An Overview of the Data Augmentation Algorithm

Grant Backlund

Department of Statistics, University of Florida

March 2017

Abstract

Markov chain Monte Carlo algorithms provide a way of approximately sampling from complicated probability distributions in high dimensions. The data augmentation algorithm is a popular MCMC method which is easy to implement but sometimes suffers from slow convergence. In this report, an overview of the data augmentation algorithm is given, along with a description of two variants that can often result in dramatic improvements in the convergence rates of the underlying Markov chains. A general method based on operator theory is presented to facilitate a theoretical comparison of the convergence rates associated with the above algorithms. The results are illustrated using the Bayesian probit regression model analyzed by Albert and Chib (1993).

1 Background and Motivation for MCMC

1.1 Classical Monte Carlo Methods

Suppose we are given a probability distribution $\pi(\cdot)$ defined on a measurable space $(\mathsf{X}, \mathcal{B})$ and we are interested in computing a particular numerical characteristic of $\pi(\cdot)$, like its mean or standard deviation, but the complexity of the expressions does not allow us to do the computations directly.

More precisely, suppose $\pi(\cdot)$ has density f with respect to a σ-finite measure $\mu(\cdot)$ on $(\mathsf{X}, \mathcal{B})$. Typically X is an open subset of $\mathbb{R}^d$ and the densities are taken with respect to Lebesgue measure. Suppose we want to estimate expectations of functions $g : \mathsf{X} \to \mathbb{R}$ with respect to $\pi(\cdot)$, i.e. we want to estimate

$$\pi(g) = E_\pi[g(x)] = \int_{\mathsf{X}} g(x)\,\pi(dx) = \int_{\mathsf{X}} g(x)\,f(x)\,\mu(dx).$$

If X is high dimensional and f is a complicated function, then direct integration (either analytic or numerical) of the integrals above is infeasible.

The "gold standard" solution to the above problem, the classical Monte Carlo method, goes as follows. Simulate *iid* random variables $X_1, X_2, \ldots, X_n \sim \pi(\cdot)$ and estimate $\pi(g)$ by $\hat{\pi}(g) = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$. Then $\hat{\pi}(g)$ is unbiased and if $\pi(|g|) < \infty$, then by the usual SLLN for *iid* sequences, $\hat{\pi}(g)$ is a strongly consistent estimator for $\pi(g)$ having standard deviation of order $O(n^{-1/2})$. Furthermore, if $\pi(g^2) < \infty$, the error $\hat{\pi}(g) - \pi(g)$ will have a limiting normal distribution by the classical CLT, which allows us to compute valid asymptotic standard errors for $\hat{\pi}(g)$. The problem with this method is that if f is complicated, then it is very difficult to directly simulate *iid* random variables from $\pi(\cdot)$.

1.2 Markov Chain Monte Carlo Methods

The Markov chain Monte Carlo (MCMC) solution is to construct a Markov chain $\{X_n\}$ on X which is easily run on a computer and which has stationary distribution $\pi(\cdot)$. That is, define easily simulated Markov chain transition probabilities $P(x, dy)$ for $x, y \in \mathsf{X}$ such that $\int_{\mathsf{X}} \pi(dx)\, P(x, dy) = \pi(dy)$. A sufficient condition for $\{X_n\}$ to have stationary distribution $\pi(\cdot)$ is for $\{X_n\}$ to be *reversible* with respect to $\pi(\cdot)$, i.e. $\pi(dx)\, P(x, dy) = \pi(dy)\, P(y, dx)$ for $x, y \in \mathsf{X}$. For $A \subseteq \mathcal{B}$, letting $P^n(x, A) = P[X_n \in A | X_0 = x]$ denote the n-step transition probabilities for the chain, the following result motivates the use of MCMC.

Theorem 1. If a Markov chain on a state space X with countably generated σ-algebra $\mathcal{B}$ is *Harris ergodic* (i.e., ϕ-irreducible, aperiodic, and Harris recurrent) and has stationary distribution $\pi(\cdot)$, then for π-*a.e.* $x \in \mathsf{X}$,

$$\lim_{n\to\infty} \|P^n(x, \cdot) - \pi(\cdot)\| = 0,$$

where $\|\nu_1(\cdot) - \nu_2(\cdot)\|$ denotes the *total variation distance* between the two probability measures $\nu_1(\cdot)$ and $\nu_2(\cdot)$. (cf. Roberts and Rosenthal, 2004).

In particular, $\lim_{n\to\infty} P^n(x, A) = \pi(A)$ for every $A \subseteq \mathcal{B}$.

Remark 1. Under the conditions of Theorem 1, if $g : \mathsf{X} \to \mathbb{R}$ with $\pi(|g|) < \infty$, then a SLLN holds (as in the classical Monte Carlo case) as follows:

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} g(X_i) = \pi(g) \;\; w.p.\,1.$$

Remark 2. Note that Theorem 1 says that the law of a well-behaved Markov chain will converge to the stationary distribution $\pi(\cdot)$ as n increases, but it gives no information about the *rate* at which the total variation distance converges to 0. There are important practical benefits to using an MCMC algorithm for which this rate is (at least) geometrically fast.

Definition 1. Formally, the chain $\{X_n\}$ is called *geometrically ergodic* if there exist a function $M : \mathsf{X} \to [0, \infty)$ and a constant $\rho \in [0, 1)$ such that for all $x \in \mathsf{X}$ and all $n = 1, 2, \ldots$,

$$\|P^n(x, \cdot) - \pi(\cdot)\| \le M(x)\, \rho^n.$$

Theorem 2. Under the assumptions of Theorem 1, if we assume in addition that $\{X_n\}$ is geometrically ergodic and $g : \mathsf{X} \to \mathbb{R}$ with $\pi(|g|^{2+\delta}) < \infty$ for some $\delta > 0$, then the following *Markov chain CLT* holds:

$$n^{-1/2} \sum_{i=1}^{n} [g(X_i) - \pi(g)] \xrightarrow{d} N(0, \sigma^2).$$

It follows that $\sigma^2 = \lim_{n\to\infty} E[(\sum_{i=1}^{n}[g(X_i) - \pi(g)])^2]$, and also $\sigma^2 = \tau \operatorname{Var}_\pi(g)$, where $\tau = \sum_{k\in\mathbb{Z}} \operatorname{Corr}(X_0, X_k)$, is the *integrated autocorrelation time.*

This result says that with a bit more than a finite second moment, if we use a geometrically ergodic chain then we may compute valid asymptotic standard errors for our MCMC based estimates. We will see a method for establishing the geometric ergodicity of a reversible Markov chain based on the properties of its corresponding self-adjoint Markov operator later in this report.

