Comput Stat (2016) 31:709–728
DOI 10.1007/s00180-015-0632-4

ORIGINAL PAPER

Improving efficiency of data augmentation algorithms using Peskun's theorem

Vivekananda Roy[1]

Received: 26 September 2014 / Accepted: 4 November 2015 / Published online: 19 November 2015

Abstract Data augmentation (DA) algorithm is a widely used Markov chain Monte Carlo algorithm. In this paper, an alternative to DA algorithm is proposed. It is shown that the modified Markov chain is always more efficient than DA in the sense that the asymptotic variance in the central limit theorem under the alternative chain is no larger than that under DA. The modification is based on Peskun's (Biometrika 60:607–612, 1973) result which shows that asymptotic variance of time average estimators based on a finite state space reversible Markov chain does not increase if the Markov chain is altered by increasing all off-diagonal probabilities. In the special case when the state space or the augmentation space of the DA chain is finite, it is shown that Liu's (Biometrika 83:681–682, 1996) modified sampler can be used to improve upon the DA algorithm. Two illustrative examples, namely the beta-binomial distribution, and a model for analyzing rank data are used to show the gains in efficiency by the proposed algorithms.

Keywords Modified data augmentation · Efficiency ordering · MCMC · Peskun ordering · Rao Blackwellization

1 Introduction

Let $f_X : \mathsf{X} \to [0, \infty)$ be a probability density function that is intractable in the sense that expectations with respect to f_X cannot be computed analytically. If direct simulation from f_X is not possible, one may resort to a Markov chain Monte Carlo (MCMC) method such as the data augmentation (DA) algorithm (Tanner and Wong

✉ Vivekananda Roy
vroy@iastate.edu

[1] Iowa State University, Ames, IA 50011, USA

1987) to estimate expectations with respect to f_X. The construction of a DA algorithm begins with finding a joint density, say $f : \mathsf{X} \times \mathsf{Y} \to [0, \infty)$, that satisfies two conditions: (i) the x-marginal of $f(x, y)$ is f_X, and (ii) it is easy to sample from the corresponding two conditional densities, $f_{X|Y}(\cdot|y)$ and $f_{Y|X}(\cdot|x)$. The Markov chain $\{X_n\}_{n=0}^{\infty}$ associated with the DA algorithm is run as follows. Given the current state, $X_n = x$, the following two steps are used to move to the new state X_{n+1}.

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

Since the Markov chain $\{X_n\}_{n\geq 0}$ is reversible with respect to f_X (Liu et al. 1994), it follows that f_X is an invariant density for $\{X_n\}_{n\geq 0}$. Suppose that $g : \mathsf{X} \to \mathbb{R}$ is a function of interest and we want to compute $E_{f_X} g := \int_{\mathsf{X}} g(x) f_X(x) \mu(dx)$. If $\{X_n\}_{n\geq 0}$ is suitably irreducible, then time averages $\bar{g}_n := \sum_{i=1}^{n} g(X_i)/n$ consistently estimate space averages $E_{f_X} g$ (Meyn and Tweedie 1993, Theorem 17.0.1). In order to evaluate the performance of $\bar{g}_n$ as an estimator of $E_{f_X} g$, like elsewhere in statistics, we consider its variance in the central limit theorem (CLT). In this paper we propose a modification of DA algorithms using Peskun ordering (Peskun 1973) that improves DA in terms of *asymptotic variance* of $\bar{g}_n$.

We now briefly describe Peskun's (1973) result. Let P and Q be two Markov transition matrices both reversible with respect to a given probability distribution. Peskun (1973) showed that if each of the off-diagonal elements of P is greater than or equal to the corresponding off-diagonal element of Q, then the asymptotic variances of time averages of any function are smaller in a chain generated using P than in one using Q. Tierney (1998) later extended this result to general state space Markov chains. The key idea behind Peskun ordering is that by moving probability off the diagonal, a Markov chain decreases probability of retaining the current state. Note that if a Markov chain is held back in the same state for succeeding times, it fails to move around the state space and thus increases autocorrelation in the observed Markov chain and hence the variance of the empirical average increases.

If we replace step 2 above in the DA algorithm with a draw from a Markov transition function that is reversible with respect to $f_{X|Y}$, we show that the resulting Markov chain $\{\tilde{X}_n\}$ is also reversible with respect to the target density $f_X(x)$. Thus, the chain $\{\tilde{X}_n\}$ can also be used to estimate expectations with respect to $f_X(x)$. Further, we establish conditions under which the Markov chain $\{\tilde{X}_n\}$ has higher off-diagonal probabilities than DA. Then as discussed above the modified chain $\{\tilde{X}_n\}$ is at least as efficient as the DA algorithm in the sense that asymptotic variances are never larger than the DA chain. Improving efficiency of the DA algorithm is practically important since if $\{\tilde{X}_n\}$ is twice as efficient as DA and if both algorithms require similar amount of time to run, then $\{\tilde{X}_n\}$ needs only half of the time the DA algorithm requires to achieve the same level of precision in the estimates. In particular, we consider the case when the state space X or the augmentation space Y is finite, and the step 2 in DA algorithm is substituted with an appropriate Metropolis Hastings (MH) step to improve upon the DA algorithm. The MH step that we use here is given in Liu (1996) who used it to increase efficiency of random scan Gibbs sampler. We call the resulting algorithm, the modified DA (MDA) algorithm. In general, the naive way of simulating Liu's (1996) sampler by repeated sampling can be very computationally expensive and hence impractical. In Sect. 4.2.1 we show that in an example involving analysis of rank data the naive way of sampling Liu's (1996) algorithm takes too long to run to be useful in practice. In Sect. 3.2.1 we develop an alternative efficient method that can be used to effectively sample from Liu's (1996) algorithm.

The remainder of this paper is organized as follows. Section 2 contains a review of results regarding efficiency ordering and Peskun ordering of Markov chains. In Sect. 3 we provide a result improving DA chains in general state space and use it to produce efficient algorithms when the state space X, or the augmentation space Y is finite. Finally in Sect. 4, we compare the DA and our proposed algorithm in the context of two specific examples. Proofs and some technical derivations are given in the Appendices.

2 Peskun's theorem and efficiency ordering

Let $P(x, dy)$ be a Markov transition function (Mtf) on X, equipped with a countably generated σ-algebra $\mathbb{B}(\mathsf{X})$. If P is reversible with respect to a probability measure π, that is, if $\pi(dx)P(x, dx') = \pi(dx')P(x', dx)$ for all $x, x' \in \mathsf{X}$, then π is invariant for P, that is,

$$\pi(A) = \int_{\mathsf{X}} P(x, A)\pi(dx) \ \text{ for all measurable set } A.$$

Let $L^2(\pi)$ be the vector space of all real valued, measurable functions on X that are square integrable with respect to π. The inner product in $L^2(\pi)$ is defined as $\langle g, h\rangle = \int_{\mathsf{X}} g(x)\, h(x)\, \pi(dx)$. The Mtf P defines an operator on $L^2(\pi)$ through, $(Pg)(x) = \int_{\mathsf{X}} g(y)P(x, dy)$. Abusing notation, we use P to denote both the Mtf and the corresponding operator. If the Mtf P is reversible with respect to π, then for all bounded functions $g, h \in L^2(\pi)$, $\langle Pg, h\rangle = \langle g, Ph\rangle$. The spectrum of the operator P is defined as

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

For reversible P, it follows from standard linear operator theory that $\sigma(P) \subseteq [-1, 1]$.

