Improving the Data Augmentation algorithm in the two-block setup

Subhadip Pal, Kshitij Khare and James P. Hobert
University of Florida

Abstract

The Data Augmentation (DA) approach to approximate sampling from an intractable probability density f_X is based on the construction of a joint density, $f_{X,Y}$, whose conditional densities, $f_{X|Y}$ and $f_{Y|X}$, can be straightforwardly sampled. However, many applications of the DA algorithm do not fall in this "single-block" setup. In these applications, X is partitioned into two components, $X = (U, V)$, in such a way that it is easy to sample from $f_{Y|X}$, $f_{U|V,Y}$ and $f_{V|U,Y}$. We refer to this alternative version of DA, which is effectively a three-variable Gibbs sampler, as "two-block" DA.

We develop two methods to improve the performance of the DA algorithm in the two-block setup. These methods are motivated by the Haar PX-DA algorithm, which has been developed in previous literature to improve the performance of the single-block DA algorithm. The Haar PX-DA algorithm, which adds a computationally inexpensive extra step in each iteration of the DA algorithm while preserving the stationary density, has been shown to be optimal among similar techniques. However, as we illustrate, the Haar PX-DA algorithm does not lead to the required stationary density f_X in the two-block setup. Our methods incorporate suitable generalizations and modifications to this approach, and work in the two-block setup. A theoretical comparison of our methods to the two-block DA algorithm, a much harder task than the single-block setup due to non-reversibility and structural complexities, is provided. We successfully apply our methods to applications of the two-block DA algorithm in Bayesian robit regression and Bayesian quantile regression.

Keywords and phrases: Data Augmentation algorithm, sandwich algorithm, two-block DA algorithm, group action, Haar measure

1 Introduction

Suppose that the random variable X has an intractable probability density f_X that we would like to explore. The standard Data Augmentation (DA) approach [Tanner and Wong (1987),

Liu, Wong and Kong (1994)] consists of constructing a random variable Y such that it is easy to sample from the conditional densities $f_{X|Y}$ and $f_{Y|X}$. This allows us to construct the DA Markov chain, whose one-step transition starting at x can be described as follows.

- Make a draw from $f_{Y|X}(\cdot \mid x)$. Call it y.
- Make a draw from $f_{X|Y}(\cdot \mid y)$. Call it x'.

This Markov chain is reversible with respect to f_X, and consequently can be used to obtain approximate samples from f_X. We will refer to this version of the DA algorithm as the "single-block DA algorithm".

The Markov chain obtained from 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), who call it "PX-DA", and Meng and van Dyk (1999), who call it "Marginal Augmentation" (MA). A more general version of the method was considered in Hobert and Marchev (2008). The basic idea behind the method is to introduce an additional step in the DA algorithm, which is much cheaper computationally than the two conditional draws. Suppose X takes values in $\mathcal{X}$, and Y takes values in $\mathcal{Y}$. Let $R(\cdot|\cdot)$ be a transition kernel on $\mathcal{Y}$ which is reversible with respect to f_Y. Construct a Markov chain, whose one-step transition starting at x can be described as follows.

- Make a draw from $f_{Y|X}(\cdot \mid x)$. Call it y.
- Make a draw from $R(\cdot \mid y)$. Call it y'.
- Make a draw from $f_{X|Y}(\cdot \mid y')$. Call it x'.

This modified version of the DA algorithm is known as the "sandwich DA algorithm" (Yu and Meng (2011), Khare and Hobert (2011)). The corresponding sandwich Markov chain can also be shown to be reversible, and have f_X as a stationary density, and consequently can be used to obtain approximate samples from f_X. The sandwich algorithm offers a "free lunch" in the sense that 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. See Liu and Wu (1999), Meng and van Dyk (1999), van Dyk and Meng (2001), Marchev and Hobert (2004) and Hobert, Roy and Robert (2011) for examples. In fact, the chain driven by R is typically reducible, living in a small subspace of $\mathcal{Y}$ that is determined by its starting value. Drawing from such an R is usually much less expensive computationally than conditional draws from $f_{Y|X}$ and $f_{X|Y}$. Recent results have theoretically established qualitative superiority of the sandwich algorithm over the DA algorithm. Hobert and Marchev (2008) establish that under farily general conditions, the sandwich algorithm is

at least as good as the corresponding DA algorithm (in terms of a suitable operator norm). The Haar PX-DA algorithm introduced by Liu and Wu (1999), and generalized by Hobert and Marchev (2008), has been shown by these authors to be the best sandwich algorithm in terms of efficiency and operator norm. Khare and Hobert (2011) establish necessary and sufficient conditions for the Haar PX-DA algorithm to be strictly better than the corresponding DA algorithm, given that the Markov operator corresponding to the DA algorithm is trace class.

Clearly, a significant amount of progress has been made in understanding and applying the sandwich method to the DA algorithm in the single-block DA case. *However, there are various applications of the DA algorithm where a "two-block" version of the DA algorithm is used instead of the single-block version (see for example Albert and Chib (1993), Kozumi and Kobayashi (2011), and Park and Casella (2008)).* The two-block DA algorithm can be described as follows. Suppose that we would like to sample from the intractable density f_X. Suppose X can be partitioned into 2 blocks or components, $X = (U, V)$, and a random variable Y can be constructed such that sampling from $f_{Y|X}$, $f_{U|V,Y}$ and $f_{V|U,Y}$ is feasible. This allows the construction of the two-block DA Markov chain, whose one-step transition starting from $x = (u, v)$ can be described as follows.

- Make a draw from $f_{Y|X}(\cdot \mid x)$. Call it y.
- Make a draw from $f_{U|V,Y}(\cdot \mid v, y)$. Call it u'.
- Make a draw from $f_{V|U,Y}(\cdot \mid u', y)$. Call it v'.

The two-block DA Markov chain has f_X as a stationary density. Consequently, under standard ergodicity assumptions, this Markov chain can be used to draw approximate samples from f_X. *However, unlike the single-block DA algorithm, the Markov chain for the two-block DA algorithm is not reversible with respect to f_X in general.*

The study of whether and how the "sandwich" idea of inserting an extra step can be usefully generalized in the two-block setup is clearly of interest, and has not been undertaken previously. In particular, can a feasible recipe, generalizing the standard recipes in the single-block case, be developed to improve the performance of the two-block DA algorithm? As we show in Section 2, the Haar PX-DA algorithm of Liu and Wu (1999) and Hobert and Marchev (2008), is not applicable in this case, as the corresponding sandwich Markov chain may not necessarily have f_X as a stationary density.

In this paper, we develop two recipes to construct an extra step in the two-block DA setting, such that the resulting sandwich Markov chains still have f_X as a stationary density. These recipes can be viewed as generalizations of the Haar PX-DA algorithm of Liu and Wu (1999) and Hobert and Marchev (2008). Here is an outline and summary of the main

results in the paper. In Section 2, we provide two recipes for constructing an extra step for the two-block DA algorithm using group actions, and show that the corresponding sandwich Markov chains have f_X as a stationary density (Lemma 1). As opposed to the single-block case, the two-block DA Markov chain and the corresponding sandwich Markov chain are both not reversible in general. Hence, a theoretical analysis and comparison of the DA and sandwich Markov chains is much more difficult in the two-block setup. In Section 3, we consider the lifted versions of the sandwich and DA Markov chains on $\mathcal{X} \times \mathcal{Y}$. The properties of the lifted versions are shown to be closely related to the properties of the corresponding DA and sandwich Markov chains (Lemma 2 and 3). We then show that the lifted version of the sandwich Markov chain is at least as good as the lifted version of the DA Markov chain in terms of the operator norm (Lemma 4). In Section 4, we illustrate the utility of our methods by considering two applications of the two-block DA algorithm: the Bayesian robit regression algorithm of Albert and Chib (1993), and the Bayesian median regression algorithm of Kozumi and Kobayashi (2011).

