Supplementary materials for this article are available online.
Please click the JCGS link at http://pubs.amstat.org.

To Center or Not to Center: That Is Not the Question—An Ancillarity–Sufficiency Interweaving Strategy (ASIS) for Boosting MCMC Efficiency

Yaming Yu and Xiao-Li Meng

For a broad class of multilevel models, there exist two well-known competing parameterizations, the centered parameterization (CP) and the non-centered parameterization (NCP), for effective MCMC implementation. Much literature has been devoted to the questions of when to use which and how to compromise between them via partial CP/NCP. This article introduces an alternative strategy for boosting MCMC efficiency via simply interweaving—but *not* alternating—the two parameterizations. This strategy has the surprising property that failure of both the CP and NCP chains to converge geometrically does not prevent the interweaving algorithm from doing so. It achieves this seemingly magical property by taking advantage of the *discordance* of the two parameterizations, namely, the sufficiency of CP and the ancillarity of NCP, to substantially reduce the Markovian dependence, especially when the original CP and NCP form a "beauty and beast" pair (i.e., when one chain mixes far more rapidly than the other). The ancillarity–sufficiency reformulation of the CP–NCP dichotomy allows us to borrow insight from the well-known Basu's theorem on the independence of (complete) sufficient and ancillary statistics, albeit a Bayesian version of Basu's theorem is currently lacking. To demonstrate the competitiveness and versatility of this ancillarity–sufficiency interweaving strategy (ASIS) for real-world problems, we apply it to fit (1) a Cox process model for detecting changes in source intensity of photon counts observed by the Chandra X-ray telescope from a (candidate) neutron/quark star, which was the problem that motivated the ASIS strategy as it defeated other methods we initially tried; (2) a probit model for predicting latent membranous lupus nephritis; and (3) an interval-censored normal model for studying the lifetime of fluorescent lights. A bevy of open questions are presented, from the mysterious but exceedingly suggestive connections between ASIS and fiducial/structural inferences to nested ASIS for further boosting MCMC efficiency. This article has supplementary material online.

Key Words: Ancillary augmentation; Basu's theorem; Centered parameterization; Data augmentation; EM; GLMM; Interval censoring; Latent variables; Missing data; Non-centered parameterization; Parameter-driven model; Poisson time series; Probit regression; Sufficient augmentation.

Yaming Yu is Associate Professor, Department of Statistics, University of California, Irvine, CA 92697 (E-mail: *yamingy@uci.edu*). Xiao-Li Meng is Whipple V. N. Jones Professor and Chair of Statistics, Department of Statistics, Harvard University, Cambridge, MA 02138 (E-mail: *meng@stat.harvard.edu*).

Journal of Computational and Graphical Statistics, Volume 20, Number 3, Pages 531–570
DOI: 10.1198/jcgs.2011.203main

1. COUPLING IS MORE PROMISING THAN COMPROMISING

As a powerful set of tools for simulating complex distributions, MCMC methods, such as the Gibbs sampler (Geman and Geman 1984; Gelfand and Smith 1990; Smith and Roberts 1993) and the Metropolis–Hastings algorithm (Metropolis et al. 1953; Hastings 1970), have revolutionized statistics, especially Bayesian statistics (Gelman et al. 2004). An early idea in the statistical literature is the data augmentation (DA) algorithm (Tanner and Wong 1987), a stochastic counterpart of the popular EM algorithm (Dempster, Laird, and Rubin 1977; Wu 1983). Both EM and DA work under the missing data formulation by specifying a joint distribution $p(Y_{mis}, Y_{obs}|\theta)$ whose marginal $p(Y_{obs}|\theta)$ is the observed-data model of interest. By purposefully introducing missing data/auxiliary variables, EM and DA achieve their goals through iterative schemes that are usually easy to implement.

Nevertheless, their potential slow convergence has been a concern for their users and a challenge for their designers. Among many developments, Meng and van Dyk (1997, 1999) and van Dyk and Meng (2001) proposed efficient data augmentation, where the key observation was that often a model under the missing data formulation can be written in a variety of ways. Mathematically, any joint model $p(Y_{obs}, Y_{mis}|\theta, \alpha)$ qualifies as a DA model if

$$\int p(Y_{obs}, Y_{mis}|\theta, \alpha)\mu(dY_{mis}) = p(Y_{obs}|\theta) \quad \text{for all } Y_{obs}, \tag{1.1}$$

where α is the "working parameter," only identifiable from $Y_{aug} = \{Y_{mis}, Y_{obs}\}$. We emphasize that it is the specification $p(Y_{aug}|\theta, \alpha)$ that constitutes a DA scheme. The common notation Y_{mis} is merely for convenience, and there are cases where it is not appropriate (see Section 4).

Each formulation in (1.1), indexed by α, corresponds to a (potentially) different EM or DA. Therefore we can choose a formulation that results in fast convergence and easy implementation. In the context of Bayesian hierarchical models, seeking efficient DA is known as a reparameterization issue (e.g., Hills and Smith 1992; Gelfand and Carlin 1995; Gelfand, Sahu, and Carlin 1995, 1996; Roberts and Sahu 1997; Papaspiliopoulos, Roberts, and Skold 2003, 2007), because both the missing data Y_{mis} and the parameter θ are viewed as "parameters." Meng and van Dyk (1998), for example, investigated several rules for choosing an appropriate parameterization for mixed-effects models for faster EM. For MCMC, Papaspiliopoulos, Roberts, and Skold (2007) discussed the importance of effective parameterizations and the strategies for constructing such parameterizations in hierarchical models.

This article introduces an ancillarity–sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency. Instead of choosing a single DA scheme, ASIS singles out two special DA schemes, the *ancillary augmentation* (AA) and the *sufficient augmentation* (SA), and couples them by "going back and forth" between them within each iteration of an MCMC sampler. Specifically, in SA, the missing data are a sufficient statistic for the parameter of interest, whereas in AA, the missing data are an ancillary statistic. It has been long observed that between AA and SA (though not using these terms; see below), if one leads to fast convergence, the other is usually slow, depending on the observed data (e.g., Gelfand, Sahu, and Carlin 1995; Meng and van Dyk 1998; Papaspiliopoulos, Roberts, and Skold 2007). Therefore, it is not surprising that by combining the two in some way, such as alternating between them, one might achieve some compromise, or at least avoid disasters.

What is surprising, at least to us initially, is that there exists an *interweaving* strategy, to be defined in Section 2, that does not compromise but *takes advantage of* this beauty-and-beast contrasting feature of the two DA algorithms; the resulting algorithm can often outperform both, sometimes very substantially. Furthermore, when the two DA schemes being interwoven form an AA–SA pair, we can show theoretically that in some cases ASIS provides the fastest converging algorithm within a general class as defined by Liu and Wu (1999).

On the application side, our motivating problem came from X-ray astrophysics: modeling changes in source intensity for photon counts. We built a parameter-driven Poisson time series model (Cox 1981), but the required computation was very challenging until the interweaving strategy was applied (Yu 2005). Here we apply our general strategy to this and another two real-world problems to demonstrate its competitiveness and flexibility.

We emphasize that the notions of SA and AA are mathematically equivalent to the *centered parameterization* (CP) and *non-centered parameterization* (NCP), respectively (see Gelfand, Sahu, and Carlin 1995, 1996; Roberts and Sahu 1997; Papaspiliopoulos, Roberts, and Skold 2003, 2007; Papaspiliopoulos and Roberts 2008). We believe the *sufficiency* and *ancillarity* terminology better captures the essence of these methods, because *centering* and *non-centering* might leave the false impression that the methods are only applicable to location families. Indeed, the awkwardness of the CP/NCP nomenclature has been noted, for example, by Christensen in his discussion of the work of Papaspiliopoulos, Roberts, and Skold (2003): “The authors use the terminology ‘non-centered’ and ‘partially non-centered’ for the latter two types of parameterisations, which suggest that they find them un-natural. I think these two types of parameterisations deserve better names.” More importantly, the connection with the classical notions of ancillarity and sufficiency reminds us of a theoretical insight suggested by Basu’s theorem (Basu 1955); see Section 2. Nevertheless, our central contribution is not about these two notions individually, but about coupling them in a particularly efficient way.

The rest of the article is organized as follows. Section 2 defines the ASIS and more generally the component-wise ASIS, and uses examples from the works of Liu and Wu (1999) and Papaspiliopoulos, Roberts, and Skold (2003, 2007) to show how, why, and when ASIS works. Section 3 describes how we use component-wise ASIS in our motivating astrophysics problem. Section 4 further demonstrates the competitiveness and versatility of ASIS using probit regression and a normal model with interval censoring. Section 5 establishes four theorems, three on the robustness of the interweaving strategy, and one on the optimality of ASIS under further conditions. Section 6 concludes with a discussion of limitations and open problems. Due to space limitation, most proofs and technical details are deferred to an on-line appendix, available at the journal website *http://www.pubs.amstat.org*.

2. DEFINING AND EXPLAINING ASIS

2.1 ANCILLARY AND SUFFICIENT DA SCHEMES

Consider the simplest two-level normal hierarchical model (Liu and Wu 1999)

$$Y_{obs}|(\theta, Y_{mis}) \sim N(Y_{mis}, 1), \tag{2.1}$$

$$Y_{mis}|\theta \sim N(\theta, V), \tag{2.2}$$

where θ is the parameter, Y_{obs} (a scalar) is the observed datum, Y_{mis} is the missing datum or latent variable, and $V > 0$ is a known constant. With a constant prior distribution on θ, the posterior is $\theta|Y_{obs} \sim N(Y_{obs}, 1 + V)$, which is our target density.

Treating Y_{mis} directly as the missing data, the standard DA algorithm iterates between drawing $Y_{mis}|(\theta, Y_{obs})$ and drawing $\theta|(Y_{mis}, Y_{obs})$:

$$Y_{mis}|(\theta, Y_{obs}) \sim N\left(\frac{\theta + VY_{obs}}{1+V}, \frac{V}{1+V}\right), \tag{2.3}$$

$$\theta|(Y_{mis}, Y_{obs}) \sim N(Y_{mis}, V). \tag{2.4}$$

Because the right side of (2.1) is free of θ, we call such a DA scheme $p(Y_{mis}, Y_{obs}|\theta)$ *sufficient augmentation* (SA), since Y_{mis} is a sufficient statistic for θ in the augmented-data model. In a Bayesian setting, this implies that the augmented-data posterior of θ depends on Y_{mis} alone.

On the other hand, if we let

$$\tilde{Y}_{mis} = Y_{mis} - \theta, \tag{2.5}$$

and treat $\tilde{Y}_{mis}$ as the missing data, then the model can be rewritten as

$$Y_{obs}|(\theta, \tilde{Y}_{mis}) \sim N(\tilde{Y}_{mis} + \theta, 1), \tag{2.6}$$

$$\tilde{Y}_{mis}|\theta \sim N(0, V), \tag{2.7}$$

which gives a different DA algorithm:

$$\tilde{Y}_{mis}|(\theta, Y_{obs}) \sim N\left(\frac{V(Y_{obs} - \theta)}{1+V}, \frac{V}{1+V}\right), \tag{2.8}$$

$$\theta|(\tilde{Y}_{mis}, Y_{obs}) \sim N(Y_{obs} - \tilde{Y}_{mis}, 1). \tag{2.9}$$

We call such a DA scheme *ancillary augmentation* (AA), because the distribution of $\tilde{Y}_{mis}$, as in (2.7), is free of θ, that is, $\tilde{Y}_{mis}$ is an ancillary statistic for θ. In Bayesian terms, $\tilde{Y}_{mis}$ and θ are independent a priori.

Though both schemes have the same target distribution $p(\theta|Y_{obs})$, their convergence rates are usually different. For EM-type algorithms, the convergence rates are governed by "the fraction of missing information" (Dempster, Laird, and Rubin 1977; Meng and Rubin 1991; Meng 1994; Meng and van Dyk 1996, 1997), which also provides insights into the convergence behavior of DA-type algorithms (Liu 1994; van Dyk and Meng 2001, 2010). In the current problem, for SA as in (2.2), the smaller the (conditional) variance V, the more informative Y_{mis} is about θ, and hence the more missing information when we treat Y_{mis} as missing data. In contrast, for (2.7), the smaller the value of V, the more we know about $\tilde{Y}_{mis}$ (as it gets closer to zero stochastically), and hence the less the missing information when we treat $\tilde{Y}_{mis}$ as missing data. Consequently, when V is small, we expect SA to be slow but AA to be fast; the situation reverses when V is large.

This intuition is indeed correct. Let us recall that for a general DA algorithm, its geometric rate of convergence is the square of the *maximal correlation* between Y_{mis} and θ under the joint posterior distribution $p(Y_{mis}, \theta | Y_{obs})$; see the articles of Liu, Wong, and Kong (1994, 1995). This is both the L_1 and L_2 geometric rate, because for time-reversible Markov chains such as the DA algorithm, geometric L_1 convergence and L_2 convergence are equivalent (see Roberts and Tweedie 2001). The usual definition of the L_p geometric rate is the smallest *constant* r such that the L_p distance between the distribution of $\theta^{(t)}$ and the target distribution is bounded above by $c(\theta^{(0)})r^t$, where c is a (nonnegative) function of $\theta^{(0)}$. When $0 \le r < 1$, we say $\{\theta^{(t)}\}$ is geometrically ergodic; otherwise it is non-geometric or sub-geometric (all with respect to the chosen L_p norm). If, in addition to $r < 1$, $c(\theta^{(0)})$ is bounded above, we say $\{\theta^{(t)}\}$ is uniformly ergodic. See the work of Meyn and Tweedie (1993) and Papaspiliopoulos and Roberts (2008). In this article we will adopt the L_2 norm mainly because our theoretical results (Section 5) are established via maximal correlations, as in the work of Liu, Wong, and Kong (1994, 1995). We note, however, that the L_2 geometric convergence implies L_1 geometric convergence in general (e.g., when we move beyond the original DA setting).

For the current case, the rate is $r_{SA} = 1/(1+V)$ for the SA chain and $r_{AA} = V/(1+V)$ for the AA chain (the larger the rate, the slower the convergence). These rates can also be visualized by noting the slope in the stochastic recursion for the SA chain (obtained by combining (2.3) and (2.4) via their stochastic representations):

$$\theta_{SA}^{(t+1)} = \frac{1}{1+V}\theta_{SA}^{(t)} + \frac{V}{1+V}Y_{obs} + \sqrt{\frac{V^2+2V}{1+V}}Z_1^{(t)}, \tag{2.10}$$

and for the AA chain (obtained by combining (2.8) and (2.9)):

$$\theta_{AA}^{(t+1)} = \frac{V}{1+V}\theta_{AA}^{(t)} + \frac{1}{1+V}Y_{obs} + \sqrt{\frac{2V+1}{1+V}}Z_2^{(t)}, \tag{2.11}$$

where $Z_1^{(t)}$ and $Z_2^{(t)}$ are i.i.d. $N(0,1)$. Because $r_{AA} + r_{SA} = 1$, when one algorithm converges fast, the other has to be slow, and either one can be arbitrarily slow on its own, depending on the value of V. Gelfand, Sahu, and Carlin (1995), Papaspiliopoulos, Roberts, and Skold (2003, 2007), and Papaspiliopoulos and Roberts (2008) discussed this issue for general classes of Bayesian hierarchical models. A key contribution of our article is to show that, by *interweaving* the two schemes, we can create a potentially much better algorithm. Indeed, for the toy example, our proposed strategy will lead to i.i.d. draws and hence the rate r is zero. This is in contrast with the simple alternating scheme, which has a rate of $r_{AA}r_{SA}$ even though the alternating scheme requires four steps (i.e., (2.3)–(2.4) and (2.8)–(2.9)) at each iteration whereas the interweaving strategy uses only three, as detailed below.