Let $\{\eta_n\}_{n\geq 0}$ denote the Markov chain driven by P starting at η_0. If $\{\eta_n\}_{n\geq 0}$ is ψ-irreducible and Harris recurrent, that is, if it is a *positive Harris* chain, then the estimator $\bar{g}_n := \sum_{i=1}^n g(\eta_i)/n$ is strongly consistent for $E_\pi g := \int_{\mathsf{X}} g(x)\pi(dx)$, no matter how the chain is started (see Meyn and Tweedie 1993, for definition of ψ-irreducibility, and Harris recurrence). In practice, this estimator is useful if it is possible to provide an associated standard error of $\bar{g}_n$. This is where a central limit theorem (CLT) for $\bar{g}_n$ is called for, that is, we need that as $n \to \infty$,

$$\sqrt{n}(\bar{g}_n - E_\pi g) \xrightarrow{\text{d}} N(0, v(g, P)), \tag{1}$$

for some positive, finite quantity $v(g, P)$. Thus if (1) holds and $\hat{v}(g, P)$ is a consistent estimator of $v(g, P)$, then an asymptotic standard error of $\bar{g}_n$ based on MCMC sample of size n is $\hat{v}(g, P)/\sqrt{n}$. If the CLT fails to hold, then we simply write $v(g, P) = \infty$. Unfortunately, even if $g \in L^2(\pi)$, and $\{\eta_n\}_{n\geq 0}$ is positive Harris, $v(g, P)$ can still be ∞. Different sufficient conditions for CLT can be found in Jones (2004) (see also Roberts and Rosenthal 2004). Let P be reversible and let ϵ_g be the *spectral decomposition measure* (Rudin 1991) of g associated with P, then from Kipnis and Varadhan (1986) we know that

$$v(g, P) = \int_{\sigma(P)} \frac{1+\lambda}{1-\lambda} \epsilon_g(d\lambda). \tag{2}$$

In the finite state space case, that is, when the cardinality of the set X, $\#\mathsf{X} = d < \infty$, P is simply a reversible Markov transition matrix (Mtm) and $\sigma(P)$ consists of its eigenvalues (see Hobert et al. 2011, for a discussion on these ideas). In this case, the asymptotic variance $v(g, P)$ can be written as (see e.g. Brémaud 1999, p. 235)

$$v(g, P) = \sum_{i=1}^{d-1} \frac{1+\lambda_i}{1-\lambda_i} \langle g, u_i \rangle^2, \tag{3}$$

where $1 = \lambda_0 > \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{d-1} \geq -1$ are the eigenvalues of P with right eigenvectors $u_i, i = 1, 2, \ldots, d-1 (u_0 = 1)$. Here $\lambda_0 = 1 > \lambda_1$ follows because P is irreducible (see e.g. Brémaud 1999, p. 204). From (3) we see that asymptotic variance is an increasing function of the eigenvalues of the Mtm P.

Now suppose that we have two reversible positive Harris Mtf's P and Q with invariant distribution π. Hence either one of them can be used to estimate $E_\pi g$. If P and Q are similar in terms of computation effort, then we prefer the chain with smaller asymptotic variance. In general, if $v(g, P) \leq v(g, Q)$ for all g then P is said to be more efficient than Q as defined below.

Definition 1 (Mira and Geyer 1999) Let P and Q be two Mtf's with the same invariant distribution π. Then P is better than Q in the efficiency ordering, written $P \succeq_E Q$, if $v(g, P) \leq v(g, Q)$ for all $g \in L^2(\pi)$.

As mentioned in the introduction that a sufficient condition for efficiency ordering of reversible Markov chains is due to Peskun (1973) which was later extended to general state space Markov chains by Tierney (1998).

Definition 2 (Tierney 1998) Let P and Q be two Mtf's with the same invariant measure π. Then P dominates Q in the Peskun sense, written $P \succeq_P Q$, if for π-almost all x we have $P(x, A\backslash\{x\}) \geq Q(x, A\backslash\{x\})$ for all $A \in \mathbb{B}(\mathsf{X})$.

Tierney's (1998) Theorem 4 show that if P and Q are reversible with respect to π and $P \succeq_P Q$ then $P \succeq_E Q$. When X is finite, from Definition 2 we see that $P \succeq_P Q$ implies that each of the off-diagonal elements of the Mtm P is greater than or equal to the corresponding element of the Mtm Q. Mira and Geyer (1999) show that $P \succeq_P Q$ implies that the (ordered) eigenvalues P are no larger than those of Q. Since smaller eigenvalues result in smaller asymptotic variance [see the expressions of $v(g, P)$ in (2) and (3)], it follows that $P \succeq_E Q$. On the other hand, the speed of convergence (of the Markov chain driven by P) to stationarity (π) is determined by its spectral radius, $\rho(P) := \sup\{|\lambda| : \lambda \in \sigma(P)\}$ (Rosenthal 2003). The smaller the spectral radius, the faster the Markov chain converges. That is, while small asymptotic variance of time average estimators is achieved by having small eigenvalues, faster convergence of the Markov chain requires having small eigenvalues in absolute value. Since $P \succeq_P Q$ does not imply an ordering on the absolute values of the eigenvalues, P may have slower convergence than Q.

3 Improving upon the DA algorithm

We begin with a result showing how Peskun ordering can be used for improving efficiency of DA chains.

3.1 A result improving general state space DA chains

Let $f_X : \mathsf{X} \to [0, \infty)$ be a probability density function with respect to a σ-finite measure μ and $f(x, y)$ be a probability density function on $\mathsf{X} \times \mathsf{Y}$ with respect to a σ-finite measure $\mu \times \nu$. The Markov transition density (Mtd) of the DA algorithm presented in the Introduction is given by

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

Let $K(x, \cdot)$ be the corresponding Mtf. As mentioned in the Introduction, K is reversible with respect to f_X and hence f_X is invariant for K. So if K is ψ-irreducible and Harris recurrent, it can be used to estimate means with respect to f_X. A simple sufficient condition for K satisfying these conditions can be found in Hobert (2011). For each $y \in \mathsf{Y}$, let $k_y(x'|x)$ be a Mtd on X with respect to μ. Define

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

Let $\tilde{k}(x'|x)$ be a Mtd with corresponding Mtf $\tilde{K}$. We have the following proposition comparing the Mtf's K and $\tilde{K}$.

Proposition 1 *1. Suppose that for all $y \in \mathsf{Y}$, and $x' \in \mathsf{X}$,*

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

that is, $f_{X|Y}(x'|y)$ is invariant for $k_y(x'|x)$. Then $f_X(x)$ is invariant for $\tilde{k}$.

2. If for all $y \in \mathsf{Y}$, and $x, x' \in \mathsf{X}$,

$$k_y(x'|x) f_{X|Y}(x|y) = k_y(x|x') f_{X|Y}(x'|y), \tag{5}$$

that is, k_y is reversible with respect to $f_{X|Y}(x|y)$, then $\tilde{k}$ is reversible with respect to $f_X(x)$.

3. *Assume that* (5) *holds, and if for all $A \in \mathbb{B}(\mathsf{X})$*

$$\int_{A\setminus\{x\}} k_y(x'|x)\mu(dx') \geq \int_{A\setminus\{x\}} f(x'|y)\mu(dx'), \tag{6}$$

for f_X almost all $x \in \mathsf{X}$ and all $y \in \mathsf{Y}$. Then $\tilde{K} \succeq_P K$ and hence $\tilde{K} \succeq_E K$.

The proof of Proposition 1 is given in "Appendix A". A sufficient condition for (6) to hold is $k_y(x'|x) \geq f(x'|y)$ for all $x' \neq x \in \mathsf{X}$. The DA algorithm requires one to be able to draw from the two conditional densities $f_{Y|X}$, and $f_{X|Y}$. Whereas simulating $\tilde{K}$ requires a draw from $f_{Y|X}$ followed by a draw from $k_y(x'|x)$. So, if drawing from $f_{X|Y}$ is difficult and we can find $k_y(x'|x)$ which satisfies (4), we can use $\tilde{K}$ for estimating $E_{f_X} g$.

Remark 1 In Proposition 1 we replace step 2 (draw from $f_{X|Y}$) of the DA algorithm with draw from another Mtd. It is important to note that if we substitute step 1 of the DA algorithm with a draw from an Mtd $k_x(y|y')$ that is reversible with respect to $f_{Y|X}(y|x)$, the resulting chain is not a Markov chain.