2 Constructing the extra step for two-block DA

Assume that $\mathcal{U}, \mathcal{V}$ and $\mathcal{Y}$ are locally compact, separable metric spaces equipped with their Borel σ-algebras. Assume further that $\mu_{\mathcal{U}}, \mu_{\mathcal{V}}$ and $\mu_{\mathcal{Y}}$ are σ-finite measures on $\mathcal{U}, \mathcal{V}$ and $\mathcal{Y}$ respectively. Let $\mathcal{X} = \mathcal{U} \times \mathcal{V}$ and $\mu_{\mathcal{X}} = \mu_{\mathcal{U}} \times \mu_{\mathcal{V}}$. Consider the random variable $X = (U, V)$ and Y, and let $f_{X,Y} = f_{U,V,Y}$, denote their joint probability density with respect to $\mu_{\mathcal{X}} \times \mu_{\mathcal{Y}}$. Let f_X, f_Y denote the associated marginal densities, and $f_{Y|X}, f_{U|V,Y}$ and $f_{V|U,Y}$ denote the associated conditional densities. Suppose we wish to simulate from f_X, but direct simulation is not possible. The two-block DA algorithm is useful in this situation, assuming it is feasible to sample from the three conditional densities listed above. For $x = (u, v), x' = (u', v') \in \mathcal{X}$, the Markov transition density (with respect to $\mu_{\mathcal{X}}$) of the corresponding Markov chain is given by

$$q(x' \mid x) = \int_{\mathcal{Y}} f_{Y|X}(y \mid x) f_{U|V,Y}(u' \mid v, y) f_{V|U,Y}(v' \mid u', y) \mu_{\mathcal{Y}}(dy). \tag{1}$$

It is known, and easy to verify, that the two-block DA markov chain has f_X as a stationary density.

To construct the extra step in the two-block setup, we require some assumptions on $\mathcal{Y}$ as in Hobert and Marchev (2008). Suppose G is a locally compact separable metric space and also a topological group. Suppose G acts topologically left to $\mathcal{Y}$, i.e., there exists a function $F : G \times \mathcal{Y} \to \mathcal{Y}$ such that $F(e, y) = y$ for all $y \in G$, where e is the identity element in G, and $F(g_1, F(g_2, y)) = F(g_1 g_2, y)$ for all $g_1, g_2 \in G$ and $y \in \mathcal{Y}$. For simplicity of notation we

write this function as $F(g, y) := gy$.

We assume the existence of a multiplier $\chi : G \to \Re^+$, i.e., χ is a continuous function, and $\chi(g_1 g_2) = \chi(g_1)\chi(g_2)$ for all $g_1, g_2 \in G$. We assume that the measure $\mu_{\mathcal{Y}}$ on $\mathcal{Y}$ is relatively left invariant with respect to the multiplier χ, i.e.,

$$\chi(g) \int_{\mathcal{Y}} h(gy)\mu_{\mathcal{Y}}(dy) = \int_{\mathcal{Y}} h(y)\mu_{\mathcal{Y}}(dy) \tag{2}$$

for all $g \in G$ and for all integrable $h : \mathcal{Y} \to \mathbb{R}$.

Since G is a locally compact separable metric space and also a topological group, there exists a left Haar measure ν_l satisfying

$$\int_G h(\tilde{g}g)\nu_l(dg) = \int_G h(g)\nu_l(dg) \tag{3}$$

for all $\tilde{g} \in G$ and $h : G \to \mathbb{R}$. The left Haar measure is unique up to a multiplicative constant. Moreover, there exists a real-valued function Δ on G called the (right) modular function of the group, with the property that $\nu_r(dg) := \Delta(g^{-1})\nu_l(dg)$ where $\nu_r(dg)$ is a right-Haar measure, which satisfies the obvious analogue of (3). Two useful properties that will be useful in the subsequent analysis are as follows.

$$\int_G h(g\tilde{g}^{-1})\nu_l(dg) = \Delta(\tilde{g}) \int_G h(g)\nu_l(dg). \tag{4}$$

$$\int_G h(g^{-1})\nu_l(dg) = \int_G h(g)\Delta(g^{-1})\nu_l(dg). \tag{5}$$

Given $x = (u, v) \in \mathcal{X}, y \in \mathcal{Y}$, we define three probability densities $f^0_{x,y}, f^1_{x,y}$ and $f^2_{x,y}$ on G (with respect to ν_l). All these densities can potentially be used to construct the extra step for the two-block DA algorithm. In fact, Liu and Wu (1999) and Hobert and Marchev (2008) use f^0 to construct the Haar PX-DA sandwich Markov chain in the single-block setup. However, we will show that using f^0 leads to an 'invalid' extra step, while using f^1 or f^2 leads to a valid extra step in the two-block setup. In this paper, the 'validity' of a proposed extra step will be synonymous to the resulting sandwich Markov chain having f_X as a stationary density. Define

$$f^0_{x,y}(g) = \frac{f_Y(gy)\chi(g)}{C_0(x, y)}\ , \tag{6}$$

$$f^1_{x,y}(g) = \frac{f_{V,Y}(v, gy)\chi(g)}{C_1(x, y)}\ , \tag{7}$$

and

$$f^2_{x,y}(g) = \frac{f_{U,V,Y}(u, v, gy)\chi(g)}{C_2(x, y)} \ , \tag{8}$$

where $C_j(x, y) = C_j(u, v, y)$ is the normalizing constant to ensure that $f^j_{x,y}$ is a probability density with respect to the left Haar measure ν_l. It is assumed here that $C_j(x, y) < \infty$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$ and $j = 0, 1, 2$. Note that $C_0(x, y)$ does not depend on x, and $C_1(x, y) = C_1(u, v, y)$ does not depend on u.

We can now construct sandwich Markov chains corresponding to each of the densities $f^j,\ j = 0, 1, 2$. One step of the sandwich Markov chain corresponding to f^j, starting at $x = (u, v)$, can be described as follows.

- Make a draw from $f_{Y|X}(\cdot \mid x)$. Call it y.
- Make a draw from $f^j_{x,y}$. Call it g.
- Make a draw from $f_{U|V,Y}(\cdot \mid v, gy)$. Call it u'.
- Make a draw from $f_{V|U,Y}(\cdot \mid u', gy)$. Call it v'.

The transition density of this Markov chain is given by

$$q_j(x' \mid x) = \int_{\mathcal{Y}} \int_G f_{Y|X}(y \mid x) f^j_{x,y}(g) f_{U|V,Y}(u' \mid v, gy) f_{V|U,Y}(v' \mid u', gy) \nu_l(dg) \mu_{\mathcal{Y}}(dy). \tag{9}$$

The following lemma establishes that the Markov transition densities q_1 and q_2 have f_X as a stationary density.

Lemma 1.

$$\int_{\mathcal{X}} f_X(x) q_1(x' \mid x) \mu_X(dx) = \int_{\mathcal{X}} f_X(x) q_2(x' \mid x) \mu_X(dx) = f_X(x') \ \forall x' \in \mathcal{X}.$$

A proof of this result is provided in the Supplemental document. It is also shown in the Supplemental document that our recipes can be extended to the k-block DA algorithm, with $k > 2$.

Remark 1. *Note that in order to use the DA or the sandwich chains to sample from f_X or to approximate linear functionals of f_X, one needs to establish Harris-ergodicity of these Markov chains. For this purpose, in addition to showing that f_X is a stationary density, one needs to prove that these Markov chains are $\mu_{\mathcal{X}}$-irreducible (see Asmussen and Glynn (2011)). A sufficient condition for $\mu_{\mathcal{X}}$-irreducibility is that the Markov transition density is strictly positive everywhere (see page 87 in Meyn and Tweedie (1993)). This is indeed the*

case in particular with all examples of DA and sandwich Markov chains considered in Section 4, and in general with most of the DA and sandwich Markov chains that we have come across in practice.

Remark 2. *In this section we have considered the setting where X can be partioned into (U, V), and Y can be constructed so that it is feasible to sample from the three conditionals. We would like to point out that other slightly different settings can be adopted to the framework above.*

(a) Suppose we want to sample from the intractable density $f_{\tilde{X}}$ of $\tilde{X}$. Suppose it is possible to construct $\tilde{Y} = (\tilde{T}, \tilde{W})$ such that it is feasible to sample from conditionals $f_{\tilde{X}|\tilde{T},\tilde{W}}, f_{\tilde{T}|\tilde{X},\tilde{W}}$ and $f_{\tilde{W}|\tilde{X},\tilde{T}}$. This setting can be easily adopted to our framework by considering $U = \tilde{X}, V = \tilde{T}$ and $Y = \tilde{W}$. Our methods will yield a sample from $f_{U,V} = f_{\tilde{X},\tilde{T}}$, which in particular will provide the desired sample from $f_{\tilde{X}}$.

