)^i}\right]\times$$
$$\left[\frac{I_{\{\rho\}}(s')\rho^{m_1-i}(1-\rho)^{m_0-j} + I_{\{1-\rho\}}(s')\rho^{m_0-j}(1-\rho)^{m_1-i}}{\rho^{m_1-i}(1-\rho)^{m_0-j} + \rho^{m_0-j}(1-\rho)^{m_1-i}}\right]\frac{r^i(1-r)^j s^{m_1-i}(1-s)^{m_0-j}}{(r+s)^{m_1}(2-r-s)^{m_0}} .$$

Now, for $k = 0, 1, 2$ define

$$w_k(\rho) =$$
$$\sum_{i=0}^{m_1}\sum_{j=0}^{m_0}\binom{m_1}{i}\binom{m_0}{j}\left[\frac{\rho^{k(m_0-j+i)}(1-\rho)^{k(m_1-i+j)}}{\left(\rho^i(1-\rho)^j + \rho^j(1-\rho)^i\right)\left(\rho^{m_1-i}(1-\rho)^{m_0-j} + \rho^{m_0-j}(1-\rho)^{m_1-i}\right)}\right] .$$

Using this notation, we can write the Mtm as follows:

$$k = \begin{bmatrix} \frac{\rho^{m_1}(1-\rho)^{m_0}}{2^m}w_0(\rho) & \frac{1}{2^m}w_1(\rho) & \frac{1}{2^m}w_1(\rho) & \frac{\rho^{m_0}(1-\rho)^{m_1}}{2^m}w_0(\rho) \\ \rho^{m_1}(1-\rho)^{m_0}w_1(\rho) & w_2(\rho) & \rho^m(1-\rho)^m w_0(\rho) & \rho^{m_0}(1-\rho)^{m_1}w_1(\rho) \\ \rho^{m_1}(1-\rho)^{m_0}w_1(\rho) & \rho^m(1-\rho)^m w_0(\rho) & w_2(\rho) & \rho^{m_0}(1-\rho)^{m_1}w_1(\rho) \\ \frac{\rho^{m_1}(1-\rho)^{m_0}}{2^m}w_0(\rho) & \frac{1}{2^m}w_1(\rho) & \frac{1}{2^m}w_1(\rho) & \frac{\rho^{m_0}(1-\rho)^{m_1}}{2^m}w_0(\rho) \end{bmatrix} .$$

We have ordered the points in the state space as follows: (ρ,ρ), $(\rho,1-\rho)$, $(1-\rho,\rho)$, and $(1-\rho,1-\rho)$. So, for example, the element in the second row, third column is the probability of moving from $(\rho, 1-\rho)$ to $(1-\rho,\rho)$. Note that all of the transition probabilities are strictly positive, which implies that the DA chain is Harris ergodic.

Of course, since k is a Mtm, it satisfies $kv_0 = \lambda_0 v_0$ where $v_0 = (1\ 1\ 1\ 1)^T$ and $\lambda_0 = 1$. However, (v_0,λ_0) does not count as an eigen-solution for us because we are using $L_0^2(f_X)$ instead of $L^2(f_X)$, and the only constant function in $L_0^2(f_X)$ is 0. For us, there are three eigen-solutions, and we write them as (v_i,λ_i), $i \in \{1,2,3\}$, where $0 \le \lambda_3 \le \lambda_2 \le \lambda_1 < 1$. Note that the first and fourth rows of k are identical, which means that $\lambda_3 = 0$. The remaining eigen-solutions follow from the general results in the Appendix. Indeed,

$$\lambda_1 = w_2(\rho) - \rho^m(1-\rho)^m w_0(\rho) ,$$

and the corresponding eigen-vector is $v_1 = (0\ 1\ -1\ 0)^T$. Finally,

$$\lambda_2 = \frac{g(\rho)w_0(\rho)}{2^m} - g(\rho)w_1(\rho)$$

and $v_2 = (\alpha\ 1\ 1\ \alpha)^T$, where $g(\rho) = \rho^{m_1}(1-\rho)^{m_0} + \rho^{m_0}(1-\rho)^{m_1}$ and

$$\alpha = \frac{g(\rho)w_0(\rho) - 2^m}{2^m g(\rho)w_1(\rho)} .$$

(The fact that $\lambda_2 \leq \lambda_1$ actually follows from our analysis of the FS chain that appears later in this subsection.) We now use these results to demonstrate that the mixture DA algorithm can perform quite poorly for the Bernoulli model.

Consider a numerical example in which $m = 10$, $\rho = 1/10$ and the data are $z_1 = \cdots = z_5 = 0$ and $z_6 = \cdots = z_{10} = 1$. The posterior mass function is as follows:

$$\pi(\rho, \rho|\boldsymbol{z}) = \pi(1-\rho, 1-\rho|\boldsymbol{z}) = 0.003 \quad \text{and} \quad \pi(\rho, 1-\rho|\boldsymbol{z}) = \pi(1-\rho, \rho|\boldsymbol{z}) = 0.497 .$$

So there are two points with exactly the same very high probability, and two points with exactly the same very low probability. As we now explain, the DA chain converges slowly due to its inability to move between the two high probability points. Indeed, the Markov transition matrix in this case is:

$$k = \begin{bmatrix} 0.10138 & 0.39862 & 0.39862 & 0.10138 \\ 0.00241 & 0.99457 & 0.00061 & 0.00241 \\ 0.00241 & 0.00061 & 0.99457 & 0.00241 \\ 0.10138 & 0.39862 & 0.39862 & 0.10138 \end{bmatrix} .$$