2.2 INTERWEAVING AA AND SA

To define *interweaving*, suppose we have a pair of DA schemes Y_{mis} and $\tilde{Y}_{mis}$ (not necessarily an SA–AA pair) such that their joint distribution $p(Y_{mis}, \tilde{Y}_{mis}|\theta, Y_{obs})$, *conditional on both θ and Y_{obs}*, is well defined. We emphasize that this joint distribution is often degenerate in the sense that $\tilde{Y}_{mis} = M(Y_{mis}; \theta)$ where $M(\cdot; \theta)$ is a one-to-one (deterministic) mapping for given θ. In our toy example, $M(Y_{mis}; \theta) = Y_{mis} - \theta$, and $M^{-1}(\tilde{Y}_{mis}; \theta) = \tilde{Y}_{mis} + \theta$. However, there are important applications that call for more general stochastic relationships. Consider the example provided by Papaspiliopoulos, Roberts, and Skold (2007), where $Y_{mis} \sim \text{Bernoulli}(\theta)$ is a latent class indicator, and is an SA because the membership probability θ is the only parameter in their model. An AA is obtained by specifying $\tilde{Y}_{mis} \sim \text{Uniform}(0, 1)$ and letting Y_{mis} be the indicator of the event $\tilde{Y}_{mis} \leq \theta$. Although Y_{mis} is a deterministic function of $\tilde{Y}_{mis}$ (for fixed θ), the inverse relationship is stochastic: given θ, $\tilde{Y}_{mis} \sim \text{Uniform}(0, \theta)$ if $Y_{mis} = 1$ and $\tilde{Y}_{mis} \sim \text{Uniform}(\theta, 1)$ if $Y_{mis} = 0$. As emphasized by Papaspiliopoulos, Roberts, and Skold (2007), such many-to-one mappings are necessary for "state-space expansion," an important technique for constructing AA (i.e., NCP) for discretely observed diffusion processes (Papaspiliopoulos, Roberts, and Skold 2003; Beskos et al. 2006).

The two DA schemes lead to two algorithms: one iterates between Step 1: Draw $Y_{mis} \sim p(Y_{mis}|\theta)$ and Step 2: Draw $\theta \sim p(\theta|Y_{mis})$; and the other between Step $\tilde{1}$: Draw $\tilde{Y}_{mis} \sim p(\tilde{Y}_{mis}|\theta)$ and Step $\tilde{2}$: Draw $\theta \sim p(\theta|\tilde{Y}_{mis})$. (For simplicity of notation, conditioning on Y_{obs} is suppressed henceforth unless otherwise noted.) A straightforward *alternating* scheme would execute one iteration from each algorithm in turn; that is, we form a combined iteration by carrying out Steps 1, 2, $\tilde{1}$, and $\tilde{2}$ in that order and treat these four steps as one iteration:

$$\left[Y_{mis}|\theta^{(t)}\right] \longrightarrow [\theta|Y_{mis}] \longrightarrow [\tilde{Y}_{mis}|\theta] \longrightarrow \left[\theta^{(t+1)}|\tilde{Y}_{mis}\right]. \tag{2.12}$$

For comparing different schemes, we only index (via the superscript) those draws that will form the chain $\{\theta^{(t)}, t = 1, 2, \ldots\}$; all un-indexed draws serve as intermediate vehicles for moving from $\theta^{(t)}$ to $\theta^{(t+1)}$. For example, the middle two steps in (2.12) facilitate the transfer from Y_{mis} to $\tilde{Y}_{mis}$. (A subtle point: (2.12) does not require $\{Y_{mis}, \tilde{Y}_{mis}\}$ to have a *joint distribution* given θ and Y_{obs}.)

Our interweaving strategy, in a nutshell, replaces Step 2 and Step $\tilde{1}$ by a single step: drawing $\tilde{Y}_{mis}$ from $p(\tilde{Y}_{mis}|Y_{mis})$ (no conditioning on θ); schematically, we have

$$\left[Y_{mis}|\theta^{(t)}\right] \longrightarrow [\tilde{Y}_{mis}|Y_{mis}] \longrightarrow \left[\theta^{(t+1)}|\tilde{Y}_{mis}\right]. \tag{2.13}$$

It is usually more convenient to draw $p(\tilde{Y}_{mis}|Y_{mis})$ by drawing $\theta \sim p(\theta|Y_{mis})$ and then drawing $\tilde{Y}_{mis} \sim p(\tilde{Y}_{mis}|Y_{mis}, \theta)$. Hence (2.13) becomes

$$\left[Y_{mis}|\theta^{(t)}\right] \longrightarrow [\theta|Y_{mis}] \longrightarrow [\tilde{Y}_{mis}|Y_{mis}, \theta] \longrightarrow \left[\theta^{(t+1)}|\tilde{Y}_{mis}\right]. \tag{2.14}$$

Viewed this way, the only difference between *interweaving* and *alternating* is that we replace the third step in (2.12) by the conditional draw $\tilde{Y}_{mis} \sim p(\tilde{Y}_{mis}|Y_{mis}, \theta)$. Ironically, although replacing $p(\tilde{Y}_{mis}|\theta)$ by $p(\tilde{Y}_{mis}|Y_{mis}, \theta)$ appears to introduce more dependence (often $\tilde{Y}_{mis}$ is completely determined by $\{Y_{mis}, \theta\}$!), the resulting algorithm usually possesses less dependence between $\theta^{(t)}$ and $\theta^{(t+1)}$ and hence improves convergence, as demonstrated in later sections.

To explain the name *interweaving*, consider (2.14) again but this time we omit the second step, $[\theta|Y_{mis}]$. The scheme then becomes

$$\left[Y_{mis}|\theta^{(t)}\right] \longrightarrow \left[\tilde{Y}_{mis}|Y_{mis}, \theta^{(t)}\right] \longrightarrow \left[\theta^{(t+1)}|\tilde{Y}_{mis}\right]. \tag{2.15}$$

But this is trivially the same as

$$\left[\tilde{Y}_{mis}|\theta^{(t)}\right] \longrightarrow \left[\theta^{(t+1)}|\tilde{Y}_{mis}\right], \tag{2.16}$$

which is just the original DA algorithm based on $\tilde{Y}_{mis}$ alone. So in this sense the new scheme (2.13)—or equivalently (2.14)—just *interweaves* (or injects) the $[\theta|Y_{mis}]$ step from the DA algorithm based on Y_{mis} into the DA algorithm based on $\tilde{Y}_{mis}$ (or vice versa).

2.3 GLOBAL INTERWEAVING STRATEGY AND ITS POTENTIAL

To summarize, each iteration of the resulting *global interweaving strategy* ("global" is used to distinguish from "component-wise" introduced later) performs the following steps, where non-integer superscripts index intermediate draws.

Global Interweaving Strategy (GIS):
Step 1. Draw Y_{mis} given θ: $Y_{mis}^{(t)}|\theta^{(t)}$.
Step 2. Draw θ given Y_{mis}: $\theta^{(t+0.5)}|Y_{mis}^{(t)}$.
Step $\tilde{2}$. Redraw θ given $\tilde{Y}_{mis}$: $\theta^{(t+1)}|\tilde{Y}_{mis}^{(t+1)}$, where $\tilde{Y}_{mis}^{(t+1)} \sim p(\tilde{Y}_{mis}|Y_{mis}^{(t)}; \theta^{(t+0.5)})$.

To verify that this GIS chain preserves the stationary density $p(\theta)$ as shared by the original two DA algorithms, we note, from (2.13), that its transition density is

$$k(\theta'|\theta) = \int\int p(\theta'|\tilde{Y}_{mis})p(\tilde{Y}_{mis}|Y_{mis})p(Y_{mis}|\theta)\,dY_{mis}\,d\tilde{Y}_{mis}. \tag{2.17}$$

For simplicity of presentation we assume all distributions involved have densities with respect to Lebesgue measure. Assuming that $p(Y_{mis}, \tilde{Y}_{mis}, \theta)$ is a well-defined joint density and its margins $p(Y_{mis}, \theta)$ and $p(\tilde{Y}_{mis}, \theta)$ are the (joint) stationary densities of the original two DA chains such that they share the same margin $p(\theta)$, we have, by Fubini's theorem,

$$\begin{aligned}\int k(\theta'|\theta)p(\theta)\,d\theta &= \int\int p(\theta'|\tilde{Y}_{mis})p(\tilde{Y}_{mis}|Y_{mis})\left[\int p(Y_{mis}|\theta)p(\theta)\,d\theta\right]dY_{mis}\,d\tilde{Y}_{mis}\\ &= \int p(\theta'|\tilde{Y}_{mis})\left[\int p(\tilde{Y}_{mis}|Y_{mis})p(Y_{mis})\,dY_{mis}\right]d\tilde{Y}_{mis}\\ &= \int p(\theta'|\tilde{Y}_{mis})p(\tilde{Y}_{mis})\,d\tilde{Y}_{mis} = p(\theta').\end{aligned}$$

Hence $p(\theta)$ is the stationary density; see the article by Tierney (1994) for a general theory for ensuring valid target distributions in MCMC. When the two DAs form an AA–SA pair, we call the resulting GIS an Ancillarity–Sufficiency Interweaving Strategy (ASIS).

We note here that the kernel expression (2.17) is closely related to the "sandwiched" kernel given by Hobert and Marchev (2008). The difference is that the kernel in the work of Hobert and Marchev (2008), using our notation, can be written as

$$k_{HM}(\theta'|\theta) = \int\int p(\theta'|Y'_{mis})p(Y'_{mis}|Y_{mis})p(Y_{mis}|\theta)\,dY_{mis}\,dY'_{mis}, \tag{2.18}$$

where (θ', Y'_{mis}) and (θ, Y_{mis}) have the *same* joint distribution; when $Y'_{mis} = Y_{mis}$, (2.18) becomes the standard DA. Such a setup is general enough to subsume both marginal augmentation and PX-DA, but is still less general than (2.17), where $(\theta', \tilde{Y}_{mis})$ and (θ, Y_{mis}) do not (necessarily) have the same distribution precisely because we want to take advantage of the "beauty-and-beast" nature of the $\tilde{Y}_{mis}$ and Y_{mis} pair. Technically, the reason that we can relax the restriction that (θ', Y'_{mis}) and (θ, Y_{mis}) have the same distribution is that we only aim to preserve the stationary distribution for the θ margin.

If $\tilde{Y}_{mis}$ and Y_{mis} are linked by a one-to-one transformation $\tilde{Y}_{mis} = M(Y_{mis}; \theta)$, then geometrically each conditional draw in Steps 1–$\tilde{2}$ can be viewed as sampling along a certain direction in the (θ, Y_{mis}) space. In this case GIS falls within the framework of alternating subspace-spanning resampling (Liu 2003). Which combination of directions produces an algorithm that is both efficient and easy to implement is a critical issue, and the current article demonstrates that ASIS is a promising and surprisingly simple recipe.

Figure 1 illustrates the three directions for the toy model. Remarkably, by sampling in these particular directions, ASIS converges immediately, that is, $r_{ASIS} = 0$, as can be verified from its stochastic recursion (derived in a similar way as for (2.10) or (2.11))

$$\theta_{ASIS}^{(t+1)} = Y_{obs} + Z_1 + \sqrt{V}Z_2, \tag{2.19}$$

where Z_1 and Z_2 are i.i.d. $N(0, 1)$. That is, the ASIS chain produces i.i.d. draws from the target distribution $N(Y_{obs}, 1 + V)$.

Contrary to what one might suspect, this phenomenon of i.i.d. draws has nothing to do with the normality assumptions in (2.1) or in (2.2), as we will show in the next section. However, the change of the distribution shape can affect substantially the convergence behaviors of the original SA chain or the AA chain, as proved by Papaspiliopoulos and Roberts (2008). In particular, if we change the normal distribution in (2.1) to a Cauchy distribution, but keep the normality for (2.2), then the SA chain is *not* geometrically ergodic, whereas the AA chain is uniformly ergodic. On the other hand, if we retain the normality

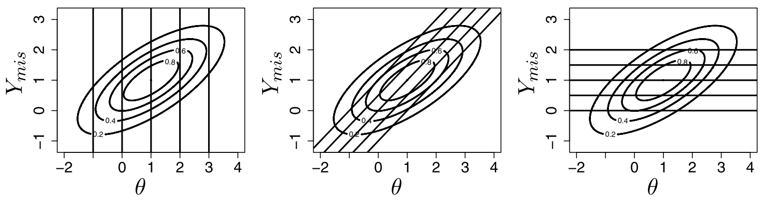

Figure 1. Sampling directions (solid lines) for ASIS in the toy model with $Y_{obs} = 1$ and $V = 1$. Left: Step 1, conditional draw along the lines $\theta = const$. Center: Step $\tilde{2}$, conditional draw along $Y_{mis} - \theta = const$. Right: Step 2, conditional draw along $Y_{mis} = const$. The ellipses are contours of the posterior distribution $p(\theta, Y_{mis}|Y_{obs})$.

for (2.1) but adopt a Cauchy distribution in (2.2), then the SA chain will be uniformly ergodic but the AA chain will fail to be geometrically ergodic. However, regardless of which original chain fails, the ASIS sampler delivers i.i.d. draws.

Most strikingly, let us consider a bivariate extension of (2.1)–(2.2), where the first component $\{Y_{obs,1}, Y_{mis,1}, \theta_1\}$ follows the Cauchy–Normal pair version of (2.1)–(2.2), and the second component $\{Y_{obs,2}, Y_{mis,2}, \theta_2\}$ follows the Normal–Cauchy pair version of (2.1)–(2.2), and the two components are independent (including independent priors $p(\theta_1, \theta_2) \propto 1$). Since this bivariate model is just two independent univariate models, obviously $\mathbf{Y}_{mis} \equiv (Y_{mis,1}, Y_{mis,2})$ is SA for $\boldsymbol{\theta} \equiv (\theta_1, \theta_2)$ and $\tilde{\mathbf{Y}}_{mis} \equiv (Y_{mis,1} - \theta_1, Y_{mis,2} - \theta_2) = \mathbf{Y}_{mis} - \boldsymbol{\theta}$ is AA for $\boldsymbol{\theta}$. Both the SA chain for θ and the AA chain for θ fail to be geometrically ergodic because each has a component that fails to be so by construction, yet the ASIS sampler will still produce i.i.d. draws for θ since each of its components does so. Although this extreme example is not indicative at all of what ASIS can achieve in practice, it shows that the power of ASIS comes not so much from the individual behaviors of SA and AA but rather from their contrasting relationship.

2.4 SO WHERE DOES THE MAGIC COME FROM?

Although achieving i.i.d. sampling is wishful thinking for general MCMC (albeit with effort even negatively associated draws are possible; see, e.g., Craiu and Meng 2005), our empirical and theoretical results demonstrate that achieving substantial improvements via ASIS is a real possibility. In particular, in Section 5 we establish the following result (the simplest among several):

Theorem 1. *Given a posterior distribution of interest $p(\theta|Y_{obs}), \theta \in \Theta$, suppose we have two augmentation schemes $Y_{mis,1}$ and $Y_{mis,2}$ such that their joint distribution, conditioning on both θ and Y_{obs}, is well defined for $\theta \in \Theta$ (almost surely with respect to $p(\theta|Y_{obs})$). Denote the geometric rate of convergence of the DA algorithm under $Y_{mis,i}$ by $r_i, i = 1, 2$, which are allowed to take value 1 (i.e., being sub-geometric). Then*

$$r_{1\&2} \leq \mathcal{R}_{1,2}\sqrt{r_1 r_2}, \tag{2.20}$$

where $r_{1\&2}$ is the geometric rate of the GIS sampler interweaving $Y_{mis,1}$ and $Y_{mis,2}$, and $\mathcal{R}_{1,2}$ is the maximal correlation between $Y_{mis,1}$ and $Y_{mis,2}$ in their joint posterior distribution $p(Y_{mis,1}, Y_{mis,2}|Y_{obs})$.