(b) Suppose we want to sample from the intractable density $f_{\tilde{X}}$ of $\tilde{X} = (\tilde{U}, \tilde{V})$. Suppose that it is feasible to sample from $f_{\tilde{V}|\tilde{U}}$. Hence, essentially the task is to sample from $f_{\tilde{U}}$. Suppose we can construct $\tilde{Y}$ such that it is easy to sample from the three conditionals $f_{\tilde{U}|\tilde{V},\tilde{Y}}, f_{\tilde{V}|\tilde{U},\tilde{Y}}$ and $f_{\tilde{Y}|\tilde{U},\tilde{V}}$.

Suppose no feasible group action can be found on $\tilde{\mathcal{Y}}$ (the range of $\tilde{Y}$). However a feasible group action based extra step can be constructed on $\tilde{\mathcal{V}}$. This setting can easily be adopted to our framework by considering $U = \tilde{U}, V = \tilde{Y}$ and $Y = \tilde{V}$. Our methods will yield a sample from $f_{U,V} = f_{\tilde{U},\tilde{Y}}$, which in particular will provide the desired sample from $f_{\tilde{U}}$.

The Markov transition density q_0 does not have f_X as a stationary density in general. We start by providing a heuristic argument to explain this, followed by a concrete example. Suppose we start with $X_0 = (U_0, V_0) \sim f_X$ and obtain $X_1 = (U_1, V_1)$ as follows.

- Make a draw from $f_{Y|X}(\cdot \mid X_0)$. Call it Y_1.
- Make a draw from $f^0_{X_0,Y_1}$. Call it g.
- Make a draw from $f_{U|V,Y}(\cdot \mid V_0, gY_1)$. Call it U_1.
- Make a draw from $f_{V|U,Y}(\cdot \mid U_1, gY_1)$. Call it V_1.

It is clear that the conditional density of X_1 given X_0 is indeed q_0. Firstly, note that $(X_0, Y_1) \sim f_{X,Y}$. Note that $f^0_{X_0,Y_1}$ does not depend on X_0. Hence, although it can be established that $gY_1 \sim f_Y$, it is not necessarily true that $(V_0, gY_1) \sim f_{V,Y}$. As a cascading effect of this fact, eventually it is not necessarily true that $(U_1, V_1) \sim f_X$.

We now provide a simple example of a two-block setup where the Markov transition density q_0 does not have $f_X = f_{U,V}$ as a stationary density.

Example 1. *Let $\mathcal{U} = \mathcal{V} = \mathcal{Y} = \{-1, 1\}$. Define each one of $\mu_{\mathcal{U}}, \mu_{\mathcal{V}}$ and $\mu_{\mathcal{Y}}$ to be the counting measure on $\{-1, 1\}$. Define the joint probability density $f_{U,V,Y}$ as*

$$f_{U,V,Y}(u, v, y) = \begin{cases} \frac{1}{4} & \text{if } u + v + y > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Consider the group $G = \{-1, 1\}$ (with multiplication as the group operation). The group G acts on $\mathcal{Y}$ through scalar multiplication, i.e., $F(g, y) = g \times y$. It is easy to check that the group G is unimodular, and the Haar measure is given by the counting measure on $\{-1, 1\}$. Consider the multiplier χ satisfying $\chi(1) = \chi(-1) = 1$. It is again easy to check that χ is left invariant with respect to the measure $\mu_{\mathcal{Y}}$. Note that $f_{U,V}(1, 1) = \frac{1}{2}$. However,

$$\begin{aligned} & \sum_{(u,v)\in\mathcal{U}\times\mathcal{V}} f_{U,V}(u, v) q_0((1, 1) \mid (u, v)) \\ = & \sum_{(u,v)\in\mathcal{U}\times\mathcal{V}} \sum_{g\in\{-1,1\}} \sum_{y\in\{-1,1\}} f_{U,V}(u, v) f_{Y|U,V}(y \mid u, v) f^0_{u,v,y}(g) f_{U|V,Y}(1 \mid v, gy) \\ & \qquad \times f_{V|U,Y}(1 \mid 1, gy) \\ = & \frac{27}{64} \end{aligned}$$

Hence, $f_{U,V}$ is not a stationary density for q_0.

3 Comparison of DA and sandwich Markov chains

Let $L^2_0(f_X)$ denote the Hilbert space of all square-integrable mean-zero functions with respect to f_X. The space $L^2_0(f_X)$ is equipped with the usual inner product and norm, defined by

$$< h_1, h_2 >_{L^2_0(f_X)} := \int_{\mathcal{X}} h_1(x) h_2(x) f_X(x) \mu_{\mathcal{X}}(dx), \ \|h\|_{L^2_0(f_X)} := \sqrt{< h, h >_{L^2_0(f_X)}}.$$

The Markov operator on $L^2_0(f_X)$ corresponding to the DA Markov transition density q in (1) will be denoted by Q. The Markov operator on $L^2_0(f_X)$ corresponding to the sandwich Markov transition density q_j in (9) will be denoted by Q_j.

In the single-block setup, the extra step corresponding to the density f^0 in (6) (which corresponds to the Haar PX-DA step of Liu and Wu (1999) and Hobert and Marchev (2008)) is a valid step, i.e., the corresponding sandwich Markov chain has the desired stationary

density. Let Q_0^{SB} denote the operator corresponding to this sandwich Markov chain (SB stands for single-block). Let Q^{SB} denote the operator corresponding to the single-block DA Markov chain. Then, Q_0^{SB} and Q^{SB} are both self adjoint operators on $L_0^2(f_X)$. In this context, Hobert and Marchev (2008) prove that $< Q_0^{SB}h, h >_{L_0^2(f_X)} \leq < Q^{SB}h, h >_{L_0^2(f_X)}$ for all $h \in L_0^2(f_X)$, which in particular implies that Q_0^{SB} has a smaller operator norm than Q^{SB}.

However, as we have shown in Section 2, in the two-block setup, the operator Q_0 (corresponding to the choice of f^0) is not a feasible option. The operators Q_1, Q_2 and Q are not self-adjoint. Also, the densities $f_{V,Y}$ and $f_{U,V,Y}$ are used to construct the extra step corresponding to Q_1 and Q_2 respectively (as opposed to the use of f_Y in constructing the extra step corresponding to Q_0 and Q_0^{SB}). All these structural differences make the task of comparing the sandwich and DA Markov chains much harder in the two-block case. However, it is desirable to have a theoretical assurance or guarantee that the proposed sandwich Markov operators are 'better' than the DA operator in terms of a reasonable criterion. This will make them a really attractive option in cases where the extra step has negligible computational cost as compared to the other steps in the DA algorithm. In this section, we will show that although it is hard to directly compare the sandwich operators Q_1 and Q_2 to the DA operator Q, one can meaningfully compare their *lifted versions* on $L_0^2(f_{U,V,Y})$.

3.1 Constructing the lifted version of the DA Markov operator Q

Consider the Markov chain on $\mathcal{U} \times \mathcal{V} \times \mathcal{Y}$ whose one step transition from (u, v, y) to $(\tilde{u}, \tilde{v}, \tilde{y})$ can be described as follows.

(i) Make a draw from $f_{U|V,Y}(\cdot \mid v, y)$. Call it $\tilde{u}$.

(ii) Make a draw from from $f_{V|U,Y}(\cdot \mid \tilde{u}, y)$. Call it $\tilde{v}$.

(iii) Make a draw from $f_{Y|U,V}(\cdot \mid \tilde{u}, \tilde{v})$. Call it $\tilde{y}$.

Let $\bar{q}$ denote the Markov transition density of this Markov chain, and $\bar{Q}$ denote the corresponding Markov Operator on $L_0^2(\mathcal{U} \times \mathcal{V} \times \mathcal{Y})$. Following the terminology introduced in Chen et al. (1999), we refer to $\bar{Q}$ as the lifted version of Q on $L_0^2(\mathcal{U} \times \mathcal{V} \times \mathcal{Y})$. We show below that Q and $\bar{Q}$ are intimately related. To see this, consider the following construction. Start with $U_0 = u$ and $V_0 = v$ for arbitrarily fixed $u \in \mathcal{U}, v \in \mathcal{V}$. Draw Y_0 from $f_{Y|U,V}(\cdot \mid u, v)$. We now construct a sequence $\{(U_i, V_i, Y_i)\}_{i>0}$, where given $(U_{i-1}, V_{i-1}, Y_{i-1})$, the triplet (U_i, V_i, Y_i) is generated by using steps (i)-(iii) (with $(u, v, y) = (U_{i-1}, V_{i-1}, Y_{i-1})$ and $(\tilde{u}, \tilde{v}, \tilde{y}) = (U_i, V_i, Y_i)$). The following lemma establishes the connection between the Markov operators Q and $\bar{Q}$.