It may be difficult to find $k_y(x'|x)$ satisfying the conditions (5) and (6). Liu (1996) proposed a modification to discrete state space random scan Gibbs sampler where a Metropolis–Hastings step is used to prevent staying at the current value of a coordinate for consecutive iterations. In the next two sections we show that when the state space X or the augmentation space Y is finite, we can use Liu's (1996) modified sampler to improve upon the DA algorithm.

3.2 When state space X is finite

Suppose the state space X has d elements. So K is a $d \times d$ Mtm. We consider the following Metropolis–Hastings (MH) algorithm with invariant density $f_{X|Y}(x|y)$.

Draw $x' \sim q_y(x'|x)$, where the proposal density $q_y(x'|x)$ is

$$q_y(x'|x) = \frac{f_{X|Y}(x'|y)}{1 - f_{X|Y}(x|y)} I(x' \neq x), \tag{7}$$

and $I(A)$ is the indicator function of the set A. Accept x' with probability

$$\alpha_y(x, x') = \min\left(1, \frac{1 - f_{X|Y}(x|y)}{1 - f_{X|Y}(x'|y)}\right),$$

otherwise remain at x. Our alternative algorithm, which we call modified DA (MDA) algorithm replaces step 2 of the DA algorithm presented in the Introduction by the above Metropolis–Hastings step. Let $\{\tilde{X}_n\}_{n\geq 0}$ denote the MDA chain. If $\tilde{X}_n = x$ is the current state, the following two steps are used to move to the new state $\tilde{X}_{n+1}$.

```
Iteration $n + 1$ of the MDA Algorithm:
1. Draw $Y \sim f_{Y|X}(\cdot|x)$, and call the observed value $y$.
2. Draw $\tilde{X}_{n+1}$ using a MH step with proposal density (7).
```

Note the MDA algorithm is a sub-chain of a conditional Metropolis–Hastings sampler defined in Jones et al. (2014). Define,

$$\begin{aligned} r_y(x) &= 1 - \sum_{x' \neq x} q_y(x'|x)\alpha_y(x, x') \\ &= 1 - \sum_{x' \neq x} \min\left(\frac{f_{X|Y}(x'|y)}{1 - f_{X|Y}(x|y)}, \frac{f_{X|Y}(x'|y)}{1 - f_{X|Y}(x'|y)} \right). \end{aligned}$$

Then the elements of the Mtm K_{MDA} are given by

$$\begin{aligned} k_{MDA}(x'|x) = &\int_{\mathsf{Y}} q_y(x'|x)\alpha_y(x, x') f_{Y|X}(y|x)\nu(dy) \\ &+ I(x' = x) \int_{\mathsf{Y}} r_y(x) f_{Y|X}(y|x)\nu(dy). \end{aligned}$$

The second term in the above expression is the probability that the algorithm remains at x. Since $q_y(x'|x)\alpha_y(x, x') \geq f(x'|y)$ for all $x' \neq x \in \mathsf{X}$ and all $y \in \mathsf{Y}$, the following corollary follows from Proposition 1.

Corollary 1 *The Mtm K_{MDA} is reversible with respect to $f_X(x)$, and $K_{MDA} \succeq_E K$.*

Liu et al. (1994) showed that the Markov operators corresponding to DA algorithms are *positive* (see e.g. Rudin 1991, for definition of positive operator). This implies that $\sigma(K) \subset [0, 1)$. Furthermore, from Mira and Geyer (1999), we have the following corollary.

Corollary 2 *Let λ_i and $\tilde{\lambda}_i$, $i = 1, 2, \ldots, d - 1$ be the eigenvalues of the Mtm's K and K_{MDA} respectively. Then $\lambda_i \in [0, 1)$, $\tilde{\lambda}_i \in (-1, 1)$, and $\tilde{\lambda}_i \leq \lambda_i$ for all i.*

Since the DA algorithms are known to have slow convergence, over the last two decades a great deal of effort has gone into modifying DA to speed up its convergence. The parameter expanded DA (PX-DA) algorithm of Liu and Wu (1999), and closely related conditional and marginal augmentation algorithms of Meng and van Dyk (1999) and van Dyk and Meng (2001) are alternatives to DA which often converges faster than DA algorithms. Generalizing these alternative algorithms, Hobert and Marchev (2008) recently introduced sandwich algorithms. Although Hobert and Marchev (2008) proved that the sandwich algorithms are at least as (asymptotically) efficient as the original DA algorithms, it was noted recently that even the optimal PX-DA algorithm could take millions of iterations before it provided any improvement over the DA algorithm (Roy 2014). The DA algorithms and also generally the sandwich algorithms that are used in practice are positive Markov chains leading to positive eigenvalues (Khare and Hobert 2011, p. 2587). On the other hand, the MDA algorithm can have negative eigenvalues and hence it can have superior performance than DA and sandwich algorithms in terms of asymptotic variance.

On the other hand, since MDA may have larger eigenvalues in absolute value than DA, it may have slower convergence than DA. In Sect. 4.1 we provide an example where MDA has faster convergence than DA. In fact, MDA results in iid samples in this example.

3.2.1 An efficient method for sampling Liu's (1996) algorithm

In order to efficiently run the MDA algorithm, we need fast method of sampling from $q_y(x'|x)$ defined in (7). The naive way of sampling from $q_y(x'|x)$ is to repeatedly draw from $f_{X|Y}(x'|y)$ until a value x' different from x is obtained. This method of sampling from $q_y(x'|x)$ can be very costly when $f_{X|Y}(x|y)$ is large (close to one). Below we describe an alternative recipe for the Metropolis–Hastings step in the MDA algorithm when $f_{X|Y}(x|y)$ is larger than $(\sqrt{5}-1)/2 \approx 0.618$. When $f_{X|Y}(x|y) \leq (\sqrt{5}-1)/2$, sampling from $q_y(x'|x)$ can be performed by the naive repeated sampling mentioned above.

Recipe for MH step when $f_{X|Y}(x|y) > (\sqrt{5}-1)/2$:

(i) Draw $x' \sim f_{X|Y}(\cdot|y)$. If $x' \neq x$ where x is the current value, go to (ii). Otherwise, make another draw from $f_{X|Y}(\cdot|y)$. If the new value is also equal to x, then return x as the result. Otherwise, continue to (ii).

(ii) We now have a value x' different from x. Accept and return x' as the result with probability

$$\beta_y(x, x') = \frac{1}{(1 - f_{X|Y}(x'|y))(1 + f_{X|Y}(x|y))}. \tag{8}$$

Otherwise return x.

We now explain why the above method works. Note that when $f_{X|Y}(x|y) \geq 1/2$,

$$\begin{aligned} \alpha_y(x, x') &= \frac{1 - f_{X|Y}(x|y)}{1 - f_{X|Y}(x'|y)} \text{ implying} \\ q_y(x'|x)\alpha_y(x, x') &= \frac{f_{X|Y}(x'|y)}{1 - f_{X|Y}(x'|y)} I(x' \neq x). \end{aligned} \tag{9}$$

The probability of obtaining x' (which is used in step (ii)) from step (i) is

$$f_{X|Y}(x'|y) + f_{X|Y}(x|y) f_{X|Y}(x'|y) = f_{X|Y}(x'|y)(1 + f_{X|Y}(x|y)).$$

So the probability of producing x' (different from x) as the final result is

$$f_{X|Y}(x'|y)(1 + f_{X|Y}(x|y))\beta_y(x, x') = \frac{f_{X|Y}(x'|y)}{1 - f_{X|Y}(x'|y)},$$

which is same as $q_y(x'|x)\alpha_y(x, x')$ given in (9). Hence the probability of staying back at x is $1 - \sum_{x' \neq x} q_y(x'|x)\alpha_y(x, x') = r_y(x)$. Finally, the above alternative method of performing the Metropolis Hastings step works as long as the expression $\beta_y(x, x')$ in (8) is <1. In order to establish this, note that for $x' \neq x$, $f_{X|Y}(x'|y) + f_{X|Y}(x|y) \leq 1$, that is, $(1 - f_{X|Y}(x'|y)) \geq f_{X|Y}(x|y)$, implying that

$$\beta_y(x, x') \leq \frac{1}{f_{X|Y}(x|y)(1 + f_{X|Y}(x|y))},$$

which is <1 since $f_{X|Y}(x|y) > (\sqrt{5} - 1)/2$.