Suppose we start the chain in the state $(\rho, 1-\rho)$. The expected number of steps before it reaches the other high probability state, $(1-\rho, \rho)$, is quite large. First, we expect the chain to remain in the state $(\rho, 1-\rho)$ for about $1/(1-0.99457) \approx 184$ iterations. Then, conditional on the chain leaving $(\rho, 1-\rho)$, the probability that it moves to (ρ, ρ) or $(1-\rho, 1-\rho)$ is about 0.89. And if it does reach (ρ, ρ) or $(1-\rho, 1-\rho)$, there is still about a 40% chance that it will jump right back to the point $(\rho, 1-\rho)$, where it will stay for (approximately) another 184 iterations. All of this translates into slow convergence. In fact, the two non-zero eigenvalues are $(\lambda_1, \lambda_2) = (0.99395, 0.19795)$. Moreover, the problem gets worse as the sample size increases. For example, if we increase the sample size to $m = 20$ (and maintain the 50:50 split of 0s and 1s in the data), then $(\lambda_1, \lambda_2) = (0.99996, 0.15195)$. Figure 1 shows how the dominant eigenvalue, λ_1, changes with sample size for several different values of ρ. We conclude that, for fixed ρ, the convergence rate deteriorates as the sample size increases. Moreover, the (negative) impact of increasing sample size is magnified as ρ gets smaller.

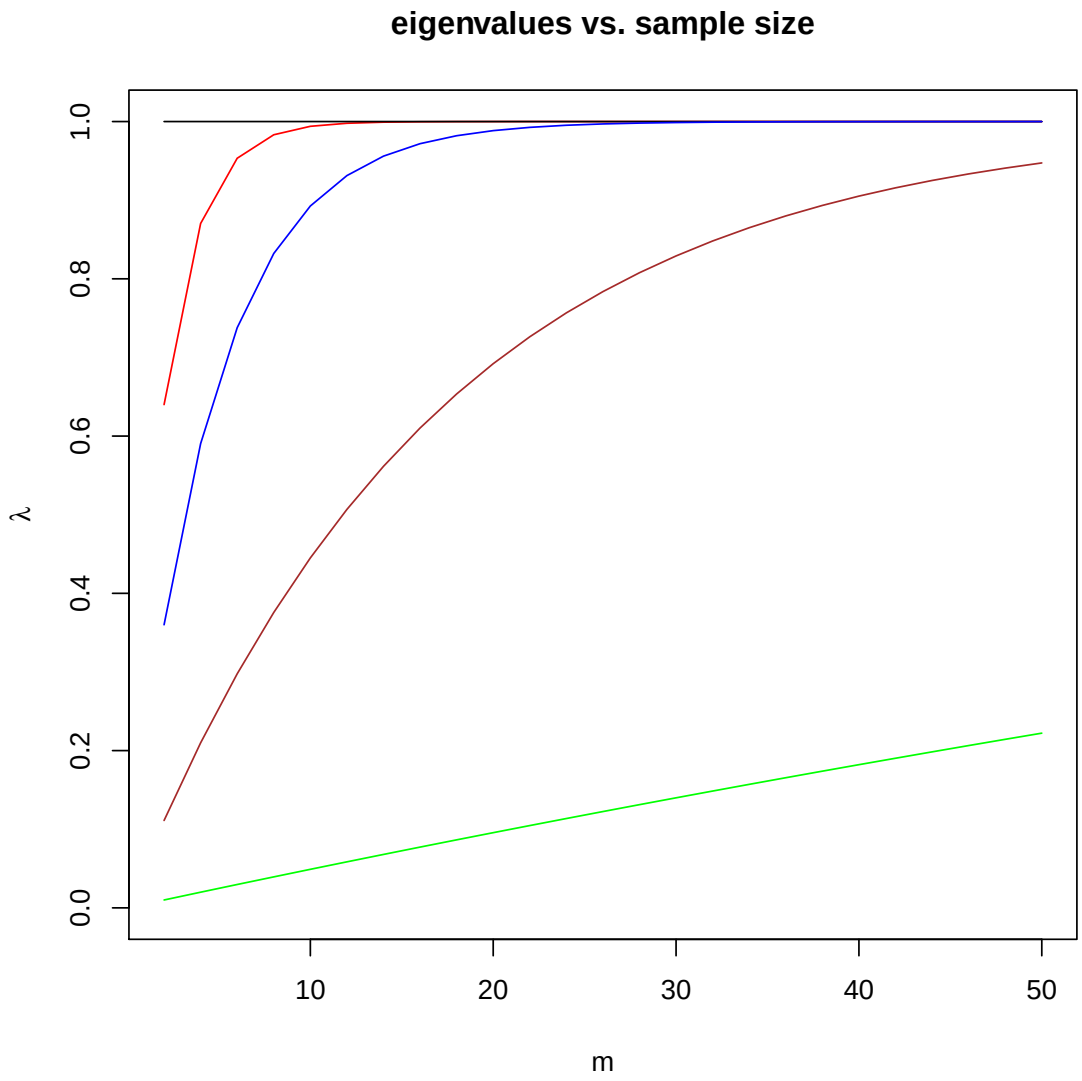

Figure 1: The behavior of the dominant eigenvalue for the mixture DA chain in the Bernoulli model. The graph shows how the dominant eigenvalue of the mixture DA chain changes with sample size, m, for several different values of ρ, in the case where half the z_is are 0 and the other half are 1. (Only even sample sizes are considered.) The red, blue, brown and green lines correspond to ρ values of $1/10$, $1/5$, $1/3$, and $9/20$, respectively.