Lemma 2. *(a) Given (U_0, V_0), the sequence $\{(U_i, V_i)\}_{i>0}$ is a Markov chain with transition density q.*

(b) Given (U_0, V_0, Y_0), the sequence $\{(U_i, V_i, Y_i)\}_{i>0}$ is a Markov chain with transition density $\bar{q}$.

(c) For any $h \in L_0^2(f_{U,V})$ and $n \geq 1$,

$$(Q^n h)(u, v) = E_{Y_0 \sim f_{Y|U,V}(\cdot|u,v)} \left[(\bar{Q}^n h^*)(u, v, Y_0) \right].$$

Here $h^* \in L_0^2(f_{U,V,Y})$ is defined by $h^*(u, v, y) = h(u, v)$ for every $u \in \mathcal{U}, v \in \mathcal{V}$ and $y \in \mathcal{Y}$.

(d) For any $h \in L_0^2(f_{U,V,Y})$ and $n \geq 1$,

$$E_{Y_0 \sim f_{Y|U,V}(\cdot|u,v)} \left[(\bar{Q}^n h)(u, v, Y_0) \right] = (Q^n \widetilde{h})(u, v).$$

Here $\widetilde{h} \in L_0^2(f_{U,V})$ is defined by $\widetilde{h}(u, v) = \int_{\mathcal{Y}} h(u, v, y) f_{Y|U,V}(y|u, v) \mu_{\mathcal{Y}}(dy)$ for every $u \in \mathcal{U}$ and $v \in \mathcal{V}$.

(e) For every $u \in \mathcal{U}, v \in \mathcal{V}$ and $n \geq 1$,

$$\left\| Q_{u,v}^n - f_{U,V} \right\|_{TV} = \left\| \bar{Q}_{u,v,f(\cdot|u,v)}^n - f_{U,V,Y} \right\|_{TV}.$$

Here $\| \cdot \|_{TV}$ denotes the total variation norm, $Q_{u,v}^n$ denotes the n-step density of the Markov chain corresponding to Q started at (u, v), and $\bar{Q}_{u,v,f(\cdot|u,v)}^n$ denotes the n-step density of the Markov chain corresponding to $\bar{Q}$ started at (u, v, Y_0) such that $Y_0 \sim f_{Y|U,V}(\cdot \mid u, v)$.

A proof of this result is provided in the Supplemental document.

3.2 Constructing the lifted version of the sandwich Markov operators Q_1 and Q_2

We now construct the lifted version of Q_1. Consider the Markov chain on $\mathcal{U} \times \mathcal{V} \times \mathcal{Y}$ whose one step transition from (u, v, y) to $(\tilde{u}, \tilde{v}, \tilde{y})$ can be described as follows.

(I) Make a draw from $f_{U,V,Y}^1$. Call it g.

(II) Make a draw from $f_{U|V,Y}(\cdot \mid v, gy)$. Call it $\tilde{u}$.

(III) Make a draw from from $f_{V|U,Y}(\cdot \mid \tilde{u}, gy)$. Call it $\tilde{v}$.

(IV) Make a draw from $f_{Y|U,V}(\cdot \mid \tilde{u}, \tilde{v})$. Call it $\tilde{y}$.

Let $\bar{q}_1$ denote the Markov transition density of this Markov chain, and $\bar{Q}_1$ denote the corresponding Markov Operator on $L_0^2(\mathcal{U} \times \mathcal{V} \times \mathcal{Y})$. As earlier, we show below that Q_1 and $\bar{Q}_1$ are intimately related, and refer to $\bar{Q}_1$ as the lifted version of Q_1 on $L_0^2(\mathcal{U} \times \mathcal{V} \times \mathcal{Y})$. To see this, consider the following construction. Start with $U_{S,0} = u$ and $V_{S,0} = v$ for arbitrarily fixed $u \in \mathcal{U}, v \in \mathcal{V}$. Draw $Y_{S,0}$ from $f_{Y|U,V}(\cdot \mid u, v)$. We now construct a sequence $\{(U_{S,i}, V_{S,i}, Y_{S,i})\}_{i>0}$, where given $(U_{S,i-1}, V_{S,i-1}, Y_{S,i-1})$, the triplet $(U_{S,i}, V_{S,i}, Y_{S,i})$ is generated from steps (I)-(IV) (with $(u, v, y) = (U_{S,i-1}, V_{S,i-1}, Y_{S,i-1})$ and $(\tilde{u}, \tilde{v}, \tilde{y}) = (U_{S,i}, V_{S,i}, Y_{S,i})$). The following lemma establishes the connection between the Markov operators Q_1 and $\bar{Q}_1$.

Lemma 3. *(a) Given $(U_{S,0}, V_{S,0})$, the sequence $\{(U_{S,i}, V_{S,i})\}_{i>0}$ is a Markov chain with transition density q_1.*

(b) Given $(U_{S,0}, V_{S,0}, Y_{S,0})$, the sequence $\{(U_{S,i}, V_{S,i}, Y_{S,i})\}_{i>0}$ is a Markov chain with transition density $\bar{q}_1$.

(c) For any $h \in L_0^2(f_{U,V})$ and $n \geq 1$,

$$(Q_1^n h)(u, v) = E_{Y_0 \sim f_{Y|U,V}(\cdot|u,v)} \left[(\bar{Q}_1^n h^*)(u, v, Y_0) \right].$$

Here $h^* \in L_0^2(f_{U,V,Y})$ is as defined in Lemma 2, part (c).

(d) For any $h \in L_0^2(f_{U,V,Y})$ and $n \geq 1$,

$$E_{Y_0 \sim f_{Y|U,V}(\cdot|u,v)} \left[(\bar{Q}_1^n h)(u, v, Y_0) \right] = (Q_1^n \widetilde{h})(u, v).$$

Here $\widetilde{h} \in L_0^2(f_{U,V})$ is as defined in Lemma 2, part (d).

(e) For every $u \in \mathcal{U}, v \in \mathcal{V}$ and $n \geq 1$,

$$\left\| (Q_1)_{u,v}^n - f_{U,V} \right\|_{TV} = \left\| (\bar{Q}_1)_{u,v,f(\cdot|u,v)}^n - f_{U,V,Y} \right\|_{TV}.$$

Here $(\bar{Q}_1)_{u,v,f(\cdot|u,v)}^n$ denotes the n-step density of the Markov chain corresponding to $\bar{Q}_1$ started at (u, v, Y_0) such that $Y_0 \sim f_{Y|U,V}(\cdot \mid u, v)$.

The proof follows exactly along the lines of the proof of Lemma 2. One can construct the lifted version $\bar{Q}_2$ of Q_2 by replacing f^1 by f^2 in the above construction.

3.3 Comparison of $\bar{Q}_j,\ j = 1, 2$ and $\bar{Q}$

We start by defining some operators on $L_0^2(f_{X,Y})$. Let $P_{U,V}$ be the projection operator into $L_0^2(f_{U,V})$, i.e., if $h \in L_0^2(f_{X,Y})$, then

$$(P_{U,V}h)(u, v) = \int_{\mathcal{Y}} h(u, v, y) f_{Y|U,V}(y \mid u, v)\mu_{\mathcal{Y}}(dy).$$

Similarly, let $P_{U,Y}$ be the projection operator into $L_0^2(f_{U,Y})$, and $P_{V,Y}$ be the projection operator into $L_0^2(f_{V,Y})$. The projection operators $P_{U,V}, P_{U,Y}$ and $P_{V,Y}$ correspond to the appropriate conditional draws in the DA algorithm. It follows from the definition of $\bar{Q}$ that $\bar{Q} = P_{U,V}P_{U,Y}P_{V,Y}$.

We now introduce the operators R_1 and R_2 corresponding to the extra step in the sandwich algorithms corresponding to Q_1 and Q_2 respectively. For $h \in L_0^2(f_{X,Y})$ and $j = 1, 2$, define

$$(R_j h)(u, v, y) = \int_G h(u, v, gy) f^j_{u,v,y}(g)\nu_l(dg).$$

It follows by the definition of Q_1 and Q_2 that for $j = 1, 2$,

$$\bar{Q}_j = R_j P_{U,V} P_{U,Y} P_{V,Y}.$$

Since R_j is a Markov operator, it follows that $\|R_j\| \leq 1$. Here, the norm of an operator A on $L_0^2(f_{X,Y})$ is defined in the standard way as $\|A\| := \sup_{h \in L_0^2(f_{X,Y})} \frac{\|Ah\|}{\|h\|}$. Hence,

$$\|\bar{Q}_j\| = \|R_j P_{U,V} P_{U,Y} P_{V,Y}\| \leq \|R_j\|\|\bar{Q}\| \leq \|\bar{Q}\|.$$

The analysis above leads to the following lemma.

