The Annals of Statistics
2011, Vol. 39, No. 5, 2585–2606
DOI: 10.1214/11-AOS916
© Institute of Mathematical Statistics, 2011

A SPECTRAL ANALYTIC COMPARISON OF TRACE-CLASS DATA AUGMENTATION ALGORITHMS AND THEIR SANDWICH VARIANTS

BY KSHITIJ KHARE AND JAMES P. HOBERT[1]

University of Florida

The data augmentation (DA) algorithm is a widely used Markov chain Monte Carlo algorithm that is easy to implement but often suffers from slow convergence. The sandwich algorithm is an alternative that can converge much faster while requiring roughly the same computational effort per iteration. Theoretically, the sandwich algorithm always converges at least as fast as the corresponding DA algorithm in the sense that $\|K^*\| \le \|K\|$, where K and K^* are the Markov operators associated with the DA and sandwich algorithms, respectively, and $\|\cdot\|$ denotes operator norm. In this paper, a substantial refinement of this operator norm inequality is developed. In particular, under regularity conditions implying that K is a trace-class operator, it is shown that K^* is also a positive, trace-class operator, and that the spectrum of K^* dominates that of K in the sense that the ordered elements of the former are all less than or equal to the corresponding elements of the latter. Furthermore, if the sandwich algorithm is constructed using a group action, as described by Liu and Wu [*J. Amer. Statist. Assoc.* **94** (1999) 1264–1274] and Hobert and Marchev [*Ann. Statist.* **36** (2008) 532–554], then there is strict inequality between at least one pair of eigenvalues. These results are applied to a new DA algorithm for Bayesian quantile regression introduced by Kozumi and Kobayashi [*J. Stat. Comput. Simul.* **81** (2011) 1565–1578].

1. Introduction.

Suppose that $f_X : \mathsf{X} \to [0, \infty)$ is an intractable probability density that we would like to explore. Consider a data augmentation (DA) algorithm [Tanner and Wong (1987), Liu, Wong and Kong (1994)] based on the joint density $f : \mathsf{X} \times \mathsf{Y} \to [0, \infty)$, which of course must satisfy

$$\int_{\mathsf{Y}} f(x, y)\nu(dy) = f_X(x).$$

We are assuming here that X and Y are two sets equipped with countably generated σ-algebras, and that $f(x, y)$ is a density with respect to $\mu \times \nu$. The Markov chain underlying the DA algorithm, which we denote by $\{X_n\}_{n=0}^{\infty}$, has Markov transition density (Mtd) given by

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

Received March 2011; revised August 2011.

[1] Supported by NSF Grant DMS-08-05860.

MSC2010 subject classifications. Primary 60J27; secondary 62F15.

Key words and phrases. Compact operator, convergence rate, eigenvalue, group action, Markov chain, Markov operator, Monte Carlo, operator norm, positive operator.

In other words, $k(\cdot|x)$ is the density of X_{n+1}, given that $X_n = x$. It is well known and easy to see that the product $k(x'|x)f_X(x)$ is symmetric in (x, x'), that is, the DA Markov chain is reversible with respect to f_X. (We assume throughout that all Markov chains on the target space, X, are Harris ergodic, that is, irreducible, aperiodic and Harris recurrent.) Of course, the DA Markov chain can be simulated by drawing alternately from the two conditional densities defined by $f(x, y)$. If the current state is $X_n = x$, then X_{n+1} is simulated in two steps: draw $Y \sim f_{Y|X}(\cdot|x)$, call the result y, and then draw $X_{n+1} \sim f_{X|Y}(\cdot|y)$.

Like its cousin the EM algorithm, the DA algorithm can be very slow to converge. A powerful method for speeding up the DA algorithm was discovered independently by Liu and Wu (1999) and Meng and van Dyk (1999). The basic idea behind the method (called "PX-DA" by Liu and Wu and "marginal augmentation" by Meng and van Dyk) 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 viable DA algorithms, some of which may converge much faster than the original. Here is a brief description of the method in the context where X and Y are Euclidean spaces, and $f(x, y)$ is a density with respect to the Lebesgue measure. Suppose that for each g in some set G, there is a function $t_g : \mathsf{Y} \to \mathsf{Y}$ that is one-to-one and differentiable. Consider a parametric family of densities (indexed by g) given by $\tilde{f}(x, y; g) = f(x, t_g(y))|J_g(y)|$, where $J_g(z)$ is the Jacobian of the transformation $z = t_g^{-1}(y)$. Note that $\int_{\mathsf{Y}} \tilde{f}(x, y; g)\,dy = f_X(x)$, so g is not identifiable in f_X. Now fix a "working prior" density on g, call it $r(g)$, and define a joint density on $\mathsf{X} \times \mathsf{Y}$ as follows:

$$f_r(x, y) = \int_G \tilde{f}(x, y; g) r(g)\,dg.$$

Clearly, the x-marginal of $f_r(x, y)$ is the target, f_X. Thus, each working prior leads to a new DA algorithm that is potentially better than the original one based on $f(x, y)$. Liu and Wu (1999), Meng and van Dyk (1999) and van Dyk and Meng (2001) find the working priors that lead to particularly fast algorithms.

Of course, one iteration of the DA algorithm based on $f_r(x, y)$ could be simulated using the usual two-step method (described above) which entails drawing from the two conditional densities defined by $f_r(x, y)$. However, Liu and Wu (1999) showed that this simulation can also be accomplished using a *three-step* procedure in which the first and third steps are draws from $f_{Y|X}(\cdot|x)$ and $f_{X|Y}(\cdot|y)$, respectively, and the middle step involves a single move according to a Markov chain on the space Y that has invariant density $f_Y(y)$. In this paper, we study a generalization of Liu and Wu's (1999) three-step procedure that was introduced by Hobert and Marchev (2008) and is now described.

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

$$k^*(x'|x) = \int_{\mathsf{Y}} \int_{\mathsf{Y}} f_{X|Y}(x'|y') R(y, dy') f_{Y|X}(y|x) \nu(dy). \tag{1}$$

It's easy to see that $k^*(x'|x)f_X(x)$ is symmetric in (x, x'), so the Markov chain defined by k^*, which we denote by $\{X_n^*\}_{n=0}^{\infty}$, is reversible with respect to f_X. If the current state of the new chain is $X_n^* = x$, then X_{n+1}^* can be simulated using the following three-steps, which are suggested by the form of k^*. Draw $Y \sim f_{Y|X}(\cdot|x)$, call the result y, then draw $Y' \sim R(y, \cdot)$, call the result y', and finally draw $X_{n+1}^* \sim f_{X|Y}(\cdot|y')$. Again, the first and third steps are exactly the two steps used to simulate the original DA algorithm. Because the draw from $R(y, \cdot)$ is "sandwiched" between the draws from the two conditional densities, Yu and Meng (2011) call this new algorithm the "sandwich algorithm" and we will follow their lead. The PX-DA/marginal augmentation method can be viewed as one particular recipe for constructing $R(y, dy')$. Another general method for building R is described in Section 4.

It is often possible to construct a sandwich algorithm that converges much faster than the underlying DA algorithm while requiring roughly the same computational effort per iteration. Examples can be found in Liu and Wu (1999), Meng and van Dyk (1999), van Dyk and Meng (2001), Marchev and Hobert (2004) and Hobert, Roy and Robert (2011). What makes this "free lunch" possible is the somewhat surprising fact that a low-dimensional perturbation on the Y space can lead to a major improvement in mixing. In fact, the chain driven by R is typically reducible, living in a small subspace of Y that is determined by its starting value. Drawing from such an R is usually much less expensive computationally than drawing from $f_{Y|X}(\cdot|x)$ and $f_{X|Y}(\cdot|y)$.

Empirical studies pointing to the superiority of the sandwich algorithm abound. Unfortunately, the development of confirmatory theoretical results has been slow. It is known that the sandwich chain always converges at least as fast as the DA chain in the operator norm sense. Indeed, Hobert and Román (2011) show that Yu and Meng's (2011) Theorem 1 can be used to show that

$$\|K^*\| \le \|R\| \|K\|, \tag{2}$$

where K, K^* and R denote the usual Markov operators defined by k, k^* and $R(y, dy')$, respectively, and $\|\cdot\|$ denotes the operator norm. (See Section 2 for more details as well as references.) Of course, we would like to be able to say that $\|K^*\|$ is strictly smaller than $\|K\|$, and this is certainly the case when $\|R\| < 1$. However, the R's used in practice typically have norm 1 (because the corresponding chains are reducible). In fact, in most applications, R is reducible and idempotent, that is, $\int_{\mathsf{Y}} R(y, dy'')R(y'', dy') = R(y, dy')$.