This theorem establishes that *minimally* the interweaving strategy leads to an algorithm that is better than the worse of the two, because (2.20) trivially implies that $r_{1\&2} \leq \max\{r_1, r_2\}$. Theorem 1 therefore sharpens Hobert and Marchev's (2008) result that the Markov chain defined by their "sandwiched" kernel (2.18) is no slower than the original chain determined by Y_{mis}. This is because when $(\theta, Y_{mis,1})$ and $(\theta, Y_{mis,2})$ have the same distribution (which is the case for (2.18)), $r_1 = r_2$, and hence (2.20) improves Hobert and Marchev's (2008) result by the factor $\mathcal{R}_{1,2} \leq 1$. Although $\mathcal{R}_{1,2}$ is typically difficult to calculate—so this improvement has little use in evaluating the bound—it is the key for understanding the power of the interweaving strategy because it brings the posterior dependence of $Y_{mis,1}$ and $Y_{mis,2}$ into the picture, which provides a new direction for boosting MCMC efficiency. For example, Theorem 1 says that as long as $\mathcal{R}_{1,2} < 1$ (see Theorem 2 of Section 5 when this fails), the interweaving strategy will be geometrically convergent, even if neither of the original two algorithms is (as illustrated in Section 2.3). The theoretical and empirical evidence in later sections demonstrates that ASIS often converges noticeably faster than the original two (issues such as time per iteration will be discussed in the context of real examples).

Theorem 1 also says that when $\mathcal{R}_{1,2} = 0$, that is, when $Y_{mis,1}$ and $Y_{mis,2}$ are a posteriori independent, then $r_{1\&2} = 0$ and hence the interwoven algorithm will provide i.i.d. draws. As (2.13) depicts, if Y_{mis} $(= Y_{mis,1})$ and $\tilde{Y}_{mis}$ $(= Y_{mis,2})$ are independent, then so are $\theta^{(t+1)}$ and $\theta^{(t)}$. Although independence is generally not achievable, Theorem 1 provides a critical insight as to which pairs of $Y_{mis,1}$ and $Y_{mis,2}$ we should seek. The classical theorem of Basu (1955) says that a complete sufficient statistic is independent of any ancillary statistic given the parameter. Although this does not imply that sufficient and ancillary DAs should be independent a posteriori (and usually they are not), it does suggest that they are good candidates for interweaving.

In many common models, Y_{mis}, the "default" DA, is the SA or CP, as emphasized by Gelfand, Sahu, and Carlin (1995, 1996) and Papaspiliopoulos, Roberts, and Skold (2003, 2007). How should we construct its partner via $\tilde{Y}_{mis} = M(Y_{mis}; \theta)$? Regardless of the choice of the (one-to-one) mapping $M(\cdot; \theta)$, the joint posterior density of $\tilde{Y}_{mis}$ and θ can be expressed as

$$p(\theta, \tilde{Y}_{mis}|Y_{obs}) \propto p(Y_{obs}|\tilde{Y}_{mis}, \theta) p(\tilde{Y}_{mis}|\theta) p_0(\theta), \tag{2.21}$$

where $p_0(\theta)$ is the prior. If θ and Y_{mis} are one-to-one given $\tilde{Y}_{mis}$, we can directly form a one-to-one transformation from $(\theta, \tilde{Y}_{mis})$ to $(\tilde{Y}_{mis}, Y_{mis})$, where $Y_{mis} = M^{-1}(\tilde{Y}_{mis}; \theta)$, with $M(\cdot; \theta)$ one-to-one but otherwise to be determined. Noting that $p(Y_{obs}|\tilde{Y}_{mis}; \theta) = p(Y_{obs}|Y_{mis}; \theta) = p(Y_{obs}|Y_{mis})$ because $\tilde{Y}_{mis} = M(Y_{mis}; \theta)$ is one-to-one and Y_{mis} is sufficient, we have

$$p(\tilde{Y}_{mis}, Y_{mis}|Y_{obs}) \propto p(Y_{obs}|Y_{mis}) p(\tilde{Y}_{mis}|\theta) p_0(\theta) J(\tilde{Y}_{mis}, Y_{mis}), \tag{2.22}$$

where $\theta = \theta(\tilde{Y}_{mis}, Y_{mis})$ is determined by $\tilde{Y}_{mis} = M(Y_{mis}; \theta)$ and

$$J(\tilde{Y}_{mis}, Y_{mis}) = \frac{|\det\{\partial M(Y_{mis}; \theta)/\partial Y_{mis}\}|}{|\det\{\partial M(Y_{mis}; \theta)/\partial \theta\}|}. \tag{2.23}$$

In all examples of Section 2.3 $p_0(\theta) \propto 1$ and $M(Y_{mis}; \theta) = Y_{mis} - \theta$, so that $J(\tilde{Y}_{mis}, Y_{mis}) = 1$. Consequently, (2.22) suggests that if $\tilde{Y}_{mis}$ is ancillary for θ, then $\tilde{Y}_{mis}$ and Y_{mis} are a posteriori independent because (2.22) then factors as a function of $\tilde{Y}_{mis}$ and Y_{mis}:

$$p(\tilde{Y}_{mis}, Y_{mis}|Y_{obs}) \propto p(Y_{obs}|Y_{mis}) p(\tilde{Y}_{mis}). \tag{2.24}$$

Hence $\mathcal{R}_{1,2} = 0$, explaining why ASIS led to i.i.d. draws in all examples of Section 2.3.

Achieving such complete factorization is of course rare. Nevertheless, as long as $\tilde{Y}_{mis}$ is an AA, we still have the factorization of the first two terms in the right side of (2.22):

$$p(\tilde{Y}_{mis}, Y_{mis}|Y_{obs}) \propto p(Y_{obs}|Y_{mis}) p(\tilde{Y}_{mis}) p_0(\theta(\tilde{Y}_{mis}, Y_{mis})) J(\tilde{Y}_{mis}, Y_{mis}). \tag{2.25}$$

Therefore, the a posteriori dependence is determined only by the form of the prior and the Jacobian of our transformation. This is largely good news, because in practice priors tend to be weak (for likelihood computation via MCMC, e.g., the prior is constant), and the transformation is for us to make, at least to a certain extent. For example, suppose θ is a scale parameter for Y_{mis}, which is an SA for θ, and we use the usual constant prior on $\log\theta$. If we choose $\tilde{Y}_{mis} = M(Y_{mis};\theta) = Y_{mis}/\theta$, which is an AA, then we have

$$\Delta(\tilde{Y}_{mis}, Y_{mis}) \equiv p_0(\theta(\tilde{Y}_{mis}, Y_{mis}))J(\tilde{Y}_{mis}, Y_{mis}) = \theta^{-1} \times \theta^{-1}/(Y_{mis}\theta^{-2}) = Y_{mis}^{-1}. \quad (2.26)$$

Hence $p(\tilde{Y}_{mis}, Y_{mis}|Y_{obs})$ factors, leading again to i.i.d. draws. We remark that it does not seem a coincidence that the location map $M(Y_{mis};\theta) = Y_{mis} - \theta$ works with a constant prior on θ and the scale map $M(Y_{mis};\theta) = Y_{mis}/\theta$ works with a constant prior on $\log\theta$; see Section 6.

A perceptive reader must wonder if there is something too good to be true here. It may seem possible to choose $M(\cdot;\theta)$ (as a *functional* of the prior p_0) such that $\Delta(\tilde{Y}_{mis}, Y_{mis})$ in (2.26) factors. However, the resulting parameterization is not guaranteed to be an AA. More critically, (2.22) is correct *only when θ and Y_{mis} are of the same dimension* (and one-to-one given $\tilde{Y}_{mis}$). Otherwise the joint density of $\tilde{Y}_{mis}$ and Y_{mis} is of a more complicated form, and much less likely to factor. Nevertheless, we shall demonstrate both empirically and theoretically that ASIS, particularly with its component-wise implementation described in the next section, holds good promise for boosting MCMC efficiency because of its simplicity, generality, and flexibility.

2.5 COMPONENT-WISE INTERWEAVING STRATEGY

When GIS is difficult to implement, it is natural to consider partitioning θ into $\theta = \{\theta_1, \ldots, \theta_J\}$. The usual Gibbs sampler draws Y_{mis} given θ, and then each θ_j in turn given Y_{mis} and all other components of θ, at each iteration. The component-wise interweaving strategy (CIS) modifies it by drawing each θ_j using interweaving, conditional on all other components. Thus we only need a *conditional AA* and a *conditional SA* for each component of θ, treating all other components as known. Specifically, let $Y_{S,j}$ and $Y_{A,j}$ be the conditional SA and conditional AA schemes for θ_j, respectively, $j = 1, \ldots, J$, such that the joint distribution of *all* $\{Y_{S,j}, Y_{A,j}; j = 1, \ldots, J\}$ given θ and Y_{obs} is well defined. Generalizing (2.14), an iteration of CIS with $J = 2$ is as follows:

$$\begin{aligned} &[Y_{A,1}|\theta^{(t)}] \longrightarrow [\theta_1|Y_{A,1};\theta_2^{(t)}] \longrightarrow [Y_{S,1}|Y_{A,1},(\theta_1,\theta_2^{(t)})] \longrightarrow [\theta_1^{(t+1)}|Y_{S,1};\theta_2^{(t)}] \\ &\longrightarrow [Y_{A,2}|Y_{S,1};(\theta_1^{(t+1)},\theta_2^{(t)})] \longrightarrow [\theta_2|Y_{A,2};\theta_1^{(t+1)}] \\ &\longrightarrow [Y_{S,2}|Y_{A,2};(\theta_1^{(t+1)},\theta_2)] \longrightarrow [\theta_2^{(t+1)}|Y_{S,2},\theta_1^{(t+1)}]. \end{aligned}$$

In general, for each j, CIS simply adds a step based on $\tilde{Y}_{mis,j}$ to the $(j+1)$th step of the original Gibbs sampler. Typically the DA Y_{mis} used by the original Gibbs-type sampler is either an AA or SA for each θ_j, and hence only its complementary $\tilde{Y}_{mis,j}$ is needed to complete the AA–SA pair. For each j, let $\theta_{<j}$ and $\theta_{>j}$ denote the components of θ before and after θ_j, respectively. Adopting the previous conventions, we can express the general CIS scheme as follows.

Component-wise Interweaving Strategy (CIS):
Step 1. Draw $Y_{mis} \sim p(Y_{mis}|\theta^{(t)})$.
For $j = 1, \dots, J$, perform the following J *cycles* in turn:
Step $(j+1)$. Draw $\theta_j^{(t+0.5)} \sim p(\theta_j|\theta_{<j}^{(t+1)}, \theta_{>j}^{(t)}; Y_{mis})$.
Step $\widetilde{(j+1)}$. Update $\tilde{Y}_{mis,j} \sim p(\tilde{Y}_{mis,j}|\theta_{<j}^{(t+1)}, \theta_j^{(t+0.5)}, \theta_{>j}^{(t)}; Y_{mis})$, and then draw

$$\theta_j^{(t+1)} \sim p\big(\theta_j|\theta_{<j}^{(t+1)}, \theta_{>j}^{(t)}; \tilde{Y}_{mis,j}\big);$$

update Y_{mis} by drawing it from $p(Y_{mis}|\theta_{<j}^{(t+1)}, \theta_j^{(t+1)}, \theta_{>j}^{(t)};\ \tilde{Y}_{mis,j})$.

In Step $\widetilde{(j+1)}$, the two draws updating $\tilde{Y}_{mis,j}$ and Y_{mis} are typically simple deterministic transformations, that is, $\tilde{Y}_{mis,j} = M_j(Y_{mis}; \theta)$ and its inverse; they are not on equal footing with Step 1, the real sampling step as called for by the original Gibbs sampler. By including these extra draws in Step $\widetilde{(j+1)}$, it becomes clear that once Step $\widetilde{(j+1)}$ is removed from the jth *cycle* for all j, CIS reduces to the original Gibbs sampler. (This simple observation is helpful for understanding Theorem 3 in Section 5, which links the convergence rate of CIS to that of the original Gibbs sampler.) Also, for simplicity, we have chosen to link all $\tilde{Y}_{mis,j}$ to a common Y_{mis}, although we could have used a more general notation $Y_{mis,j}$.

The CIS is valid because every step maintains the invariance of the target density; the proof is the same as the one for GIS in Section 2.3. Specifically, if we denote the output of θ immediately after Step $\widetilde{(j+1)}$ by

$$\theta^{(t+j/J)} = \big(\theta_{<j}^{(t+1)}, \theta_j^{(t+1)}, \theta_{>j}^{(t)}\big), \tag{2.27}$$

then $\theta^{(t+j/J)}$ follows the target distribution for all $j = 1, \dots, J$ as long as $\theta^{(t)}$ does. This holds regardless of whether the DA schemes form ancillary-sufficient pairs. As a consequence, CIS is amenable to the so-called Metropolis-within-Gibbs strategy, which replaces certain intractable conditional draws by Metropolis–Hastings (M-H) steps. (A subtle point: the final draw of Y_{mis} in Step $\widetilde{(J+1)}$ is redundant if Step 1 is exact, but is needed if a Metropolis-within-Gibbs strategy is implemented in Step 1, so that the Metropolis–Hastings moves are correctly updated.)

Like the Gibbs sampler, the convergence rate of CIS may depend on the ordering of the steps. If we switch Step $(j+1)$ with Step $\widetilde{(j+1)}$ for certain values of j, then the convergence rate (but not the validity) may be affected. Which ordering is more efficient is theoretically interesting. However, our experiences in this and the related ECM-type algorithms (e.g., van Dyk and Meng 1997) indicate that the difference is often minor, at least relative to the difference between using and not using the interweaving strategy. We shall therefore not pursue this issue of ordering, but refer interested readers to the article by Amit and Grenander (1991) for investigations in the context of Gibbs sampling.

The theoretical insight from Theorem 1 for GIS suggests that CIS may achieve efficiency by reducing, in turn, the dependence between $\theta_j^{(t+1)}$ and $\theta_j^{(t)}$ *conditional* on the rest of the components. Theorem 3 of Section 5 supports this intuition. Theorem 3 uses the notion of the *minimal speed* of a Gibbs sampler or more generally an MCMC algorithm, which circumvents certain theoretical difficulties in dealing with the geometric rate of convergence of non-reversible Markov chains (e.g., Gibbs samplers with more than two components). Empirical evidence will be provided in Section 3 to demonstrate the flexibility and power of the component-wise ASIS.

2.6 THE SIMPLICITY, GENERALITY, AND FLEXIBILITY OF ASIS AND CIS

A central advantage of ASIS, or more generally CIS, is its simplicity. It is a simpler construction than the marginal DA algorithm of Meng and van Dyk (1999), or the PX-DA algorithm of Liu and Wu (1999), where the model is expanded to include a parameter α that is unidentifiable from observed data. Under certain conditions, Liu and Wu (1999) proved that the optimal prior on the expansion parameter α (i.e., the prior that produces the fastest sampler) is the Haar measure. Meng and van Dyk (1999) showed that while a proper prior on α always gives a valid algorithm, an improper prior (excluding the Haar measure) may result in an algorithm that does not have the correct target distribution. In contrast, ASIS does not require an expansion parameter or a prior associated with it; one simply identifies an SA and a corresponding AA and samples the parameter under both schemes, in the way as described in Section 2.3 and more generally in Section 2.5.