3.3 When augmentation space Y is finite

Next, we consider the case when the parameter space X is uncountable, but the augmentation space Y is finite. An example of this situation is the Bayesian mixture models as discussed in Hobert et al. (2011). In this case, we consider the so-called conjugate Markov chain that lives on Y and makes transition $y \to y'$ with probability

$$k^*(y'|y) = \int_{\mathsf{X}} f_{Y|X}(y'|x) f_{X|Y}(x|y)\mu(dx).$$

Straightforward calculations show that f_Y is the invariant density of k^*, where f_Y is the y-marginal density of $f(x, y)$. Hobert et al. (2011) showed that if $|\mathsf{X}| = \infty$, and $|\mathsf{Y}| = d < \infty$, then $\sigma(K)$ consists of the points $\{0\}$ together with the $d - 1$ eigenvalues of the Mtm K^* associated with the conjugate chain. Since Y is finite, we can use Liu's (1996) modified algorithm, as in Sect. 3.2, to construct an MDA Mtm K^*_{MDA} which is more efficient than K^*. That is, to estimate means with respect to f_Y we prefer K^*_{MDA} over K^*. Below we show that a Rao-Blackwellized estimator based on K^*_{MDA} is more efficient than the time average estimator based on the DA algorithm K.

As before suppose we are interested in estimating $E_{f_X} g$ for some function $g : \mathsf{X} \to \mathbb{R}$. Now

$$E_{f_X} g = E_{f_Y}[E_{f_{X|Y}}(g(X)|y)] = E_{f_Y} h,$$

where $h(y) := E_{f_{X|Y}}(g(X)|y)$. If h is available in closed form, then we can use K^*_{MDA} to estimate $E_{f_Y} h$, that is, $E_{f_X} g$.

Proposition 2 *The Markov chain driven by K^*_{MDA} is more efficient than the DA algorithm K, for estimating $E_{f_X} g$, that is, $v(h, K^*_{MDA}) \leq v(g, K)$, for all $g \in L^2(f_X)$ where $h(y) = E_{f_{X|Y}}(g(X)|y)$.*

Proof From Liu et al. (1994) we know that $v(h, K^*) \leq v(g, K)$. Then the proposition follows since $v(h, K^*_{MDA}) \leq v(h, K^*)$ by Proposition 1. □

Remark 2 It is known that Peskun's criterion as defined in Definition 2 can not be used for comparing Mtf's for which $P(x, \{x\}) = 0$ for every x in the state space. For example, as mentioned in (Mira and Geyer 1999, p. 14), Gibbs samplers with continuous full conditionals cannot be compared using Peskun ordering. In Proposition 2 we have constructed more efficient estimators than time averages based on the DA chain even when the state space X is continuous.

4 Examples

In this section we consider two examples—the beta-binomial model, and a model for analyzing rank data. In the first example we consider two situations—the state space X is finite, the augmentation space Y is infinite and $|\mathsf{X}| = \infty$, but $|\mathsf{Y}| < \infty$. In the second example, X is finite and the augmentation space Y is infinite.

4.1 Beta-binomial model

Consider the following beta-binomial model

$$f(x, y) \propto \binom{n}{x} y^{x+\alpha-1}(1-y)^{n-x+\beta-1}, \; x = 0, 1, \ldots, n; 0 \leq y \leq 1,$$

from Casella and George (1992) who were interested in calculating some characteristics of the marginal distribution f_X based on the DA chain. The two conditionals used in the DA chain are standard distributions. In fact, $f_{X|Y}$ is Binomial (n, y) and $f_{Y|X}$ is Beta $(x + \alpha, n - x + \beta)$. The transition probabilities of the DA chain are given by

$$\begin{aligned} k(x'|x) &= \int_0^1 \binom{n}{x'} y^{x'}(1-y)^{n-x'} \frac{y^{x+\alpha-1}(1-y)^{n-x+\beta-1}}{B(x+\alpha, n-x+\beta)} dy \\ &= \frac{\binom{n}{x'} B(x+x'+\alpha, 2n-(x+x')+\beta)}{B(x+\alpha, n-x+\beta)}, \end{aligned}$$

where $B(\cdot, \cdot)$ is the beta function.

Liu (1996) in an associated Technical report (Metropolized Gibbs Sampler: An Improvement) considered the above example and by comparing autocorrelation plots in the case $n = 1 = \alpha = \beta$, he conjectured that the MDA algorithm is more efficient than the standard Gibbs sampler. Below we show that it is indeed the case. Since $n = 1$, the state space of the DA chain is $\{0, 1\}$ and $f_X(0) = 1/2 = f_X(1)$. Simple calculations show that the Mtm's of the DA and the MDA algorithms are given by

$$K = \begin{pmatrix} 2/3 & 1/3 \\ 1/3 & 2/3 \end{pmatrix} \text{ and } K_{MDA} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}.$$

So, the MDA algorithm produces iid draws from the invariant distribution in this case. This explains why the autocorrelations for MDA chain "dropped quickly to zero in two iterations" as observed in the above mentioned Technical report. Note that in this example $\tilde{\lambda}_1 = 0 < \lambda_1 = 1/3$. Suppose we want to estimate $E(X)$. Since MDA results in iid samples from f_X, $v(X, K_{MDA}) = \mathrm{Var}_{f_X}(X) = 1/4$. On the other hand, we have (see e.g. Brémaud 1999, p. 233)

$$v(X, K) = \mathrm{Var}_{f_X}(X) + 2\langle x, (\mathbf{Z} - I)x\rangle, \tag{10}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in $L^2(f_X)$, $\mathbf{Z} \equiv (I - (K - F))^{-1}$, with $F = (1, 1)^T (f_X(0), f_X(1))$. Since

$$K - F = \begin{pmatrix} 1/6 & -1/6 \\ -1/6 & 1/6 \end{pmatrix} \;\Rightarrow\; Z = \begin{pmatrix} 5/4 & -1/4 \\ -1/4 & 5/4 \end{pmatrix},$$

we have $\langle x, (\mathbf{Z} - I)x\rangle = 1/8$. Then from (10) we have $v(X, K) = 1/4 + 2/8 = 1/2$, and so $v(X, K)/v(X, K_{MDA}) = 2$. Thus MDA algorithm in this case is twice as efficient as the DA algorithm for estimating $E(X)$.