Hobert, Roy and Robert (2011) provided a refinement of (2) for the case in which Y is finite and R is reducible and idempotent. These authors showed that, in this case, K and K^* both have pure eigenvalue spectra that are subsets of $[0, 1)$, and that at most $m - 1$ of the eigenvalues are nonzero, where $|\mathsf{Y}| = m < \infty$. They also showed that the spectrum of K^* dominates that of K in the sense that $0 \le \lambda_i^* \le \lambda_i < 1$ for all i, where λ_i and λ_i^* denote the ith largest eigenvalues of K and K^*, respectively. Note that taking $i = 1$ yields $\|K^*\| = \lambda_1^* \le \lambda_1 = \|K\|$.

In this paper we develop results that hold in the far more common situation where $|\mathsf{Y}| = \infty$. First, we generalize Hobert, Roy and Robert's (2011) result by showing that the assumption that Y is finite can be replaced by the substantially weaker assumption that $\int_{\mathsf{X}} k(x|x)\mu(dx) < \infty$. In this more general case, K and K^* still have pure eigenvalue spectra that are subsets of $[0, 1)$ and an analogous domination holds, but the number of nonzero eigenvalues is no longer necessarily finite. Second, we show that if R is constructed using a group action, as described by Liu and Wu (1999) and Hobert and Marchev (2008), then the domination is strict in the sense that there exists at least one i such that $0 \le \lambda_i^* < \lambda_i < 1$. Finally, we apply our results to a new DA algorithm for Bayesian quantile regression that was recently introduced by Kozumi and Kobayashi (2011).

The remainder of this paper is organized as follows. Section 2 contains a brief review of the relationship between the spectral properties of Markov operators and the convergence properties of the corresponding Markov chains. The DA and sandwich algorithms are formally defined and compared in Section 3. The construction of R using group actions is discussed in Section 4, and our analysis of Kozumi and Kobayashi's DA algorithm is described in Section 5.

2. Brief review of self-adjoint Markov operators. Let $P(x, dx')$ be a generic Mtf on X that is reversible with respect to $f_X(x)\mu(dx)$. Denote the Markov chain driven by P as $\Phi = \{\Phi_n\}_{n=0}^{\infty}$. (Note that Φ is not necessarily a DA Markov chain.) The convergence properties of Φ can be expressed in terms of a related operator that is now defined. Let $L_0^2(f_X)$ be the space of real-valued functions with domain X that are square integrable and have mean zero with respect to f_X. In other words, $g \in L_0^2(f_X)$ if $\int_{\mathsf{X}} g^p(x) f_X(x)\mu(dx)$ is finite when $p = 2$, and vanishes when $p = 1$. This is a Hilbert space where the inner product of $g, h \in L_0^2(f_X)$ is defined as

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

and the corresponding norm is, of course, given by $\|g\| = \langle g, g\rangle^{1/2}$. 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

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

Note that $(Pg)(x)$ is simply the conditional expectation of $g(\Phi_{n+1})$ given that $\Phi_n = x$. Reversibility of the Mtf $P(x, dx')$ is equivalent to the operator P being self-adjoint. The (operator) norm of P is defined as

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

where $L_{0,1}^2(f_X)$ in the subset of $L_0^2(f_X)$ that contains the functions g satisfying $\int_{\mathsf{X}} g^2(x) f_X(x)\mu(dx) = 1$. It's easy to see that $\|P\| \in [0, 1]$. Roberts and Rosenthal

(1997) show that $\|P\| < 1$ if and only if Φ is geometrically ergodic. Moreover, in the geometrically ergodic case, $\|P\|$ can be viewed as the asymptotic rate of convergence of Φ [see, e.g., Rosenthal (2003), page 170].

If P satisfies additional regularity conditions, much more can be said about the convergence of the corresponding Markov chain. Assume that the operator P is compact and positive, and let $\{\alpha_i\}$ denote its eigenvalues, all of which reside in $[0, 1)$. The number of eigenvalues may be finite or countably infinite (depending on the cardinality of X), but in either case there is a largest one and it is equal to $\|P\|$. [For a basic introduction to the spectral properties of Markov operators, see Hobert, Roy and Robert (2011).] If $\operatorname{tr}(P) := \sum_i \alpha_i < \infty$, then P is a *trace-class* operator [see, e.g., Conway (1990), page 267]. As explained in Diaconis, Khare and Saloff-Coste (2008), when P is positive and trace-class, the chain's χ^2-distance to stationarity can be written explicitly as

$$\int_{\mathsf{X}} \frac{|p^n(x'|x) - f_X(x')|^2}{f_X(x')} \mu(dx') = \sum_i \alpha_i^{2n} e_i^2(x), \tag{3}$$

where $p^n(\cdot|x)$ denotes the density of Φ_n given that $\Phi_0 = x$, and $\{e_i\}$ is an orthonormal basis of eigen-functions corresponding to $\{\alpha_i\}$. Of course, the χ^2-distance serves as an upper bound on the total variation distance. Assume that the eigenvalues are ordered so that $\alpha_i \geq \alpha_{i+1}$, and let $i^* = \max\{i \in \mathbb{N} : \alpha_i = \alpha_1\}$. Asymptotically, the term $\alpha_1^{2n}(e_1^2(x) + \cdots + e_{i^*}^2(x))$ will dominate the sum on the right-hand side of (3). Hence, in this context it is certainly reasonable to call $\|P\| = \alpha_1$ the "asymptotic rate of convergence." Our focus in this paper will be on DA algorithms whose Markov operators are trace-class.

3. Spectral comparison of the DA and sandwich algorithms. As in Section 1, let $K : L_0^2(f_X) \to L_0^2(f_X)$, $K^* : L_0^2(f_X) \to L_0^2(f_X)$ and $R : L_0^2(f_Y) \to L_0^2(f_Y)$ denote the (self-adjoint) Markov operators defined by the DA chain, the sandwich chain and $R(y, dy')$, respectively. We will exploit the fact that K and K^* can be represented as products of simpler operators. Indeed, let $P_X : L_0^2(f_Y) \to L_0^2(f_X)$ map $h \in L_0^2(f_Y)$ to

$$(P_X h)(x) = \int_{\mathsf{Y}} h(y) f_{Y|X}(y|x) \nu(dy)$$

and, analogously, let $P_Y : L_0^2(f_X) \to L_0^2(f_Y)$ map $g \in L_0^2(f_X)$ to

$$(P_Y g)(y) = \int_{\mathsf{X}} g(x) f_{X|Y}(x|y) \mu(dx).$$

It is easy to see that $K = P_X P_Y$ and $K^* = P_X R P_Y$. This representation of K was used in Diaconis, Khare and Saloff-Coste (2008).

Again, as in Section 1, let $f : \mathsf{X} \times \mathsf{Y} \to [0, \infty)$ be the joint density that defines the DA Markov chain. Throughout the next two sections, we assume that f satisfies the following condition:

$$\int_{\mathsf{X}}\int_{\mathsf{Y}} f_{X|Y}(x|y) f_{Y|X}(y|x)\nu(dy)\mu(dx) < \infty. \tag{4}$$

Buja (1990) shows that (4) implies that K is a trace-class operator. It is clear that (4) holds if X and/or Y has a finite number of elements. However, (4) can also hold in situations where $|\mathsf{X}| = |\mathsf{Y}| = \infty$. Indeed, in Section 5 we establish that (4) holds for a DA algorithm for Bayesian quantile regression where X and Y are both uncountable. On the other hand, (4) certainly does not hold for all DA algorithms. For example, (4) cannot hold if the DA chain is not geometrically ergodic (because subgeometric chains cannot be trace-class). Simple examples of subgeometric DA chains can be found in Papaspiliopoulos and Roberts (2008) and Tan (2008), Chapter 4.

Condition (4) has appeared in the Markov chain Monte Carlo literature before. It is exactly the bivariate version of Liu, Wong and Kong's (1995) "Condition (b)" and it was also employed by Schervish and Carlin (1992). Unfortunately, there does not appear to be any simple, intuitive interpretation of (4) in terms of the joint density $f(x, y)$ or the corresponding Markov chain. In fact, referring to their Condition (b), Liu, Wong and Kong (1995) state that "It is standard but not easy to check and understand."

Our analysis of the DA and sandwich algorithms rests heavily upon a useful *singular value decomposition* of $f(x, y)$ whose existence is implied by (4). In particular, Buja (1990) shows that if (4) holds, then

$$\frac{f(x,y)}{f_X(x)f_Y(y)} = \sum_{i=0}^{\infty} \beta_i g_i(x) h_i(y), \tag{5}$$

where:

- $\beta_0 = 1$, $g_0 \equiv 1$, $h_0 \equiv 1$.
- $\{g_i\}_{i=0}^{\infty}$ and $\{h_i\}_{i=0}^{\infty}$ form orthonormal bases of $L^2(f_X)$ and $L^2(f_Y)$, respectively.
- $\beta_i \in [0, 1]$, and $\beta_i \le \beta_{i-1}$ for all $i \in \mathbb{N}$.
- $\int_{\mathsf{X}}\int_{\mathsf{Y}} g_i(x)h_j(y) f(x, y)\nu(dy)\mu(dx) = 0$ if $i \neq j$.