Another advantage is its generality. Clearly GIS or CIS is as generic as EM or the Gibbs sampler, in that the recipe is by no means limited to any particular problem. Incidentally, GIS/CIS is also analogous to EM in another sense. Before its general formulation by Dempster, Laird, and Rubin (1977), special cases of EM had already existed (see Meng and van Dyk 1997, for a historical review). For particular problems, the interweaving strategy was used independently in at least two Ph.D. theses (Yu 2005; Kypraios 2007), suggesting how easy it is to "stumble upon" such a method even without the general formulation and theory provided in the current article.

To identify an SA is usually easy, much like identifying sufficient statistics (more precisely, minimal sufficient statistics) in classical settings. It is common to seek exponential family models as our *augmented-data* model—without altering the observed-data model—for easy implementation of EM and DA. Hence (minimal) *complete-data* sufficient statistics are readily available, especially when we use the component-wise strategy; see Sections 3 and 4 for real-data examples. This easiness was also emphasized by Papaspiliopoulos, Roberts, and Skold (2007), who provided about a dozen examples, ranging from repeated measurements models to diffusion processes models.

Once an SA scheme is identified, the construction of a corresponding AA scheme is typically even easier, often by "standardizing/pivotizing" the identified SA. This is the same strategy used by Liu and Wu (1999), Meng and van Dyk (1997, 1999), van Dyk and Meng (2001), and Papaspiliopoulos, Roberts, and Skold (2007). "Standardizing/pivotizing" is not restricted to continuous missing data. For example, when the SA is a latent homogeneous Poisson process $\{X(t), t > 0\}$ with rate θ, as in certain diffusion processes problems

(Papaspiliopoulos, Roberts, and Skold 2007), we can construct an AA process by setting $\tilde{X}(t) = X(t/\theta)$, for all t, which has unit rate and is therefore ancillary. In the simpler case of a univariate and continuous $Y_{mis} \sim F(Y_{mis}|\theta)$, the CDF transformation $\tilde{Y}_{mis} = F(Y_{mis}|\theta)$ leads to an AA, because $\tilde{Y}_{mis} \sim \text{Uniform}(0, 1)$.

The third advantage of the interweaving strategy is its *flexibility*; there are many variations and modifications at the user's disposal for reaching a desirable balance between theoretical speed, computational efficiency (e.g., CPU time), and ease of programming. The full CIS as described in Section 2.5 has $2J + 1$ steps within each iteration. However, if some of the steps are difficult to implement, or costly in terms of CPU time, then one can easily omit them. (We shall call the resulting algorithm a *partial* CIS algorithm.) In general, the *space-filling condition* of Meng and Rubin (1993) is satisfied if we include Step 1 and, for each j, at least one of the Steps $(j + 1)$ and $\widetilde{(j + 1)}$ (since both update θ_j). The space-filling condition simply means that the entire space of $\{\theta, Y_{mis}\}$ lies in the span of the directions searched collectively by the steps executed within one iteration. These directions are not required to be orthogonal; indeed, as illustrated in Figure 1, ASIS purposefully searches for an "over-complete representation" (borrowing a term from the wavelets literature) or an "over-saturated design" (borrowing a term from experimental design) to achieve rapid mixing. With the space-filling condition satisfied, it is typically as easy to verify the irreducibility of any partial CIS as that of the original Gibbs-type sampler, even if some steps themselves become "incomplete"; see Section 4.2. We emphasize that, after accounting for CPU time or programming cost, partial CIS algorithms may be more desirable than the full CIS, as the real examples in Section 4 will demonstrate.

3. COMPONENT-WISE ASIS FOR A POISSON TIME SERIES MODEL

3.1 A Parameter-Driven Poisson Time Series Model

The scientific problem that motivated our work came from X-ray astrophysics, where it is common to model (binned) photon counts observed from a source by a Poisson distribution. Questions of scientific interest are whether source intensity is changing over time, and if so, how. The dataset plotted in Figure 2 is one of a few for which we were asked to determine if there is any statistical evidence for the change of intensity over the observation period.

For such data, we consider the following time series model (subscript "t" indexes time):

$$Y_t|(\xi_t, \beta) \overset{\text{ind}}{\sim} \text{Poisson}(d_t e^{X_t\beta+\xi_t}), \tag{3.1}$$

$$\begin{aligned} \xi_t|(\xi_{<t}, \rho, \delta) &\sim N(\rho\xi_{t-1}, \delta^2), \qquad t \geq 2, \qquad \text{and} \\ \xi_1 &\sim N(0, \delta^2/(1-\rho^2)), \end{aligned} \tag{3.2}$$

where $Y = \{Y_t\}$ denotes observed counts at time $t, t = 1, \ldots, T$, X_t is a $1 \times p$ vector of covariates, d_t is a known positive constant (e.g., the width of bin t), $\xi = \{\xi_t\}$ is a latent stationary AR(1) process, and $\xi_{<t} = \{\xi_j, j = 1, \ldots, t-1\}$. Marginally $\xi_t \sim N(0, \tau^2)$, where

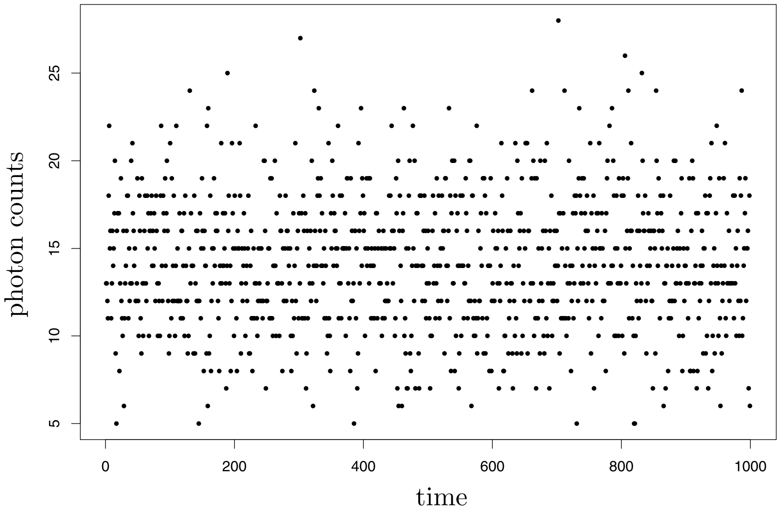

Figure 2. Photon counts observed by the Chandra X-ray telescope High Resolution Camera from a point source, the isolated neutron star/quark star candidate RX J1856.5-3754. The total exposure time is 55,476 seconds, which is divided into 1000 equal bins.

$\tau^2 = \delta^2/(1-\rho^2)$. The parameters of interest are $\beta = (\beta_1, \ldots, \beta_p)^\top$, ρ, and δ. In our case $p = 2$ and $X_t = (1, t/T)$, and the key interest is to infer whether the trend parameter β_2 is away from zero.

This model belongs to a broad class of *parameter-driven* time series models (Cox 1981). Perhaps due to its flexibility and ease of interpretation, it is commonly used in medical and social studies (e.g., Zeger 1988; Chan and Ledolter 1995; Frühwirth-Schnatter and Wagner 2006). However, the analytically intractable likelihood is a serious challenge for estimating model parameters; the MLE admits no simple formula. See the article by Davis and Rodriguez-Yam (2005) for numerical methods to approximate such a likelihood. Whereas Zeger (1988) used generalized estimating equations, Chan and Ledolter (1995) adopted a Monte Carlo EM approach. For Bayesian estimation, Frühwirth-Schnatter and Wagner (2006) proposed an interesting (approximate) Gibbs sampler.

To demonstrate the effectiveness of ASIS, we shall perform a Bayesian analysis with both synthetic and real data. Further illustrations using a dataset of Zeger (1988) can be found in an earlier version of this article (Yu and Meng 2007).

3.2 DATA AUGMENTATION AND ALGORITHMS

The standard Gibbs sampler for simulating from the posterior, assuming a prior $p(\beta, \rho, \tau) \propto 1$, has three blocks of steps. In our implementation, ρ is constrained between -0.99 and 0.99 to avoid numerical problems. Conditions derived by Michalak (2001) can be adapted to verify posterior propriety in all our examples. Technically we do not have a Gibbs sampler because we use Metropolis–Hastings to carry out some of its steps, but we nevertheless refer to it as "the standard Gibbs sampler" to emphasize that its key structure is still drawing from all full conditionals.

Standard Gibbs Sampler:

Step 1. Draw $\xi|(\beta, \rho, \delta, Y)$. Because ξ's are autocorrelated, we update ξ_t given ξ_{t-1} and ξ_{t+1} for $t = 1, \dots, T$ in turn via a one-dimensional M-H algorithm. The details of this and the following steps can be found in Appendix A.

Step 2_A. Draw $\beta|(\xi, \rho, \delta, Y)$, or equivalently $\beta|(\xi, Y)$ due to conditional independence. This step is equivalent to posterior sampling of a Poisson generalized linear model (GLM), and we use an M-H move similar to that in Step 1.

Step 3_S. Draw $(\rho, \delta)|(\xi, \beta, Y)$, or equivalently $(\rho, \delta)|\xi$. This step, which is in closed form, is equivalent to a Bayesian fitting of an AR(1) model, treating ξ as observed data.

As shall be evident from simulations as well as real-data examples, the standard sampler may perform very poorly. To design more efficient algorithms, we first identify the sufficient and ancillary augmentation schemes for the parameters. It is easy to check that the standard augmentation, ξ, is an AA for β but an SA for (ρ, δ)—hence the subscripts in "Step 2_A" and "Step 3_S." Therefore we only need to find an SA for β and an AA for (ρ, δ).

SA for β: Observe that if we treat $\eta = \{\eta_t\}$, where

$$\eta_t = \xi_t + X_t\beta, \qquad t = 1, \dots, T, \tag{3.3}$$

as the missing data, then the model can be rewritten as

$$Y_t \sim \text{Poisson}(d_t e^{\eta_t}); \tag{3.4}$$

$$\begin{aligned} \eta_t|\eta_{<t} &\sim N(\rho\eta_{t-1} + (X_t - \rho X_{t-1})\beta, \delta^2), \qquad t \geq 2, \qquad \text{and} \\ \eta_1 &\sim N\left(X_1\beta, \frac{\delta^2}{1-\rho^2}\right). \end{aligned} \tag{3.5}$$

Although the posterior distribution of (β, ρ, δ) remains the same, the augmented-data model has changed. See Figure 3 for a comparison of the two hierarchical model structures, that is, (3.1)–(3.2) versus (3.4)–(3.5). In particular, $p(Y|\eta, \beta, \rho, \delta)$ is now free of β

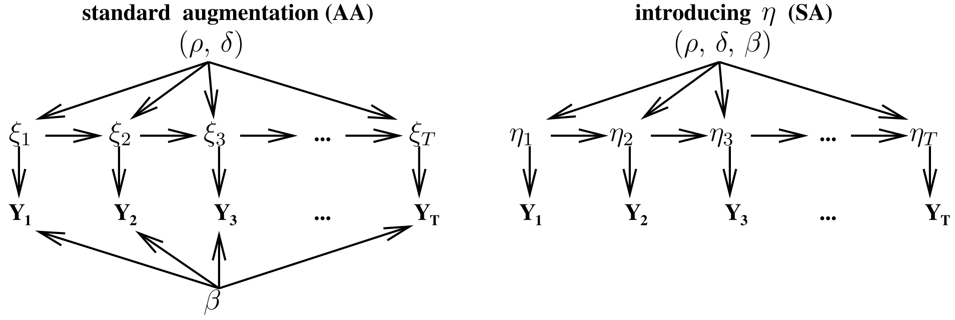

Figure 3. Change in dependence structure among the variables after introducing η. Left: structure of the original model, where ξ are the missing data and are an AA for β. Right: structure of the transformed model (3.4)–(3.5), where η are the missing data and are an SA for β. The arrows indicate how data could have been generated by sampling from the relevant conditional distributions (e.g., Y is generated conditional on $\xi_1, \dots, \xi_T$ and β in the standard augmentation).

(and in fact free of (β, ρ, δ)). The new missing data η are therefore an SA for β (and for (ρ, δ)).

Since the standard Gibbs sampler uses the AA for β, we may improve it by adding a step that samples β under the SA:

Step 2_S. Draw $\beta|(\eta, \rho, \delta, Y)$. (In addition to Steps 1, 2_A, and 3_S.)

This step happens to be simple and inexpensive, because only linear regression is involved; see Appendix A. Since Step 2_S treats η as missing data, we need to invert (3.3) to keep ξ updated, that is, we set $\xi_t^{new} = \eta_t - X_t\beta^{new}$ at the end of Step 2_S.

AA for (ρ, δ): Since the standard Gibbs sampler draws (ρ, δ) jointly using an SA, here we seek a joint AA for (ρ, δ), via "standardizing/pivotizing." Let

$$\kappa_1 = \frac{\sqrt{1-\rho^2}}{\delta}\xi_1 \qquad \text{and} \qquad \kappa_t = \frac{\xi_t - \rho\xi_{t-1}}{\delta}, \qquad t \geq 2. \tag{3.6}$$

Because κ_t's are i.i.d. $N(0, 1)$, $\kappa = \{\kappa_t, t \geq 1\}$ is an AA for both β and (ρ, δ). We therefore add

Step 3_A. Draw $(\rho, \delta)|(\kappa, \beta, \ Y)$, via a random-walk type M-H step.

Similarly to Step 2_S, we need to invert (3.6) at the end of Step 3_A to keep ξ updated:

$$\xi_1^{new} = \delta^{new}\kappa_1/\sqrt{1-(\rho^{new})^2}, \qquad \xi_t^{new} = \rho^{new}\xi_{t-1}^{new} + \delta^{new}\kappa_t, \qquad t \geq 2. \tag{3.7}$$

A variation of Step 3_A is to further break it into two Gibbs steps:

Step $3'_A$. Draw $\rho|(\kappa, \beta, \delta, Y)$ via a random-walk type Metropolis step.

Step $3''_A$. Draw $\delta|(\kappa, \beta, \rho, Y)$ via a random-walk type Metropolis step on the scale of $\log(\delta)$.

With all the steps defined above, we can design a variety of Gibbs-type samplers, among which Schemes A–E below will receive special attention:

Scheme A. The standard Gibbs sampler, that is, using only Steps 1, 2_A, and 3_S.

Scheme B. Each iteration cycles through Steps 1, 2_S, and 3_S.

Scheme C. Each iteration cycles through Steps 1, 2_A, 2_S, and 3_S.

Scheme D. Each iteration uses all five Steps: 1, 2_A, 2_S, 3_A, and 3_S.

Scheme E. Same as Scheme D except that Step 3_A is further split into Steps $3'_A$ and $3''_A$.

Table 1 summarizes the DA schemes for each parameter used by the five samplers. By comparing Scheme C with Scheme A or Scheme B, we can assess the effect of interweaving SA and AA on β; by comparing Scheme D with Scheme C, we can assess this effect on ρ and δ. The most comprehensive among the first four is Scheme D, which uses both SA and AA for every parameter. Scheme E is a variation of Scheme D, with a trade-off between computational load (which is a bit less for Scheme E because of the one-dimensional M-H for drawing ρ and δ) and convergence rate (Scheme D may converge faster as it draws (ρ, δ) together). In our simulations Scheme E turns out to be a good compromise, and is therefore included in the comparisons.

3.3 COMPUTATIONAL PERFORMANCE WITH SIMULATED AND REAL DATA

Two datasets are simulated, each with $T = 200$ observations, $p = 2$, and $X_t = (1, t/T)$. The five schemes are run for 15,000 iterations each, with the first 5000 as burn-in, and

Table 1. The DA schemes used by five samplers for each parameter.

	Scheme A	Scheme B	Scheme C	Scheme D	Scheme E
β	AA	SA	SA & AA	SA & AA	SA & AA
ρ	SA	SA	SA	SA & AA	SA & AA
δ	SA	SA	SA	SA & AA	SA & AA

the remaining 10,000 draws for displaying their trajectories and autocorrelations for all parameters of interest.