Lemma 4. *The lifted versions $\bar{Q}_1$ and $\bar{Q}_2$ of the sandwich operators are at least as good in operator norm as the lifted version $\bar{Q}$ of the DA operator.*

Remark 3. *Lemma 2 and Lemma 3 establish the equivalence of the the DA and sandwich operators with their respective lifted versions in total variation norm. On the other hand, the comparison in Lemma 4 is in terms of the operator norms of the lifted versions. Hence, our analysis does not provide a direct comparsion between the operator norms of the original DA and sandwich operators. However, the equivalence of total variation norms does imply that the mixing properties of the DA and sandwich Markov chains are intimately related with their lifted versions. Hence, the analysis does provide evidence in favor of the conjecture that in standard situations, the sandwich operators are at least as good as the DA operator in terms of the operator norm.*

Remark 4. *Note that in most practical situations, the extra step corresponds to a draw from a low dimensional subspace of $\mathcal{Y}$. Hence, typically $\|R_1\| = \|R_2\| = 1$. So, the current method of proof only establishes that $\bar{Q}_1$ or $\bar{Q}_2$ is at least as good as $\bar{Q}$ in operator norm. In the much more amenable single-block case, Khare and Hobert (2011) provide conditions under which the sandwich operator is strictly better than the DA operator in terms of the operator norm. A similar analysis in the two-block setup seems to be much more complicated and is a topic of current research.*

4 Applications

In this section, we consider two different applications of the two-block DA algorithm, namely, Albert and Chib's (1993) robit regression algorithm and Kozumi and Kobayashi's (2011) Bayesian quantile regression algorithm. For each application, we identify an appropriate group G acting on the relevant space $\mathcal{Y}$. We then investigate the feasibility of the sandwich algorithms corresponding to Q_1 and Q_2. It turns out that for one of these applications, the extra step corresponding to Q_2 is feasible and computationally inexpensive (Q_1 is not feasible), and for the other application, the extra step corresponding to Q_1 is feasible and computationally inexpensive (Q_2 is not feasible). We then compare the computational cost and efficieny of the DA algorithm and the respective sandwich algorithms on standard datasets.

4.1 Bayesian robit regression

The standard Bayesian robit regression model is formulated as follows. Let $R_1, R_2, \ldots, R_n$ be independent Bernoulli random variables such that $P(R_i = 1 \mid \beta) = H_\nu(s_i^T\beta)$ where $s_i \in \mathbb{R}^p$ is a vector of known covariates associated with R_i, $\beta \in \mathbb{R}^p$ is a vector of unknown regression coefficients and $H_\nu(\cdot)$ is the distribution function of a t random variable with ν degrees of freedom. A commonly used prior for β is the improper flat prior. For inferential purposes, one would like to sample from the posterior density of β. Let $r = (r_1, r_2, \cdots, r_n)$ be the vector of observed responses. The posterior density of β is given by

$$f_{\beta|R}(\beta' \mid r) \propto \prod_{i=1}^{n} \left[H_\nu(s_i^T\beta')\right]^{r_i} \left[1 - H_\nu(s_i^T\beta')\right]^{1-r_i}. \tag{10}$$

Let $c(r)$ denote the normalizing constant; that is

$$c(r) = \int_{\mathbb{R}^p} \prod_{i=1}^{n} \left[H_\nu(s_i^T\beta')\right]^{r_i} \left[1 - H_\nu(s_i^T\beta')\right]^{1-r_i} d\beta'. \tag{11}$$

Assuming that $c(r) < \infty$ (a sufficient condition is provided in the Supplemental document), the posterior density in (10) is a proper probability density. The widely-used DA algorithm for exploring this highly intractable density was introduced by Albert and Chib (1993). It is based on introducing $\Lambda \in \mathbb{R}^n_+$ and $Z \in \mathbb{R}^n$, so that the joint posterior density of β, Λ, Z is given by

$$\begin{aligned} f_{\beta,\Lambda,Z|R}(\beta', \lambda, z \mid r) &= \frac{1}{c(r)} \Pi_{i=1}^n \left[\{I(z_i > 0)I(r_i = 1) + I(z_i \leq 0)I(r_i = 0)\} \sqrt{\frac{\lambda_i}{2\pi}} e^{-\frac{\lambda_i}{2}(z_i - s_i^T\beta')^2} \right] \\ &\quad \times \left[\left\{ \lambda_i^{\frac{\nu}{2}-1} e^{-\frac{\nu\lambda_i}{2}} \right\} \right]. \end{aligned} \tag{12}$$

A straight forward calculation, partly using the fact that the t-distribution is a scale mixture of normal distributions, shows that

$$\int_{\mathbb{R}^n} \int_{\mathbb{R}^n_+} f_{\beta,\Lambda,Z|R}(\beta', \lambda, z \mid r) d\lambda dz = f_{\beta|R}(\beta' \mid r).$$

Albert and Chib (1993) show that the full conditional distributions of β, Λ and Z can be specified as follows.

- $$\beta \mid \Lambda = \lambda, Z = z, R = r \sim \mathcal{N}_p \left(\hat{\beta}(z, \lambda), (S^t W(\lambda) S)^{-1} \right),$$

  where $S := ((s_{i,j}))_{1 \leq i \leq n, 1 \leq j \leq p}$ is the covariate matrix, $W(\lambda)$ is a diagonal matrix with diagonal entries $\lambda_1, \lambda_2, \ldots, \lambda_n$, and $\hat{\beta}(z, \lambda) := (S^T W(\lambda) S)^{-1} S^T W(\lambda) z$.

- $Z_1, Z_2, \ldots, Z_n$ are conditionally independent with

  $$Z_i \mid \beta = \beta', \Lambda = \lambda, R = r \sim \begin{cases} \mathcal{N}^+(s_i^T\beta', \frac{1}{\lambda_i}) & \text{if } r_i = 1 \\ \mathcal{N}^-(s_i^T\beta', \frac{1}{\lambda_i}) & \text{if } r_i = 0, \end{cases}$$

  for $i = 1, 2, \ldots, n$. Here $\mathcal{N}^+$ is the truncated normal distribution on $(0, \infty)$ and $\mathcal{N}^-$ is the truncated normal distribution on $(-\infty, 0)$.

- $\Lambda_1, \Lambda_2, \ldots, \Lambda_n$ are conditionally independent with

  $$\Lambda_i \mid \beta = \beta', Z = z, R = r \sim Gamma\left(\frac{\nu + 1}{2}, \frac{\nu + (z_i - s_i^T\beta')^2}{2} \right)$$

  for $i = 1, 2, \ldots, n$. The notation $Gamma(a, b)$ corresponds to the density $f(x) = \frac{1}{b^a \Gamma(a)} x^{a-1} e^{-bx}$.

The DA algorithm in Albert and Chib (1993) based on the above conditional densities can be adapted into the two-block setup considered in Section 2, with $U = \beta, V = \lambda$ and $Y = Z$ (see Remark 2 (a)). See Roy (2012) for an alternate DA approach. Consider the topological group $G = \mathbb{R}_+$, which acts on $\mathbb{R}^n$ (the sample space of Z) through scalar multiplication. The left Haar measure for G is given by $\nu_l(dg) = \frac{dg}{g}$. Consider the multiplier χ on G defined by $\chi(g) = g^n$. It can be easily seen that the Lebesgue measure on $\mathbb{R}^n$ is relatively invariant with respect to χ. We now derive the form of the extra step density f^1 on G defined in (7), for this particular example. Fix $\beta' \in \mathbb{R}^p, \lambda \in \mathbb{R}^n_+, z \in \mathbb{R}^n$. Using (7) and (12), it follows that

$$\begin{aligned} f^1_{\beta',\lambda,z}(g) &\propto f_{\Lambda,Z|R}(\lambda, gz \mid r)\chi(g)\nu_l(dg) \\ &\propto \left(\int_{\mathbb{R}^p} e^{-\sum_{i=1}^n \frac{\lambda_i}{2}(gz_i - s_i^T\beta'')^2} d\beta''\right) g^{n-1}dg. \end{aligned} \tag{13}$$