The remainder of this report is organized as follows. Section 2 contains an introduction to the data augmentation algorithm and introduces the running example of the Bayesian probit regression model studied by Albert and Chib (1993). The marginal augmentation (or PX-DA) algorithm of Meng and Van Dyk (1999) is discussed in Section 3 and applied to our running example. The theory and application of the Haar PX-DA algorithm of Liu and Wu (1999) is discussed in Section 4. Finally, in Section 5 a brief overview of the relationship between the spectral properties of Markov operators and the convergence properties of the corresponding Markov chains is presented and used to compare the three algorithms studied.

2 The Data Augmentation Algorithm

2.1 Introduction

The approaches of Hobert (2011) and Khare and Hobert (2011) are followed throughout the next three sections. Throughout the rest of this report, we assume that all Markov chains on the target space are Harris ergodic.

Let $(\mathsf{X}, \mathcal{B}, \mu)$ and $(\mathsf{Y}, \mathcal{A}, \nu)$ be two probability spaces and assume that $\mathcal{B}$ and $\mathcal{A}$ are countably generated. Suppose that $f_X : \mathsf{X} \to [0, \infty)$ is a density with respect to μ which is intractable in the sense that expectations with respect to f_X cannot be computed analytically. In such situations it is often possible to find a joint density $f : \mathsf{X} \times \mathsf{Y} \to [0, \infty)$ with respect to $\mu \times \nu$ that satisfies two properties:

1. The x-marginal is f_X, that is $\int_{\mathsf{Y}} f(x, y)\, \nu(dy) = f_X(x)$.
2. Simulating from the associated conditional densities $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$ is straightforward.

The data augmentation (DA) algorithm is based on this joint density. To specify the Markov chain underlying the DA algorithm, consider the function

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

Proposition 1. (i.) $k(x'|x)$ is a Markov transition density (Mtd).

(ii.) The Markov chain $\{X_n\}$ induced by k is reversible with respect to f_X.

Proof. First, the integrand defining k is the product of two conditional densities. Hence k is nonnegative for each $x \in \mathsf{X}$. Next, fix $x \in \mathsf{X}$ and apply Fubini's theorem to obtain

$$\begin{aligned}
\int_{\mathsf{X}} k(x'|x)\,\mu(dx') &= \int_{\mathsf{X}} \left[\int_{\mathsf{Y}} f_{X|Y}(x'|y)\, f_{Y|X}(y|x)\,\nu(dy) \right] \mu(dx') \\
&= \int_{\mathsf{Y}} f_{Y|X}(y|x) \left[\int_{\mathsf{X}} f_{X|Y}(x'|y)\,\mu(dx') \right] \nu(dy) \\
&= \int_{\mathsf{Y}} f_{Y|X}(y|x)\,\nu(dy) \\
&= 1.
\end{aligned}$$

To prove that $\{X_n\}$ is reversible with respect to f_X, let $f_Y(y) = \int_{\mathsf{X}} f(x,y)\,\mu(dx)$ and note that

$$k(x'|x)\,f_X(x) = f_X(x) \int_{\mathsf{Y}} f_{X|Y}(x'|y)\, f_{Y|X}(y|x)\,\nu(dy) = \int_{\mathsf{Y}} \frac{f(x',y)f(x,y)}{f_Y(y)}\,\nu(dy),$$

which is *symmetric* in (x,x'). Hence for all $(x,x') \in \mathsf{X}$, we have $k(x'|x)\,f_X(x) = k(x|x')\,f_X(x')$. □

With the above result in hand, we may describe the dynamics of the DA Markov chain in the following way. If the current state of the chain is $X_n = x$, then the density of the next state, X_{n+1}, is $k(\cdot|x)$. Also, recalling a result from Section 1.2, Proposition 1 implies that f_X is an *invariant* density for the chain $\{X_n\}$. If the current state of the chain is $X_n = x$, then we may simulate X_{n+1} as follows.

One iteration of the DA algorithm:

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

Note that the above procedure amounts to drawing (x,y) from the joint density $f(x,y)$ using a two variable Gibbs sampler and ignoring the y coordinate.

2.2 Bayesian Probit Regression Example

We now present Albert and Chib's (1993) widely used DA algorithm for Bayesian Probit Regression.

Let $Z_1, \dots, Z_n$ be independently distributed Bernoulli random variables with success probability $P(Z_i = 1) = \Phi(x_i^T \beta)$ where $\Phi(\cdot)$ denotes the standard normal CDF, $x_i \in \mathbb{R}^p$ is a $p \times 1$ vector of known covariates associated with Z_i, and $\beta \in \mathbb{R}^p$ is a $p \times 1$ vector of unknown regression coefficients.

Letting $z = (z_1, \dots, z_n)^T$ denote the observed data, the likelihood is given by

$$P(Z_1 = z_1, \dots, Z_n = z_n | \beta) = \prod_{i=1}^n \left[\Phi(x_i^T \beta)\right]^{z_i} \left[1 - \Phi(x_i^T \beta)\right]^{1-z_i}.$$

To analyze this model within the Bayesian paradigm, consider putting a flat prior on the vector of regression coefficients β, i.e. the prior density is $\pi(\beta) \propto 1$. The resulting marginal density is

$$m(z) = \int\limits_{\mathbb{R}^p} \prod_{i=1}^n \left[\Phi(x_i^T \beta)\right]^{z_i} \left[1 - \Phi(x_i^T \beta)\right]^{1-z_i} \, d\beta.$$

We assume throughout this report that the posterior is proper, i.e. that $m(z) < \infty$. The posterior density is given by

$$\pi(\beta | z) = \frac{1}{m(z)} \prod_{i=1}^n \left[\Phi(x_i^T \beta)\right]^{z_i} \left[1 - \Phi(x_i^T \beta)\right]^{1-z_i}.$$

This posterior is quite intractable: it involves a p-dimensional integral of a product of normal distribution functions. Therefore it is a good candidate for MCMC based analysis.

Albert and Chib's (1993) idea was to introduce *missing data* $y = (y_1, \dots, y_n)^T \in \mathbb{R}^n$ to facilitate the computation of a DA algorithm for $\pi(\beta|z)$. (Similar to the familiar EM algorithm).