A few remarks about notation are in order. First, we state all results for the case $|\mathsf{X}| = |\mathsf{Y}| = \infty$, and leave it to the reader to make the obvious, minor modifications that are required when one or both of the spaces are finite. For example, in the singular value decomposition above, if one or both of the spaces are finite, then one or both of the orthonormal bases would have only a finite number of elements, etc. Second, we will let $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ do double duty as inner product and norm on both $L_0^2(f_X)$ and $L_0^2(f_Y)$. However, the norms of operators whose domains and ranges differ will be subscripted. The following result can be gleaned from calculations in Buja (1990), but we present a proof in Appendix A for completeness.

LEMMA 1. *Assume that* (4) *holds and let* $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots$ *denote the eigenvalues of* K, *which reside in the set* $[0, 1)$. *For each* $i \in \mathbb{N}$, $P_X h_i = \beta_i g_i$ *and* $P_Y g_i = \beta_i h_i$. *Moreover*,

$$\|P_X\|_{L_0^2(f_Y)\to L_0^2(f_X)} = \|P_Y\|_{L_0^2(f_X)\to L_0^2(f_Y)} = \beta_1$$

and $\lambda_i = \beta_i^2$.

Here is the first of our two main results.

THEOREM 1. *Assume that* (4) *holds and that* R *is idempotent with* $\|R\| = 1$. *Define* $l = \max\{i \in \mathbb{N} : \beta_i = \beta_1\}$ *and* $N = \{i \in \mathbb{N} : \beta_i > 0\}$. *Then*:

(1) K^* *is a positive*, *trace-class operator.*

(2) $\lambda_i^* \leq \lambda_i$ *for all* $i \in \mathbb{N}$, *where* $\{\lambda_i\}_{i=1}^\infty$ *and* $\{\lambda_i^*\}_{i=1}^\infty$ *denote the* (*ordered*) *eigenvalues of* K *and* K^*, *respectively.*

(3) $\lambda_i^* = \lambda_i$ *for all* $i \in \mathbb{N}$ *if and only if* $Rh_i = h_i$ *for every* $i \in N$.

(4) *A necessary and sufficient condition for* $\|K^*\| < \|K\|$ *is that the only* $a = (a_1, \dots, a_l) \in \mathbb{R}^l$ *for which*

$$R\sum_{i=1}^{l} a_i h_i = \sum_{i=1}^{l} a_i h_i \tag{6}$$

is the zero vector in $\mathbb{R}^l$.

REMARK 1. Part (3) can be rephrased as follows: $\operatorname{tr}(K^*) = \operatorname{tr}(K)$ if and only if $Rh_i = h_i$ for every $i \in N$. Also, note that $\|K^*\| = \lambda_1^*$ and $\|K\| = \lambda_1$.

PROOF OF THEOREM 1. We begin by noting that for $g \in L_0^2(f_X)$ and $h \in L_0^2(f_Y)$, we have $\langle P_X h, g\rangle = \langle h, P_Y g\rangle$. Hence,

$$\langle K^* g, g\rangle = \langle P_X R P_Y g, g\rangle = \langle R P_Y g, P_Y g\rangle = \langle R^{1/2} P_Y g, R^{1/2} P_Y g\rangle \geq 0,$$

which shows that K^* is positive. Since K is trace-class, it follows from Lemma 1 that $\operatorname{tr}(K) = \sum_{i=1}^\infty \beta_i^2 < \infty$. Now, since $\{g_i\}_{i=1}^\infty$ is an orthonormal basis for $L_0^2(f_X)$, we have

$$\begin{aligned}\operatorname{tr}(K^*) &= \sum_{i=1}^\infty \langle K^* g_i, g_i\rangle = \sum_{i=1}^\infty \langle P_X R P_Y g_i, g_i\rangle = \sum_{i=1}^\infty \langle R P_Y g_i, P_Y g_i\rangle \\ &= \sum_{i=1}^\infty \beta_i^2 \langle R h_i, h_i\rangle \leq \sum_{i=1}^\infty \beta_i^2 = \operatorname{tr}(K),\end{aligned}$$

where the inequality follows from the fact that $\langle Rh_i, h_i\rangle \leq 1$. Thus, K^* is trace-class. Moreover, it is clear that $\operatorname{tr}(K^*) = \operatorname{tr}(K)$ if and only if $\langle Rh_i, h_i\rangle = 1$ whenever $\beta_i > 0$. Since R is idempotent with norm 1, it is a projection [Conway

(1990), page 37]. Thus, for any $h \in L_0^2(f_Y)$, $\langle Rh, h\rangle = \langle h, h\rangle \Rightarrow Rh = h$. [Indeed, $\langle h, h\rangle = \langle Rh, h\rangle + \langle (I-R)h, h\rangle$, so $\langle Rh, h\rangle = \langle h, h\rangle \Rightarrow \langle (I-R)h, h\rangle = \langle (I-R)h, (I-R)h\rangle = 0$.] Consequently, $\operatorname{tr}(K^*) = \operatorname{tr}(K)$ if and only if $Rh_i = h_i$ for every i such that $\beta_i > 0$. This takes care of (3).

Now, note that $K - K^* = P_X(I-R)P_Y$ is positive since

$$\langle P_X(I-R)P_Y g, g\rangle = \langle (I-R)P_Y g, P_Y g\rangle = \langle (I-R)P_Y g, (I-R)P_Y g\rangle \geq 0.$$

Therefore, for any nonnull $g \in L_0^2(f_X)$, we have

$$\frac{\langle K^* g, g\rangle}{\langle g, g\rangle} \leq \frac{\langle K g, g\rangle}{\langle g, g\rangle}.$$

Now, for any $i \in \mathbb{N}$, the Courant–Fischer–Weyl minmax characterization of eigenvalues of compact, positive, self-adjoint operators [see, e.g., Voss (2003)] yields

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

where V denotes a subspace of $L_0^2(f_X)$, and $\dim(V)$ is its dimension. This proves (2).

All that remains is (4). Assume there exists a nonzero a such that (6) holds. We will show that $\|K^*\| = \|K\|$. Since we know that $\|K^*\| \leq \|K\| = \beta_1^2$, it suffices to identify a function $g \in L_0^2(f_X)$ such that $\|K^* g\| = \beta_1^2 \|g\|$. If we take $g = a_1 g_1 + \cdots + a_l g_l$, then

$$\|K^* g\| = \left\| K^* \sum_{i=1}^{l} a_i g_i \right\| = \left\| P_X R P_Y \sum_{i=1}^{l} a_i g_i \right\| = \left\| P_X R \sum_{i=1}^{l} a_i \beta_i h_i \right\|.$$

But $\beta_1 = \cdots = \beta_l$, and, hence,

$$\|K^* g\| = \beta_1 \left\| P_X R \sum_{i=1}^{l} a_i h_i \right\| = \beta_1 \left\| P_X \sum_{i=1}^{l} a_i h_i \right\| = \beta_1^2 \left\| \sum_{i=1}^{l} a_i g_i \right\| = \beta_1^2 \|g\|.$$

The second half of the proof is by contradiction. Assume that the only $a \in \mathbb{R}^l$ for which (6) holds is the zero vector, and assume also that $\|K^*\| = \|K\| = \beta_1^2$. By completeness of the Hilbert space, $L_0^2(f_X)$, there exists a nontrivial function $g \in L_0^2(f_X)$ such that $\|K^* g\| = \beta_1^2 \|g\|$. The rest of the argument differs depending upon whether g is in the span of $\{g_1, \ldots, g_l\}$ or not.

Case I: Assume that $g = \sum_{i=1}^{l} a_i g_i$ for some nonzero $a \in \mathbb{R}^l$. Using the results above, we have

$$\|K^* g\| = \|P_X R P_Y g\| \leq \beta_1 \|R P_Y g\| = \beta_1^2 \left\| R \sum_{i=1}^{l} a_i h_i \right\|.$$

But R is a projection, so $Rh \neq h \Rightarrow \|Rh\| \neq \|h\|$. Hence, $R\sum_{i=1}^{l} a_i h_i \neq \sum_{i=1}^{l} a_i h_i$ in conjunction with $\|R\| = 1$ yields

$$\left\| R\sum_{i=1}^{l} a_i h_i \right\| < \left\| \sum_{i=1}^{l} a_i h_i \right\| = \sqrt{\sum_{i=1}^{l} a_i^2} = \|g\|.$$

Thus, $\|K^* g\| < \beta_1^2 \|g\|$, which is a contradiction.