Now consider implementing the FS algorithm for the Bernoulli mixture. Because the mixture has only two components, the random label switching step, $\boldsymbol{y} \to \boldsymbol{y}'$, is quite simple. Indeed, we simply flip a fair coin. If the result is heads, then we take $\boldsymbol{y}' = \boldsymbol{y}$, and if the result is tails, then we take $\boldsymbol{y}' = \overline{\boldsymbol{y}}$, where $\overline{\boldsymbol{y}}$ denotes $\boldsymbol{y}$ with its 1s and 2s flipped. The Mtm of the FS chain has entries given by

$$\tilde{k}(r', s'|r, s) = \frac{1}{2} \sum_{\boldsymbol{y} \in \mathsf{Y}} \pi(r', s'|\overline{\boldsymbol{y}}, \boldsymbol{z})\, \pi(\boldsymbol{y}|r, s, \boldsymbol{z}) + \frac{1}{2} \sum_{\boldsymbol{y} \in \mathsf{Y}} \pi(r', s'|\boldsymbol{y}, \boldsymbol{z})\, \pi(\boldsymbol{y}|r, s, \boldsymbol{z}) \ .$$

It follows that

$$\tilde{k} = \begin{bmatrix} \frac{\rho^{m_1}(1-\rho)^{m_0}}{2^m} w_0(\rho) & \frac{1}{2^m} w_1(\rho) & \frac{1}{2^m} w_1(\rho) & \frac{\rho^{m_0}(1-\rho)^{m_1}}{2^m} w_0(\rho) \\ \rho^{m_1}(1-\rho)^{m_0} w_1(\rho) & \frac{w_2(\rho)+\rho^m(1-\rho)^m w_0(\rho)}{2} & \frac{w_2(\rho)+\rho^m(1-\rho)^m w_0(\rho)}{2} & \rho^{m_0}(1-\rho)^{m_1} w_1(\rho) \\ \rho^{m_1}(1-\rho)^{m_0} w_1(\rho) & \frac{w_2(\rho)+\rho^m(1-\rho)^m w_0(\rho)}{2} & \frac{w_2(\rho)+\rho^m(1-\rho)^m w_0(\rho)}{2} & \rho^{m_0}(1-\rho)^{m_1} w_1(\rho) \\ \frac{\rho^{m_1}(1-\rho)^{m_0}}{2^m} w_0(\rho) & \frac{1}{2^m} w_1(\rho) & \frac{1}{2^m} w_1(\rho) & \frac{\rho^{m_0}(1-\rho)^{m_1}}{2^m} w_0(\rho) \end{bmatrix} .$$

Note that this matrix differs from k only in the middle four elements. Indeed, the $(2, 2)$ and $(2, 3)$ elements in k have both been replaced by their average in $\tilde{k}$, and the same is true of the $(3, 2)$ and $(3, 3)$ elements. The matrix $\tilde{k}$ has rank at most two, so there is at most one non-zero eigenvalue to find. Using the results in the Appendix along with the eigen-analysis of k performed earlier, it is easy to see that the non-trivial eigen-solution of $\tilde{k}$ is $(\tilde{v}_1, \tilde{\lambda}_1) = (v_2, \lambda_2)$. So, the effect on the spectrum of adding the random label switching step is to replace the dominant eigenvalue with 0! (Note that Theorem 1 implies that $\lambda_2 = \tilde{\lambda}_1 \leq \lambda_1$, which justifies our ordering of the eigenvalues of k.) Consider again the simple numerical example with the 50:50 split of 0s and 1s. In the case $m = 10$, the result of adding the extra step is to replace the dominant eigenvalue, 0.99395, by 0.19795. When $m = 20$, 0.99996 is replaced by 0.15195. This suggests that, in contrast to the mixture DA algorithm, increasing sample size does not adversely affect the FS algorithm. More evidence for this is provided in Figure 2, which is the analogue of Figure 1 for the FS algorithm. Note that the dominant eigenvalues are now substantially smaller, and no longer converge to 1 as the sample size increases. In fact, based on experimental evidence, it appears that, for a fixed value of ρ, λ_2 hits a maximum and then decreases with sample size. It is surprising that such a minor change in the mixture DA algorithm could result in such a huge improvement. In the next section, we consider a mixture of normal densities.