Next we consider estimating $E_{f_Y}(Y)$. In this case f_Y plays the role of the target density f_X from the Introduction and the DA chain is denoted by $\{Y_n\}_{n\geq 0}$. Here, the marginal density f_Y is simply a uniform distribution in (0, 1). Note that, $v(Y, K) = \mathrm{Var}_{f_Y}(Y) + 2\sum_{k=1}^{\infty} \mathrm{Cov}(Y_0, Y_k)$. We can calculate the lag-k autocovariances using the following formula given in Liu et al. (1994)

$$\mathrm{Cov}(Y_0, Y_k) = \mathrm{Var}(E(\cdots E(E(Y|X)|Y)\cdots)),$$

where the expression in the right hand side has k conditional expectations alternately with respect to $f_{Y|X}$ and $f_{X|Y}$. Then $\mathrm{Cov}(Y_0, Y_1) = \mathrm{Var}(E(Y|X)) = \mathrm{Var}((X + 1)/3) = (1/3^2)(1/4)$, $\mathrm{Cov}(Y_0, Y_2) = \mathrm{Var}(E(E(Y|X)|Y)) = \mathrm{Var}((Y + 1)/3) = (1/3^2)(1/12)$, so on. In general, $\mathrm{Cov}(Y_0, Y_{2k-1}) = (1/3^{2k})(1/4)$ and $\mathrm{Cov}(Y_0, Y_{2k}) = (1/3^{2k})(1/12)$ for $k = 1, 2, \ldots$. So

$$v(Y, K_{DA}) = \frac{1}{12} + 2\left[\frac{1}{3^2}\frac{1}{4} + \frac{1}{3^2}\frac{1}{12} + \frac{1}{3^4}\frac{1}{4} + \frac{1}{3^4}\frac{1}{12} + \cdots\right] = \frac{1}{12} + \frac{1}{12} = \frac{1}{6}.$$

In this case the support of the target density f_Y is (0, 1), which is not finite. So we can not use the approach mentioned in Sect. 3.2 to improve the time average estimator $\sum_{i=1}^{n} Y_i/n$. On the other hand, since $h(x) = E(Y|X = x) = (x + 1)/3$, is available in closed form and the augmentation space {0, 1} is finite, we can use the Rao-Blackwellized MDA estimator discussed in Sect. 3.3 to estimate $E_{f_Y}(Y)$. Since MDA results in iid draws, $v(h, K^*_{MDA}) = \mathrm{Var}_{f_X}((X + 1)/3) = (1/3^2)(1/4) = 1/36$. So, $v(Y, K)/v(h, K^*_{MDA}) = 6$, that is, our proposed estimator is six times more efficient than the standard estimator $\sum_{i=1}^{n} Y_i/n$ of $E_{f_Y}(Y)$ based on the DA chain. On the other hand, using (10) for the function $h(X)$, we see that the asymptotic variance of the Rao-Blackwellized estimator of Y based on the conjugate chain is $v(h, K^*) = 1/18$, that is, MDA is only twice more efficient than this estimator.

4.2 Bayesian analysis of rank data

In many examples of rank data, like ranking of students for their proficiency in a given subject, one may use some objective criterion like the marks scored in an appropriately designed test for determining their rank. However, in other situations, like the case of evaluating a job applicant, usually different attributes are considered. Typically, a panel of judges evaluate the candidates (items) on a variety of relevant aspects, some of which may be objective, and others are subjective. The whole group of experts then tries to arrive at a consensus ranking based on all the rankings by the individual experts through some subjective decision making process. Laha and Dongaonkar (2009) call this generally accepted rank as the "true rank". Consider a situation in which p items are ranked by a random sample of m judges from a population. Let $\mathfrak{S}_p$ be the set (group) of permutations of the integers $1, 2, \ldots, p$. Laha and Dongaonkar (2009) assume that the observed ranks z_i's are "perturbed" versions of the true rank π of the p items. Formally, $z_i = \sigma_i \circ \pi$, where $\sigma_i \in \mathfrak{S}_p$, for $i = 1, 2, \ldots, m$ and $\circ$ denotes the composition operation on $\mathfrak{S}_p$, that is, the observed ranks z_i's are considered permutations of the true rank π. The permutation σ_i plays the role of "error" analogous to ϵ in the linear model $z = \mu + \epsilon$.

Often there are covariates on the experts and the true rank depends on the value of the covariate. Recently, Laha et al. (2013) generalized the above model to incorporate covariates. They assume that $z_i = \sigma_i \circ \pi(x_i)$, where $\pi(x_i)$ is the true rank when the covariate is x_i and x_i falls in one of the c categories numbered 1 through c, that is, $x_i \in \{1, 2, \ldots, c\}$. Denoting the $p!$ possible rankings as $\zeta_1, \zeta_2, \ldots, \zeta_{p!}$, ($\zeta_1$ being the identity permutation) and assuming that the errors σ_i's are iid having multinomial distribution $\sigma_i \sim \text{Mult}(1; \theta_1\theta_2, \ldots, \theta_{p!})$ with $\theta_i \equiv \theta_{\zeta_i}$, the likelihood function is given by

$$\ell(\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{p!}\prod_{j=1}^{c} \theta_{\zeta_i \circ \pi(j)^{-1}}^{m_{ij}},$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_{p!})$, $\boldsymbol{\pi} = (\pi(1), \pi(2), \ldots, \pi(c))$, and m_{ij} is the number of times ranking ζ_i is given by respondents in the jth category. A conjugate prior for the parameter $\boldsymbol{\theta}$ is the Dirichlet distribution. Let $p(\boldsymbol{\theta}) \propto \prod_{i=1}^{p!} \theta_i^{a_i-1}$ be the prior on $\boldsymbol{\theta}$ with some suitably chosen hyperparameters a_i's. We assume that $a_i \propto 2^{-d_C(\zeta_i, \zeta_1)}$, where $d_C(\cdot, \cdot)$ is the Cayley's distance in $\mathfrak{S}_p$. (See Laha et al. 2013, for a more general choice for this hyperparameter.) Assume that the prior on $\boldsymbol{\pi}$ is $p(\boldsymbol{\pi}) = \prod_{j=1}^{c} p(\pi(j))$ where a uniform prior is specified on $\pi(j)$'s. Then the posterior distribution is given by

$$p(\boldsymbol{\theta}, \boldsymbol{\pi}|z) \propto \ell(\boldsymbol{\theta}, \boldsymbol{\pi}) p(\boldsymbol{\theta}) \propto \prod_{i=1}^{p!}\prod_{j=1}^{c} \theta_{\zeta_i \circ \pi(j)^{-1}}^{m_{ij}} \prod_{i=1}^{p!} \theta_i^{a_i-1} = \prod_{k=1}^{p!} \theta_k^{M_k(\boldsymbol{\pi})+a_k-1},$$

where $M_k(\boldsymbol{\pi}) = \sum_{i=1}^{p!}\sum_{j=1}^{c} m_{ij} I(\zeta_i \circ \pi(j)^{-1} = \zeta_k)$. Since the conditional density $p(\boldsymbol{\pi}|\boldsymbol{\theta}, z)$ is product of multinomial distributions and $p(\boldsymbol{\theta}|\boldsymbol{\pi}, z)$ is Dirichlet distribution, we can construct a DA algorithm. Indeed from Laha et al. (2013) we have that the conditional distribution of $\boldsymbol{\theta}$ given $\boldsymbol{\pi}$ and z is given by

$$p(\boldsymbol{\theta}|\boldsymbol{\pi}, z) \propto \prod_{k=1}^{p!} \theta_k^{M_k(\boldsymbol{\pi})+a_k-1},$$

and conditional on $\boldsymbol{\theta}$ and z, $\pi(1), \ldots, \pi(c)$ are independent with

$$p(\pi(j) = \zeta_r|\boldsymbol{\theta}, z) \propto \prod_{k=1}^{p!} \theta_k^{\sum_i m_{ij}\mathbb{I}(\zeta_i=\zeta_k \circ \zeta_r)}, \quad r = 1, \ldots, p!.$$

Note that here the roles of x and y from the Introduction, are being played by $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$, respectively. For an integer r larger than 1, denote the standard $(r-1)$ simplex by

$$S_r := \left\{(t_1, t_2, \ldots, t_r) \in \mathbb{R}^r : t_i \in [0, 1] \text{ and } t_1 + \cdots + t_r = 1\right\}. \tag{11}$$

Then the transition probabilities of the DA chain are given by

$$k(\boldsymbol{\pi}'|\boldsymbol{\pi}) = \int_{S_{p!}} p(\boldsymbol{\pi}'|\boldsymbol{\theta}, z)p(\boldsymbol{\theta}|\boldsymbol{\pi}, z)d\boldsymbol{\theta},$$

Both using a simulation study and real data examples, Laha et al. (2013) showed that the DA algorithm converges slowly especially when sample size m is large. Due to intractability of the marginal posterior density $p(\boldsymbol{\theta}|z)$, a computationally efficient sandwich algorithm as described in Hobert and Marchev (2008) is not available in this example. We now show that the technique proposed in Sect. 3.2 can be used to improve the efficiency of the DA chain.