The first simulated dataset, DATA1, is intended to show the poor performance of the standard Gibbs sampler. DATA1 is generated according to $d_t = 5000$ and $(\beta_1, \beta_2, \rho, \delta) = (0, 1, 0.5, 0.1)$. The counts generated are large, that is, in the order of thousands. From Figures 4 (trajectories of the draws) and A.1 (autocorrelations; Appendix A), it is clear that Scheme C is a dramatic improvement. The improvement is dramatic even after accounting for CPU time, as hinted by Table 2, which is for DATA2 discussed next but the pattern is similar for DATA1.

Since adding Step 2_S to Scheme A performs so well, and with little increase in computational load, one is tempted to simply substitute Step 2_S for Step 2_A in the standard Gibbs sampler, which gives Scheme B. Indeed, for DATA1 the performance of Scheme B is almost identical to that of Scheme C (hence the plots are not shown).

To show that Scheme B does not always work well, a second dataset (DATA2) is generated according to $d_t = 10$ and $(\beta_1, \beta_2, \rho, \delta) = (0, 1, 0, 0.01)$. The counts are much smaller

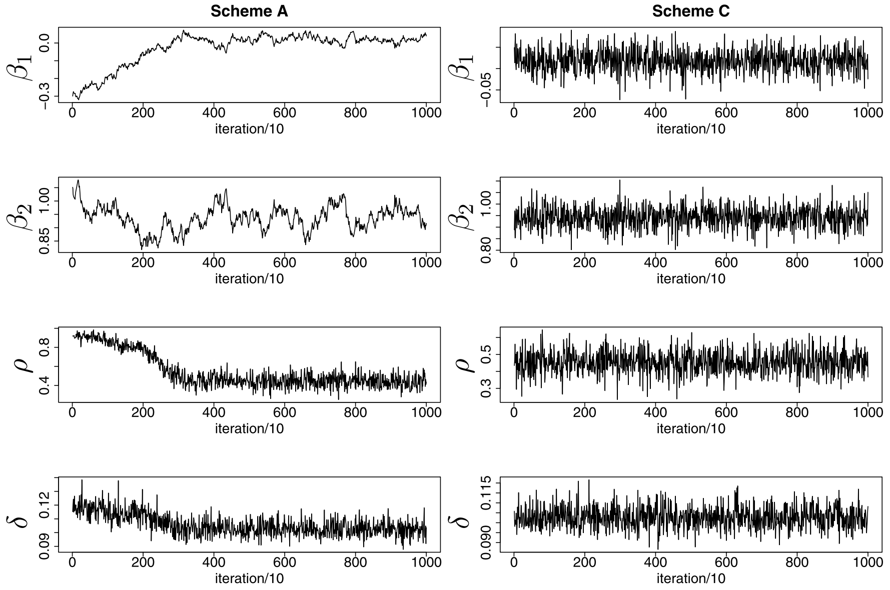

Figure 4. Comparing Scheme A with Scheme C, which adds Step 2_S to each iteration of Scheme A. Displayed are the trajectories of the Monte Carlo draws (excluding the burn-in period and keeping every tenth iteration) produced by the two schemes on DATA1.

Table 2. Time consumed to complete 15,000 iterations for Schemes A–E on DATA2.

	Scheme A	Scheme B	Scheme C	Scheme D	Scheme E
CPU time (secs)	42	31	42	56	47

(in tens). The first two columns of Figures 5 (trajectories) and A.2 (autocorrelations; Appendix A) compare Scheme B with Scheme C, which can also be viewed as adding Step 2_A to Scheme B. Although draws for ρ and δ behave equally badly for both schemes, convergence of β_1 and β_2 is improved considerably after adding Step 2_A, that is, when Scheme C is used.

Scheme C is far from perfect: for DATA2 the draws of ρ and δ still exhibit heavy autocorrelations. The draws of δ, in particular, tend to stay near zero for long periods of time. The third columns of Figures 5 and A.2 (Appendix A) display the performance of Scheme D, which adds Step 3_A (corresponding to the AA for (ρ, δ)) to Scheme C. Observe that Scheme D does as well as Scheme C for β, but improves considerably for ρ and δ. Again using both SA and AA pays off.

Scheme E (not shown) performs slightly more poorly than Scheme D, but it is computationally faster per iteration, as shown in Table 2. Scheme B is the least costly per iteration, but overall Scheme E is arguably the most efficient strategy, taking into account both the convergence rate and the computing time per iteration. (Scheme A, not shown, performs no better than Scheme C.)

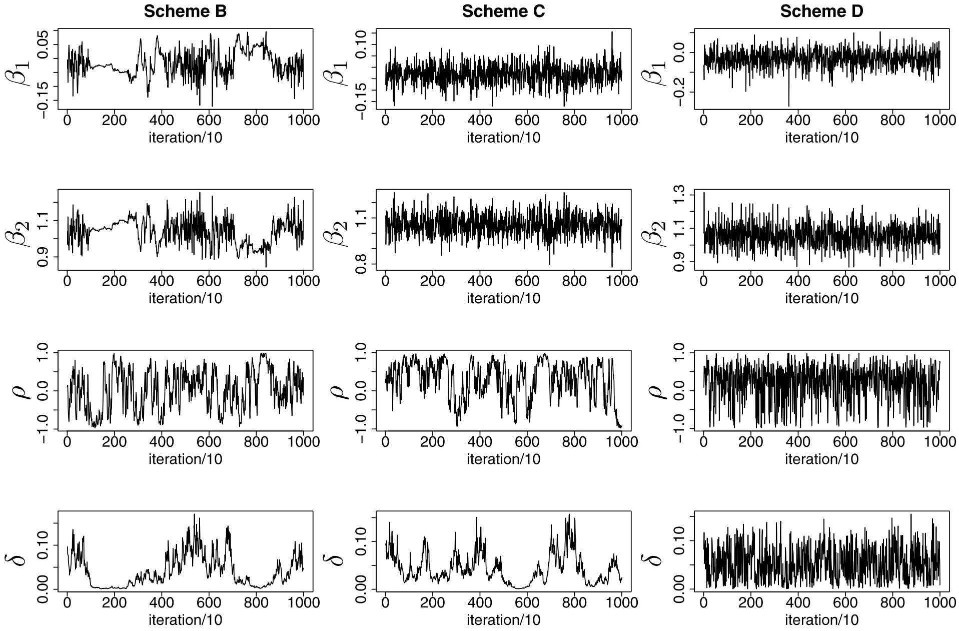

Figure 5. Trajectories under Schemes B, C, and D on DATA2, after the burn-in period.

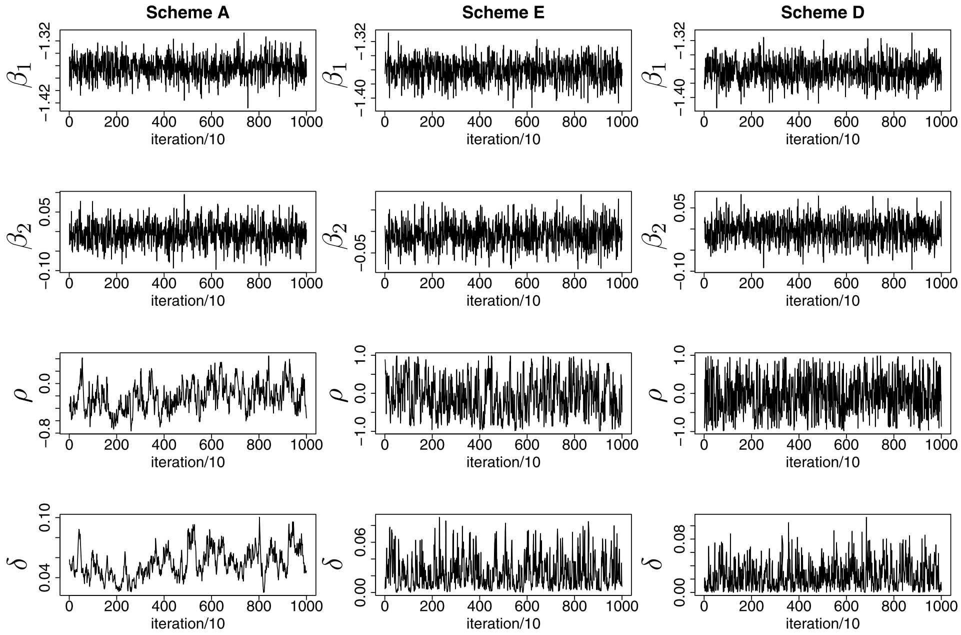

Figure 6. Trajectories under Schemes A, D, and E on the Chandra X-ray data, after the burn-in period.

Finally, we model our motivating real data (Figure 2) according to (3.1)–(3.2) with $T = 1000$. We take d_t as the width of the time bin (in seconds), and specify the covariates as $X_t = (1, t/1000)$, which allows for a linear trend in the log Poisson intensity. Figures 6 (trajectories) and A.3 (autocorrelations; Appendix A) compare MCMC Scheme A with Schemes D and E. Evidently, draws for β parameters have little autocorrelations under all three schemes; draws for ρ and δ, however, mix very slowly under Scheme A, but their convergence is much improved under Scheme D or Scheme E. (The results were confirmed through replications.)

Figure 7 displays some plots of the posterior distribution calculated by Scheme E. The marginal posterior of β_2, the trend parameter, is centered around zero with a small posterior variance, and the posterior of δ, the parameter that captures extra-Poisson variation, is very close to zero. Our inference, therefore, has to be that there is little evidence in this dataset to suggest that source intensity varies, given our modeling assumptions.

As a closing remark on this application, we note that the improvement for each component appears to be brought in by a specifically designed ASIS for that component, which may or may not have much impact on other components. That is, CIS seems to work in a piecemeal fashion. Mathematically this is difficult to formalize, because an algorithm converges only if all its components converge. Nevertheless, Theorem 3 of Section 5 provides a theoretical insight that this piecemeal phenomenon perhaps can be expected in many applications because of the multiplicative nature of a lower bound on the speed of CIS. Roughly speaking, CIS is effective if (i) interweaving is effective for each component conditional on other components, and (ii) after marginalizing over the missing data, the dependence between components of θ is not extreme.

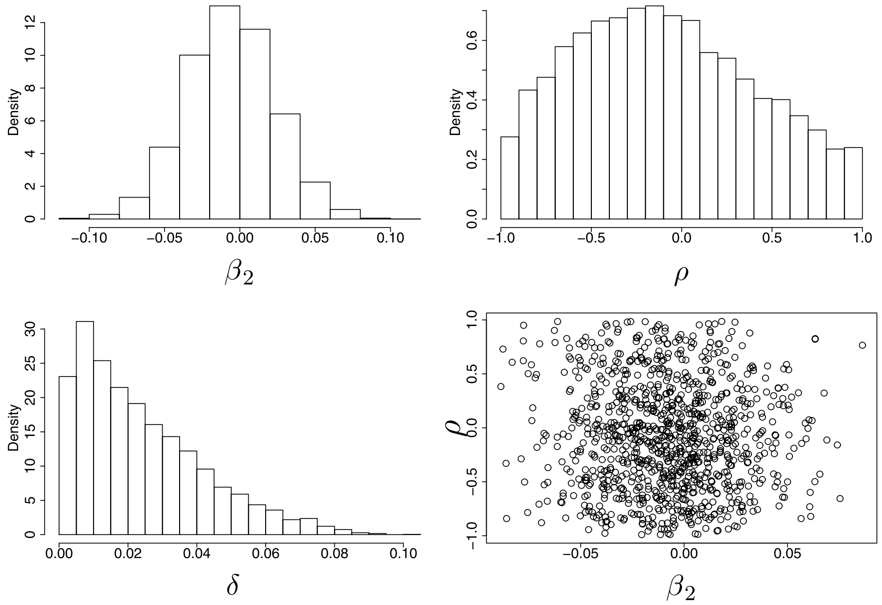

Figure 7. Posterior inference of the Chandra X-ray data, with the draws produced by Scheme E.

4. COMPETITIVENESS AND FLEXIBILITY OF PARTIAL ASIS/CIS

4.1 APPLICATION: PROBIT REGRESSION

Consider the widely used probit regression model

$$Y_i = \operatorname{sgn}(\phi_i), \qquad \phi_i|(\theta, X) \overset{\text{ind}}{\sim} N(X_i\theta, 1), \tag{4.1}$$

where Y_i is the observed binary (± 1) outcome, the sign of the latent score ϕ_i, X_i is a $1 \times p$ vector of covariates, and θ is a $p \times 1$ vector of regression coefficients. Write $Y = (Y_1, \ldots, Y_n)^\top$, $\phi = (\phi_1, \ldots, \phi_n)^\top$, and $X^\top = (X_1^\top, \ldots, X_n^\top)$. For Bayesian inference, a standard noninformative prior is $p(\theta) \propto 1$. The model formulation (4.1) leads to

$$\phi_i|(\theta, Y) \sim \mathrm{TN}(X_i\theta, 1, Y_i), \tag{4.2}$$

$$\theta|(\phi, Y) \sim N(\hat{\theta}, (X^\top X)^{-1}), \tag{4.3}$$

where $\hat{\theta} = (X^\top X)^{-1} X^\top \phi$, and $\mathrm{TN}(\mu, \sigma^2, Y_i)$ denotes a $N(\mu, \sigma^2)$ distribution truncated to the interval $(0, \infty)$ if $Y_i = 1$ and to $(-\infty, 0)$ if $Y_i = -1$. Although in this example it is more appropriate to express the latent variable ϕ as Y_{aug} instead of Y_{mis}, because Y is determined by ϕ, we retain the Y_{mis} notation for simplicity.

The standard DA algorithm (Albert and Chib 1993), which treats ϕ as Y_{mis}, iteratively draws $\phi|\theta$ according to (4.2) and $\theta|\phi$ according to (4.3). Though convenient, this algorithm is sometimes very slow. To apply ASIS, we first note that ϕ is already an SA for θ in (4.1). To find a corresponding AA, we simply set $\eta_i = \phi_i - X_i\theta$ and treat $\eta = (\eta_1, \ldots, \eta_n)^\top$

as $\tilde{Y}_{mis}$, using our generic notation. Our model then becomes

$$Y_i = \operatorname{sgn}(\eta_i + X_i\theta), \qquad \eta_i|\theta \sim N(0, 1). \tag{4.4}$$

We can then follow the ASIS recipe defined in Section 2.2.

This simple construction is only slightly compromised by the fact that drawing from $p(\theta|\eta, Y)$, albeit a uniform distribution on the convex set $\{\theta : Y_i = \operatorname{sgn}(\eta_i + X_i\theta), i = 1, \dots, n\}$, might require some bookkeeping because of potentially complicated boundaries. This problem can be dealt with, perhaps with a slight loss of mixing efficiency, by implementing a "nested Gibbs." That is, we draw each coordinate of θ uniformly conditional on the other coordinates and subject to the boundary constraints. The resulting variation of the ASIS sampler then becomes:

Step 1. Draw ϕ according to (4.2).

Step 2_S. Draw θ according to (4.3).

Step 2_A. Compute $\eta = \phi - X\theta$, and draw in turn each θ_j, $j = 1, \dots, p$, from a uniform distribution on the interval $\{\theta_j : Y_i = \operatorname{sgn}(\eta_i + X_i\theta), i = 1, \dots, n\}$ with η and the rest of θ fixed.