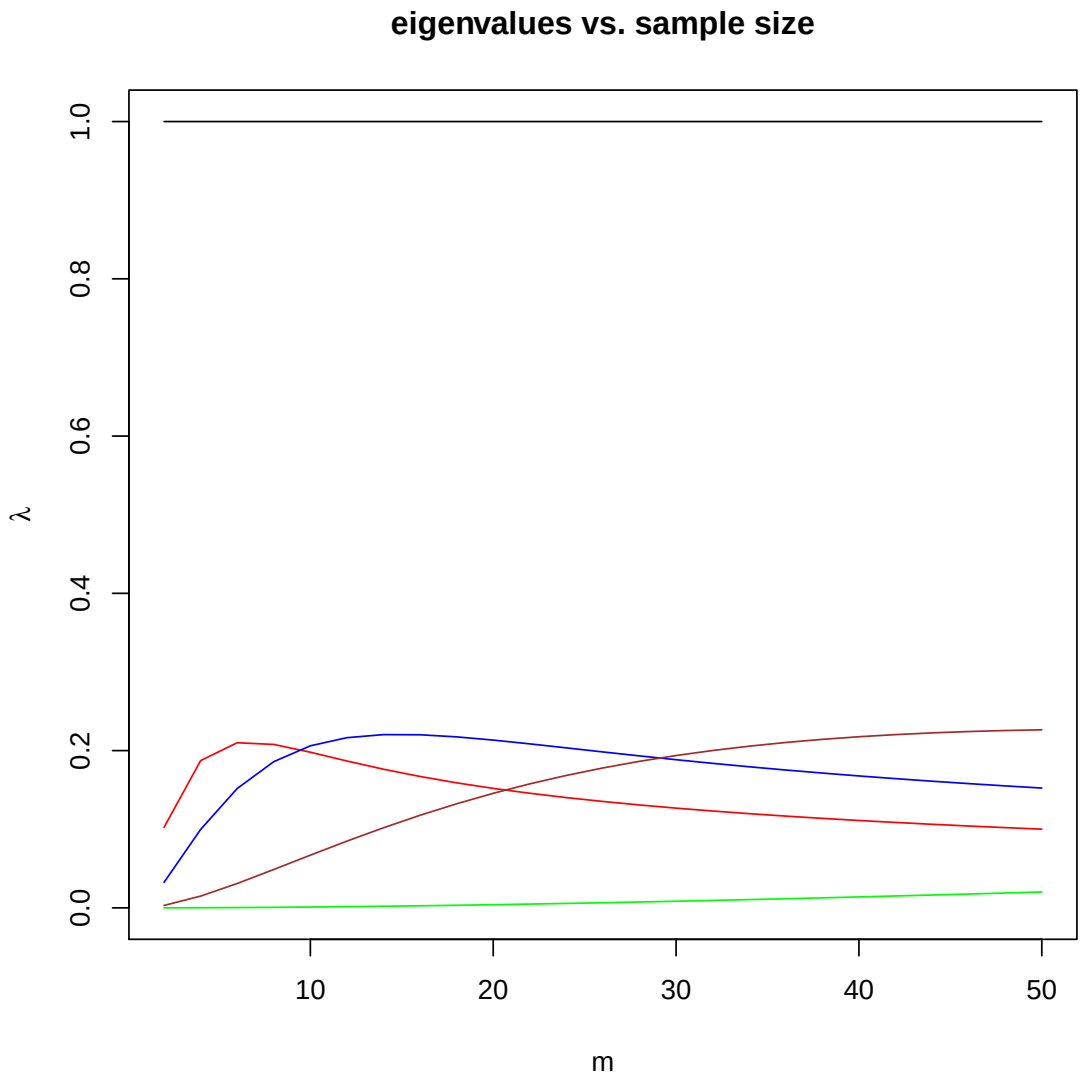

Figure 2: The behavior of the dominant eigenvalue for the FS chain in the Bernoulli model. The graph shows how the dominant eigenvalue of the FS chain changes with sample size, m, for several different values of ρ, in the case where half the z_is are 0 and the other half are 1. (Only even sample sizes are considered.) The red, blue, brown and green lines correspond to ρ values of $1/10$, $1/5$, $1/3$, and $9/20$, respectively.

5.2 The normal mixture

Assume that $Z_1, \ldots, Z_m$ are iid from the density

$$f(z|\mu, \tau^2, p) = p\frac{1}{\tau_1}\phi\Big(\frac{z-\mu_1}{\tau_1}\Big) + (1-p)\frac{1}{\tau_2}\phi\Big(\frac{z-\mu_2}{\tau_2}\Big) \ ,$$

where $p \in [0,1]$, $\mu = (\mu_1, \mu_2) \in \mathbb{R}^2$, $\tau^2 = (\tau_1^2, \tau_2^2) \in \mathbb{R}_+^2$, and $\phi(\cdot)$ denotes the standard normal density function. The prior for p is Uniform$(0,1)$, and the prior for (μ, τ^2) takes the form $\pi(\mu_1, \tau_1^2)\,\pi(\mu_2, \tau_2^2)$. As for π, we use the standard (conditionally conjugate) prior given by

$$\pi(\mu_1, \tau_1^2) = \pi(\mu_1|\tau_1^2)\pi(\tau_1^2) \ ,$$

where $\pi(\mu_1|\tau_1^2) = \mathrm{N}(0, \tau_1^2)$ and $\pi(\tau_1^2) = \mathrm{IG}(2, 1/2)$ (Robert and Casella, 2004, Section 9.1). By $W \sim \mathrm{IG}(\alpha, \gamma)$, we mean that W is a random variable with density function proportional to $w^{-\alpha-1}\exp\{-\gamma/w\}I_{\mathbb{R}_+}(w)$. In contrast with the Bernoulli example from the previous subsection, the posterior density associated with the normal mixture is quite intractable and has a complicated (and uncountable) support given by $\mathsf{X} = \mathbb{R}^2 \times \mathbb{R}_+^2 \times [0,1]$.

The mixture DA algorithm is based on the complete-data posterior density, which we denote here by $\pi(\mu, \tau^2, p, \boldsymbol{y}|\boldsymbol{z})$. Again, the development in Subsection 4.1 implies that, given $(\mu, \tau^2, p, \boldsymbol{z})$, the elements of $\boldsymbol{y}$ are independent multinomials and the probability that the ith coordinate equals 1 (which is one minus the probability that it equals 2) is given by

$$\frac{p\frac{1}{\tau_1}\phi\Big(\frac{z_i-\mu_1}{\tau_1}\Big)}{p\frac{1}{\tau_1}\phi\Big(\frac{z_i-\mu_1}{\tau_1}\Big) + (1-p)\frac{1}{\tau_2}\phi\Big(\frac{z_i-\mu_2}{\tau_2}\Big)} \ . \tag{17}$$

We sample $\pi(\mu, \tau^2, p|\boldsymbol{y}, \boldsymbol{z})$ via sequential sampling from $\pi(p|\boldsymbol{y}, \boldsymbol{z})$ and $\pi(\mu, \tau^2|p, \boldsymbol{y}, \boldsymbol{z})$. The results in Subsection 4.1 show that $p|\boldsymbol{y}, \boldsymbol{z} \sim \mathrm{Beta}(c_1+1, c_2+1)$. Moreover, it's easy to show that, given $(p, \boldsymbol{y}, \boldsymbol{z})$, (μ_1, τ_1^2) and (μ_2, τ_2^2) are independent. Routine calculations show that

$$\mu_1|\tau_1^2, p, \boldsymbol{y}, \boldsymbol{z} \sim \mathrm{N}\bigg(\frac{c_1}{c_1+1}\overline{z}_1, \frac{\tau_1^2}{(c_1+1)}\bigg)$$

and

$$\tau_1^2|p, \boldsymbol{y}, \boldsymbol{z} \sim \mathrm{IG}\bigg(\frac{c_1+4}{2}, \frac{1}{2}\Big(s_1^2 + \frac{c_1\overline{z}_1^2}{(c_1+1)} + 1\Big)\bigg) \ ,$$

where $\overline{z}_1 = \frac{1}{c_1}\sum_{i=1}^m I_{\{1\}}(y_i)z_i$ and $s_1^2 = \sum_{i=1}^m I_{\{1\}}(y_i)(z_i - \overline{z}_1)^2$. Of course, the distribution of (μ_2, τ_2^2) given $(p, \boldsymbol{y}, \boldsymbol{z})$ has an analogous form.