We consider a special case where $c = 2 = p$, that is, there are two categories, two items to rank and we assume that $m_{ij} = m/4$ for all i, j. Doing some algebra, it can be shown that in this case $k_{ij} = 1/4$ for all i, j for any value of the hyper parameters a_i's. That is, the DA algorithm produces iid draws from the marginal posterior density $p(\boldsymbol{\pi}|y)$, which is, in this special case, a uniform distribution on $\{(\zeta_i, \zeta_j) : i, j = 1, 2\}$. Since the state space is finite with cardinality $(p\,!)^c$, we can construct the MDA algorithm in this example. Note that the cardinality of the augmentation space here is infinite. Using the formula for the elements of K_{MDA}, that is, $\tilde{k}_{ij}$ given in "Appendix B", we observe that

$$K_{MDA} = (1/3)J - (1/3)I,$$

where J is the 4×4 matrix of ones. Note that, the Mtm K_{MDA} can be obtained by moving the diagonal probabilities of the DA Mtm K uniformly to the three off-diagonal elements. Suppose we want to estimate $P(\pi(1) = \zeta_1) = E_{p(\boldsymbol{\pi}|y)}[g(\boldsymbol{\pi})]$, where $g(\boldsymbol{\pi}) = I(\pi(1) = \zeta_1)$. Since $\tilde{\lambda}_i = -1/3$ for $i = 1, 2, 3$, the spectral radius of K_{MDA} is $1/3$. Hence, MDA has slower convergence than DA, which produces iid draws in this special case. But, as we show now, K_{MDA} is twice as efficient as K for estimating $E_{p(\boldsymbol{\pi}|y)}[g(\boldsymbol{\pi})]$. Since DA results in iid draws, $v(g, K) = \text{Var}(g(\boldsymbol{\pi})|y) = 1/4$. From (10) we know that $v(g, K_{MDA}) = \text{Var}(g(\boldsymbol{\pi})|y) + 2\langle g, (\mathbf{Z} - I)g\rangle$, where $\langle \cdot, \cdot \rangle$ denotes the inner product in $L^2(p(\boldsymbol{\pi}|y))$, and $\mathbf{Z} \equiv (I - (K_{MDA} - P(\boldsymbol{\pi}|y)))^{-1}$, with $P(\boldsymbol{\pi}|y) = (1/4)\mathbf{11}^T$. Doing some algebra we see that

$$Z = (1/16)\mathbf{11}^T + (3/4)I \quad \Rightarrow \quad 2\langle g, (\mathbf{Z} - I)g\rangle = -1/8.$$

So $v(g, K_{MDA}) = 1/4 - 1/8 = 1/8$ and hence $v(g, K)/v(g, K_{MDA}) = 2$. Thus as in the previous section, MDA algorithm in this special case is twice as efficient as the DA algorithm.

We could not carry out closed form calculations for $v(g, K_{DA})$ and $v(g, K_{MDA})$ when $m_{ij} \neq m/4$ for some i, j. In this case, we compare DA and MDA chains using numerical approximation. Laha et al. (2013) used simulation study to demonstrate the slow convergence of DA chains. In their simulation study, they considered $p = 2$, and $c = 2$, as above and they let the number of observations, m, vary. The true ranks for the two categories are assumed to be ζ_1 and ζ_2 respectively. The true value of $\boldsymbol{\theta}$ is taken to be $(0.7, 0.3)$. The small values of p and c allow us to compute the Markov transition probabilities in closed form. We let the sample size m vary between 20 and 60 in increments of 10, and equal size of sample is taken from the two categories. For example, if $m = 20$, then we simulate 10 observations from category 1 and 10 observations from category 2. For each fixed sample size, the simulation is repeated 1000 times. From Sect. 2 we know that the second largest eigenvalue (in absolute value) of a Mtm shows the speed of convergence of the chain. In fact, Laha et al. (2013) used the box plot of the 1000 λ_1 values (corresponding to 1000 Mtm's based on repeated simulations) to show that the largest eigenvalues of the DA chain tends to one as the sample size increases, that is, the DA algorithm slows down with increasing sample size.

For the same simulation setting mentioned above, we calculate the eigenvalues of the K_{MDA}. We calculate the entries of the K_{MDA} by numerical integration using the expressions given in the "Appendix B". Figure 1 shows the boxplots of the eigenvalues of K_{MDA} matrices corresponding 1000 simulated data. From the top plot in Fig. 1 we see that the largest eigenvalues tends to one as the sample size increases, that is, both DA and MDA algorithms slow down with increasing sample size. But, the MDA results in smaller (even negative) eigenvalues resulting in smaller asymptotic variance.

Next we compare the performance of DA and MDA algorithms in a real data example.

4.2.1 Tyne–Wear metropolitan district council election data

It is of interest to see whether the position of a candidate's name on the ballot paper has any effect in terms of the number of votes which he receives. We consider a study presented in Brook and Upton (1974, p. 415) regarding local government election in the Tyne–Wear area. Consider a particular party fielding three candidates for this election and label them as a, b and c in the order in which these candidates appear on the ballot paper. A particular outcome in terms of votes received can be expressed as

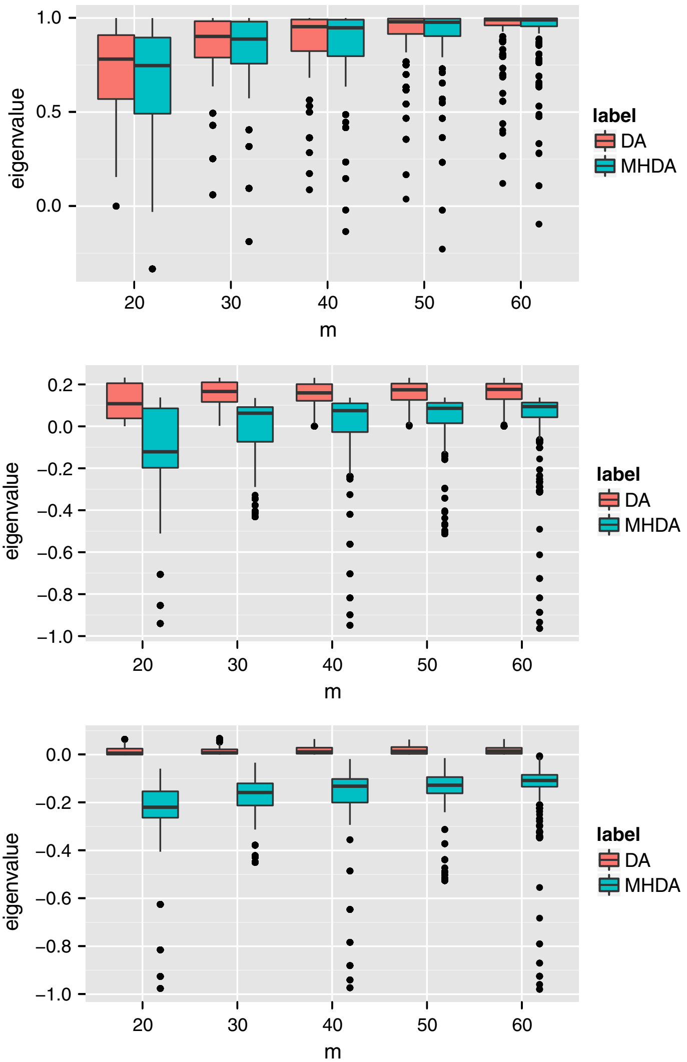

Fig. 1 The behavior of the eigenvalues for the DA and MDA chains. The graph shows how the eigenvalues of these chains change with sample size, *m* and also how MDA results in smaller eigenvalues

a permutation such as *bac* which means that candidate b has received the maximum number of votes and c the least. The data aggregated over all parties with three candidates and over all wards in the Tyne–Wear area is given in Brook and Upton (1974). We reproduce the data in Table 1 for ready reference.