Case II: Assume that g is not in the span of $\{g_1, \dots, g_l\}$. In other words, $g = \sum_{i=1}^{\infty} b_i g_i$ where at least one term in the sequence $\{b_{l+1}, b_{l+2}, \dots\}$ is nonzero. Then,

$$\|P_Y g\| = \left\| P_Y \sum_{i=1}^{\infty} b_i g_i \right\| = \left\| \sum_{i=1}^{\infty} b_i \beta_i h_i \right\| = \sqrt{\sum_{i=1}^{\infty} b_i^2 \beta_i^2} < \sqrt{\beta_1^2 \sum_{i=1}^{\infty} b_i^2} = \beta_1 \|g\|.$$

It follows that

$$\|K^* g\| \le \|P_X\|_{L_0^2(f_Y) \to L_0^2(f_X)} \|R\| \|P_Y g\| < \beta_1^2 \|g\|,$$

and, again, this is a contradiction. □

4. Using a group action to construct *R*.

Following Liu and Wu (1999) and Liu and Sabatti (2000), Hobert and Marchev (2008) introduced and studied a general method for constructing practically useful versions of $R(y, dy')$ using group actions. For the remainder of this section, assume that X and Y are locally compact, separable metric spaces equipped with their Borel σ-algebras. Suppose that G is a third locally compact, separable metric space that is also a topological group. As usual, let e denote the identity element of the group. Also, let $\mathbb{R}_+ = (0, \infty)$. Any continuous function $\chi : G \to \mathbb{R}_+$ such that $\chi(g_1 g_2) = \chi(g_1)\chi(g_2)$ for all $g_1, g_2 \in G$ is called a *multiplier* [Eaton (1989)]. Clearly, a multiplier must satisfy $\chi(e) = 1$ and $\chi(g^{-1}) = 1/\chi(g)$. One important multiplier is the *modular function*, Δ, which relates the left-Haar and right-Haar measures on G. Indeed, if we denote these measures by $\omega_l(\cdot)$ and $\omega_r(\cdot)$, then $\omega_r(dg) = \Delta(g^{-1})\omega_l(dg)$. Groups for which $\Delta \equiv 1$ are called unimodular groups.

An example (that will be used later in Section 5) is the multiplicative group, $\mathbb{R}_+$, where group composition is defined as multiplication, the identity element is $e = 1$ and $g^{-1} = 1/g$. This group is unimodular with Haar measure given by $\omega(dg) = dg/g$ where dg denotes the Lebesgue measure on $\mathbb{R}_+$.

Let $F : G \times \mathsf{Y} \to \mathsf{Y}$ be a continuous function satisfying $F(e, y) = y$ and $F(g_1 g_2, y) = F(g_1, F(g_2, y))$ for all $g_1, g_2 \in G$ and all $y \in \mathsf{Y}$. The function F represents G acting topologically on the left of Y and, as is typical, we abbreviate $F(g, y)$ with gy. Now suppose there exists a multiplier, χ, such that

$$\chi(g) \int_{\mathsf{Y}} \phi(gy)\nu(dy) = \int_{\mathsf{Y}} \phi(y)\nu(dy)$$

for all $g \in G$ and all integrable $\phi : \mathsf{Y} \to \mathbb{R}$. Then the measure ν is called *relatively* (*left*) *invariant* with multiplier χ. For example, suppose that $\mathsf{Y} = \mathbb{R}^m$, $\nu(dy)$ is the Lebesgue measure, G is the multiplicative group described above, and the group action is defined to be scalar multiplication, that is, $gy = (gy_1, gy_2, \dots, gy_m)$. Then $\nu(dy)$ is relatively invariant with multiplier $\chi(g) = g^m$. Indeed,

$$g^m \int_{\mathbb{R}^m} \phi(gy)\nu(dy) = \int_{\mathbb{R}^m} \phi(y)\nu(dy).$$

We now explain how the group action is used to construct R. Define

$$m(y) = \int_G f_Y(gy)\chi(g)\omega_l(dg).$$

Assume that $m(y)$ is positive for all $y \in \mathsf{Y}$ and finite for ν-almost all $y \in \mathsf{Y}$. For the remainder of this section, we assume that $R : L_0^2(f_Y) \to L_0^2(f_Y)$ is the operator that maps $h(y)$ to

$$(Rh)(y) = \frac{1}{m(y)} \int_G h(gy) f_Y(gy)\chi(g)\omega_l(dg).$$

Hobert and Marchev (2008) show that R is a self-adjoint, idempotent Markov operator on $L_0^2(f_Y)$. The corresponding Markov chain on Y evolves as follows. If the current state is y, then the distribution of the next state is that of gy, where g is a random element from G whose density is

$$\frac{f_Y(gy)\chi(g)}{m(y)}\omega_l(dg). \tag{7}$$

Therefore, we can move from $X_n^* = x$ to X_{n+1}^* as follows: draw $Y \sim f_{Y|X}(\cdot|x)$, call the result y, then draw g from the density (7) and set $y' = gy$, and finally draw $X_{n+1}^* \sim f_{X|Y}(\cdot|y')$.

Hobert and Marchev (2008) also show that, if $\{Y_n\}_{n=0}^\infty$ denotes the Markov chain defined by R, then conditional on $Y_0 = y$, $\{Y_n\}_{n=1}^\infty$ are i.i.d. Thus, either $\{Y_n\}_{n=1}^\infty$ are i.i.d. from f_Y, or the chain is reducible.

LEMMA 2. *If $\|R\| = 1$, then the Markov operator R is a projection onto the space of functions that are invariant under the group action, that is, h is in the range of R if and only if $h(gy) = h(y)$ for all $g \in G$ and all $y \in \mathsf{Y}$.*

PROOF. First, assume that $h(gy) = h(y)$ for all $g \in G$ and all $y \in \mathsf{Y}$. Then

$$\begin{aligned}(Rh)(y) &= \frac{1}{m(y)} \int_G h(gy) f_Y(gy)\chi(g)\omega_l(dg) \\ &= \frac{h(y)}{m(y)} \int_G f_Y(gy)\chi(g)\omega_l(dg) = h(y).\end{aligned}$$

To prove the necessity, we require two results that were used repeatedly by Hobert and Marchev (2008). First,

$$\chi(g)m(gy) = \Delta(g^{-1})m(y). \tag{8}$$

Second, if $\tilde{g} \in G$ and $\phi : G \to \mathbb{R}$ is integrable with respect to ω_l, then

$$\int_G \phi(g\tilde{g}^{-1})\omega_l(dg) = \Delta(\tilde{g}) \int_G \phi(g)\omega_l(dg). \tag{9}$$

Now, fix $h \in L_0^2(f_Y)$ and $g' \in G$, and note that

$$\begin{aligned}
(Rh)(g'y) &= \frac{1}{m(g'y)} \int_G h(gg'y) f_Y(gg'y)\chi(g)\omega_l(dg) \\
&= \frac{1}{\chi(g')m(g'y)} \int_G h(gg'y) f_Y(gg'y)\chi(gg')\omega_l(dg) \\
&= \frac{\Delta(g'^{-1})}{\chi(g')m(g'y)} \int_G h(gy) f_Y(gy)\chi(g)\omega_l(dg) \\
&= \frac{1}{m(y)} \int_G h(gy) f_Y(gy)\chi(g)\omega_l(dg) \\
&= (Rh)(y),
\end{aligned}$$

where the third and fourth equalities are due to (9) and (8), respectively. □

We now show that when R is constructed using the group action recipe described above, there is at least one eigenvalue of K^* that is strictly smaller than the corresponding eigenvalue of K. To get a strict inequality, we must rule out trivial cases in which the DA and sandwich algorithms are the same. For example, if we take G to be the subgroup of the multiplicative group that contains only the point $\{1\}$, then element-wise multiplication of $y \in \mathbb{R}^m$ by g has no effect and the sandwich algorithm is the same as the DA algorithm. More generally, if

$$f_{X|Y}(x|y) = f_{X|Y}(x|gy) \qquad \forall g \in G, x \in \mathsf{X}, y \in \mathsf{Y}, \tag{10}$$

then the Mtd of the sandwich chain can be expressed as

$$\begin{aligned}
k^*(x'|x) &= \int_{\mathsf{Y}} \int_G f_{X|Y}(x'|gy) \left[\frac{f_Y(gy)\chi(g)}{m(y)} \omega_l(dg) \right] f_{Y|X}(y|x)\nu(dy) \\
&= \int_{\mathsf{Y}} \int_G f_{X|Y}(x'|y) \left[\frac{f_Y(gy)\chi(g)}{m(y)} \omega_l(dg) \right] f_{Y|X}(y|x)\nu(dy) \\
&= \int_{\mathsf{Y}} f_{X|Y}(x'|y) f_{Y|X}(y|x)\nu(dy) \\
&= k(x'|x).
\end{aligned}$$

Thus, (10) implies that the DA and sandwich algorithms are *exactly the same* and, consequently, $\operatorname{tr}(K^*) = \operatorname{tr}(K)$. In fact, as the next result shows, (10) is also necessary for $\operatorname{tr}(K^*) = \operatorname{tr}(K)$.