The results developed in Section 2 imply that the spectrum of the operator associated with the mixture DA chain consists of the point $\{0\}$ and the eigenvalues of the Mtm of the conjugate chain, which lives on $\mathsf{Y} = \{1, 2\}^m$. Unfortunately, the Mtm of the conjugate chain is also intractable. Indeed, a generic element of the Mtm of the conjugate chain has the following form:

$$\hat{k}(\boldsymbol{y}'|\boldsymbol{y}) = \int_0^1 \int_{\mathbb{R}_+^2} \int_{\mathbb{R}^2} \pi(\boldsymbol{y}'|\mu, \tau^2, p, \boldsymbol{z})\, \pi(\mu, \tau^2, p|\boldsymbol{y}, \boldsymbol{z})\, d\mu\, d\tau^2\, dp\ .$$

This integral cannot be computed in closed form. In particular, $\pi(\boldsymbol{y}'|\mu, \tau^2, p, \boldsymbol{z})$ is the product of m probabilities of the form (17), and the sums in the denominators of these probabilities render the integral intractable. However, note that $\hat{k}(\boldsymbol{y}'|\boldsymbol{y})$ can be interpreted as the expected value of $\pi(\boldsymbol{y}'|\mu, \tau^2, p, \boldsymbol{z})$ with respect to the density $\pi(\mu, \tau^2, p|\boldsymbol{y}, \boldsymbol{z})$. Of course, for fixed $\boldsymbol{z}$, we know how to draw from $\pi(\mu, \tau^2, p|\boldsymbol{y}, \boldsymbol{z})$, and we have $\pi(\boldsymbol{y}'|\mu, \tau^2, p, \boldsymbol{z})$ in closed form. We therefore have the ability to estimate $\hat{k}(\boldsymbol{y}'|\boldsymbol{y})$ using classical Monte Carlo. Once we have an estimate of the entire $2^m \times 2^m$ Mtm, we can calculate its eigenvalues.

The same idea can be used to approximate the eigenvalues of the FS chain. The results in Section 3 show that we can express the FS algorithm as a DA algorithm with respect to an alternative complete-data posterior density, which we write as $\pi^*(\mu, \tau^2, p, \boldsymbol{y}|\boldsymbol{z})$. The eigenvalues of the operator defined by the FS chain are the same as those of the Mtm in which the probability of the transition $\boldsymbol{y} \to \boldsymbol{y}'$ is given by

$$\int_0^1 \int_{\mathbb{R}_+^2} \int_{\mathbb{R}^2} \pi^*(\boldsymbol{y}'|\mu, \tau^2, p, \boldsymbol{z})\, \pi^*(\mu, \tau^2, p|\boldsymbol{y}, \boldsymbol{z})\, d\mu\, d\tau^2\, dp\ .$$

It is straightforward to simulate from $\pi^*(\mu, \tau^2, p|\boldsymbol{y}, \boldsymbol{z})$, and $\pi^*(\boldsymbol{y}'|\mu, \tau^2, p, \boldsymbol{z})$ is available in closed form.

To use our classical Monte Carlo idea to estimate the spectra associated with the mixture DA and FS chains, we must specify the data, $\boldsymbol{z}$. Furthermore, the Bernoulli example in the previous subsection showed that the convergence rates of the two algorithms can depend heavily on the sample size, m. Thus, we would like to explore how an increasing sample size affects the convergence rates of the mixture DA and FS chains in the current context. To generate data, we simulated a random sample of size 10 from a 50:50 mixture of a $\mathrm{N}(0, .55^2)$ and a $\mathrm{N}(3, .55^2)$, and this resulted in the following observations:

$$\begin{aligned}\boldsymbol{z} &= (z_1, \ldots, z_{10}) \\ &= (0.2519, 2.529, -0.2930, 2.799, 3.397, 0.5596, 2.810, 2.541, 2.487, -0.1937)\ .\end{aligned}$$

We considered 10 different data sets ranging in size from $m = 1$ to $m = 10$. The first data set contained the single point $z_1 = 0.25192$, the second contained the first two observations $(z_1, z_2) = (0.25192, 2.5287)$, the third contained $(z_1, z_2, z_3) = (0.25192, 2.5287, -0.29303)$, and so on up to the tenth data set, which contained all ten observations. For each of these 10 data sets, we used the classical Monte Carlo technique described above to estimate the Mtm for both the mixture DA and FS algorithms. In particular, for each row of the Mtm we used a single Monte Carlo sample of size 200,000 (from $\pi(\mu, \tau^2, p|\boldsymbol{y}, \boldsymbol{z})$ for DA, and from $\pi^*(\mu, \tau^2, p|\boldsymbol{y}, \boldsymbol{z})$ for FS) to estimate each of the entries in that row. We then calculated the eigenvalues of the estimated Mtms and recorded the largest one. The results are shown in Figure 3, which has some interesting features. Note that the dominant eigenvalues of mixture DA chain are much closer to 1 than the corresponding dominant eigenvalues of the FS chain. Even at $m = 5$, the dominant eigenvalue of the mixture DA chain is already above 0.99. As in the previous example, the convergence rate of the mixture DA chain deteriorates as m increases. It is not clear whether the FS chain slows down as m increases. It may be the case that the FS eigenvalue would eventually level off, or perhaps the FS chain would eventually begin to speed up, as in the Bernoulli example. Note that, as proven in Subsection 4.2, when $m = 1$, the FS eigenvalue is 0. (To ascertain the accuracy of our estimates, we repeated the entire simulation 6 times, with different random number seeds, and based on this, we believe that our eigenvalue estimates are correct up to three decimal places.)