Recall that S is the covariate matrix, $W(\lambda)$ is a diagonal matrix with diagonal entries $\lambda_1, \lambda_2, \ldots, \lambda_n$, and $\hat{\beta}(\lambda, z) = (S^TW(\lambda)S)^{-1}S^TW(\lambda)z$. Let $\Omega(\lambda) := W(\lambda)S(S^TW(\lambda)S)^{-1}S^TW(\lambda)$. It follows by standard algebraic manipulations that

$$\begin{aligned} \sum_{i=1}^n \left[\lambda_i(z_i - s_i^T\beta'')^2\right] &= \sum_{i=1}^n \left[\lambda_i(z_i^2 - 2z_is_i^T\beta'' + \beta''^T s_i s_i^T\beta'')\right] \\ &= z^TW(\lambda)z - 2z^TW(\lambda)S\beta'' + \beta''^T(S^TW(\lambda)S)\beta'' \\ &= (\beta'' - \hat{\beta}(\lambda, z))^T(S^TW(\lambda)S)(\beta'' - \hat{\beta}(\lambda, z)) - z^T(\Omega(\lambda) - W(\lambda))z. \end{aligned} \tag{14}$$

It follows by (14) that

$$\int_{\mathbb{R}^p} e^{-\sum_{i=1}^n \frac{\lambda_i}{2}(z_i - s_i^T\beta'')^2} d\beta'' = \left[(2\pi)^p det(S^TW(\lambda)S)\right]^{\frac{1}{2}} e^{-\frac{z^T(W(\lambda)-\Omega(\lambda))z}{2}}.$$

Combining this with (13), we get that

$$f^1_{\beta',\lambda,z}(g) \propto g^{n-1}e^{-g^2\frac{z^T(W(\lambda)-\Omega(\lambda))z}{2}}. \tag{15}$$

Note that

$$W(\lambda) - \Omega(\lambda) = W(\lambda)^{\frac{1}{2}}\left(I - W(\lambda)^{\frac{1}{2}}S(S^TW(\lambda)S)^{-1}S^TW(\lambda)^{\frac{1}{2}}\right)W(\lambda)^{\frac{1}{2}}.$$

Since $W(\lambda)^{\frac{1}{2}}S(S^TW(\lambda)S)^{-1}S^TW(\lambda)^{\frac{1}{2}}$ is the projection matrix for the column space of $W(\lambda)^{\frac{1}{2}}S$, and $\lambda_i > 0$ for $1 \leq i \leq n$, it follows that $W(\lambda) - \Omega(\lambda)$ is a positive definite matrix. It fol-

lows by (15) that $f^1_{\beta',\lambda,z}$ is the density of the square root of a $Gamma\left(\frac{n}{2}, \frac{z^T(W(\lambda)-\Omega(\lambda))z}{2}\right)$ random variable. Hence, a draw from f^1 is a univariate, computationally inexpensive and straightforward draw.

We now derive the expression for the extra step density f^2 (on G) defined in (8). Fix $\beta' \in \mathbb{R}^p, \lambda \in \mathbb{R}^n_+, z \in \mathbb{R}^n$. Using (8) and (12), it follows that

$$f^2_{\beta',\lambda,z}(g) \propto f_{\beta,\Lambda,Z|R}(\beta', \lambda, gz \mid r)\chi(g)\nu_l(dg) \propto e^{-\sum_{i=1}^n \frac{\lambda_i}{2}(gz_i - s_i^T\beta')^2} g^{n-1} dg.$$

Clearly, this is a non-standard density. However, a reasonably efficient rejection sampler can be derived to generate samples from it. The details of this rejection sampler are provided in the Supplemental document.

We implement the DA algorithm and the two sandwich algorithms (with extra step governed by f^1 and f^2) on van Dyk and Meng's (2001) lupus data, which consists of triplets $\{(r_i, s_{i1}, s_{i2})\}_{i=1}^{55}$, where s_{i1} and s_{i2} are covariates indicating the levels of certain antibodies in the i^{th} individual and r_i is an indicator for latent membranous lupus nephritis (1 for presence and 0 for absence). We consider fitting a Bayesian robit regression model on this data with $\nu = 20$ degrees of freedom. Note that $p = 3$ (including the intercept term). It can be easily verified (using the sufficient conditions for posterior propriety in the Supplmental document) that the posterior density is a proper probability density. For all three Markov chains, we choose the inital value of β to be the vector of all ones. The initial entries of λ are generated independently from the prior $Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$ density.

The DA algorithm on the lupus data (in the single-block probit setting) is known to be very slow (see for example Roy and Hobert (2007)). We expect similar behaviour in the two-block robit setting. Hence, following Roy and Hobert (2007), we ran all the three Markov chains for a burn-in period of 10^6 iterations. The next 10^5 iterations were used to obtain the autocorrelations for the function $h(\beta) = (S\beta)^T(S\beta)$ for all three Markov chains. This function was a natural choice as it is the *drift function* used to prove geometric ergodicity of the closely related probit DA Markov chain (Roy and Hobert, 2007). Also, arguments similar to Lemma S3 in the Supplemental document imply that in the current setting, $E_{f_{\beta|R=r}}[h^2(\beta)] < \infty$. Table 1 provides the first five autocorrelations for the DA chain and the two sandwich chains. Clearly, the sandwich chain with f^1 leads to a substantial reduction in autocorrelations, whereas the sandwich chain with f^2 only provides a nominal reduction in autocorrelations (and also takes more time per iteration than the sandwich chain with f^1). Figure 1 provides a comparison of the first 50 autocorrelations for the DA chain and the sandwich chain with f^1. We also note that the time difference between the DA chain and the sandwich chain with f^1 (for 1100000 iterations) is only 3 minutes. Hence, the extra step (corresponding

Lag	1	2	3	4	5
Autocorrelation (DA)	0.99968	0.99935	0.99903	0.9987	0.99838
Autocorrelation (Sandwich f^1)	0.9357	0.87852	0.82743	0.78168	0.74012
Autocorrelation (Sandwich f^2)	0.9996	0.99921	0.99882	0.99842	0.9804

Table 1: First five autocorrelations for DA and sandwich Markov chains for the Bayesian robit regression model applied to lupus data

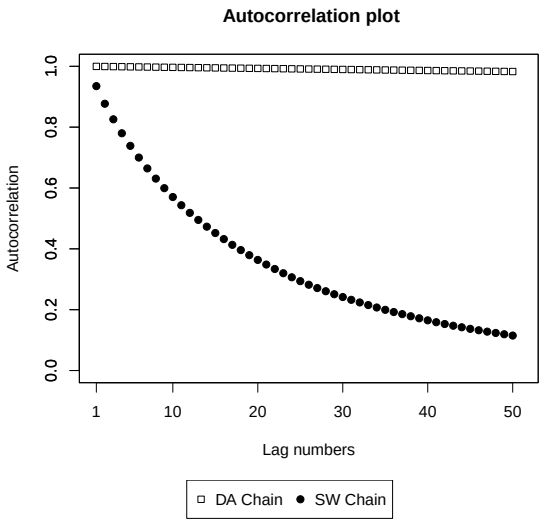

Figure 1: Autocorrelation plot for $h(\beta) = (S\beta)^T(S\beta)$ for DA Markov chain and sandwich Markov chain with f^1 for Bayesian robit regression applied to lupus data

to f^1) takes on average only 0.7% more time per iteration. However, the improvement in the autocorrelations is tremendous. This strongly indicates that it is worthwhile to use the sandwich chain (with f^1) over the DA Markov chain in this setting. The Supplemental document contains a table with the first 40 autocorrelations for the DA chain and both the sandwich chains.

4.2 Bayesian quantile regression

Consider a linear model $R_i = s_i^T\beta + \epsilon_i; i = 1, 2, ..., n$ subject to the condition that the α^{th} quantile of the distribution of ϵ_i is 0. Here $\beta \in \mathbb{R}^p$, and $s_1, s_2, \ldots, s_n \in \mathbb{R}^p$ are the covariate vectors. A standard method to estimate the regression parameter β is by minimizing the function

$$\sum_{i=1}^{n} \left(\rho_\alpha(R_i - s_i^T\beta)\right) \tag{16}$$

where $\rho_\alpha(x) = x\left(\alpha - I(x < 0)\right)$ with $I(\cdot)$ denoting the standard indicator function. Linear

programming methods are commonly used, as it is not easy to minimize the non-differentiable objective function.