THEOREM 2. *If* (10) *does not hold, then* $\operatorname{tr}(K^*) < \operatorname{tr}(K)$, *so at least one eigenvalue of* K^* *is strictly smaller than the corresponding eigenvalue of* K.

PROOF. It is enough to show that $\operatorname{tr}(K^*) = \operatorname{tr}(K)$ implies (10). Recall that $N = \{i \in \mathbb{N} : \beta_i > 0\}$. By Theorem 1, $\operatorname{tr}(K^*) = \operatorname{tr}(K)$ implies that $Rh_i = h_i$ for every $i \in N$. By Lemma 2, if $Rh_i = h_i$ for every $i \in N$, then every member of the set $\{h_i : i \in N\}$ is invariant under the group action. Now, using the singular value decomposition, we see that for every $g \in G$, $x \in \mathsf{X}$, $y \in \mathsf{Y}$, we have

$$\begin{aligned} f_{X|Y}(x|y) &= \sum_{j=0}^{\infty} \beta_j g_j(x) h_j(y) f_X(x) \\ &= \sum_{j=0}^{\infty} \beta_j g_j(x) h_j(gy) f_X(x) \\ &= f_{X|Y}(x|gy). \end{aligned}$$

□

In practice, $f_{X|Y}(x|y)$ is known exactly and it's easy to verify that (10) does not hold. An example is given in the next section.

It is important to note that, while Theorem 2 guarantees strict inequality between at least one pair of eigenvalues of K and K^*, it does not preclude equality of λ_1 and λ_1^*. Thus, we could still have $\|K\| = \|K^*\|$. We actually believe that one would have to be quite unlucky to end up in a situation where $\|K\| = \|K^*\|$. To keep things simple, suppose that the largest eigenvalue of K is unique. According to Theorem 1 and Lemma 2, $\|K\| = \|K^*\|$ if and only if h_1 [from (5)] is invariant under the group action. This seems rather unlikely given that the choice of group action is usually based on simplicity and convenience. This is borne out in the toy examples analyzed by Hobert, Roy and Robert (2011) where there is strict inequality among *all* pairs of eigenvalues.

Recall from Section 1 that the PX-DA/marginal augmentation algorithm is based on a class of transformations $t_g : \mathsf{Y} \to \mathsf{Y}$, for $g \in G$. This class can sometimes be used [as the function $F(g, y)$] to construct an R as described above, and when this is the case, the resulting sandwich algorithm is the same as the optimal limiting PX-DA/marginal augmentation algorithm [Liu and Wu (1999), Meng and van Dyk (1999), Hobert and Marchev (2008)].

5. A DA algorithm for Bayesian quantile regression. Suppose $Z_1, Z_2, \dots, Z_m$ are independent random variables such that $Z_i = x_i^T \beta + \varepsilon_i$ where $x_i \in \mathbb{R}^p$ is a vector of known covariates associated with Z_i, $\beta \in \mathbb{R}^p$ is a vector of unknown regression coefficients, and $\varepsilon_1, \dots, \varepsilon_m$ are i.i.d. errors with common density given by

$$d(\varepsilon; r) = r(1-r)\big[e^{(1-r)\varepsilon} I_{\mathbb{R}_-}(\varepsilon) + e^{-r\varepsilon} I_{\mathbb{R}_+}(\varepsilon)\big],$$

where $r \in (0, 1)$. This error density, called the asymmetric Laplace density, has rth quantile equal to zero. Note that when $r = 1/2$, it is the usual Laplace density with location and scale equal to 0 and $1/2$, respectively.

If we put a flat prior on β, then the product of the likelihood function and the prior is equal to $r^m(1-r)^m s(\beta, z)$, where

$$s(\beta, z) := \prod_{i=1}^{m} \big[e^{(1-r)(z_i - x_i^T\beta)} I_{\mathbb{R}_-}(z_i - x_i^T\beta) + e^{-r(z_i - x_i^T\beta)} I_{\mathbb{R}_+}(z_i - x_i^T\beta)\big].$$

If $s(\beta, z)$ is normalizable, that is, if

$$c(z) := \int_{\mathbb{R}^p} s(\beta, z)\, d\beta < \infty,$$

then the posterior density is well defined (i.e., proper), intractable and given by

$$\pi(\beta|z) = \frac{s(\beta, z)}{c(z)}.$$

For the time being, we assume that the posterior is indeed proper.

Let U and V be independent random variables such that $U \sim \mathrm{N}(0, 1)$ and $V \sim \mathrm{Exp}(1)$. Also, define $\theta = \theta(r) = \frac{1-2r}{r(1-r)}$ and $\tau^2 = \tau^2(r) = \frac{2}{r(1-r)}$. Routine calculations show that the random variable $\theta V + \tau\sqrt{V}U$ has the asymmetric Laplace distribution with parameter r. Kozumi and Kobayashi (2011) exploit this representation to construct a DA algorithm as follows. For $i = 1, \dots, m$, let (Z_i, Y_i) be independent pairs such that $Z_i | Y_i = y_i \sim \mathrm{N}(x_i^T\beta + \theta y_i, y_i\tau^2)$ and, marginally, $Y_i \sim \mathrm{Exp}(1)$. Then $Z_i - x_i^T\beta$ has the asymmetric Laplace distribution with parameter r, as in the original model. Combining this model with the flat prior on β yields the augmented posterior density defined as

$$\pi(\beta, y|z) = \frac{1}{c'(z)}\left[\prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\tau^2 y_i}} \exp\left\{-\frac{1}{2\tau^2 y_i}(z_i - x_i^T\beta - \theta y_i)^2\right\} e^{-y_i} I_{\mathbb{R}_+}(y_i)\right],$$

where $c'(z) = r^m(1-r)^m c(z)$. Of course, $\int_{\mathbb{R}_+^m} \pi(\beta, y|z)\, dy = \pi(\beta|z)$. This leads to a DA algorithm based on the joint density $\pi(\beta, y|z)$, which is viable because, as we now explain, simulation from $\pi(\beta|y, z)$ and $\pi(y|\beta, z)$ is straightforward.

As usual, define X to be the $m \times p$ matrix whose ith row is the vector x_i^T. We assume throughout that $m \ge p$ and that X has full column rank, p. Also, let D denote an $m \times m$ diagonal matrix whose ith diagonal element is $(\tau^2 y_i)^{-1}$. A straightforward calculation shows that

$$\beta | y, z \sim \mathrm{N}_p(\mu, \Sigma),$$

where $\Sigma = \Sigma(y, z) = (X^T DX)^{-1}$, and, letting l denote an $m \times 1$ vector of ones,

$$\mu = \mu(y, z) = (X^T DX)^{-1}\left(X^T Dz - \frac{\theta}{\tau^2} X^T l\right).$$

Also, it's clear from the form of $\pi(\beta, y|z)$ that, given (β, z), the y_i's are independent, and y_i has density proportional to

$$\frac{1}{\sqrt{y_i}} \exp\left\{-\frac{1}{2\tau^2}\left[y_i(2\tau^2 + \theta^2) + \frac{(z_i - x_i^T \beta)^2}{y_i}\right]\right\} I_{\mathbb{R}_+}(y_i). \tag{11}$$

This is the density of the reciprocal of an inverse Gaussian random variable with parameters $2 + \theta^2/\tau^2$ and $\sqrt{2\tau^2 + \theta^2}/|z_i - x_i^T \beta|$. Thus, one iteration of the DA algorithm requires one draw from a p-variate normal distribution, and m independent inverse Gaussian draws. Note that in this example $\mathsf{X} = \mathbb{R}^p$ and $\mathsf{Y} = \mathbb{R}_+^m$, so both spaces have uncountably many points.

From this point forward, we restrict ourselves to the special case where $r = 1/2$, that is, to median regression. The proof of the following result, which is fairly nontrivial, is provided in Appendix B.

PROPOSITION 1. *If $r = 1/2$ and X has full column rank, then the joint density upon which Kozumi and Kobayashi's DA algorithm is based satisfies* (4). *Thus, the corresponding Markov operator is trace class.*

REMARK 2. Proposition 1 implies that, if $r = 1/2$ and X has full column rank, then the posterior is proper, that is, $c(z) < \infty$. First, by construction, the function $s(\beta, z)$ is an invariant density for the DA Markov chain, whether it is integrable (in β) or not. Now, the fact that the DA Markov operator is trace class implies that the DA Markov chain is geometrically ergodic, which in turn implies that the chain is positive recurrent. Hence, the chain cannot admit a nonintegrable invariant density [Meyn and Tweedie (1993), Chapter 10], so $s(\beta, z)$ must be integrable, that is, the posterior must be proper.