Let $\phi(v; \mu, \sigma^2)$ denote the $N(\mu, \sigma^2)$ density function evaluated at the point $v \in \mathbb{R}$. Consider the function

$$\pi(\beta, y | z) = \frac{1}{m(z)} \prod_{i=1}^n \left\{ I_{\mathbb{R}_+}(y_i) \, I_{\{1\}}(z_i) + I_{\mathbb{R}_-}(y_i) \, I_{\{0\}}(z_i) \right\} \, \phi(y_i; x_i^T \beta, 1),$$

where $I_A(\cdot)$ is the indicator function of the set A, $\mathbb{R}_+ = (0, \infty)$, and $\mathbb{R}_- = (-\infty, 0)$. Our goal is to realize $\pi(\beta, y|z)$ as a joint density in (β, y) whose β-marginal is $\pi(\beta|z)$, and then derive the conditionals of this joint density to produce a DA algorithm for $\pi(\beta|z)$.

Indeed, integrating y out of $\pi(\beta, y|z)$ we obtain:

$$
\begin{aligned}
\int\limits_{\mathbb{R}^n} \pi(\beta, y|z)\, dy &= \frac{1}{m(z)} \int\limits_{\mathbb{R}^n} \prod_{i=1}^{n} \left\{ I_{\mathbb{R}_+}(y_i)\, I_{\{1\}}(z_i) + I_{\mathbb{R}_-}(y_i)\, I_{\{0\}}(z_i) \right\} \, \phi(y_i; x_i^T\beta, 1)\, dy_n \cdots dy_2\, dy_1 \\
&= \frac{1}{m(z)} \prod_{i=1}^{n} \int\limits_{\mathbb{R}} \left\{ I_{\mathbb{R}_+}(y_i)\, I_{\{1\}}(z_i) + I_{\mathbb{R}_-}(y_i)\, I_{\{0\}}(z_i) \right\} \, \phi(y_i; x_i^T\beta, 1)\, dy_i \\
&= \frac{1}{m(z)} \prod_{i=1}^{n} \left\{ I_{\{1\}}(z_i) \int_0^\infty \phi(y_i; x_i^T\beta, 1)\, dy_i + I_{\{0\}}(z_i) \int_{-\infty}^0 \phi(y_i; x_i^T\beta, 1)\, dy_i \right\} \\
&= \frac{1}{m(z)} \prod_{i=1}^{n} \left\{ I_{\{1\}}(z_i)\Phi(x_i^T\beta) + I_{\{0\}}(z_i) \left[1 - \Phi(x_i^T\beta\right] \right\} \\
&= \frac{1}{m(z)} \prod_{i=1}^{n} \left[\Phi(x_i^T\beta)\right]^{z_i} \left[1 - \Phi(x_i^T\beta\right]^{1-z_i} \\
&= \pi(\beta|z).
\end{aligned}
$$

In the standard notation of linear models, let $X \in \mathbb{R}^{n \times p}$ denote the $n \times p$ matrix whose ith row is x_i^T, let $\hat{\beta}(y) = (X^TX)^{-1}X^Ty \in \mathbb{R}^p$ denote the usual least squares estimator of β for fixed y, and let $P = X(X^TX)^{-1}X^T \in \mathbb{R}^{n \times n}$ denote the projection onto the column space of X. It follows that

$$
\pi(\beta|y, z) = \prod_{i=1}^{n} \phi(y_i; x_i^T\beta, 1) = (2\pi)^{-n/2} \exp\left\{ -\frac{y^T(I-P)y}{2} \right\} \exp\left\{ -\frac{1}{2}(\beta - \hat{\beta}(y))^T\, X^TX\, ((\beta - \hat{\beta}(y)) \right\},
$$

i.e. $\beta|\, y, z \sim N_p(\hat{\beta}(y), (X^TX)^{-1})$.

Finally, let $TN(\mu, \sigma^2, v)$ denote a $N(\mu, \sigma^2)$ distribution that is *truncated* to be positive if $v = 1$ and negative if $v = 0$. It is clear looking at the form of the complete data posterior density $\pi(\beta, y|z)$ that $Y_i|\beta, z \sim TN(x_i^T\beta, 1, z_i)$ independently. Note that the $TN(\mu, \sigma^2, v)$ family of distributions may be easily simulated using a simple rejection sampler.

We may now implement a DA algorithm for $\pi(\beta|z)$ as follows. Given the current state $X_n = \beta$, we simulate the next state, X_{n+1} by performing the following two steps:

One iteration of the DA algorithm for $\pi(\beta|z)$:

1. Draw $Y_1, \ldots, Y_n$ independently such that $Y_i \sim TN(x_i^T\beta, 1, z_i)$, and call the observed value $y = (y_1, \ldots, y_n)^T$.

2. Draw $X_{n+1} \sim N_p(\hat{\beta}(y), (X^TX)^{-1})$.

Roy and Hobert (2007) showed that this particular DA algorithm is geometrically ergodic by establishing a *geometric drift condition* (cf. Hobert 2011).

3 The Marginal Augmentation / PX-DA Algorithm

3.1 Sandwich Algorithms

We have seen thus far that one iteration of the DA algorithm based on $f(x, y)$ can be simulated using the two-step procedure described above, which entails drawing from the two conditional densities defined by $f(x, y)$. Khare and Hobert (2011) showed that this simulation can also be accomplished using a *three-step* procedure in which the first and third steps are draws from $f_{Y|X}(y|x)$ and $f_{X|Y}(x|y)$ respectively, and the middle step involves a single move according to a Markov chain on the space Y that has invariant density $f_Y(y)$.

Suppose that $R(y, dy')$ is any Markov transition function on Y that is reversible with respect to $f_Y(y)\nu(dy)$, i.e. $R(y, dy')f_Y(y)\nu(dy) = R(y', dy)f_Y(y')\nu(dy')$. Consider a new Mtd given by

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

It is straightforward to show using the same technique in the proof of Proposition 1 that $k^*(x'|x)f_X(x)$ is symmetric in (x, x'), so the Markov chain defined by k^*, denoted by $\{X_n^*\}$, is reversible with respect to f_X. The form of k^* suggests that if the current state is $X_n^* = x$, then X_{n+1}^* can be simulated using the following three steps:

One iteration of the Sandwich algorithm:

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

Note that the draw from $R(y, \cdot)$ is "sandwiched" between the draws from the two conditional densities. The reasoning behind this extra step is that it is often possible to construct a sandwich algorithm that converges much faster than the original DA algorithm while requiring roughly the same computational effort per iteration. It turns out that this low-dimensional perturbation on the Y space can lead to a major improvement in mixing. The next subsection provides a general recipe for building the function R.