To see its effectiveness, we compare it with the optimal marginal augmentation algorithm of van Dyk and Meng (2001), or equivalently the PX-DA algorithm of Liu and Wu (1999) (PX stands for "Parameter Expanded"; see Liu, Rubin, and Wu 1998). This algorithm introduces a working parameter α:

$$Y_i = \operatorname{sgn}(w_i), \qquad w_i|(\theta, X) \sim N(X_i\theta\alpha, \alpha^2). \tag{4.5}$$

The key is that the working parameter is only identifiable from the augmented-data model; the observed-data model $p(Y|\theta)$ remains the same. One then assigns a working prior on α and alternatingly draws $(\theta, \alpha)|(w, Y)$ and $(w, \alpha)|(\theta, Y)$. This is in theory equivalent to drawing $\theta|(w, Y)$ and $w|(\theta, Y)$ alternatingly, with α integrated out.

The choice of the working prior clearly will affect the rate of convergence, and indeed it can even affect the validity of the resulting algorithm when improper working priors are involved (see Meng and van Dyk 1999). Liu and Wu (1999) showed that the optimal choice is the improper prior $p(\alpha) \propto \alpha^{-1}$, optimal in the sense of achieving the fastest geometric rate among a class (which includes the standard DA) as defined by Liu and Wu (1999). Each iteration of this optimal algorithm consists of the following three steps:

1. Draw ϕ according to (4.2).
2. Draw $\alpha \sim \sqrt{\mathrm{RSS}/\chi_n^2}$, where $\mathrm{RSS} = \sum_i (\phi_i - X_i\hat{\theta})^2$ and $\hat{\theta} = (X^\top X)^{-1} X^\top \phi$.
3. Draw θ according to $N(\hat{\theta}/\alpha, (X^\top X)^{-1})$.

Van Dyk and Meng (2001) provided a real-data example showing considerable improvement of this optimal algorithm over the algorithm of Albert and Chib (1993). The dataset considered by van Dyk and Meng (2001) (their table 1) concerns two clinical measurements (i.e., covariates) that are used to predict the occurrence of latent membranous lupus nephritis. We extend this comparison by including the interwoven sampler described earlier. Figures 8 and 9 display the trajectories and autocorrelations of the draws of θ_1, θ_2,

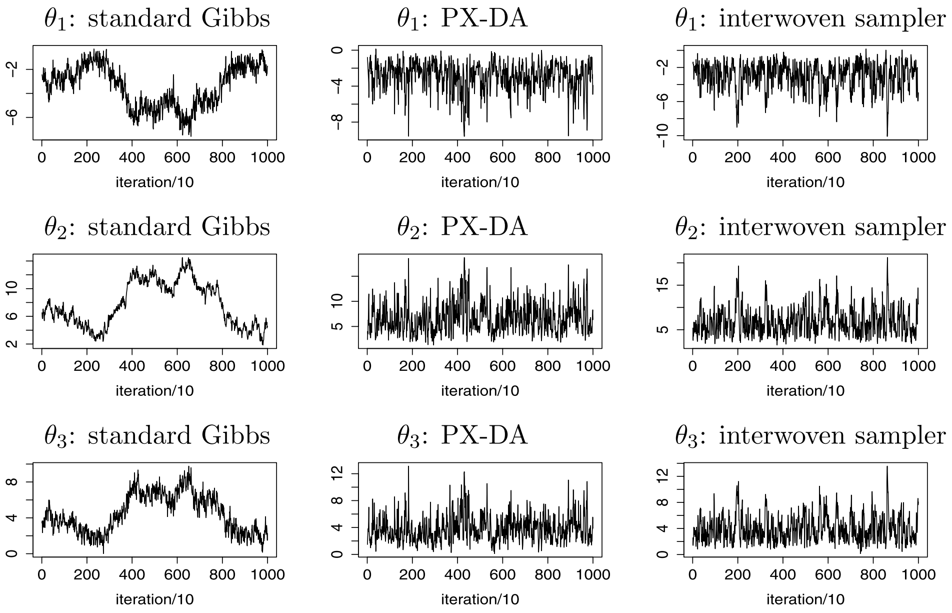

Figure 8. Comparing standard Gibbs with the optimal PX-DA and the interwoven sampler for the lupus nephritis data: trajectories of the draws.

and θ_3 for standard Gibbs, the optimal PX-DA, and the ASIS sampler to fit (4.1) with both covariates, that is, $p = 3$. Each algorithm is run for 11,000 iterations and the autocorrelations are calculated from the last 10,000 iterations.

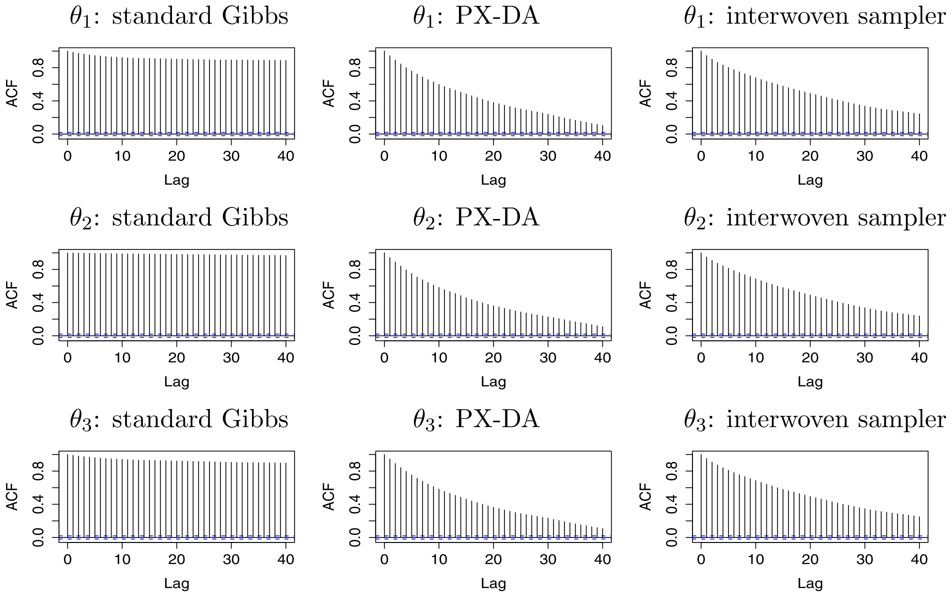

Figure 9. Comparing standard Gibbs with the optimal PX-DA and the interwoven sampler for the lupus nephritis data: autocorrelations of the draws. The online version of this figure is in color.

Both optimal PX-DA and ASIS offer considerable improvement over the standard Gibbs for all three parameters. While the optimal PX-DA performs slightly better in terms of the autocorrelation, the ASIS sampler is also doing very well. In terms of the CPU time, the ASIS sampler, as we implemented it, is about 40% more costly than the optimal PX-DA, though each is within a total of 0.8 seconds for all 11,000 iterations, and hence the difference in CPU time is not a practical concern. The major cost of ASIS appears to be our brute force way of determining the boundary of the convex set in Step 2_A, a problem to which others may have more efficient solutions. However, even without such efficient implementations, in terms of improving standard Gibbs, ASIS is very competitive. The significance of this competitiveness lies in that ASIS boosts MCMC efficiency without going beyond interweaving two readily available SA and AA algorithms—the ASIS algorithm does not require any steps that are not required by the original SA or AA algorithm, other than the (often trivial) transformations between $\tilde{Y}_{mis}$ and Y_{mis}, as in all our applications.

4.2 APPLICATION: NORMAL REGRESSION WITH INTERVAL CENSORING

Consider the following regression model:

$$Y_i|(\beta,\sigma^2) \overset{\text{ind}}{\sim} N(X_i\beta,\sigma^2), \qquad Y_i \in (Y_i^l, Y_i^r), \qquad i = 1,\dots,n, \tag{4.6}$$

where the intervals $(Y_i^l, Y_i^r), i = 1,\dots,n$, are observed data, Y_i's are the latent response, $X = (X_1^\top,\dots,X_n^\top)^\top$ is the $n \times p$ matrix of explanatory variables, and $\{\beta,\sigma^2\}$ are the parameters. This model was used by Hamada and Wu (1995) to analyze an experiment of improving the lifetime of fluorescent lights (Liu and Sun 2000; Liu 2003); data can be found in table 1 of the article by Liu (2003).

Following Liu (2003), let $Y_i = \log(u_i) - 3$, where u_i is the lifetime (in days) of unit i, for $i = 1,\dots,n$. The prior on $\{\beta,\sigma^2\}$ is specified by $p(\sigma^2) = \text{Inv}\chi^2(\nu_0, \nu_0 s_0)$, that is, a scaled inverse χ^2 with the density proportional to $(\sigma^{-2})^{\nu_0/2+1}\exp[-\nu_0 s_0/(2\sigma^2)]$, and by $p(\beta|\sigma^2) = N(0, \sigma^2 I_p/\tau_0)$. The hyperparameters are $\nu_0 = 1, s_0 = 0.01$, and $\tau_0 = 0.0001$, as in the work of Liu (2003).

We emphasize that for (4.6) to make scientific sense, we have assumed that the interval censoring mechanism is ignorable (Rubin 1976). This point is also important for properly defining what is missing. Here interval censoring occurs because inspection of the working status of a light can only be done at a finite number of times, say $T_i \equiv \{T_{i,1},\dots,T_{i,m_i}\}$. The ignorability assumption requires that the selection mechanism for T_i does not depend on any knowledge of the unobserved lifetime itself, conditional on the covariate X. In the actual experiment, the inspection was every other day over 20 days, and hence the ignorability assumption is reasonable.

Under such an assumption, the correct likelihood can be specified by conditioning on the value of T_i, and the observed data are given by the binary vector $W_i = \{W_{i,1},\dots,W_{i,m_i}\}$, where $W_{i,j} = 1_{\{Y_i \ge T_{i,j}\}}$, $j = 1,\dots,m_i$. Clearly, W_i can be summarized by

$$Y_i^l = \max_j\{T_{i,j} : W_{i,j} = 1\} \qquad \text{and} \qquad Y_i^r = \min_j\{T_{i,j} : W_{i,j} = 0\}, \tag{4.7}$$

with no loss of information. In this sense we can treat Y^l and Y^r as the observed data. We can then formulate a DA scheme by treating $Y = (Y_1, \ldots, Y_n)^\top$ as the missing data, because $p(Y^l, Y^r, Y|T, X, \beta, \sigma)$ is a well-defined DA model. Under this model, the standard DA algorithm consists of the following two steps (Liu 2003):

Standard Gibbs Sampler:

Step 1. Draw $Y|(Y_{obs}, \beta, \sigma)$ where Y_{obs} denotes all observed data, including (Y_i^l, Y_i^r) and X. This conditional distribution is a product of n independent truncated normal distributions, each Y_i having a mean parameter $X_i\beta$, a variance σ^2, and truncation bounds Y_i^l and Y_i^r.

Step 2_S. Draw $(\beta, \sigma)|Y$, which amounts to a Bayesian linear regression of Y on X.

Although both steps are in closed form, the standard sampler is extremely slow for the dataset considered by Hamada and Wu (1995); see the left column of Figure 10, which displays the trajectories of the draws of $\{\beta_1, \beta_2, \beta_3\}$ for standard Gibbs. The right column displays corresponding plots from the interwoven algorithm described below. The autocorrelation plots (not shown to save space) confirm that the new sampler produces essentially uncorrelated draws. The interwoven sampler costs about 10% more time per iteration, which is well compensated for by the dramatically improved mixing rate. Liu (2003) proposed a covariance adjustment step and reported similar improvements in his figure 1. Indeed, because both Liu's algorithm and our algorithm improve

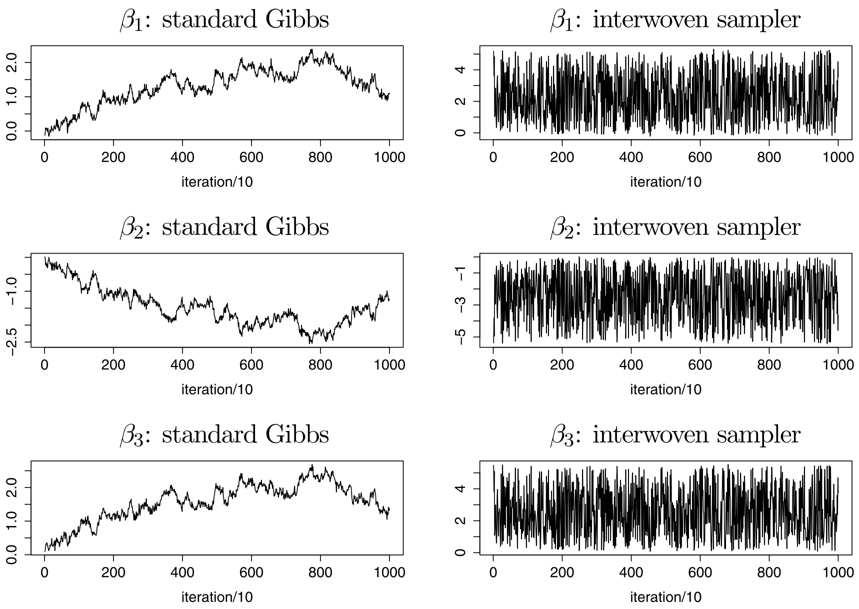

Figure 10. Comparing standard Gibbs with the interwoven sampler for the fluorescent light lifetime data: trajectories of the draws.

the standard Gibbs sampler so much, visually we are not able to tell which one improves more.

The construction of an ASIS sampler for this problem is a great illustration of its flexibility and power. First, the constructions of both SA and AA are straightforward, as in many other problems. Similarly to the probit regression example, the missing data Y for the standard DA are naturally an SA for (β, σ) (hence the subscript for Step 2_S). By standardizing it in the usual way,

$$\eta_i = \frac{Y_i - X_i\beta}{\sigma}, \qquad i = 1, \ldots, n,$$

we can re-express (4.6) as

$$\eta_i \sim N(0,1), \qquad \eta_i \in \left(\frac{Y_i^l - X_i\beta}{\sigma}, \frac{Y_i^r - X_i\beta}{\sigma}\right), \qquad i = 1, \ldots, n, \tag{4.8}$$

and hence $\eta = (\eta_1, \ldots, \eta_n)^\top$ is an AA for (β, σ) because η is free of (β, σ) marginally. Note here the truncation on η_i is not a part of its *marginal distribution*, but rather it defines the relationship between Y_{obs} and η given the model parameter (β, σ).

Once the AA is in place, the GIS of Section 2.3 would call for a Step 2_A, which samples $\{\beta, \sigma\}$ given η (and Y_{obs}), in addition to Steps 1 and 2_S. Analogous to Step 2_S, we may consider drawing σ given η and then drawing β given σ and η. Unfortunately, because of the restrictions on η as in (4.8), the density of the conditional distribution of σ given η is complicated, involving intractable multiple integrals. In contrast, the two conditionals $p(\beta|\sigma, \eta)$ and $p(\sigma|\beta, \eta)$ are easier to handle, since the former is a multivariate truncated normal ($i = 1, \ldots, n$):

$$p(\beta|\eta, Y_{obs}, \sigma) \propto \exp\left\{-\frac{\tau_0\beta^\top\beta}{2\sigma^2}\right\}, \qquad X_i\beta \in (Y_i^l - \sigma\eta_i, Y_i^r - \sigma\eta_i), \tag{4.9}$$

and the latter is a truncated scaled-inverse-χ:

$$\begin{gathered} p(\sigma|\eta, Y_{obs}, \beta) \propto \sigma^{-\nu_0-p-1} \exp\left\{-\frac{\nu_0 s_0 + \tau_0\beta^\top\beta}{2\sigma^2}\right\}, \\ \sigma\eta_i \in (Y_i^l - X_i\beta, Y_i^r - X_i\beta). \end{gathered} \tag{4.10}$$

The availability of these full conditionals leads to a number of choices, including (I) we could implement a nested Gibbs sampler within the planned Step 2_A by iterating between (4.9) and (4.10) until convergence, or more practically, for a fixed but large number of iterations (e.g., 100); (II) we could just iterate between (4.9) and (4.10) once; (III) we do not even need to use both (4.9) and (4.10); we can just use one of them, say (4.9); (IV) because sampling from multivariate truncated normal distributions is already a nontrivial problem, we draw each coordinate of β in turn in a Gibbs sampling fashion, as we did for θ in Section 4.1.