We now construct a sandwich algorithm for this problem. Let G be the multiplicative group, $\mathbb{R}_+$. Given $y \in \mathsf{Y} = \mathbb{R}_+^m$ and $g \in \mathbb{R}_+$, define gy to be scalar multiplication of each element in y by g, that is, $gy = (gy_1, gy_2, \dots, gy_m)$. Clearly, $ey = y$ and $(g_1 g_2)y = g_1(g_2 y)$, so the compatibility conditions described in Section 4 are satisfied. It is also easy to see that the Lebesgue measure on Y is relatively invariant with multiplier $\chi(g) = g^m$. When $r = 1/2$, $\pi(y|z)$ is proportional to

$$\frac{e^{-\sum_{i=1}^m y_i}}{|X^T DX|^{1/2}} \exp\left\{-\frac{1}{2} z^T D^{1/2}[I - D^{1/2}X(X^T DX)^{-1}X^T D^{1/2}]D^{1/2}z\right\}$$
$$\times \prod_{i=1}^m y_i^{-1/2} I_{\mathbb{R}_+}(y_i).$$

Therefore, in this case, the density (7) takes the form

$$\begin{aligned}&\frac{\pi(gy|z)g^m}{m(y)}\omega_l(dg)\\&\qquad\propto g^{(m-p-2)/2}e^{-g\sum_{i=1}^m y_i}\\&\qquad\quad\times\exp\biggl\{-\frac{1}{2g}z^T D^{1/2}[I-D^{1/2}X(X^TDX)^{-1}X^TD^{1/2}]D^{1/2}z\biggr\}\,dg.\end{aligned}$$

So at the middle step of the three-step procedure for simulating the sandwich chain, we draw a g from the density above and move from $y=(y_1,y_2,\dots,y_m)$ to $(gy_1,gy_2,\dots,gy_m)$, which is a random point on the ray that emanates from the origin and passes through the point y. If m happens to equal $p+1$, then this density has the same form as (11), so we can draw from it using the inverse Gaussian distribution as described earlier. Otherwise, we can employ a simple rejection sampler based on inverse Gaussian and/or gamma candidates. In either case, making one draw from this density is relatively inexpensive.

Recall that $\pi(\beta|y,z)$ is a normal density. It's easy to see that, if $g\neq 1$, then $\pi(\beta|gy,z)$ is a different normal density, which implies that (10) does not hold. Therefore, Theorems 1 and 2 are applicable and they imply that the ordered eigenvalues of the sandwich chain are all less than or equal to the corresponding eigenvalues of the DA chain, and at least one is strictly smaller. As far as we know, this sandwich algorithm has never been implemented in practice.

APPENDIX A: PROOF OF LEMMA 1

Fix $i\in\mathbb{N}$. Since $f(x,y)=f_X(x)f_Y(y)\sum_{j=0}^\infty \beta_j g_j(x)h_j(y)$, we have

$$(P_Xh_i)(x)=\int_{\mathsf{Y}} h_i(y)\Biggl(\sum_{j=0}^\infty \beta_j g_j(x)h_j(y)\Biggr)f_Y(y)\nu(dy)=\beta_i g_i(x).$$

A similar calculation shows that $P_Y g_i=\beta_i h_i$. Now, fix $h\in L_0^2(f_Y)$. Because $\{h_i\}_{i=1}^\infty$ forms an orthonormal basis for $L_0^2(f_Y)$, we have $h=\sum_{i=1}^\infty a_i h_i$. Thus,

$$\|P_Xh\|=\Biggl\|\sum_{i=1}^\infty a_i(P_Xh_i)\Biggr\|=\Biggl\|\sum_{i=1}^\infty a_i\beta_i g_i\Biggr\|=\sqrt{\sum_{i=1}^\infty a_i^2\beta_i^2}\le\beta_1\|h\|,$$

and we have equality if $h(y)=h_1(y)$. Hence, $\|P_X\|_{L_0^2(f_Y)\to L_0^2(f_X)}=\beta_1$. An analogous argument shows that $\|P_Y\|_{L_0^2(f_X)\to L_0^2(f_Y)}=\beta_1$. Now, for each $i\in\mathbb{N}$, we have

$$Kg_i=P_XP_Yg_i=\beta_iP_Xh_i=\beta_i^2g_i.$$

But $\{g_i\}_{i=1}^\infty$ form an orthonormal basis of $L_0^2(f_X)$, which proves that K has eigenvalues $\{\beta_i^2\}_{i=1}^\infty$.

APPENDIX B: PROOF OF PROPOSITION 1

Here we show that the joint density underlying Kozumi and Kobayashi's (2011) DA algorithm for median regression satisfies (4). That is, we will show that

$$\int_{\mathbb{R}_+^m} \int_{\mathbb{R}^p} \pi(\beta|y,z)\pi(y|\beta,z)\,d\beta\,dy < \infty.$$

PROOF OF PROPOSITION 1. First,

$$\pi(y|\beta,z) = c\frac{e^{-ay./2}}{\sqrt{\hat{y}}}\exp\left\{\frac{\sqrt{a}}{\tau}\sum_{i=1}^m |z_i - x_i^T\beta| - \frac{(z-X\beta)^T D(z-X\beta)}{2}\right\},$$

where $a = (2\tau^2+\theta^2)/\tau^2$, $y. = \sum_{i=1}^m y_i$, $\hat{y} = \prod_{i=1}^m y_i$, and c is a constant (that does not involve y or β). Now let

$$\mathcal{W} = \{w \in \mathbb{R}^m : w_i \in \{-1,1\} \text{ for } i = 1,2,\dots,m\}.$$

For any $\beta \in \mathbb{R}^p$ and any $\sigma > 0$, we have

$$\exp\left\{\sigma\sum_{i=1}^m |z_i - x_i^T\beta|\right\} \le \exp\left\{\sigma\sum_{i=1}^m |z_i|\right\}\sum_{w\in\mathcal{W}} \exp\{\sigma w^T X\beta\}.$$

Thus, it suffices to show that, for every $w \in \mathcal{W}$,

$$\int_{\mathbb{R}_+^m} \frac{e^{-ay./2}}{\sqrt{\hat{y}}}\left[\int_{\mathbb{R}^p}\exp\left\{\frac{\sqrt{a}}{\tau}w^T X\beta - \frac{(z-X\beta)^T D(z-X\beta)}{2}\right\}\pi(\beta|y,z)\,d\beta\right]dy$$

is finite. We start by analyzing the inner integral. First, recall that $\pi(\beta|y,z)$ is a multivariate normal density with mean $\mu = (X^T DX)^{-1}X^T Dz$ and variance $\Sigma = (X^T DX)^{-1}$. Now,

$$(z-X\beta)^T D(z-X\beta) = z^T Dz + (\beta-\mu)^T\Sigma^{-1}(\beta-\mu) - \mu^T\Sigma^{-1}\mu.$$

Therefore, the integrand (of the inner integral) can be rewritten as

$$\exp\left\{-\frac{1}{2}(z^T Dz - \mu^T\Sigma^{-1}\mu)\right\}\exp\left\{\frac{\sqrt{a}}{\tau}w^T X\beta\right\}\frac{|2X^T DX|^{1/2}}{(2\pi)^{p/2}2^{p/2}}$$
$$\times\exp\left\{-\frac{1}{2}(\beta-\mu)^T 2\Sigma^{-1}(\beta-\mu)\right\},$$

so the inner integral can be expressed as

$$\exp\left\{-\frac{1}{2}(z^T Dz - \mu^T\Sigma^{-1}\mu)\right\}\frac{1}{2^{p/2}}\int_{\mathbb{R}^p}\exp\left\{\frac{\sqrt{a}}{\tau}w^T X\beta\right\}\tilde{\pi}(\beta|y,z)\,d\beta, \tag{12}$$

where $\tilde{\pi}(\beta|y,z)$ is a multivariate normal density with mean μ and variance $\Sigma/2$. But the integral in (12) is just the moment generating function of β evaluated at the point $\sqrt{a}w^T X/\tau$. Hence, (12) is equal to

$$2^{-p/2}\exp\left\{-\frac{1}{2}(z^T Dz - \mu^T\Sigma^{-1}\mu) + \frac{\sqrt{a}}{\tau}(w^T X\mu) + \frac{a}{4\tau^2}(w^T X\Sigma X^T w)\right\}.$$

Now, straightforward manipulation yields

$$z^T Dz - \mu^T \Sigma^{-1}\mu = z^T D^{1/2}\big(I - D^{1/2}X(X^T DX)^{-1}X^T D^{1/2}\big)^2 D^{1/2}z \geq 0.$$

It follows that $e^{-(z^T Dz - \mu^T \Sigma^{-1}\mu)/2} \leq 1$. A similar calculation reveals that $w^T X\Sigma \times X^T w \leq w^T D^{-1}w = \tau^2 y.$. Hence, (12) is bounded above by

$$2^{-p/2}\exp\left\{\frac{\sqrt{a}}{\tau}(w^T X(X^T DX)^{-1}X^T Dz) + \frac{ay.}{4}\right\}.$$

Thus, it only remains to show that, for any $w \in \mathcal{W}$,

$$\int_{\mathbb{R}_+^m} \frac{1}{\sqrt{\hat{y}}}\exp\left\{-\frac{ay.}{4} + \frac{\sqrt{a}}{\tau}(w^T X(X^T DX)^{-1}X^T Dz)\right\}dy < \infty.$$

We will prove this by demonstrating that $w^T X(X^T DX)^{-1}X^T Dz$ is uniformly bounded in y.