3.2 Basic Theory of Marginal Augmentation / PX-DA

A powerful method to speed up the DA algorithm was developed independently by Liu and Wu (1999), who referred to it as "PX-DA", and Meng and Van Dyk (1999), who referred to it as "marginal augmentation". The basic idea is to introduce a low-dimensional parameter into the joint density $f(x, y)$ that is not identifiable in the target f_X. This allows for the construction of an entire class of possible DA algorithms, indexed by this "working parameter."

For simplicity, we assume for the rest of this report that X and Y are Euclidean spaces and $f(x, y)$ is a density with respect to Lebesgue measure. Let $\mathsf{G} \subset \mathbb{R}^d$, let $\{t_g : \mathsf{Y} \to \mathsf{Y}\}_{g \in \mathsf{G}}$ be a class of one-to-one differentiable functions indexed by $g \in \mathsf{G}$, and let $J_g(z)$ denote the Jacobian of the transformation $z = t_g^{-1}(y)$. Now fix a "working prior" density on G, call it $w(g)$ and define a joint density on $\mathsf{X} \times \mathsf{Y} \times \mathsf{G}$ by $f^{(w)}(x, y, g) = f(x, t_g(y))\,|J_g(y)|\,w(g)$. Then

$$\int_{\mathsf{Y}} \int_{\mathsf{G}} f^{(w)}(x, y, g)\,dg\,dy = \int_{\mathsf{Y}} \int_{\mathsf{G}} f(x, t_g(y))\,|J_g(y)|\,w(g)\,dg\,dy$$

$$= \int_{\mathsf{Y}} f(x, t_g(y))\,|J_g(y)|\,dy = \int_{\mathsf{Y}} f(x, z)dz = f_X(x).$$

We have shown that the x-marginal of $f^{(w)}(x, y, g)$ is $f_X(x)$. Hence we may define a joint density on $\mathsf{X} \times \mathsf{Y}$ having x-marginal $f_X(x)$ by $f^{(w)}(x, y) = \int_{\mathsf{G}} f^{(w)}(x, y, g)\,dg$. Thus if it is easy to sample from $f^{(w)}_{X|Y}(x|y)$ and $f^{(w)}_{Y|X}(y|x)$, then we have a new DA algorithm that can be compared with the one based on $f(x, y)$.

The problem is that in real examples, it will often be impossible to sample directly from (or even compute) these conditionals. However, it is possible to develop *indirect* methods of drawing from $f^{(w)}_{X|Y}$ and $f^{(w)}_{Y|X}$, that use only draws from $f_{X|Y}$, $f_{Y|X}$, $w(g)$, and one other density. First consider $f^{(w)}_{Y|X}(y|x)$.

$$\begin{aligned} f^{(w)}_{Y|X}(y|x) &= \frac{\int_{\mathsf{G}} f^{(w)}(x, y, g)\,dg}{f_X(x)} \\ &= \int_{\mathsf{G}} \frac{f(x, t_g(y))}{f_x(x)}\,|J_g(y)|\,w(g)\,dg \\ &= \int_{\mathsf{G}} f_{Y|X}(t_g(y)|x)\,|J_g(y)|\,w(g)\,dg. \end{aligned}$$

Draw $Y' \sim f_{Y|X}(\cdot|x)$ and $G \sim w(\cdot)$ and suppose Y' and G are independent. Then this last integrand can be expressed as the joint density of (G, Y) where $Y = t_g^{-1}(Y')$. Hence Y has density $f^{(w)}_{Y|X}(\cdot|x)$.

Next consider $f^{(w)}_{X|Y}(x|y)$. Write $f^{(w)}_{X|Y}(x|y) = \int_{\mathsf{G}} f^{(w)}_{X,G|Y}(x, g|y)\, dg = \int_{\mathsf{G}} f^{(w)}_{X|Y,G}(x|y, g)\, f^{(w)}_{G|Y}(g|y)\, dg$. Thus using the two-step technique in Section 2.1, we may simulate from $f^{(w)}_{X|Y}(x|y)$ as follows. First, draw $G \sim f^{(w)}_{G|Y}(\cdot|y)$ and call the result g. Then draw $X \sim f^{(w)}_{X|Y,G}(\cdot|y, g)$.

It suffices to show we can draw from $f^{(w)}_{G|Y}$ and $f^{(w)}_{X|Y,G}$. First,

$$f^{(w)}_{X|Y,G}(x|y, g) = \frac{f^{(w)}(x, y, g)}{\int_{\mathsf{X}} f^{(w)}(x, y, g)\, dx} = \frac{f(x, t_g(y))|J_g(y)|w(g)}{f_Y(t_g(y))|J_g(y)|w(g)} = f_{X|Y}(x|t_g(y)),$$

i.e. drawing from $f^{(w)}_{X|Y,G}(\cdot|y, g)$ is equivalent to drawing from $f_{X|Y}(\cdot|t_g(y))$, which we are assuming is straightforward.

Next,

$$f^{(w)}_{G|Y}(g|y) = \frac{\int_{\mathsf{X}} f^{(w)}(x, y, g)\, dx}{\int_{\mathsf{G}} \int_{\mathsf{X}} f^{(w)}(x, y, g)\, dx\, dg} \propto \int_{\mathsf{X}} f^{(w)}(x, y, g)\, dx = f_Y(t_g(y))|J_g(y)|w(g).$$

It appears that sampling from $f^{(w)}_{G|Y}$ would be quite challenging. Fortunately in most applications G is low-dimensional, making it possible to sample from $f^{(w)}_{G|Y}$ despite the fact that f_Y is intractable.

To summarize,

- To draw from $f^{(w)}_{Y|X}(\cdot|x)$, first draw Y' and G independently from $f_{Y|X}(\cdot|x)$ and $w(\cdot)$, respectively, and then take $Y = t_g^{-1}(Y')$.
- To draw from $f^{(w)}_{X|Y}(\cdot|y)$, draw $G \sim f^{(w)}_{G|Y}(\cdot|y)$ and call the result g, then draw $X \sim f_{X|Y}(\cdot|t_g(y))$.

Finally, we may use the Sandwich structure of Section 3.1 to simulate the PX-DA Markov chain. Indeed, if the current state of the chain is $X_n = x$, then we can simulate X_{n+1} as follows.

One iteration of the PX-DA algorithm:

1. Draw $Y \sim f_{Y|X}(\cdot|x)$, and call the observed value y.
2. Draw $G \sim w(\cdot)$, call the result g, then draw $G' \sim f^{(w)}_{G|Y}(\cdot|t_g^{-1}(y))$, call the result g', and finally set $y' = t_{g'}(t_g^{-1}(y))$.
3. Draw $X_{n+1} \sim f_{X|Y}(\cdot|y')$.