Yu and Moyeed (2001) pointed out that the minimization problem in (16) is exactly same as finding the MLE for β if the errors are assumed to be i.i.d. with the asymmetric Laplace density given by

$$g_\alpha(\epsilon) = \alpha(1-\alpha)e^{-\rho_\alpha(\epsilon)}.$$

It is easy to see that this error density has α^{th} quantile equal to zero. (When $\alpha = 1/2$, g_α becomes the standard Laplace density with location and scale equal to 0 and 1/2, respectively). Using this analogy as a motivation, Yu and Moyeed (2001) suggested the following Bayesian approach. Let $R_i = s_i^T\beta + \sigma\epsilon_i; i = 1, 2, ..., n$ where $\{\epsilon_i\}_{i=1}^n$ are i.i.d. with common density g_α, and $\sigma > 0$ is an unknown scale parameter. Assume that β and σ are apriori independent with $\beta \sim \mathcal{N}_p(\beta_0, B_0)$ and $\sigma \sim IG\left(\frac{n_0}{2}, \frac{t_0}{2}\right)$. The notation $IG(a, \eta)$ corresponds to the density $f(x) = \frac{\eta^a}{\Gamma(a)}x^{-a-1}e^{-\frac{\eta}{x}}$. It turns out that the posterior density $f_{\beta,\sigma|R=r}$ is intractable. Hence, Yu and Moyeed (2001) use the Metropolis-Hastings algorithm to sample from the posterior.

As an alternative to the ineffecient Metropolis-Hastings algorithm, Kozumi and Kobayashi (2011) develop a Data Augmentation approach. Their algorithm exploits a latent data formulation of the quantile regression model that is based on a normal/exponential mixture representation of the asymmetric Laplace distribution.
Define $\theta := \theta(\alpha) = \frac{1-2\alpha}{\alpha(1-\alpha)}$ and $\tau^2 := \tau^2(\alpha) = \frac{2}{\alpha(1-\alpha)}$. Let $\{(R_i, Z_i)\}_{i=1}^n$ be random pairs such that $R_i \mid Z_i = z_i, \beta, \sigma \sim \mathcal{N}(s_i^T\beta + \theta z_i, z_i\sigma\tau^2)$ and $Z_i \mid \beta, \sigma \sim Exp(\sigma)$. It can be easily verified that the marginal density of $\frac{R_i - s_i^T\beta}{\sigma}$ given (β, σ) is indeed g_α. If we use the normal and inverse gamma priors for β and σ specified above, the joint posterior density of β, Z, σ given $R = r$ is given by

$$\begin{aligned} & f_{\beta,Z,\sigma|R=r}(\beta', z, \sigma') \\ \propto \; & \frac{1}{(\sigma')^{\frac{n}{2}}\prod_{i=1}^n\sqrt{z_i}} e^{-\frac{(r-S\beta'-\theta z)^T D_z^{-1}(r-S\beta'-\theta z)}{2\tau^2\sigma'}} \frac{e^{-\sum_{i=1}^n \frac{z_i}{\sigma'}}}{(\sigma')^n} \\ & \times e^{-\frac{(\beta'-\beta_0)^T B_0^{-1}(\beta'-\beta_0)}{2}} (\sigma')^{-\frac{n_0}{2}-1} e^{-\frac{t_0}{2\sigma'}}. \end{aligned}$$

Here D_z denotes the diagonal matrix with diagonal entries $z_1, z_2, \ldots, z_n$ and S denotes the matrix of covariates. Straightforward calculations (see Khare and Hobert (2012)) show that

$$\int_{\mathbb{R}_+^n} f_{\beta,\sigma,Z|R=r}(\beta', \sigma', z)dz = f_{\beta,\sigma|R=r}(\beta', \sigma').$$

Kozumi and Kobayashi (2011) show that the full conditional (posterior) distributions of β, σ

and Z can be specified as follows.

- $$\beta \mid \sigma = \sigma', Z = z, R = r \sim \mathcal{N}_p \left(\Omega(z, \sigma')^{-1} \left(\frac{S^T D_z^{-1}(r - \theta z)}{\tau^2 \sigma'} + B_0^{-1} \beta_0 \right), \Omega(z, \sigma')^{-1} \right),$$

  where $\Omega(z, \sigma') := \frac{S^T D_z^{-1} S}{\tau^2 \sigma'} + B_0^{-1}$.

- $$\sigma \mid \beta = \beta', Z = z, R = r \sim IG \left(\frac{\tilde{n}_0}{2}, \frac{\tilde{s}(\beta', z)}{2} \right),$$

  where $\tilde{n}_0 := n_0 + 3n$ and

  $$\tilde{s}(\beta', z) := \frac{(r - \theta z - S\beta')^T D_z^{-1}(r - \theta z - S\beta')}{\tau^2} + 2 \sum_{i=1}^{n} z_i + t_0.$$

- $z_i,\ i = 1, 2, \ldots, n$ are conditionally independent, with

  $$z_i \mid \beta = \beta', \sigma = \sigma', R = r \sim GIG \left(\frac{1}{2}, \frac{1}{\sigma'} \left(\frac{\theta^2}{\tau^2} + 2 \right), \frac{(r_i - s_i^T \beta')^2}{\tau^2 \sigma'} \right),$$

  where the density of a $GIG(p, a, b)$ random variable (at $x > 0$) is proportional to $x^{p-1} e^{-\frac{ax}{2} - \frac{b}{2x}}$.

It is straightforward to establish that the posterior density of σ given β is an inverse gamma density, which is easy to sample from. Hence to sample from the target posterior density of (β, σ), it suffices to devise a mechanism to sample from the posterior density of β. Using Remark 2 (b), we adapt the DA algorithm based on the above conditional densities into the two-block DA setup considered in Section 2, with $U = \beta, V = Z$ and $Y = \sigma$. Our methods will then yield a sample from $f_{U,V}$, i.e., the posterior density of (β, Z). This in particular yields the desired sample from the posterior density of β.

Consider the topological group $G = \mathbb{R}_+$, which acts on $\mathbb{R}_+$ (the sample space of σ) through scalar multiplication. The left Haar measure for G is given by $\nu_l(dg) = \frac{dg}{g}$. Consider the multiplier χ on G defined by $\chi(g) = g$. It can be easily seen that the Lebesgue measure on $\mathbb{R}_+$ is relatively invariant with respect to χ. We now derive the form of the extra step density f^1 (on G) defined in (7). Let $\beta' \in \mathbb{R}^p, \sigma' \in \mathbb{R}_+$ and $z \in \mathbb{R}_+^n$. Note that

$$\begin{aligned} f^1_{\beta', \sigma', z}(g) &\propto f_{Z, \sigma \mid R}(z, g\sigma' \mid r) \chi(g) \nu_l(dg) \\ &\propto \left(\int_{\mathbb{R}^p} f_{\beta, Z, \sigma \mid R}(\beta'', z, g\sigma' \mid r) d\beta'' \right) \chi(g) \nu_l(dg). \end{aligned}$$

After routine computations involving the multivariate normal density, it can be proved that

$$f^1_{\beta',\sigma',z}(g) \propto \left|\frac{S^T D_z^{-1} S}{\tau^2 g\sigma'} + B_0^{-1}\right|^{-\frac{1}{2}} g^{-\frac{3n+n_0}{2}-1} \times e^{-\frac{1}{g}\left(\frac{(r-\theta z)^T D_z^{-1}(r-\theta z)}{2\tau^2\sigma'} + \sum_{i=1}^n \frac{z_i}{\sigma'} + \frac{t_0}{2\sigma'}\right)} \times e^{\frac{1}{2}\mu^T_{z,\sigma'}\left(\frac{S^T D_z^{-1} S}{\tau^2 g\sigma'} + B_0^{-1}\right)^{-1}\mu_{z,\sigma'}},$$

where $\mu_{z,\sigma} := \frac{S^T D_z^{-1}(r-\theta z)}{\tau^2 g\sigma'} + B_0^{-1}\beta_0$. Clearly, this is a non-standard density. However, if $\beta_0 = 0$, then ignoring B_0^{-1} in the above density gives an upper bound which corresponds to an inverse gamma density (up to proportionality). This fact can be used to derive a simple and efficient rejection sampler for the density f^1 in the current setting.