In this example, $p = 3$, $c = 6$, and thus the cardinality of the state space is $(3!)^6 = 46,656$. Laha et al. (2013) noted that for all areas except the second (Wigan, Bolton, Bury, Rochdale), $\zeta_1 = abc$ is the mode of the posterior distributions of the true ranks. The mode of the posterior distribution of $\pi(2)$ is $\zeta_6 = cba$. We consider the functions $g(\boldsymbol{\pi}) = I(\pi(2) = \zeta_6, \pi(i) = \zeta_1; i = 1, 3, 4, 5, 6)$. We ran both DA and MDA

Table 1 The Tyne–Wear metropolitan district council election data from Brook and Upton (1974, pp. 415)

		Areas						
Lex. order	Order	1	2	3	4	5	6	Total
ζ_1	abc	46	32	38	27	50	39	232
ζ_2	acb	22	20	20	25	27	22	136
ζ_3	bac	32	15	29	32	27	39	174
ζ_4	bca	23	32	23	24	26	23	151
ζ_5	cab	17	25	13	13	23	23	114
ζ_6	cba	25	26	13	16	35	26	141
Total		165	150	136	137	188	172	948

Areas are as follows:
1. Tyneside and Wearside
2. Wigan, Bolton, Bury, Rochdale
3. Salford, Manchester, Oldham
4. Trafford Stockport, tameside
5. Kent and Worcestershire (June 10th, 1973)
6. An assortment of results from various sources including Birmingham, Doncaster and Merseyside

algorithms and used batch means method for estimating $v(g, K)$ and $v(g, K_{MDA})$. Based on 50,000 iterations of these algorithms started at $\pi(2) = 6, \pi(j) = 1$, for $j = 1, 3, 4, 5, 6$, we observed $\hat{v}(g, K)/\hat{v}(g, K_{MDA}) = 1.47$.

Next we assess the performance of the alternative method presented in Sect. 3.2.1 for sampling from Liu's (1996) sampler. We observe that in the analysis of Tyne–Wear election data, the MDA algorithm, using the naive repeated sampling for MH step takes similar amount of time to run as the MDA algorithm using the method in Sect. 3.2.1. The reason that the repeated sampling method does not perform poorly in this example is because the conditional density $p(\boldsymbol{\pi}|\boldsymbol{\theta}, z)$ never took any large value. Indeed, it never took a value larger than 0.71 in 50,000 iterations. We then consider a fictitious data to assess the performance of the sampler in Sect. 3.2.1 when the conditional density takes larger values. As in the above data set, we take $p = 3, c = 6$. Let the data be $m_{1j} = 50, m_{2j} = 40, m_{3j} = 20, m_{4j} = 10 = m_{5j}$, and $m_{6j} = j$ for $j = 1, 2, \ldots, 6$. Starting at $\pi(j) = 1$ for all $j = 1, 2, \ldots, 6$, it took 23.9 and 30.2 s on an old Intel Q9550 2.83GHz machine running Windows 7 to run 50,000 iterations of DA and MDA (using the method in Sect. 3.2.1) respectively. (The codes were written in R (R Development Core Team 2011).) Whereas, the MDA using repeated sampling did not complete 50,000 iterations even after 30 hours. Indeed in one iteration when the value of the conditional density $p(\boldsymbol{\pi}|\boldsymbol{\theta}, z)$ was 0.999978, the repeated sampling made 196,760 draws before producing a vector $\boldsymbol{\pi}'$ different from the current $\boldsymbol{\pi}$.

5 Discussions

Each iteration of a DA algorithm consists of draws from two conditional distributions. In this paper, it is shown that if the draw from the second conditional distribution is replaced with a draw from an appropriate Mtf, the resulting Markov chain is always at least as efficient as the DA chain. When either the state space or the augmentation space is finite, using Liu's (1996) sampler an algorithm, called MDA, is constructed that is more efficient than the DA algorithm. Since the naive method for implementing Liu's (1996) sampler can be impractical, an efficient alternative method is proposed. This alternative method for Liu's (1996) sampler is not specific to the MDA algorithm, and it can be used anywhere Liu's (1996) algorithm is implemented. It would be interesting to construct an improved algorithm following Proposition 1 when both the state space and the augmentation space are infinite.

Acknowledgments The author thanks Radford Neal for suggesting the algorithm presented in Sect. 3.2.1 for the case $f_{X|Y}(x|y) > (\sqrt{5}-1)/2$. The author also thanks two reviewers for their suggestions that have improved the paper.

Appendices

Appendix A: Proof of Proposition 1

Proof To show $f_X(x)$ is invariant for $\tilde{K}$, note that

$$\begin{aligned}\int_{\mathsf{X}} \tilde{k}(x'|x) f_X(x)\mu(dx) &= \int_{\mathsf{X}}\int_{\mathsf{Y}} k_y(x'|x) f_{Y|X}(y|x)\nu(dy) f_X(x)\mu(dx)\\ &= \int_{\mathsf{Y}}\int_{\mathsf{X}} k_y(x'|x) f_{X|Y}(x|y)\mu(dx) f_Y(y)\nu(dy)\\ &= \int_{\mathsf{Y}} f_{X|Y}(x'|y) f_Y(y)\nu(dy) = f_X(x')\end{aligned}$$

where the third equality follows from (4).

Next assume that k_y is reversible with respect to $f_{X|Y}(x|y)$, that is, (5) holds. Then

$$\begin{aligned}\tilde{k}(x'|x) f_X(x) &= \int_{\mathsf{Y}} k_y(x'|x) f_{Y|X}(y|x)\nu(dy) f_X(x)\\ &= \int_{\mathsf{Y}} k_y(x'|x) f_{X|Y}(x|y) f_Y(y)\nu(dy)\\ &= \int_{\mathsf{Y}} k_y(x|x') f_{X|Y}(x'|y) f_Y(y)\nu(dy)\\ &= \int_{\mathsf{Y}} k_y(x|x') f_{Y|X}(y|x')\nu(dy) f_X(x') = \tilde{k}(x|x') f_X(x'),\end{aligned}$$

that is, $\tilde{k}$ is is reversible with respect to f_X.

Since (6) is in force, for all $A \in \mathbb{B}(\mathsf{X})$ and for f_X almost all $x \in \mathsf{X}$ we have

$$\tilde{K}(x, A\backslash\{x\}) = \int_{A\backslash\{x\}} \tilde{k}(x'|x)\mu(dx') = \int_{A\backslash\{x\}}\int_{\mathsf{Y}} k_y(x'|x) f_{Y|X}(y|x)\nu(dy)\mu(dx')$$

$$\begin{aligned}&\geq \int_{\mathsf{Y}} \int_{A\backslash\{x\}} f_{X|Y}(x'|y)\mu(dx') f_{Y|X}(y|x)\nu(dy)\\ &= K(x, A\backslash\{x\}),\end{aligned}$$

that is, $\tilde{K} \succeq_P K$. □

Appendix B: The Mtm K_{MDA} when $p = 2, c = 2$

We order the points in the state space as follows: (ζ_1, ζ_1), (ζ_1, ζ_2), (ζ_2, ζ_1), and (ζ_2, ζ_2). We denote the entries of K_{MDA} by $\tilde{k}_{ij}$. So, for example, the element $\tilde{k}_{23}$ is the probability of moving from (ζ_1, ζ_2) to (ζ_2, ζ_1). In order to write down the expressions for $\tilde{k}_{ij}$ we need to introduce some notations. Recall that m_{ij} denotes the number of observations in the jth category with rank ζ_i for $i, j = 1, 2$. Let $m_{i.} = m_{i1} + m_{i2}$ for $i = 1, 2$, $m_d = m_{11} + m_{22}$, and $m_{od} = m_{12} + m_{21}$. Finally, for fixed $w \in (0, 1)$, let