In the case where all 10 observations are considered, the dimension of the Mtms is 1024×1024, and each element must be estimated by classical Monte Carlo. Thus, while it would be very interesting to consider larger sample sizes (beyond 10), and even mixtures with more than 2 components, the matrices in these cases are simply too big to handle.

We simulated a second set of 10 observations from the same 50:50 mixture and repeated the entire process for the purpose of validation. The second simulation resulted in the following data:

$$\begin{aligned}\boldsymbol{z} &= (z_1, \ldots, z_{10}) \\ &= (0.6699, 3.408, 0.1093, 3.289, -0.1407, 3.525, 2.454, 0.2716, -0.7443, 3.570)\ .\end{aligned}$$

Figure 4 is the analogue of Figure 3 for the second simulation. The results are nearly identical to those from the first simulation.

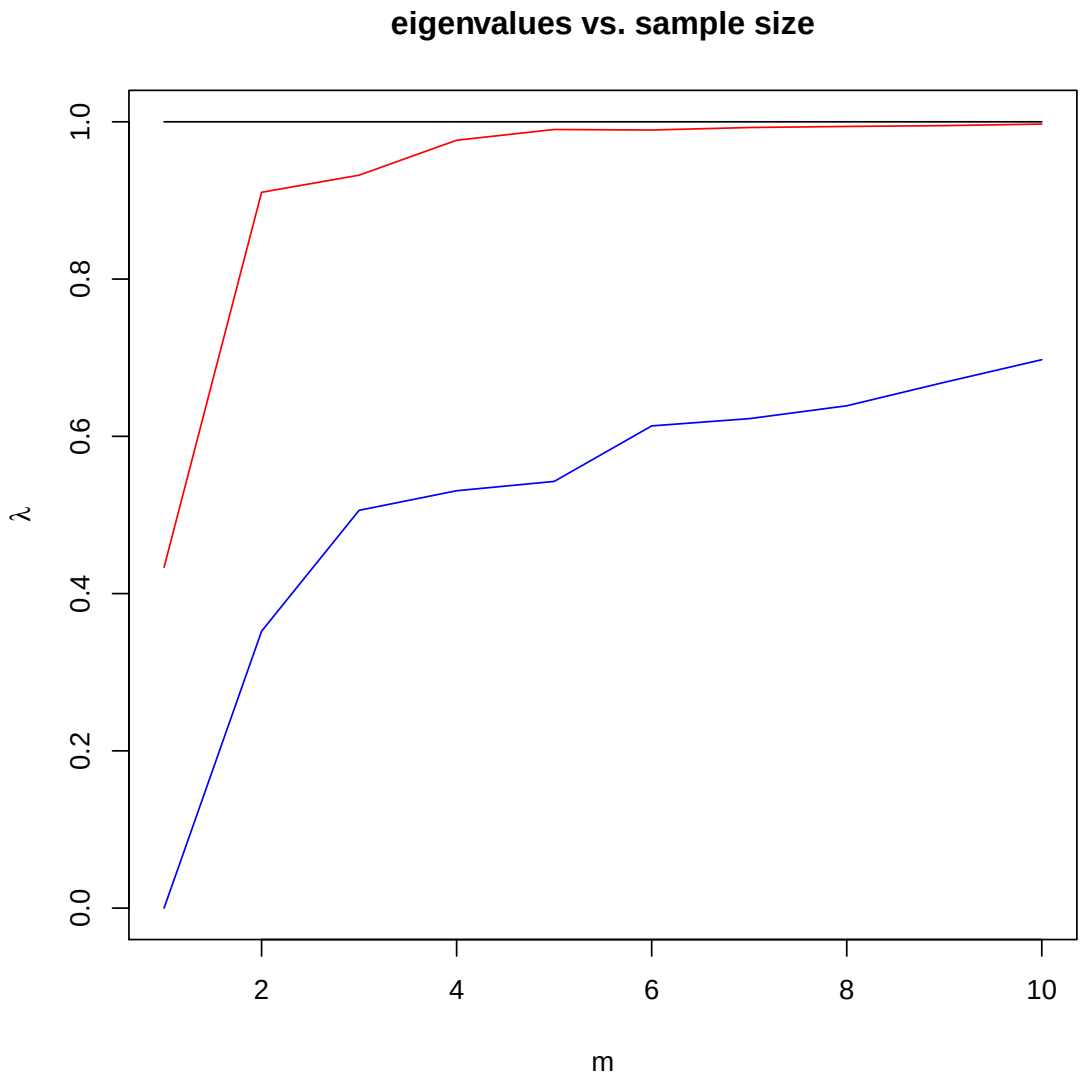

Figure 3: The behavior of the dominant eigenvalue for the mixture DA and FS chains in the normal model. The graph is based on the first simulated data set and shows how the dominant eigenvalue changes with sample size, m, for the mixture DA algorithm (red line) and the FS algorithm (blue line).