Interweaving (I) with the standard Gibbs sampler amounts to implementing GIS almost exactly. But this can be costly, and hence it defeats the purpose of boosting computational efficiency. As discussed in Section 2.6, however, it is not necessary that we carry out the full interweaving strategy in order to have substantial improvement. Interweaving any one of (II)–(IV) with the standard Gibbs may still produce substantial gains, and indeed may be better than performing the full version (I) if it is too costly to implement (see below).

We emphasize that the decision of not carrying out full interweaving is largely governed by computational cost. For example, option (IV) could be very ineffective, if the mass of the truncated normal for β is highly concentrated in a lower-dimensional subset of $\mathbf{R}^p$. In such a case, a rotation $\tilde{\beta} = Z\beta$, where Z is a judiciously chosen invertible matrix, may alleviate the problem. That is, it would be better to alternate between the components of $\tilde{\beta}$, and then transform back to $\beta = Z^{-1}\tilde{\beta}$ in the end. Liu and Rubin (1996) advocated a Markov-normal analysis, based on a preliminary run, to identify slowly converging components of β. This may help in choosing Z.

Intriguingly, for our real data, an inspection of the covariance matrix of the draws of β reveals that $\beta_2, \beta_5, \beta_7$, and β_8 have pairwise correlation coefficients nearly 1, and yet they are heavily negatively correlated (with correlation coefficients nearly -1) with $\beta_1, \beta_3, \beta_4, \beta_6$. This suggests a modification to (IV), that is, we do not even need to draw all components of β because β effectively lives in a one-dimensional space. We therefore reduce the full Step 2_A to a very simple one: propose a conditional move of β_1 given $(\beta_1 + \beta_2, \beta_1 - \beta_3, \beta_1 - \beta_4, \beta_1 + \beta_5, \beta_1 - \beta_6, \beta_1 + \beta_7, \beta_1 + \beta_8)$ and η, Y_{obs}, σ. Equivalently, letting $V = (1, -1, 1, 1, -1, 1, -1, -1)^\top$, we draw δ from

$$p(\delta|\beta, \eta, Y_{obs}, \sigma) \propto \exp\left\{-\frac{\tau_0(\beta + V\delta)^\top(\beta + V\delta)}{2\sigma^2}\right\},$$

$$X_i(\beta + V\delta) \in (Y_i^l - \sigma\eta_i, Y_i^r - \sigma\eta_i),$$

and then set $\beta \leftarrow \beta + V\delta$.

Although this clearly is not identical to the draw called for by the original AA (e.g., we have ignored (4.10)), it is an extremely effective strategy for this dataset, as seen in the right column of Figure 10. It also greatly simplifies the algorithm because drawing δ only involves a one-dimensional truncated normal. This is also an example of *partial* GIS/CIS because if we use it *in place of* the original Step 2_S, the resulting algorithm will not be space filling. However, when it is used as an *added step* to the original Step 2_S, the whole algorithm remains valid.

To examine the difference between using the full and the partial Step 2_A, we also carry out a small simulation study comparing standard Gibbs with two interweaving strategies, ASIS 1 and ASIS 2. For ASIS 1, in addition to Steps 1 and 2_S, each iteration uses a Step 2_A that (i) draws each coordinate of β in turn given (η, Y_{obs}, σ) according to (4.9), and (ii) draws σ given (η, Y_{obs}, β) according to (4.10). In contrast, each iteration of ASIS 2 performs 100 nested iterations of the substeps (i) and (ii). A major goal is to evaluate the trade-off between a nearly exact but expensive Step 2_A (Scheme ASIS 2) and a cheap surrogate of it (Scheme ASIS 1). We emphasize that both ASIS 1 and ASIS 2 are legitimate samplers, but ASIS 2 is closer to the original global ASIS, and would be exactly the global ASIS sampler if we run the nested iteration until convergence; this is analogous to the ECM algorithm (Meng and Rubin 1993), which will approach the original EM algorithm if we perform a sufficient number of nested CM-steps.

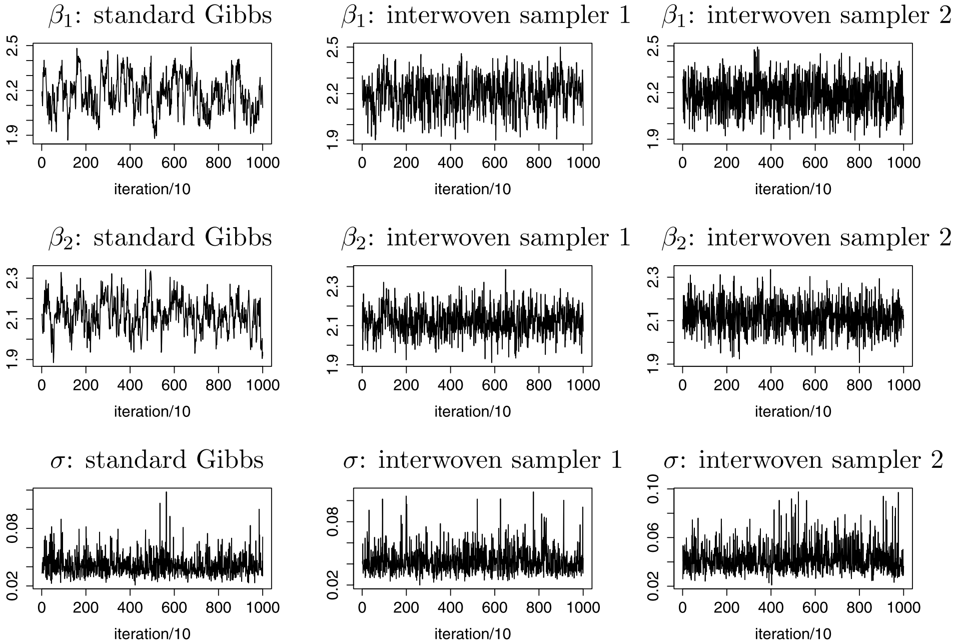

Figure 11. Comparing standard Gibbs with two versions of interweaving for simulated data: trajectories of the draws.

A dataset is generated with $n = 16$ observations and $p = 8$ covariates. The true parameters are $\beta_i = 2, i = 1, \ldots, p$, and $\sigma = 0.1$. All entries of the design matrix X except for those in the first column, which are all ones, are simulated as i.i.d. $N(0, 1)$ variables. For the censoring mechanism we set $Y_i^l = \lfloor Y_i \rfloor$ and $Y_i^r = Y_i^l + 1$, where $\lfloor x \rfloor$ denotes the integer part of x. The same prior on $\{\beta, \sigma\}$ as for the real data is assumed.

The trajectories of the draws of β_1, β_2, σ are displayed as Figure 11. (Repeated experiments show similar patterns.) It is clear that both versions of the interwoven sampler have considerably better autocorrelations than standard Gibbs. Autocorrelation plots (not shown) indicate that ASIS 2 achieves near zero autocorrelation at lag 1, while ASIS 1 does so at about lag 10. After taking into account the computing time per iteration, however, ASIS 2 is not nearly as impressive as ASIS 1. Table 3 displays the total CPU time for completing all 11,000 iterations for the three algorithms. While ASIS 1 costs about 60% more time per iteration than standard Gibbs, ASIS 2 costs 38 times more time than ASIS 1. This indicates that ASIS 1 is a good trade-off for this example.

Table 3. Comparing total CPU time using the simulated data.

	Standard Gibbs	ASIS 1	ASIS 2
CPU time (secs)	0.62	1.00	38.55

5. ROBUSTNESS AND OPTIMALITY OF INTERWEAVING STRATEGY

5.1 RATE OF CONVERGENCE AND SPEED OF CONVERGENCE

For theoretical investigations, we first focus on robustness properties of GIS and CIS. These results (Theorems 1–3) are established under the most general setting, that is, they are not restricted to ASIS nor do they require the mapping between the DAs being interwoven be one-to-one. We then prove an optimality result for global ASIS in a more restrictive setting (Theorem 4). We emphasize that these results, given their general nature, are more useful for gaining qualitative theoretical insights than for quantifying actual gains in mixing rates in particular applications. The latter assessments, as with the vast majority of studies in MCMC literature, are done via empirical investigations, as presented in Sections 3 and 4.

Before we present our main findings, however, it is important to realize that combining different transition rules for Markov chain samplers does not always improve the convergence rate, and in fact, even the irreducibility requirement, which is necessary for convergence, is not automatically preserved. See Appendix B.1 for a simple example where alternating two irreducible chains with the same stationary distribution actually results in a reducible chain. Fortunately, for the interweaving strategy, it is typically as easy to check its irreducibility as directly checking that of either of the original two schemes. In the following discussion we will therefore assume irreducibility, just as with most theoretical results on MCMC.

For a Markov chain $\{U^{(t)}, t = 1, 2, \ldots\}$ with invariant distribution π, recall that the L^2 geometric rate of convergence is the spectral radius of the forward conditional expectation operator $\mathbf{F}$ in $L_0^2(\pi) = \{h(u) : \mathrm{E}_\pi[h(U)] = 0; \mathrm{V}_\pi[h(U)] < \infty\}$, where E_π and V_π denote expectation and variance with respect to π, and $\mathbf{F}h(U^{(1)}) = \mathrm{E}[h(U^{(2)})|U^{(1)}]$, with its L^2 norm defined as

$$\|\mathbf{F}\| = \sup_{h\in L_0^2(\pi)} \sqrt{\frac{\mathrm{V}_\pi[\mathrm{E}[h(U^{(2)})|U^{(1)}]]}{\mathrm{V}_\pi[h(U^{(2)})]}}. \tag{5.1}$$

The spectral radius of $\mathbf{F}$ is governed by its norm because

$$\mathrm{spec}(\mathbf{F}) = \lim_{t\to\infty} \|\mathbf{F}^t\|^{1/t} \le \|\mathbf{F}\|, \tag{5.2}$$

where equality holds when $\mathbf{F}$ is self-adjoint, that is, when $\mathrm{cov}(\mathbf{F}h(U), g(U)) = \mathrm{cov}(h(U), \mathbf{F}g(U))$ for $U \sim \pi$ and all $h, g \in L_0^2(\pi)$. It follows that, if the Markov chain is *time reversible*, then its geometric rate is exactly $\|\mathbf{F}\|$ (see Amit 1991; Liu, Wong, and Kong 1994, 1995).

For nonreversible MCMC, the norm $\|\mathbf{F}\|$ is typically easier to handle than the spectral radius. Note that the right side of (5.1) is an equivalent expression of the maximal correlation coefficient (MCC) between $U^{(1)}$ and $U^{(2)}$ under stationarity, that is,

$$\mathcal{R}\big(U^{(1)}, U^{(2)}\big) = \sup_{g,h\in L_0^2(\pi)} \mathrm{corr}\big\{g\big(U^{(1)}\big), h\big(U^{(2)}\big)\big\}. \tag{5.3}$$

It turns out that we need a more general notion of the *maximal partial correlation* (MPC). Given three sub-σ-algebras $\mathcal{A}_1, \mathcal{A}_2$, and $\mathcal{M}$ on the same probability space, the MPC between $\mathcal{A}_1$ and $\mathcal{A}_2$ given $\mathcal{M}$ is defined as

$$\mathcal{R}_{\mathcal{M}}(\mathcal{A}_1, \mathcal{A}_2) = \sup_{\sigma(X)\subset\mathcal{A}_1, \sigma(Z)\subset\mathcal{A}_2} \frac{\mathrm{cov}(X - \mathrm{E}[X|\mathcal{M}], Z - \mathrm{E}[Z|\mathcal{M}])}{\sqrt{\mathrm{V}(X - \mathrm{E}[X|\mathcal{M}])\mathrm{V}(Z - \mathrm{E}[Z|\mathcal{M}])}}, \tag{5.4}$$

where the supremum is taken over all random variables X and Z such that X is $\mathcal{A}_1$-measurable, Z is $\mathcal{A}_2$-measurable, $0 < \mathrm{V}(X - \mathrm{E}[X|\mathcal{M}]) < \infty$, and $0 < \mathrm{V}(Z - \mathrm{E}[Z|\mathcal{M}]) < \infty$. It is convenient to define $R_{\mathcal{M}}(\mathcal{A}_1, \mathcal{A}_2) = 0$ if $\mathcal{A}_1 \subset M$ or $\mathcal{A}_2 \subset M$. A special case involving random variables (X, Y, Z) (having a well-defined joint distribution) is

$$\mathcal{R}_Y(X, Z) = \mathcal{R}_{\sigma(Y)}(\sigma(X), \sigma(Z)).$$

If (X, Y, Z) is a trivariate normal, then $\mathcal{R}_Y(X, Z)$ reduces to the usual partial correlation, that is, the correlation between the residuals of the regressions of X on Y and Z on Y.

The following inequality on MPC is instrumental to our general bounds on $\|\mathbf{F}\|$.

Lemma 1. *Let $\mathcal{A}_1, \mathcal{A}_2, \mathcal{M}$, and $\mathcal{N}$ be sub-σ-algebras on the same probability space. Assume $\mathcal{M} \subset \mathcal{N}$. Then*

$$\mathcal{R}_{\mathcal{M}}(\mathcal{A}_1, \mathcal{A}_2) \le \mathcal{R}_{\mathcal{N}}(\mathcal{A}_1, \mathcal{A}_2) + [1 - \mathcal{R}_{\mathcal{N}}(\mathcal{A}_1, \mathcal{A}_2)]\mathcal{R}_{\mathcal{M}}(\mathcal{A}_1, \mathcal{N})\mathcal{R}_{\mathcal{M}}(\mathcal{A}_2, \mathcal{N}). \tag{5.5}$$

A special case for random variables X, Y, Z is

$$\mathcal{R}(X, Z) \le \mathcal{R}_Y(X, Z) + (1 - \mathcal{R}_Y(X, Z))\mathcal{R}(X, Y)\mathcal{R}(Z, Y).$$

The inequality is sharp; for example, equality holds when (X, Y, Z) follows a trivariate normal with a common correlation. This generalizes lemma 1 of the article by Yu (2008), which states that

$$\mathcal{R}(X, Z) \le \mathcal{R}(X, Y)\mathcal{R}(Y, Z) \tag{5.6}$$

if X and Z are conditionally independent given Y. Since we are unable to locate a proof in the literature for Lemma 1 (nor the notion of MPC), we provide one in Appendix B.