3.3 Application to Bayesian Probit Regression

Recall that in this example $\pi(\beta|z)$ plays the role of $f_X(x)$ and $\pi(\beta, y|z)$ plays the role of $f(x, y)$. Let $\mathsf{G} = \mathbb{R}_+$ and take $t_g(y)$ to be the transformation that translates $y \in \mathsf{Y}$ by a fixed element $g \in \mathsf{G}$. That is, $t_g(y) = gy = (gy_1, \dots, gy_n)$. Suppose the working prior density w is given by

$$w(g; \alpha, \lambda) = \frac{2\lambda^\alpha}{\Gamma(\alpha)}\, g^{2\alpha-1}\, e^{-g^2\lambda}\, I_{\mathbb{R}_+}(g), \quad \text{where } \alpha, \lambda \in \mathbb{R}_+.$$

Note that if $U \sim \text{Gamma}(\alpha, \lambda)$, where λ is a *rate* parameter, then $G = \sqrt{U} \sim w(g; \alpha, \lambda)$, i.e. w is the density of the square root of a Gamma random variable. Using the conditional densities we computed in Section 2.2, it is easy to show that

$$\pi(y|z) = \frac{\exp\left\{-\frac{y^T(I-P)y}{2}\right\}}{|X^TX|^{\frac{1}{2}}\, m(z)(2\pi)^{\frac{m-p}{2}}} \prod_{i=1}^{n} \left\{ I_{\mathbb{R}_+}(y_i)\, I_{\{1\}}(z_i) + I_{\mathbb{R}_-}(y_i)\, I_{\{0\}}(z_i) \right\}.$$

Hence,

$$\begin{aligned} f^{(w)}_{G|Y}(g|y) &\propto \pi(t_g(y)|z)|J_g(z)|w(g) \\ &\propto \left[\exp\left\{-\frac{1}{2}(gy)^T(I-P)(gy)\right\}\right](g^n)\left[g^{2\alpha-1}\exp\{-g^2\lambda\}I_{\mathbb{R}_+}(g)\right] \\ &= \exp\left\{-g^2\left[\frac{y^T(I-P)y}{2} + \lambda\right]\right\} g^{n+2\alpha-1} I_{\mathbb{R}_+}(g). \end{aligned}$$

Observe that $f^{(w)}_{G|Y}(g|y)$ has the same form as $w(g; \alpha, \lambda)$, so we may simulate $f^{(w)}_{G|Y}(g|y)$ by drawing from a gamma density and taking its square root. We now have all the information we need to write down a PX-DA algorithm for this example. If the current state of the PX-DA Markov chain is $X_n = \beta$, then we simulate the next state, X_{n+1}, by performing the following three-step procedure:

One iteration of the PX-DA algorithm for $\pi(\beta|z)$:

1. Draw $Y_1, \dots, Y_n$ independently such that $Y_i \sim TN(x_i^T\beta, 1, z_i)$, and call the result $y = (y_1, \dots, y_n)^T$.
2. Draw $U \sim \text{Gamma}(\alpha, \lambda)$, call the result u, and set $\tilde{y} = \frac{y}{\sqrt{u}}$.
   Draw $V \sim \text{Gamma}\left(\frac{m}{2} + \alpha, \frac{\tilde{y}^T(I-P)\tilde{y}}{2} + \lambda\right)$, call the result v, and set $y' = \sqrt{v}\tilde{y}$.
3. Draw $X_{n+1} \sim N_p(\hat{\beta}(y'), (X^TX)^{-1})$.

---

4 The Haar PX-DA Algorithm

4.1 Some Group Theory

In this section, following the treatment of Eaton(1989), we present the background in group theory needed to understand the Haar PX-DA algorithm.

Definition 2. A *group* is a set G equipped with a binary operation $* : \mathsf{G} \times \mathsf{G} \to \mathsf{G}$ such that

(i). $g_1, g_2 \in \mathsf{G}$ implies $g_1 * g_2 \in \mathsf{G}$.

(ii). $(g_1 * g_2) * g_3 = g_1 * (g_2 * g_3)$ for $g_1, g_2, g_3 \in \mathsf{G}$.

(iii). There exists an element $e \in \mathsf{G}$ such that $e * g = g * e = g$ for $g \in \mathsf{G}$.

(iv). For each $g \in \mathsf{G}$, there exists a unique element $g^{-1} \in \mathsf{G}$ such that $g * g^{-1} = g^{-1} * g = e$.

We often omit $*$ and write $g_1 g_2$ to mean $g_1 * g_2$. The element $e \in \mathsf{G}$ is called the *identity* and g^{-1} is called the *inverse* of g. If the set G is a locally compact topological space whose topology has a countable base and the functions $(g_1, g_2) \mapsto g_1 g_2$ and $g \mapsto g^{-1}$ are both continuous, then G is called a *topological group.*

Definition 3. Let G be a group and let Y be a set. A function $F : \mathsf{G} \times \mathsf{Y} \to \mathsf{Y}$ satisfying

(i). $F(e, y) = y, \quad y \in \mathsf{Y},$

(ii). $F(g_1 g_2, y) = F(g_1, F(g_2, y)), \quad g_1, g_2 \in \mathsf{G},\ y \in \mathsf{Y},$

specifies G *acting on the left* of Y. If G is a topological group and Y is a topological space, we say the group G acts *topologically on the left of* Y if G acts on the left of Y and if the action of G, $F : \mathsf{G} \times \mathsf{Y} \to \mathsf{Y}$ is continuous.

Definition 4. Let G be a topological group. Let $K(\mathsf{G})$ denote the real vector space of all continuous functions with compact support defined on G.

- A function $\chi : \mathsf{G} \to \mathbb{R}_+$ is called a *multiplier* if χ is continuous, and $\chi(g_1 g_2) = \chi(g_1)\chi(g_2)$ for $g_1, g_2 \in \mathsf{G}$.
- A measure m on $K(\mathsf{G})$ is *left invariant* if for all $f \in K(\mathsf{G})$, (and hence for all m-integrable f), $\int_{\mathsf{G}} f(g^{-1}x)\, m(dx) = \int_{\mathsf{G}} f(x)\, m(dx)$ for $g \in \mathsf{G}$.
- A measure m on $K(\mathsf{G})$ is *relatively left invariant with multiplier* χ if for each $g \in \mathsf{G}$, $\int_{\mathsf{G}} f(g^{-1}x)\, m(dx) = \chi(g) \int_{\mathsf{G}} f(x)\, m(dx)$.