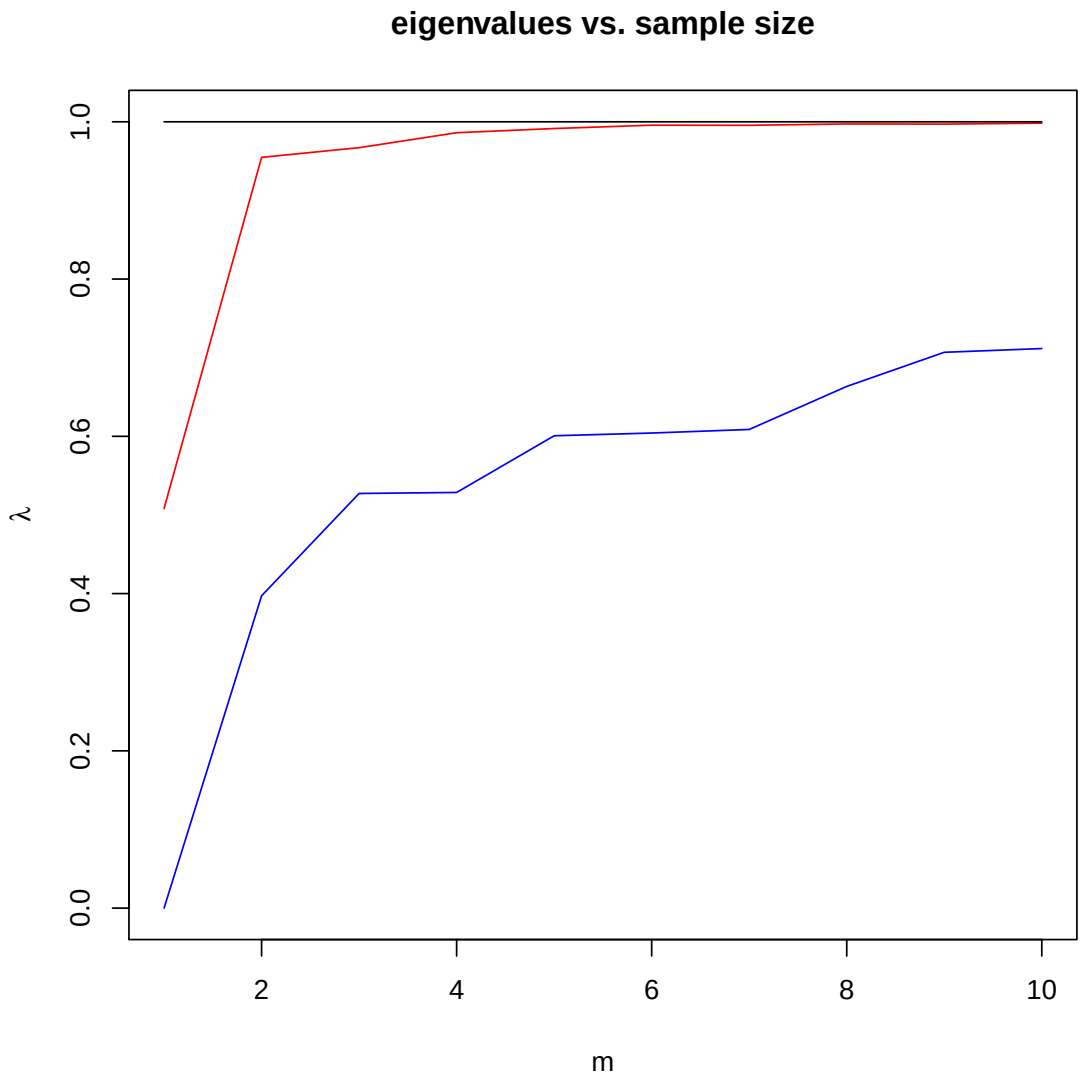

Figure 4: The behavior of the dominant eigenvalue for the mixture DA and FS chains in the normal model. The graph is based on the second simulated data set and shows how the dominant eigenvalue changes with sample size, m, for the mixture DA algorithm (red line) and the FS algorithm (blue line).

6 Appendix

Consider a Mtm of the form:

$$M = \begin{bmatrix} a & b & b & c \\ d & e & f & g \\ d & f & e & g \\ a & b & b & c \end{bmatrix} .$$

We assume that all of the elements are strictly positive, so the corresponding Markov chain is irreducible and aperiodic. Note that the two Mtms in Section 5.1 are both of this form. Of course, since M is a Mtm, it satisfies $kv_0 = \lambda_0 v_0$ where $v_0 = (1\ 1\ 1\ 1)^T$ and $\lambda_0 = 1$. Furthermore, since the first and fourth rows are equal, there is at least one eigenvalue equal to zero. Here we identify the other two eigen-solutions of M. Let $v_1 = (0\ 1\ -1\ 0)^T$ and note that

$$Mv_1 = (e - f)v_2 ,$$

so $\lambda_1 = (e - f)$ is eigenvalue. If $e = f$, then the middle two rows of M are also equal and there will be at least two eigenvalues equal to 0.

Now, let $v_2 = (\alpha\ 1\ 1\ \alpha)^T$, where α is a constant to be determined, and note that

$$Mv_2 = \begin{bmatrix} \alpha a + 2b + \alpha c \\ \alpha d + e + f + \alpha g \\ \alpha d + e + f + \alpha g \\ \alpha a + 2b + \alpha c \end{bmatrix} .$$