It follows from the general matrix result established in Appendix C that, for each $i \in \{1, 2, \dots, m\}$ and all $(y_1, y_2, \dots, y_m) \in \mathbb{R}_+^m$,

$$x_i^T\left(x_i x_i^T + \sum_{j\in\{1,2,\dots,m\}, j\neq i}\frac{y_i}{y_j}x_j x_j^T\right)^{-2} x_i \leq C_i(X),$$

where $C_i(X)$ is a finite constant. Thus,

$$\begin{aligned}
\|(X^T DX)^{-1}X^T Dz\|_2 &= \left\|\sum_{i=1}^m (X^T DX)^{-1}\frac{x_i z_i}{\tau^2 y_i}\right\|_2 \\
&\leq \sum_{i=1}^m \left\|(X^T DX)^{-1}\frac{x_i z_i}{\tau^2 y_i}\right\|_2 \\
&= \sum_{i=1}^m \left\|\left(\sum_{j=1}^m \frac{x_j x_j^T}{\tau^2 y_j}\right)^{-1}\frac{x_i z_i}{\tau^2 y_i}\right\|_2 \\
&= \sum_{i=1}^m |z_i|\left\|\left(x_i x_i^T + \sum_{j\in\{1,2,\dots,m\}, j\neq i}\frac{y_i}{y_j}x_j x_j^T\right)^{-1} x_i\right\|_2 \\
&= \sum_{i=1}^m |z_i|\sqrt{x_i^T\left(x_i x_i^T + \sum_{j\in\{1,2,\dots,m\}, j\neq i}\frac{y_i}{y_j}x_j x_j^T\right)^{-2} x_i} \\
&\leq \sum_{i=1}^m |z_i| C_i(X).
\end{aligned}$$

Hence,

$$\begin{aligned}
|w^T X(X^T DX)^{-1}X^T Dz| &= \|w^T X(X^T DX)^{-1}X^T Dz\|_2 \\
&\leq \|w^T X\|_2\|(X^T DX)^{-1}X^T Dz\|_2
\end{aligned}$$

is uniformly bounded in y. This completes the proof. $\square$

APPENDIX C: A MATRIX RESULT

Fix $x_1, x_2, \dots, x_n \in \mathbb{R}^p$ where n and p are arbitrary positive integers. Now define

$$C_{p,n}(x_1; x_2, \dots, x_n) = \begin{cases} \sup_{a \in \mathbb{R}_+} x_1^T (x_1 x_1^T + a I_p)^{-2} x_1, & \text{if } n = 1, \\ \sup_{a \in \mathbb{R}_+^n} x_1^T \left(x_1 x_1^T + \sum_{i=2}^n a_i x_i x_i^T + a_1 I_p \right)^{-2} x_1, & \text{if } n \geq 2. \end{cases}$$

LEMMA 3. $C_{p,n}(x_1; x_2, \dots, x_n) < \infty$.

PROOF. We use induction on p. Note that when $p = 1$, we have

$$C_{1,n}(x_1; x_2, \dots, x_n) = \sup_{a \in \mathbb{R}_+^n} \frac{x_1^2}{(x_1^2 + \sum_{i=2}^n a_i x_i^2 + a_1)^2} = \begin{cases} 0, & x_1 = 0, \\ \dfrac{1}{x_1^2}, & x_1 \neq 0, \end{cases}$$

which is finite in either case. Thus, the result is true for $p = 1$.

Now assume that for any $n \in \mathbb{N}$ and any $x_1, \dots, x_n \in \mathbb{R}^{p-1}$,

$$C_{p-1,n}(x_1; x_2, \dots, x_n) < \infty.$$

We will complete the argument by showing that, for any $n \in \mathbb{N}$ and any $x_1, \dots, x_n \in \mathbb{R}^p$, $C_{p,n}(x_1; x_2, \dots, x_n) < \infty$. The result is obviously true when $x_1 = 0$. Suppose that $x_1 \neq 0$, and let P be an orthogonal matrix such that $P x_1 = \|x_1\|_2 e_1$, where $e_1 = (1, 0, 0, \dots, 0)^T \in \mathbb{R}^p$. For $i = 2, 3, \dots, n$, define $b_i = P x_i$. Then we have

$$\begin{aligned} & x_1^T \left(x_1 x_1^T + \sum_{i=2}^n a_i x_i x_i^T + a_1 I_p \right)^{-2} x_1 \\ & = x_1^T \left(P^T P x_1 x_1^T P^T P + \sum_{i=2}^n a_i P^T P x_i x_i^T P^T P + a_1 P^T P \right)^{-2} x_1 \\ & = x_1^T \left(P^T \left(\|x_1\|_2^2 e_1 e_1^T + \sum_{i=2}^n a_i b_i b_i^T + a_1 I_p \right) P \right)^{-2} x_1 \\ & = x_1^T P^{-1} \left(\|x_1\|_2^2 e_1 e_1^T + \sum_{i=2}^n a_i b_i b_i^T + a_1 I_p \right)^{-1} (P^T)^{-1} \end{aligned}$$

$$\times P^{-1}\left(\|x_1\|_2^2 e_1 e_1^T + \sum_{i=2}^n a_i b_i b_i^T + a_1 I_p\right)^{-1} (P^T)^{-1} x_1$$

$$= \|x_1\|_2^2 e_1^T \left(\|x_1\|_2^2 e_1 e_1^T + \sum_{i=2}^n a_i b_i b_i^T + a_1 I_p\right)^{-2} e_1.$$

Now let $A = \{i \in \{2,\dots,n\} : b_i^T e_1 = 0\}$ and let $B = \{2,\dots,n\} \setminus A$, that is, $B = \{i \in \{2,\dots,n\} : b_i^T e_1 \neq 0\}$. If $i \in A$, then there exists a $v_i \in \mathbb{R}^{p-1}$ such that

$$b_i = \begin{bmatrix} 0 \\ v_i \end{bmatrix}$$

and, if $i \in B$, then there exists a nonzero real number u_i and $v_i \in \mathbb{R}^{p-1}$ such that

$$b_i = \begin{bmatrix} u_i \\ u_i v_i \end{bmatrix}.$$

Thus, we have

$$\begin{aligned}
&x_1^T \left(x_1 x_1^T + \sum_{i=2}^n a_i x_i x_i^T + a_1 I_p\right)^{-2} x_1 \\
&\quad = \|x_1\|_2^2 e_1^T \left(\|x_1\|_2^2 e_1 e_1^T + \sum_{i=2}^n a_i b_i b_i^T + a_1 I_p\right)^{-2} e_1 \\
&\quad = \|x_1\|_2^2 e_1^T \left(\begin{bmatrix} \|x_1\|_2^2 & 0^T \\ 0 & 0 \end{bmatrix} + \sum_{i \in A} a_i \begin{bmatrix} 0 & 0^T \\ 0 & v_i v_i^T \end{bmatrix}\right. \\
&\qquad\qquad \left. + \sum_{i \in B} a_i u_i^2 \begin{bmatrix} 1 & v_i^T \\ v_i & v_i v_i^T \end{bmatrix} + a_1 I_p\right)^{-2} e_1 \\
&\quad = \|x_1\|_2^2 e_1^T \begin{bmatrix} u & v^T \\ v & W \end{bmatrix}^{-2} e_1,
\end{aligned}$$

where $u := \|x_1\|_2^2 + \sum_{i \in B} a_i u_i^2 + a_1$, $v := \sum_{i \in B} a_i u_i^2 v_i$ and

$$W := \sum_{i \in A} a_i v_i v_i^T + \sum_{i \in B} a_i u_i^2 v_i v_i^T + a_1 I_{p-1}.$$

If B is empty, then v is taken to be the zero vector in $\mathbb{R}^{p-1}$. The formula for the inverse of a partitioned matrix yields

$$\begin{aligned}
\begin{bmatrix} u & v^T \\ v & W \end{bmatrix}^{-1} &= \frac{1}{u - v^T W^{-1} v} \\
&\quad \times \begin{bmatrix} 1 & -v^T W^{-1} \\ -W^{-1} v & (u - v^T W^{-1} v) W^{-1} + W^{-1} v v^T W^{-1} \end{bmatrix}.
\end{aligned}$$

It follows that

$$e_1^T \begin{bmatrix} u & v^T \\ v & W \end{bmatrix}^{-2} e_1 = \frac{1 + v^T W^{-2} v}{(u - v^T W^{-1} v)^2}.$$

If $n = 1$ or B is empty, then

$$C_{p,n}(x_1; x_2, \dots, x_n) = \|x_1\|_2^2 \sup_{a \in \mathbb{R}_+^n} \frac{1}{(\|x_1\|_2^2 + a_1)^2} = \frac{1}{\|x_1\|_2^2} < \infty,$$

so the result holds. In the remainder of the proof, we assume that $n \geq 2$ and B is not empty.