Theorem 3. On a topological group G, there exists a left invariant measure ν_l, called the *left Haar measure* on G such that for every $g \in \mathsf{G}$,

$$\int_{\mathsf{G}} f(x)\, \nu_l(dx) = \int_{\mathsf{G}} f(g^{-1}x)\, \nu_l(dx).$$

The *right Haar measure* is defined analogously. When the left Haar measure is the same as the right Haar measure, the group G is said to be *unimodular.* In most applications, this measure is improper, that is, $\int_{\mathsf{G}} \nu_l(dg) = \infty$.

4.2 Haar PX-DA

Liu and Wu (1999) showed that if the set G from section 3 is given the structure of a topological group, then it is possible to construct a valid PX-DA-*like* algorithm with an improper *Haar density* in place of the working prior density w. Furthermore, it is possible to show using the results of the next section that this *Haar PX-DA algorithm* is better than any PX-DA algorithm based on a proper w.

To this end, suppose G is a topological group. In keeping with the notation of Section 3, we assume that the transformation $t_g(y)$ represents G acting topologically on the left of Y. Further, we assume that Lebesgue measure on Y is relatively left invariant with multiplier χ, i.e. for any $g \in \mathsf{G}$, $\int_{\mathsf{Y}} h(y)\,dy = \chi(g) \int_{\mathsf{Y}} h(t_g(y))\,dy$ for every integrable $h : \mathsf{Y} \to \mathbb{R}$.

Finally, assume that the function $q : \mathsf{Y} \to \mathbb{R}$ defined by

$$q(y) = \int\limits_{\mathsf{G}} f_Y(t_g(y))\,\chi(g)\,\nu_l(g)\,dg$$

is strictly positive for all $y \in \mathsf{Y}$ and finite for almost all $y \in \mathsf{Y}$.

We will again take advantage of the sandwich methodology of Section 3.1 to state the Haar PX-DA algorithm. Note that in this case we have used a group action to construct the function R in the "sandwich step." If the current state of the chain is $X_n^* = x$, we simulate X_{n+1}^* as follows.

One iteration of the Haar PX-DA algorithm:

1. Draw $Y \sim f_{Y|X}(\cdot|x)$, and call the observed value y.
2. Draw G from the density proportional to $f_Y(t_g(y))\,\chi(g)\,\nu_l(g)$, call the result g, and set $y' = t_g(y)$.
3. Draw $X_{n+1}^* \sim f_{X|Y}(\cdot|y')$.

Note that Step 2 of the Haar PX-DA algorithm involves only one draw from a density on G, whereas the regular PX-DA algorithm calls for two such draws in Step 2. Hence the Haar PX-DA algorithm is *not* a PX-DA algorithm, but it is much simpler from a computational standpoint.

For completeness, we note that the Mtd of the Haar PX-DA algorithm is given by

$$k_H(x'|x) = \int\limits_{\mathsf{Y}}\int\limits_{\mathsf{Y}} f_{X|Y}(x'|y')\,l_H(y'|y)\,f_{Y|X}(y|x)\,dy\,dy',$$

where $l_H(y'|y)$ denotes the Mtd of the Markov chain on Y that is simulated at Step 2. Hobert and Marchev (2008) show that $k_H(x'|x)$ is reversible with respect to f_X, which proves that f_X is an invariant density for the Markov chain $\{X_n^*\}$.

4.3 Application to Bayesian Probit Regression

Take $\mathsf{G} = \mathbb{R}_+$, an infinite unimodular abelian group with the group operation given by multiplication, identity element $e = 1$, and inverses specified by $g^{-1} = \frac{1}{g}$, $g \in \mathbb{R}_+$. With $\mathsf{Y} = \mathbb{R}_+^n$ and $t_g(y) = gy = g(y_1, \ldots, y_n)$, it is clear that for any $y \in \mathsf{Y}$ and any $g_1, g_2 \in \mathsf{G}$, we have $t_e(y) = y$ and $t_{g_1 g_2}(y) = g_1 g_2 y = g_1(g_2 y) = t_{g_1}(t_{g_2}(y))$. Hence, G acts topologically on the left of Y. Further, for any $g \in \mathsf{G}$ and any integrable $h : \mathsf{Y} \to \mathbb{R}$, we have

$$\int\limits_{\mathsf{Y}} h(t_g(y))\, dy = \int\limits_{\mathbb{R}_+^n} h(gy)\, dy = g^{-n} \int\limits_{\mathbb{R}_+^n} h(y)\, dy.$$

which shows that Lebesgue measure on Y is relatively left invariant with multiplier $\chi(g) = g^n$.

For any $g' \in \mathsf{G}$ and any integrable $h : \mathsf{Y} \to \mathbb{R}$, we have

$$\int_0^\infty h(g'\, g)\, \frac{1}{g}\, dg = \int_0^\infty h(g)\, \frac{1}{g}\, dg,$$

which shows $\frac{dg}{g}$ is a left Haar measure for G. Then

$$\pi(t_g(y)|z)\, \chi(g)\, \nu_l(g) \propto g^{n-1} \text{exp}\left\{ -g^2 \left[\frac{y^T(I-P)y}{2} \right] \right\} I_{\mathbb{R}_+}(g),$$

and so

$$q(y) \propto \int_0^\infty g^{n-1} \text{exp}\left\{ -g^2 \left[\frac{y^T(I-P)y}{2} \right] \right\} dg = \frac{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})}{[y^T(I-P)y]^{\frac{n}{2}}},$$

recognizing the integrand as the kernel of a $\Gamma(\frac{n}{2}, \frac{y^T(I-P)y}{2})$ distribution. Hence $q(y)$ is strictly positive for all $y \in \mathsf{Y}$ and finite for almost all $y \in \mathsf{Y}$.

We may now state the Haar PX-DA algorithm for this example. Given the current state, $X_n^* = \beta$, we simulate the next state, X_{n+1}^*, as follows.

One iteration of the Haar PX-DA algorithm for $\pi(\beta|z)$:

1. Draw $Y_1, \ldots, Y_n$ independently such that $Y_i \sim TN(x_i^T\beta, 1, z_i)$, and call the result $y = (y_1, \ldots, y_n)^T$.
2. Draw $V \sim$ Gamma $\left(\frac{n}{2}, \frac{y^T(I-P)y}{2}\right)$, call the result v, and set $y' = \sqrt{v}\, y$.
3. Draw $X_{n+1}^* \sim N_p(\hat{\beta}(y'), (X^TX)^{-1})$.