It turns out that the extra step density f^2 (on G) defined in (8) is easy to sample from. Let $\beta' \in \mathbb{R}^p, \sigma' \in \mathbb{R}_+$ and $z \in \mathbb{R}^n_+$. Note that

$$\begin{aligned} f^2_{\beta',\sigma',z}(g) &\propto f_{\beta,Z,\sigma|R}(\beta', z, g\sigma' \mid r)\chi(g)\nu_l(dg) \\ &\propto \frac{1}{(g\sigma')^{\frac{n}{2}}} e^{-\frac{(r-S\beta'-\theta z)^T D_z^{-1}(r-S\beta'-\theta z)}{2\tau^2 g\sigma'}} \frac{e^{-\sum_{i=1}^n \frac{z_i}{g\sigma'}}}{(g\sigma')^n} \\ &\quad \times e^{-\frac{(\beta'-\beta_0)^T B_0^{-1}(\beta'-\beta_0)}{2}} (g\sigma')^{-\frac{n_0}{2}-1} e^{-\frac{t_0}{2g\sigma'}} dg \\ &\propto g^{-\frac{3n+n_0}{2}-1} e^{-\frac{1}{g}\left(\frac{(r-S\beta'-\theta z)^T D_z^{-1}(r-S\beta'-\theta z)}{2\tau^2\sigma'} + \sum_{i=1}^n \frac{z_i}{\sigma'} + \frac{t_0}{2\sigma'}\right)} dg. \end{aligned}$$

It follows that $f^2_{\beta',\sigma',z}$ is the density of an $IG\left(\frac{3n+n_0}{2}, \frac{(r-S\beta'-\theta z)^T D_z^{-1}(r-S\beta'-\theta z)}{2\tau^2\sigma'} + \sum_{i=1}^n \frac{z_i}{\sigma'} + \frac{t_0}{2\sigma'}\right)$ random variable.

We implemented the DA algorithm and the sandwich algorithms (with extra step governed by f^1 and f^2) on Wang et al's (1998) patent data. The data set contains information about the number of patent applications from 70 pharmaceutical and biomedical companies in 1976. A more detailed explanation of the data can be found in Hall et al. (1988). In Tsionas (2003) and Kozumi and Kobayashi (2011), the relationship between patents and research and development (R & D) spending is analyzed through the following model.

$$\log(1+N) = \beta_1 + \beta_2 \log(RD) + \beta_3 (\log(RD))^2 + \beta_4 \log\left(\frac{RD}{SALE}\right) + \epsilon,$$

where N is the number of patent applications, RD is R & D spending, and $\frac{RD}{SALE}$ is the ratio of R & D spending to sales. We consider fitting a Bayesian quantile regression model to this data with $r = \frac{1}{2}$. The prior specifications are $\beta \sim \mathcal{N}_4(0, 100 I_4)$ and $\sigma \sim IG(1,1)$. For both the DA and sandwich Markov chains, we generate the initial values of β and z from the respective prior densities.

Lag	1	2	3	4	5
Autocorrelation (DA)	0.528	0.298	0.175	0.101	0.064
Autocorrelation (Sandwich with f^1)	0.504	0.277	0.156	0.096	0.058
Autocorrelation (Sandwich with f^2)	0.519	0.289	0.159	0.091	0.043

Table 2: First five autocorrelations for DA and sandwich Markov chains for the Bayesian quantile regression model applied to patent data

We ran all the three Markov chains for a burn-in period of 5000 iterations. The next 25000 iterations were used to obtain autocorrelations for the function $h^*(\beta, \sigma) = (r - S\beta)^T(r - S\beta) + \sigma$ (see Table 2). This function was a natural choice as it involves both parameters of interest (β and σ), and is closely related to the *drift function* used to prove geometric ergodicity of a reordered version of the DA Markov chain (Khare and Hobert, 2012). The time difference between the DA Markov chain and the sandwich Markov chain (with f^1) for 30000 iterations is 38 seconds, and the extra step takes on average 3% more time per iteration. Using the autocorrelations from Table 2 (and noting that the autocorrelations not included in the table are comparitively negligible), it follows that the effective sample size obtained after one iteration of the DA Markov chain is roughly 1/3.332, and the effective sample size obtained in the same time by using the sandwich Markov chain (with f^1) is $1/(1.03 \times 3.182)$ (a 1.6% increase). Similarly, the extra step for the sandwich Markov chain (with f^2) takes 1.6% more time per iteration than the DA Markov chain. Again, using the autocorrelations from Table 2, it follows that the effective sample size obtained by using the sandwich Markov chain (with f^2) (during the time required for one DA iteration) is roughly $1/(1.016 \times 3.202)$ (a 2.4% increase). Hence, for both sandwich chains, the improvement in autocorrelations supercedes the extra time needed for the additional step in the sandwich algorithm.

Remark 5. *We would like to point out that the amount of extra time needed and the improvement in performance for the sandwich algorithm clearly depends on the model, and the dataset in hand. A deeper investigation to understand and quantify this dependence is the next goal in this line of research. A natural starting point is to establish conditions under which the performance of the sandwich algorithm is strictly better than the DA algorithm (analogous to the results in Khare and Hobert (2011) for the single-block DA algorithm).*

Supplementary material: a) A Supplemental document which includes (among other things) proofs of the technical results in the paper b) A folder containing the R code and datasets used in the paper.

References

[1] Albert, J. H. and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, **88**, pp. 669-679.

[2] Asmussen, S. and Glynn, P. W. (2011), "A new proof of convergence of MCMC via the ergodic theorem," *Statistics and Probability Letters*, **81**, pp. 1482-1485.

[3] Chen, F., Lovasz, L. and Pak, I. (1999). Lifting Markov chains to speed up mixing, *Proceedings of the thirty-first annual ACM symposium on Theory of computing*, pp. 275-281.

[4] Chen, M.H. and Shao, Q.M. (2000), "Propriety of posterior distribution for dichotomous quantal response models," *Proceedings of the American Mathematical Society*, **129**, pp. 293-302.

[5] Hall, B. H., Cummins, C., Laderman, E., and Mundy, J. (1988), "The R & D master file documentation," Technical Working Paper 72, National Bureau of Economic Research, Cambridge, MA.

[6] Hobert, J. P. and Marchev, D. (2008), "A theoretical comparison of the data augmentation, marginal augmentation and PX-DA algorithms," *Annals of Statistics*, **36**, pp. 532-554.

[7] Hobert, J. P., Roy , V. and Robert, C.P. (2011), "Improving the convergence properties of the data augmentation algorithm with an application to Bayesian mixture modeling," *Statistical Science*, **26**, pp. 332-351.

[8] Khare, K. and Hobert, J.P. (2011), "A spectral analytic comparison of trace-class data augmentation algorithms and their sandwich variants, *Annals of Statistics*, **86**, pp. 301-320.

[9] Khare, K. and Hobert, J.P. (2012), "Geometric ergodicity of the Gibbs sampler for Bayesian quantile regression, *Journal of Multivariate Analysis*, **112**, pp. 108-116.

[10] Kozumi, H. and Kobayashi, G. (2011), "Gibbs sampling methods for Bayesian quantile regression," *Journal of Statistical Computation and Simulation*, **81**, pp. 1565-1578.

[11] Liu, J. and Wu, Y.N. (1999), "Parameter Expansion for Data Augmentation," *Journal of the American Statistical Association*, **94**, pp. 1264-1274.

[12] Marchev, D. and Hobert, J.P. (2004), "Geometric ergodicity of van Dyk and Meng's algorithm for the multivariate Student's-t model," *Journal of the American Statistical Association*, **99**, pp. 228-238.

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

[14] Meyn, S. P. and Tweedie, R. L. (1993), *Markov Chains and Stochastic Stability*, Springer-Verlag, London.

[15] Park, T. and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, **103**, pp. 681-686.

[16] Roy, V. (2012). Convergence rates for MCMC algorithms for a robust Bayesian binary regression model, *Electronic Journal of Statistics* **6**, 2463-2485.

[17] Tanner, M.A and Wong, W.H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, **82**, pp. 528-540.

[18] 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**, pp. 27-40.

[19] Tsioans, E. (2003), "Bayesian quantile inference," *Journal of Statistical Computation and Simulation*, **73**, pp. 659-674.

[20] van Dyk, D. A. and Meng, X.L (2001), "The art of data augmentation (with discussion)," *Journal of Computational and Graphical Statistics*, **10**, pp. 1-111.

[21] Wang, P., Cockburn, I., and Puterman, M. K. (1998), "Analysis of patent data: A mixed-Poisson-regression-model approach," *Journal of Business and Economic Statistics*, **16**, pp. 27-41.

[22] 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)," *Journal of Computational and Graphical Statistics*, **20** , pp. 531-570.