$$A(w) = [w^{m_{1.}}(1-w)^{m_{2.}} + w^{m_{2.}}(1-w)^{m_{1.}} + w^{m_d}(1-w)^{m_{od}} + w^{m_{od}}(1-w)^{m_d}],$$

and $c = 1/\mathrm{B}(m_{1.} + a_1, m_{2.} + a_2)$. Recall that a_1, a_2 are the hyper parameters of the prior of θ. Below, we provide the the expressions for $\tilde{k}_{1j}$, for $j = 1, \ldots, 4$. The other rows of K_{MDA} can be found similarly. From Sect. 3.2 we know that

$$\begin{aligned}\tilde{k}_{12} = &\int_{S_2} \frac{p(\pi = (\zeta_1, \zeta_2)|\theta, y)}{1 - p(\pi = (\zeta_1, \zeta_1)|\theta, y)}\\ &\times \min\left(1, \frac{1 - p(\pi = (\zeta_1, \zeta_1)|\theta, y)}{1 - p(\pi = (\zeta_1, \zeta_2)|\theta, y)}\right) p(\theta|\pi = (\zeta_1, \zeta_1), y) d\theta\end{aligned}$$

Straightforward calculations show that if $m_{12} \geq m_{22}$ then

$$p(\pi = (\zeta_1, \zeta_1)|\theta, y) > p(\pi = (\zeta_1, \zeta_2)|\theta, y) \;\Leftrightarrow\; \theta_1 > 1/2.$$

On the other hand, if $m_{12} < m_{22}$ then

$$p(\pi = (\zeta_1, \zeta_1)|\theta, y) > p(\pi = (\zeta_1, \zeta_2)|\theta, y) \;\Leftrightarrow\; \theta_1 < 1/2.$$

Simple calculations show that if $m_{12} \geq m_{22}$, then

$$\begin{aligned}\tilde{k}_{12} = c\Bigg[&\int_0^{1/2} \frac{w^{m_d + m_{1.} + a_1 - 1}(1-w)^{m_{od} + m_{2.} + a_2 - 1}}{A(w) - w^{m_{1.}}(1-w)^{m_{2.}}} dw\\ &+ \int_{1/2}^{1} \frac{w^{m_d + m_{1.} + a_1 - 1}(1-w)^{m_{od} + m_{2.} + a_2 - 1}}{A(w) - w^{m_d}(1-w)^{m_{od}}} dw\Bigg].\end{aligned}$$

In the case of $m_{12} < m_{22}$, the range of integration in the above two terms are interchanged. Similarly, we find that the expression for $\tilde{k}_{13}$ depends on whether $m_{11} \geq m_{21}$

or $m_{11} < m_{21}$. If $m_{11} \geq m_{21}$,

$$\tilde{k}_{13} = c\left[\int_0^{1/2} \frac{w^{m_{od}+m_{1.}+a_1-1}(1-w)^{m_d+m_{2.}+a_2-1}}{A(w) - w^{m_{1.}}(1-w)^{m_{2.}}}dw \right.$$
$$\left. + \int_{1/2}^{1} \frac{w^{m_{od}+m_{1.}+a_1-1}(1-w)^{m_d+m_{2.}+a_2-1}}{A(w) - w^{m_{od}}(1-w)^{m_d}}dw\right],$$

and the ranges of integration in the above two terms are interchanged when $m_{11} < m_{21}$. Lastly, if $m_{1.} \geq m_{2.}$,

$$\tilde{k}_{14} = c\left[\int_0^{1/2} \frac{w^{m+a_1-1}(1-w)^{m+a_2-1}}{A(w) - w^{m_{1.}}(1-w)^{m_{2.}}}dw \right.$$
$$\left. + \int_{1/2}^{1} \frac{w^{m+a_1-1}(1-w)^{m+a_2-1}}{A(w) - w^{m_{2.}}(1-w)^{m_{1.}}}dw\right],$$

where $m = m_{1.} + m_{2.}$ is the number of observations and as before the ranges of integration are interchanged when $m_{1.} < m_{2.}$. Finally, $\tilde{k}_{11}$ is set to $1 - \sum_{j=2}^{4} \tilde{k}_{1j}$.

References

Brémaud P (1999) Markov chains Gibbs fields, Monte Carlo simulation, and queues. Springer, New York

Brook D, Upton GJG (1974) Biases in local government elections due to position on the ballot paper. Appl Stat 23:414–419

Casella G, George E (1992) Explaining the Gibbs sampler. Am Stat 46:167–174

Hobert JP (2011) The data augmentation algorithm: theory and methodology. In: Brooks S, Gelman A, Jones GL, Meng X-L (eds) Handbook of Markov chain Monte Carlo. CRC Press, Boca Raton, pp 253–293

Hobert JP, Marchev D (2008) A theoretical comparison of the data augmentation, marginal augmentation and PX-DA algorithms. Ann Stat 36:532–554

Hobert JP, Roy V, Robert CP (2011) Improving the convergence properties of the data augmentation algorithm with an application to Bayesian mixture modelling. Stat Sci 26:332–351

Jones GL (2004) On the Markov chain central limit theorem. Probab Surv 1:299–320

Jones GL, Roberts GO, Rosenthal J (2014) Convergence of conditional Metropolis–Hastings samplers. Adv Appl Probab 46:422–445

Khare K, Hobert JP (2011) A spectral analytic comparison of trace-class data augmentation algorithms and their sandwich variants. Ann Stat 39:2585–2606

Kipnis C, Varadhan SRS (1986) Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. Commun Math Phys 104:1–19

Laha A, Dutta S, Roy V (2013) A novel sandwich algorithm for empirical Bayes analysis of rank data. Tech. rep. Indian Institute of management, Ahmedabad

Laha A, Dongaonkar S (2009) Bayesian analysis of rank data using SIR. In: Sengupta A (ed) Advances in multivariate statistical methods. World Scientific Publishers, Singapore, pp 327–335

Liu JS (1996) Peskun's theorem and a modified discrete-state Gibbs sampler. Biometrika 83:681–682

Liu JS, Wong WH, Kong A (1994) Covariance structure of the Gibbs sampler with applications to comparisons of estimators and augmentation schemes. Biometrika 81:27–40

Liu JS, Wu YN (1999) Parameter expansion for data augmentation. J Am Stat Assoc 94:1264–1274

Meng X-L, van Dyk DA (1999) Seeking efficient data augmentation schemes via conditional and marginal augmentation. Biometrika 86:301–320

Meyn SP, Tweedie RL (1993) Markov chains and stochastic stability. Springer, London

Mira A, Geyer CJ (1999) Ordering Monte Carlo Markov chains. Tech. rep. no. 632. School of Statistics, University of Minnesota

Peskun PH (1973) Optimum Monte Carlo sampling using Markov chains. Biometrika 60:607–612

R Development Core Team (2011) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. ISBN 3-900051-07-0, http://www.R-project.org

Roberts GO, Rosenthal JS (2004) General state space Markov chains and MCMC algorithms. Probab Surv 1:20–71

Rosenthal JS (2003) Asymptotic variance and convergence rates of nearly-periodic Markov chain Monte Carlo algorithms. J Am Stat Assoc 98:169–177

Roy V (2014) Efficient estimation of the link function parameter in a robust Bayesian binary regression model. Comput Stat Data Anal 73:87–102

Rudin W (1991) Functional analysis, 2nd edn. McGraw-Hill, New York

Tanner MA, Wong WH (1987) The calculation of posterior distributions by data augmentation (with discussion). J Am Stat Assoc 82:528–550

Tierney L (1998) A note on Metropolis–Hastings kernels for general state spaces. Ann Appl Probab 8:1–9

van Dyk DA, Meng X-L (2001) The art of data augmentation (with discussion). J Comput Graph Stat 10:1–50