5 A Theoretical Comparison of the DA, PX-DA, and Haar PX-DA Algorithms

5.1 Theory of Self-adjoint Markov Operators

Spectral theory is a useful tool to analyze reversible Markov chains. In this section we present the ideas needed to facilitate a comparison between the three DA algorithms discussed in this report.

Let $(\mathsf{X}, \mathcal{B}, \mu)$ be a probability space and suppose $f_X : \mathsf{X} \to [0, \infty)$ is a density with respect to μ. Define

$$L^2(f_X) = \left\{ h : \mathsf{X} \to \mathbb{R} \, : h \text{ is measurable and } \int_{\mathsf{X}} h^2(x)\, f_X(x)\, \mu(dx) < \infty \right\},$$

and let

$$L_0^2(f_X) = \left\{ h \in L^2(f_X) \, : \int_{\mathsf{X}} h(x)\, f_X(x)\, \mu(dx) = 0 \right\}.$$

Then $L_0^2(f_X)$ is a Hilbert space, with inner product given by $\langle g, h \rangle = \int_{\mathsf{X}} h(x)\, g(x)\, f_X(x)\, \mu(dx)$ and corresponding norm given by $\|g\| = \sqrt{\langle g, g \rangle}$.

Let $P(x, dx')$ denote a generic Mtf on X that is reversible with respect to $f_X(x)\mu(dx)$ and denote the Markov chain driven by P as $\{X_n\}$. We may express the convergence properties of $\{X_n\}$ in terms of a related operator that is now defined. Let $P : L_0^2(f_X) \to L_0^2(f_X)$ denote the operator that maps $g \in L_0^2(f_X)$ to

$$(Pg)(x) = E\left[g(X_{n+1}) | X_n = x\right] = \int_{\mathsf{X}} g(x')\, P(x, dx').$$

The *operator norm* of P is defined as $\|P\| = \sup_{g \in L_{0,1}^2(f_X)} \|Pg\|$, where $L_{0,1}^2(f_X)$ is the subset of $L_0^2(f_X)$ that contains the functions g satisfying $\|g\| = 1$.

Proposition 2. (i). If $g \in L_0^2(f_X)$, then Pg is indeed an element of $L_0^2(f_X)$.

(ii). The operator P is self adjoint with respect to the inner product on $L_0^2(f_X)$.

(iii). $\|P\| \in [0, 1]$.

(iv). The Markov chain $\{X_n\}$ is geometrically ergodic if and only if $\|P\| < 1$.

Proof. (i). To show Pg is square integrable with respect to f_X, we compute

$$\begin{aligned}\int_{\mathsf{X}} [(Pg)(x)]^2\, f_X(x)\,\mu(dx) &= \int_{\mathsf{X}} \left[\int_{\mathsf{X}} g(x')\,P(x,dx')\right]^2 f_X(x)\mu(dx) && (1)\\ &\le \int_{\mathsf{X}} \left[\int_{\mathsf{X}} g^2(x')\,P(x,dx')\right] f_X(x)\mu(dx) && (2)\\ &= \int_{\mathsf{X}} g^2(x') \left[\int_{\mathsf{X}} P(x',dx)\,f_X(x)\right] \mu(dx') && (3)\\ &= \int_{\mathsf{X}} g^2(x')\,f_X(x')\,\mu(dx') < \infty, && (4)\end{aligned}$$

where (2) is Jensen's Inequality, (3) is Fubini's theorem, and (4) follows from the invariance of f_X.

Similarly, to show Pg has mean zero, we compute

$$\begin{aligned}\int_{\mathsf{X}} (Pg)(x)\,f_X(x)\,\mu(dx) &= \int_{\mathsf{X}} \left[\int_{\mathsf{X}} g(x')\,P(x,dx')\right] f_X(x)\,\mu(dx)\\ &= \int_{\mathsf{X}} g(x') \left[\int_{\mathsf{X}} P(x',dx)\,f_X(x)\right] \mu(dx') = \int_{\mathsf{X}} g(x')\,f_X(x')\,\mu(dx') = 0.\end{aligned}$$

(ii). Let $g, h \in L_0^2(f_X)$. Using the fact that $P(x, dx')$ is reversible with respect to f_X, we compute

$$\begin{aligned}\langle Pg, h\rangle &= \int_{\mathsf{X}} (Pg)(x)\,h(x)\,f_X(x)\,\mu(dx) = \int_{\mathsf{X}} \left[\int_{\mathsf{X}} g(x')\,P(x,dx')\right] h(x)\,f_X(x)\,\mu(dx)\\ &= \int_{\mathsf{X}}\int_{\mathsf{X}} g(x')\,h(x)\,P(x,dx')\,f_X(x)\,\mu(dx) = \int_{\mathsf{X}} g(x') \left[\int_{\mathsf{X}} h(x)\,P(x',dx)\right] f_X(x')\,\mu(dx')\\ &= \int_{\mathsf{X}} g(x')\,(Ph)(x')\,f_X(x')\,\mu(dx') = \langle g, Ph\rangle.\end{aligned}$$

(iii). Since $\|\cdot\|$ is a norm, we must have $\|P\| \ge 0$. Further, $\|Pg\|^2 = \int_{\mathsf{X}} [(Pg)(x)]^2\, f_X(x)\,\mu(dx)$ and the above calculations imply that $\|Pg\|^2 \le \|g\|^2 = 1$.

(iv). For a proof of (iv), see Roberts and Rosenthal, 1997.

□

Remark 3. Loosely speaking, the closer $\|P\|$ is to 0, the faster $\{X_n\}$ converges to its stationary distribution. Because of this, Monte Carlo Markov chains are sometimes ordered according to their operator norms. In particular, if there are two different chains available that are both reversible with respect to the target, we prefer the one with the smaller operator norm.

5.2 Comparison Results

Recall that the Mtd of the PX-DA algorithm is given by

$$k_w(x'|x) = \int_{\mathsf{Y}} f^{(w)}_{X|Y}(x'|y)\, f^{(w)}_{Y|X}(y|x)\, \nu(dy) = \int_{\mathsf{Y}}\int_{\mathsf{Y}} f_{X|Y}(x'|y')\, l_w(y'|y)\, f_{Y|X}(y|x)\, \nu(dy)\, \nu(dy'),$$

where we have used the sandwich representation of Section 3.2, and $l_w(y'|y$ denotes the Mtd of the Markov chain in Step 2 of the algorithm with state space Y. Liu and Wu's (1999) Theorem 1 implies that f_Y is an invariant density for l_w, i.e. $\int_{\mathsf{Y}} l_w(y'|y)\, f_Y(y) = f_Y(y')$. Further, since k_w is the Mtd of a DA algorithm, we have by Proposition 1 that k_w is reversible with respect to f_X, and hence that f_X is invariant for k_w.