Note that the matrix

$$\begin{bmatrix} u - \|x_1\|_2^2 & v^T \\ v & W \end{bmatrix} = \sum_{i=2}^{n} a_i b_i b_i^T + a_1 I_p$$

is positive definite, which implies that it's determinant is strictly positive, that is,

$$|W|(u - \|x_1\|_2^2 - v^T W^{-1} v) > 0.$$

Since W is also positive definite, $u - v^T W^{-1} v > \|x_1\|_2^2$. Moreover,

$$v^T W^{-2} v = \|W^{-1} v\|_2^2 = \left\| W^{-1} \left(\sum_{i \in B} a_i u_i^2 v_i \right) \right\|_2^2 \leq \left[\sum_{i \in B} \|W^{-1}(a_i u_i^2 v_i)\|_2 \right]^2.$$

Therefore,

$$\frac{1 + v^T W^{-2} v}{(u - v^T W^{-1} v)^2} \leq \frac{1 + [\sum_{i \in B} \|W^{-1}(a_i u_i^2 v_i)\|_2]^2}{\|x_1\|_2^4}.$$

Putting all of this together yields

$$x_1^T \left(x_1 x_1^T + \sum_{i=2}^{n} a_i x_i x_i^T + a_1 I_p \right)^{-2} x_1 \leq \frac{1 + [\sum_{i \in B} \|W^{-1}(a_i u_i^2 v_i)\|_2]^2}{\|x_1\|_2^2}.$$

Recall that $A \cup B = \{2, 3, \dots, n\}$. For fixed $i \in B$, let $k_{i,1}, k_{i,2}, \dots, k_{i,n-2}$ denote the $n - 2$ elements of the set $\{2, 3, \dots, n\} \setminus \{i\}$. Then we have

$$\begin{aligned} &\|W^{-1}(a_i u_i^2 v_i)\|_2^2 \\ &\quad = v_i^T \left(v_i v_i^T + \sum_{j \in A} \frac{a_j}{a_i u_i^2} v_j v_j^T + \sum_{j \in B, j \neq i} \frac{a_j u_j^2}{a_i u_i^2} v_j v_j^T + \frac{a_1}{a_i u_i^2} I_{p-1} \right)^{-2} v_i \\ &\quad \leq C_{p-1,n-1}(v_i; v_{k_{i,1}}, v_{k_{i,2}}, \dots, v_{k_{i,n-2}}). \end{aligned}$$

Thus, using the induction hypothesis, we have

$$\begin{aligned} C_{p,n}(x_1; x_2, \dots, x_n) &= \sup_{a \in \mathbb{R}_+^n} x_1^T \left(x_1 x_1^T + \sum_{i=2}^n a_i x_i x_i^T + a_1 I_p \right)^{-2} x_1 \\ &\le \frac{1 + [\sum_{i \in B} \sqrt{C_{p-1,n-1}(v_i; v_{k_{i,1}}, v_{k_{i,2}}, \dots, v_{k_{i,n-2}})}]^2}{\|x_1\|_2^2}, \end{aligned}$$

which is finite. This completes the proof of the lemma. □

REMARK 3. Note that if $x_1 x_1^T + \sum_{i=2}^n a_i x_i x_i^T$ is invertible for every $(a_2, \dots, a_n) \in \mathbb{R}_+^{n-1}$, then

$$C_{p,n}(x_1; x_2, \dots, x_n) = \sup_{(a_2,\dots,a_n) \in \mathbb{R}_+^{n-1}} x_1^T \left(x_1 x_1^T + \sum_{i=2}^n a_i x_i x_i^T \right)^{-2} x_1.$$

Acknowledgments. The authors thank Jorge Román and two anonymous reviewers for helpful comments and suggestions.

REFERENCES

BUJA, A. (1990). Remarks on functional canonical variates, alternating least squares methods and ACE. *Ann. Statist.* **18** 1032–1069. MR1062698

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

DIACONIS, P., KHARE, K. and SALOFF-COSTE, L. (2008). Gibbs sampling, exponential families and orthogonal polynomials (with discussion). *Statist. Sci.* **23** 151–200. MR2446500

EATON, M. L. (1989). *Group Invariance Applications in Statistics. NSF-CBMS Regional Conference Series in Probability and Statistics* **1**. IMS, Hayward, CA. MR1089423

HOBERT, J. P. and MARCHEV, D. (2008). A theoretical comparison of the data augmentation, marginal augmentation and PX–DA algorithms. *Ann. Statist.* **36** 532–554. MR2396806

HOBERT, J. P. and ROMÁN, J. C. (2011). Discussion of "To center or not to center: That is not the question—An ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency," by Y. Yu and X.-L. Meng. *J. Comput. Graph. Statist.* **20** 571–580.

HOBERT, J. P., ROY, V. and ROBERT, C. P. (2011). Improving the convergence properties of the data augmentation algorithm with an application to Bayesian mixture modelling. *Statist. Sci.* **26** 332–351.

KOZUMI, H. and KOBAYASHI, G. (2011). Gibbs sampling methods for Bayesian quantile regression. *J. Stat. Comput. Simul.* **81** 1565–1578.

LIU, J. S. and SABATTI, C. (2000). Generalised Gibbs sampler and multigrid Monte Carlo for Bayesian computation. *Biometrika* **87** 353–369. MR1782484

LIU, J. S., WONG, W. H. and KONG, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81** 27–40. MR1279653

LIU, J. S., WONG, W. H. and KONG, A. (1995). Covariance structure and convergence rate of the Gibbs sampler with various scans. *J. Roy. Statist. Soc. Ser. B* **57** 157–169. MR1325382

LIU, J. S. and WU, Y. N. (1999). Parameter expansion for data augmentation. *J. Amer. Statist. Assoc.* **94** 1264–1274. MR1731488

MARCHEV, D. and HOBERT, J. P. (2004). Geometric ergodicity of van Dyk and Meng's algorithm for the multivariate Student's t model. *J. Amer. Statist. Assoc.* **99** 228–238. MR2054301

MENG, X.-L. and VAN DYK, D. A. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika* **86** 301–320. MR1705351

MEYN, S. P. and TWEEDIE, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer, London. MR1287609

PAPASPILIOPOULOS, O. and ROBERTS, G. (2008). Stability of the Gibbs sampler for Bayesian hierarchical models. *Ann. Statist.* **36** 95–117. MR2387965

ROBERTS, G. O. and ROSENTHAL, J. S. (1997). Geometric ergodicity and hybrid Markov chains. *Electron. Commun. Probab.* **2** 13–25 (electronic). MR1448322

ROSENTHAL, J. S. (2003). Asymptotic variance and convergence rates of nearly-periodic Markov chain Monte Carlo algorithms. *J. Amer. Statist. Assoc.* **98** 169–177. MR1965683

SCHERVISH, M. J. and CARLIN, B. P. (1992). On the convergence of successive substitution sampling. *J. Comput. Graph. Statist.* **1** 111–127. MR1268760

TAN, A. (2008). Analysis of Markov chain Monte Carlo algorithms for random effects models. Ph.D. thesis, Dept. Statistics, Univ. Florida, Gainesville, FL.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82** 528–550. MR0898357

VAN DYK, D. A. and MENG, X.-L. (2001). The art of data augmentation (with discussions). *J. Comput. Graph. Statist.* **10** 1–50.

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

YU, Y. and MENG, X.-L. (2011). To center or not to center: That is not the question—An ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency (with discussion). *J. Comput. Graph. Statist.* **20** 531–615.

DEPARTMENT OF STATISTICS
UNIVERSITY OF FLORIDA
GAINESVILLE, FLORIDA 32611
USA
E-MAIL: kdkhare@stat.ufl.edu
jhobert@stat.ufl.edu