If v_2 is an eigenvector with corresponding eigenvalue λ_2, then the first element of Mv_2 must equal $\alpha\lambda_2$; that is,

$$\alpha a + 2b + \alpha c = \alpha\lambda_2 .$$

Now, using the fact that $2b = 1 - a - c$, we have

$$(\alpha - 1)(a + c) + 1 = \alpha\lambda_2 ,$$

and it follows that

$$\lambda_2 = \frac{(\alpha - 1)(a + c) + 1}{\alpha} . \tag{18}$$

Again, if v_2 is an eigenvector with corresponding eigenvalue λ_2, then the second element of Mv_2 must equal λ_2, or

$$\lambda_2 = \alpha d + e + f + \alpha g .$$

Now, using the fact that $e = 1 - d - f - g$, we have

$$\lambda_2 = (\alpha - 1)(d + g) + 1 \ .$$

Setting our two expressions for λ_2 equal yields:

$$\alpha(\alpha - 1)(d + g) + \alpha = (\alpha - 1)(a + c) + 1 \ .$$

This quadratic in α has two roots: $\alpha = 1$ and

$$\alpha = \frac{a + c - 1}{d + g} \ .$$

Of course, when $\alpha = 1$, $v_2 = v_0$ and we are back to $kv_0 = \lambda_0 v_0$. If $(a + c - 1)/(d + g) \neq 1$, then $v_2 = (\alpha \ 1 \ 1 \ \alpha)^T$ is a nontrivial eigenvector, and the corresponding nontrivial eigenvalue is

$$\lambda_2 = a + c - (d + g) \ .$$

This eigenvalue could be zero, but, if it is not; that is, if $a + c \neq d + g$, then we have found the last eigen-solution.

Acknowledgments

The third author spoke at length with Professor Richard Tweedie about the convergence rate of the mixture DA algorithm during a visit to Colorado State University in 1993. Although the present work is not directly related to those conversations, the third author wants to acknowledge here his admiration for Professor Tweedie's insights and his gratitude for his support. The first author's work was supported by NSF Grant DMS-08-05860. The third author's work was supported by Agence Nationale de la Recherche (ANR, 212, rue de Bercy 75012 Paris) through the 2009-2012 project ANR-08-BLAN-0218 Big'MC. The first author thanks the Université Paris Dauphine for partial travel support that funded visits to Paris in 2008 and 2009. The second author thanks the Agence Nationale de la Recherche through the 2005-2009 project Ecosstat for support that funded a visit to Paris in 2008.

References

CELEUX, G., HURN, M. A. and ROBERT, C. (2000). Computational and inferential difficulties with mixtures posterior distribution. *Journal of the American Statistical Association*, **95(3)** 957–979.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society,* Series B, **39** 1–38.

DIACONIS, P., KHARE, K. and SALOFF-COSTE, L. (2008). Gibbs sampling, exponential families and orthogonal polynomials (with discussion). *Statistical Science*, **23** 151–200.

DIEBOLT, J. and ROBERT, C. P. (1994). Estimation of finite mixture distributions by Bayesian sampling. *Journal of the Royal Statistical Society,* Series B, **56** 363–375.

FRÜHWIRTH-SCHNATTER, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, **96** 194–209.

HOBERT, J. P. (2009). The data augmentation algorithm: Theory and methodology. In *Handbook of Markov Chain Monte Carlo* (S. Brooks, A. Gelman, G. Jones and X.-L. Meng, eds.). Chapman & Hall/CRC Press.

HOBERT, J. P. and MARCHEV, D. (2008). A theoretical comparison of the data augmentation, marginal augmentation and PX-DA algorithms. *The Annals of Statistics*, **36** 532–554.

JASRA, A., HOLMES, C. C. and STEPHENS, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, **20** 50–67.

LEE, K., MARIN, J.-M., MENGERSEN, K. L. and ROBERT, C. (2008). Bayesian inference on mixtures of distributions. In *Platinum Jubilee of the Indian Statistical Institute* (N. N. Sastry, ed.). Indian Statistical Institute, Bangalore.

LIU, J. S. and SABATTI, C. (2000). Generalised Gibbs sampler and multigrid Monte Carlo for Bayesian computation. *Biometrika*, **87** 353–369.

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

LIU, J. S., WONG, W. H. and KONG, A. (1995). Covariance structure and convergence rate of the Gibbs sampler with various scans. *Journal of the Royal Statistical Society,* Series B, **57** 157–169.

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

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

MIRA, A. and GEYER, C. J. (1999). Ordering Monte Carlo Markov chains. Tech. Rep. No. 632, School of Statistics, University of Minnesota.

PAPASPILIOPOULOS, O., ROBERTS, G. O. and SKÖLD, M. (2007). A general framework for the parametrization of hierarchical models. *Statistical Science*, **22** 59–73.

RETHERFORD, J. R. (1993). *Hilbert Space: Compact Operators and the Trace Theorem*. Cambridge University Press, Cambridge.

ROBERT, C. P. and CASELLA, G. (2004). *Monte Carlo Statistical Methods*. 2nd ed. Springer, New York.

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

ROBERTS, G. O. and ROSENTHAL, J. S. (2001). Markov chains and de-initializing processes. *Scandinavian Journal of Statistics*, **28** 489–504.

ROSENTHAL, J. S. (2003). Asymptotic variance and convergence rates of nearly-periodic MCMC algorithms. *Journal of the American Statistical Association*, **98** 169–177.

ROY, V. and HOBERT, J. P. (2007). Convergence rates and asymptotic standard errors for Markov chain Monte Carlo algorithms for Bayesian probit regression. *Journal of the Royal Statistical Society,* Series B, **69** 607–623.

RUDIN, W. (1991). *Functional Analysis*. 2nd ed. McGraw-Hill, New York.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, **82** 528–550.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions (with discussion). *The Annals of Statistics*, **22** 1701–1762.

VAN DYK, D. A. and MENG, X.-L. (2001). The art of data augmentation (with discussion). *Journal of Computational and Graphical Statistics*, **10** 1–50.

VOSS, H. (2003). Variational characterizations of eigenvalues of nonlinear eigenproblems. In *Proceedings of the International Conference on Mathematical and Computer Modelling in Science and Engineering* (M. Kocandrlova and V. Kelar, eds.). Czech Technical University, Prague, 379–383.

YU, Y. and MENG, X.-L. (2009). To center or not to center: That is not the question - an ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency. Tech. rep., University of California, Irvine.