More generally, let $l : \mathsf{Y} \times \mathsf{Y} \to [0, \infty)$ be any Mtd that has $f_Y(y)$ as an invariant density. Let $k_l : \mathsf{X} \times \mathsf{X} \to [0, \infty)$ be defined as follows.

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

Then k_l is an Mtd that defines a Markov chain on X with invariant density f_X. Before stating the main result, we require some definitions.

Definition 5. If there exists a joint pdf $f^*(x, y)$ (with respect to $\mu\times\nu$) with $\int_{\mathsf{Y}} f^*(x, y)\, \nu(dy) = f_X(x)$ such that

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

then we say that k_l is *representable.*

Note that if k_l is representable, then it is also reversible with respect to $f_X(x)$, and also that k_w as above is representable with $f^{(w)}(x, y)$ playing the role of $f^*(x, y)$.

Definition 6. Let $\{X_n\}$ denote the Markov chain underlying the original DA algorithm based on $f(x, y)$. Suppose $g \in L^2(f_X)$ and let $\bar{g}_n = \frac{1}{n}\sum_{i=0}^{n-1} g(X_i)$. If $\bar{g}_n$ satisfies a CLT, then let κ_g^2 denote the corresponding asymptotic variance. If there is no CLT for $\bar{g}_n$ then set $\kappa_g^2 = \infty$. Now let $\{X_n^*\}$ denote the Markov chain associated with $k_l(x'|x)$, and define κ_g^{*2} analogously using $\bar{g}_n^* = \frac{1}{n}\sum_{i=0}^{n-1} g(X_i^*)$ in place of $\bar{g}_n$. If $\kappa_g^{*2} \leq \kappa_g^2$ for every $g \in L^2(f_X)$, we say that k_l is *more efficient than* k.

Theorem 4. (Hobert and Marchev 2008)

(i). If k_l is reversible with respect to f_X, then k_l is more efficient than k.

(ii). If k_l is representable, then $\|K_l\| \leq \|K\|$, where K_l and K are the operators on $L_0^2(f_X)$ associated with k_l and k respectively.

With regards to the PX-DA algorithm, since k_w is representable, both the results from Theorem 4 apply and we may conclude that every PX-DA algorithm is better than the corresponding DA algorithm in terms of both convergence rate and ARE. In particular, $\|K_w\| \leq \|K\|$, where K_w is the operator corresponding to k_w. Further, by a result from the previous subsection, a reversible Markov chain is geometrically ergodic if and only if the norm of the corresponding operator is strictly less than 1. Therefore, if we can prove that the DA Markov chain $\{X_n\}$ is geometrically ergodic then it follows that $\|K_w\| \leq \|K\| < 1$, which implies that $\{X_n^*\}$ is also geometrically ergodic. Hobert and Marchev (2008) provide simple sufficient conditions for k_l to be reversible with respect to f_X and a simple sufficient condition on $l(y'|y)$ for representability of k_l.

With regards to the Haar PX-DA algorithm, in the notation of Section 4.2, Hobert and Marchev (2008) show that $l_H(y'|y)$ is reversible with respect to f_Y and that k_H is representable. Hence the comparison results are applicable and imply that the Haar PX-DA algorithm is better than the DA algorithm in terms of both convergence rate and ARE. In particular, $\|K_H\| \leq \|K\|$, where K_H is the operator corresponding to k_H.

The real question is how Haar PX-DA compares to PX-DA. Hobert and Marchev (2008) show that, for any fixed proper pdf $w(\cdot)$, k_H can be re-expressed as

$$k_H(x'|x) = \int_{\mathsf{Y}} \int_{\mathsf{Y}} f_{X|Y}^{(w)}(x'|y')\, l^{(w)}(y'|y)\, f_{Y|X}^{(w)}(y|x)\, \nu(dy)\, \nu(dy'),$$

where $l^{(w)}(y'|y)$ is an Mtd on Y that is reversible with respect to $f_Y^{(w)} = \int_{\mathsf{X}} f^{(w)}(x,y)\, \mu(dx)$. Since the PX-DA algorithm is driven by $f^{(w)}(x,y)$, and is itself a DA algorithm, we see that k_H is related to k_w in exactly the same way that k_l is related to k. In view of the above results, since k_H is representable, we may conclude that Haar PX-DA is better than every PX-DA algorithm in terms of both convergence rate and ARE. In particular, $\|K_H\| \leq \|K_W\|$.

5.3 Conclusion of Bayesian Probit Regression Example

In Section 3.3, we constructed a family of PX-DA algorithms for this example, one for each $(\alpha, \lambda) \in \mathbb{R}_+ \times \mathbb{R}_+$. Using the results of the previous subsection, we may conclude that every member of this family is better than the original DA algorithm based on $f(x,y)$. Furthermore, the same comparison result implies that the Haar PX-DA algorithm is better than every member of this family of PX-DA algorithms.

Recall that the original DA algorithm for this problem is geometrically ergodic. Appealing to the fact that $\|K_H\| \leq \|K_l\| \leq \|K\| < 1$, we see that the PX-DA algorithms and the Haar PX-DA algorithm for this problem are all geometrically ergodic as well.

ACKNOWLEDGMENTS

The author would like to thank Dr. Jim Hobert for supervising this project and Dr. Larry Winner for serving on the MStat committee.

References

[1] Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association* **88**: 669-679.

[2] Eaton, M. L. (1989). Group Invariance Applications in Statistics. Institute of Mathematical Statistics and the American Statistical Association, Hayward, California and Alexandria, Virginia.

[3] 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.

[4] Hobert, J.P. (2011). The Data Augmentation Algorithm: Theory and Methodology, *Handbook of Markov Chain Monte Carlo* **Chapter 10**: 253-291.

[5] Khare, K. and Hobert, J. P. (2011). A spectral analytic comparison of trace-class data augmentation algorithms and their sandwich variants, *Annals of Statistics* **39**: 2585-2606.

[6] Liu, J. S. and Wu, Y. N. (1999). Parameter expansion for data augmentation, *Journal of the American Statistical Association* **94**: 1264-1274.

[7] Meng, X. L. and van Dyk, D. A. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika* **86**: 301-320.

[8] Roberts, G. O. and Rosenthal, J. S. (1997). Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability*, **2**: 13-25.

[9] Roberts, G.O. and Rosenthal, J.S. (2004). General State Space Markov Chains and MCMC Algorithms, *Probability Surveys* **1**: 20-71.

[10] 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.