Because it is easier to work with the MCC (and more generally MPC), and because the MCC provides a conservative measure of convergence (i.e., for the MCC to be smaller than a certain threshold, the worst autocorrelation has to be below that threshold), we define the *minimal speed* (minS) of an MCMC sequence (more precisely, any Markov chain with stationary transition probabilities) $U = \{U^{(t)}, t = 1, 2, \ldots\}$ as

$$\mathrm{minS}_U \equiv \mathcal{S}\big(U^{(1)}, U^{(2)}\big) = 1 - \mathcal{R}\big(U^{(1)}, U^{(2)}\big). \tag{5.7}$$

Because of (5.2), minS is a lower bound of the actual speed, which may be defined as $S_U = 1 - \mathrm{spec}(\mathbf{F})$; the bound is attained for a time-reversible chain. This definition of speed can be appreciated through the so-called mixing time, defined as $\tau = -[\log \gamma]^{-1} \approx (1 - \gamma)^{-1}$, where the approximation holds when $\gamma \equiv \mathrm{spec}(\mathbf{F})$ is close to 1, which is the case where most efforts are needed, as emphasized by Papaspiliopoulos, Roberts, and Skold (2007). It is then natural to view $\tau^{-1} \approx 1 - \gamma = S_U$ as a measure of speed, because it is inversely proportional to time. A deeper reason for us to use this notion is that component-wise interweaving exerts its impact in a multiplicative fashion on the (minimal) speed scale, not on the γ or $\mathcal{R}$ scale, as we shall show in the next section. This resembles results on the speed of EM and ECM algorithms; see the work of Meng (1994) and Meng and Rubin (1993). This "multiplicative effect" is already hinted by the minS version of (5.5):

$$\begin{aligned} \mathcal{S}_{\mathcal{M}}(\mathcal{A}_1, \mathcal{A}_2) \geq \mathcal{S}_{\mathcal{N}}(\mathcal{A}_1, \mathcal{A}_2)[&\mathcal{S}_{\mathcal{M}}(\mathcal{A}_1, \mathcal{N}) + \mathcal{S}_{\mathcal{M}}(\mathcal{A}_2, \mathcal{N}) \\ &- \mathcal{S}_{\mathcal{M}}(\mathcal{A}_1, \mathcal{N})\mathcal{S}_{\mathcal{M}}(\mathcal{A}_2, \mathcal{N})], \end{aligned} \tag{5.8}$$

where all $\mathcal{S}$ quantities are one minus their $\mathcal{R}$ counterparts.

5.2 GENERAL RESULTS ON THE ROBUSTNESS OF THE INTERWEAVING STRATEGY

Theorem 1 (Section 2) establishes a robustness property of GIS, and provides the key insight about how it boosts MCMC efficiency, as detailed in Section 2.3. Whereas Theorem 1 is simple, it is less useful when $\mathcal{R}(Y_{mis,1}, Y_{mis,2}) = 1$. This becomes a concern when the dimension of the missing data is higher than that of the parameter, which is typical in practice. Then there may exist a function of $Y_{mis,1}$ that coincides with another function of $Y_{mis,2}$, which implies $\mathcal{R}(Y_{mis,1}, Y_{mis,2}) = 1$. Hence we study the MPC of $Y_{mis,1}$ and $Y_{mis,2}$, after regressing out their common component. Theorem 2, proved in Appendix B, formalizes this idea.

Theorem 2. *Given a posterior distribution $p(\theta|Y_{obs})$ of interest, suppose we have two augmentation schemes $Y_{mis,1}$ and $Y_{mis,2}$ such that their joint distribution is well defined conditional on θ and Y_{obs}. Let $\mathcal{N} = \sigma(Y_{mis,1}) \cap \sigma(Y_{mis,2})$, that is, the intersection of the σ-algebras generated by $Y_{mis,1}$ and $Y_{mis,2}$ in the joint posterior of $(\theta, Y_{mis,1}, Y_{mis,2})$. Then $r_{1\&2}$, the geometric rate of convergence of GIS, satisfies*

$$\begin{aligned} r_{1\&2} \leq\ &\mathcal{R}^2(\theta, \mathcal{N}) \\ &+ (1 - \mathcal{R}^2(\theta, \mathcal{N}))\mathcal{R}_{\mathcal{N}}(\theta, Y_{mis,1})\mathcal{R}_{\mathcal{N}}(Y_{mis,1}, Y_{mis,2})\mathcal{R}_{\mathcal{N}}(\theta, Y_{mis,2}). \end{aligned} \tag{5.9}$$

Theorem 2, though more difficult to digest than Theorem 1, is nevertheless interpretable. When $\mathcal{N}$ is trivial, that is, when $Y_{mis,1}$ and $Y_{mis,2}$ have no common component, Theorem 2 reduces to Theorem 1. (In general it is unclear how the bounds given by Theorems 1 and 2 compare.) Intuitively, (5.9) says that the interweaving strategy is particularly effective when (i) $\mathcal{R}(\theta, \mathcal{N})$ is small, that is, the common component between $Y_{mis,1}$ and $Y_{mis,2}$ untouched by GIS does not depend heavily on θ, and (ii) either $\mathcal{R}_{\mathcal{N}}(Y_{mis,1}, Y_{mis,2})$ is small, that is, the two DAs $Y_{mis,1}$ and $Y_{mis,2}$ are not heavily dependent after taking out their common component, or one of $\mathcal{R}_{\mathcal{N}}(\theta, Y_{mis,i})$, $i = 1, 2$, is small, that is, one of the original DA schemes converges fast (after taking out the common component).

It is worth emphasizing that the bound in (5.9) never exceeds 1 and can reach 1 if and only if at least one of the following holds:

R1. $\mathcal{R}(\theta, \mathcal{N}) = 1$;

R2. $\mathcal{R}_{\mathcal{N}}(\theta, Y_{mis,1}) = \mathcal{R}_{\mathcal{N}}(Y_{mis,1}, Y_{mis,2}) = \mathcal{R}_{\mathcal{N}}(\theta, Y_{mis,2}) = 1$.

Because R1 and R2 are very strong conditions, especially when $Y_{mis,1}$ and $Y_{mis,2}$ form a "beauty-and-beast" pair, we conjecture that under mild conditions GIS or at least ASIS will be geometrically ergodic, regardless of the individual convergence behavior of the two chains being interwoven.

For CIS, we can apply Lemma 1 to each of the components conditional on the rest. To make this statement precise, we need the notion of *minimum partial speed*. Following the notation in Section 2.5, let σ_j, $j = 0, \dots, J$, denote the sub-σ-algebra generated by $\theta^{(t+j/J)}$ in the *joint space* of $\{\theta^{(t)}, \theta^{(t+1)}\}$, where $\theta^{(t+j/J)}$, defined in (2.27), is the output of the chain right after its jth component is updated from $\theta_j^{(t)}$ to $\theta_j^{(t+1)}$. The iteration index t is immaterial because our calculations below assume the chain has reached stationarity. The *minimum partial speed* for the jth component is defined as

$$\mathcal{S}_j = 1 - \mathcal{R}_{\sigma_{j-1}\cap\sigma_j}(\sigma_{j-1}, \sigma_j). \tag{5.10}$$

The difference between σ_{j-1} and σ_j is only in the jth component. In this sense, $\mathcal{S}_j$ measures the minimum speed for the jth component because it takes out the impact of $\sigma_{j-1} \cap \sigma_j$.

Given a Gibbs sampler involving Y_{mis} and multiple components of θ, we can, at least in theory, integrate out all the missing data Y_{mis} (be they a single set or multiple sets introduced by interweaving), resulting in a Gibbs sampler alternating between $p(\theta_j|\theta_{\neq j}; Y_{obs})$, $j = 1, \dots, J$. In the terminology of Liu, Wong, and Kong (1994, 1995), this is a *collapsed* algorithm. Let $\mathcal{S}_G$ be the minimum speed of this algorithm; if there is only one component, then $\mathcal{S}_G = 1$, since the collapsed version then produces i.i.d. draws from the target $p(\theta|Y_{obs})$. In a sense, the CIS aims to make its minimal speed as close to $\mathcal{S}_G$ as possible, just as the GIS strives to reach $\mathcal{S}_G = 1$, that is, to produce i.i.d. draws. The following theorem is our best result to support this intuition. Its proof, which uses (5.8) repeatedly, is given in Appendix B.

Theorem 3. *Suppose the target density is $\pi = p(\theta_1, \dots, \theta_J|Y_{obs})$. Let $\mathcal{S}_{CIS}$ be the minimal speed of CIS as defined in Section 2.5, let $\mathcal{S}_j$ be the minimal partial speed for the jth component as defined in (5.10), and let $\mathcal{S}_G$ be the minimal speed of the Gibbs sampler that samples from the J full conditionals $p(\theta_j|\theta_{\neq j}; Y_{obs})$, $j = 1, \dots, J$, in the same order as in CIS. Then*

$$\mathcal{S}_{CIS} \geq \left(\prod_{j=1}^{J} \mathcal{S}_j\right) \tilde{\mathcal{S}}_G, \tag{5.11}$$

where

$$\tilde{\mathcal{S}}_G = \prod_{j=1}^{J-1} \mathcal{S}_{\sigma_{j-1}\cap\sigma_J}(\sigma_{j-1} \cap \sigma_j,\ \sigma_j \cap \sigma_J) \tag{5.12}$$

is a lower bound on $\mathcal{S}_G$ and is completely determined by π, and it equals $\mathcal{S}_G$ when $J = 2$.

The right side of (5.11) is in an appealing product form. The term $\prod_{j=1}^{J} \mathcal{S}_j$ may be viewed as the "within-component minimal speed," a measure of the effectiveness of interweaving. In particular, if all $\mathcal{S}_j = 1$, then at iteration t, the jth interweaving substep is such that $\theta_j^{(t)}$ and $\theta_j^{(t+1)}$ are conditionally independent given other components of θ (and Y_{obs}). In other words, CIS reduces to collapsed Gibbs, with the missing data integrated out. The term $\tilde{\mathcal{S}}_G$ is a natural lower bound for the minimal speed of this collapsed Gibbs sampler, and, in the case of two components, the bound is attained. Hence $\tilde{\mathcal{S}}_G$ can be viewed as (a lower bound of) a measure of the "between-component minimal speed."

We remark that the bound in (5.11) is sharp in the sense that it is achievable in certain nontrivial cases, for example, when $J = 2$ and $\{Y_{mis}, \theta_1, \theta_2\}$ is a trivariate normal with a common correlation. We, however, have no theoretical result to rule out $\mathcal{S}_{CIS} > \mathcal{S}_G$. That is, we do not preclude the possibility that CIS can actually outperform the collapsed sampler; this is analogous to the phenomenon that ECM can outperform EM (Meng 1994).

The multiplicative form of $\prod_{j=1}^J \mathcal{S}_j$ suggests that we should make $\mathcal{S}_j$ as close to 1 as possible, separately for each j. That is, when updating $\theta_j^{(t)}$ to $\theta_j^{(t+1)}$, it is desirable to make $Y_{mis,j}$ and $\tilde{Y}_{mis,j}$ as independent as possible (conditioning on $\theta_{\neq j}$). But this is the same strategy as global interweaving, except that it is applied to each component in turn, lending us greater flexibility.

Theorem 3 is also intimately connected with the following bound on $\mathcal{S}_G$, which comes from the theory of alternating projections (see Deutsh 2001 or Diaconis, Khare, and Saloff-Coste 2007) but is rephrased in our terms:

$$\mathcal{S}_G \ge 1 - \left(1 - \prod_{j=1}^{J-1}\left[1 - \mathcal{R}^2_{\sigma_{j-1}\cap\sigma_J}(\sigma_{j-1}\cap\sigma_j, \sigma_j\cap\sigma_J)\right]\right)^{1/2}. \tag{5.13}$$

Although the right side of (5.13) coincides with $\tilde{\mathcal{S}}_G$ for $J = 2$, it is sharper for $J \ge 3$ because of the inequality

$$1 - \left[1 - \prod_j (1 - c_j^2)\right]^{1/2} \ge \prod_j (1 - c_j), \qquad \forall c_j \in [0, 1]. \tag{5.14}$$

However, we are unable to prove a version of (5.11) with the lower bound in (5.13) in place of $\tilde{\mathcal{S}}_G$; perhaps this is the price for the *product* form in (5.11). Since CIS is in the form of alternating projections, we can also apply the generic projection inequality underlying (5.13). The end result, however, turns out to be rather messy and does not allow us to separate the "within-component speed," as represented by $\prod_{j=1}^J \mathcal{S}_j$, and the "between-component speed," as represented by $\tilde{\mathcal{S}}_G$ or more ideally by $\mathcal{S}_G$. We therefore settle for (5.11), given the clearer theoretical insight it provides. There may well be room for improvement for (5.11), and we look forward to such results from domain experts. Here we simply point out that Theorem 3 actually is not restricted to CIS, but is applicable to any MCMC for updating $\theta^{(t)}$ to $\theta^{(t+1)}$ such that $\theta_{\le j}^{(t+1)}$ is independent of $\theta_{<j}^{(t)}$ when *conditional* on $\theta^{(t+(j-1)/J)}$, for all j (see Appendix B).

5.3 AN OPTIMALITY RESULT FOR THE GLOBAL ASIS

The next result says that not only is ASIS robust, under certain conditions, it also produces the *optimal* algorithm among a broad class of DA schemes. Specifically, in Theorem 4, we establish this optimality result by drawing a correspondence between (global)

ASIS and the optimal DA algorithm obtained under the working/expanded parameter approach (Liu and Wu 1999; Meng and van Dyk 1999; van Dyk and Meng 2001; Hobert and Marchev 2008), even though the interweaving strategy does not involve any working or expanded parameter.

To establish this correspondence, let us first review the working or expanded parameter approach already mentioned in Section 4.1. Suppose the original DA model is $p(Y_{mis}, Y_{obs}|\theta)$, and we introduce a working parameter α so that the augmented model is $p(Y_{mis}^{\alpha}, Y_{obs}|\theta, \alpha)p(\alpha)$, where $p(\alpha)$ is our *working prior*. (The notation for the missing data is switched from Y_{mis} to Y_{mis}^{α} to highlight the dependence on α.) Any choices of Y_{mis}^{α} and $p(\alpha)$ are legitimate as long as the marginal model $p(Y_{obs}|\theta)$ is preserved. Then we can implement the standard Gibbs sampler on the expanded space, yielding the so-called PX-DA algorithm:

1. Draw $(\alpha, Y_{mis}^{\alpha})|(\theta, Y_{obs})$.
2. Draw $(\alpha, \theta)|(Y_{mis}^{\alpha}, Y_{obs})$.

An intriguing result is that the optimal choice of $p(\alpha)$ typically is an improper prior, in which case the PX-DA chain itself is not positive recurrent, but the sub-chain on θ is positive recurrent, with the fastest convergence rate among a broad class of DA chains. (See Meyn and Tweedie (1993) for a definition of positive recurrence.) Such results have generated some general interests (e.g., Hobert 2001a, 2001b; Lavine 2003; Marchev and Hobert 2004; Hobert and Marchev 2008). In particular, using group-theoretic arguments, Liu and Wu (1999) showed that under mild conditions the Haar measure (typically improper) is the prior that results in the fastest convergence. This is the result which our Theorem 4 will link to. We emphasize that the link is established when the working parameter used in PX-DA corresponds to the map between SA and AA in the way as defined below. This is the reason that the PX-DA and ASIS in Section 4.1 do not coincide with each other because there the working parameter is introduced in (4.5) as a scale parameter, whereas the map from SA to AA is formed by subtracting a location parameter, as in (4.8).

Let us assume that the parameter space Θ is a locally compact Euclidean space. Moreover, suppose $G_{\Theta} \equiv \{M_{\theta}, \theta \in \Theta\}$ is a collection of one-to-one mappings that forms a group (the group operator being the composition of mappings). Let Θ be endowed with the same group structure as G_{Θ}, that is, its operator "$\cdot$" and inverse are specified by $M_{\theta \cdot \theta'} = M_{\theta}(M_{\theta'})$ and $M_{\theta^{-1}} = M_{\theta}^{-1}$, respectively. Assume that both the operator $(\theta, \theta') \to \theta \cdot \theta'$ and the inverse $\theta \to \theta^{-1}$ are continuous functions, that is, Θ is a topological group. A *right Haar measure* $H(d\theta)$ on Θ is a measure that is invariant under group acting on the right, that is, for any $\theta_0 \in \Theta$ and any measurable set $B \subset \Theta$,

$$H(B) = \int_{B} H(d\theta) = \int_{B\theta_0} H(d\theta) = H(B\theta_0),$$

where $B\theta_0 = \{\theta \cdot \theta_0 : \theta \in B\}$, that is, the set obtained by "multiplying" θ_0 from the right to every element in B. A *left Haar measure* is defined similarly. If the right Haar measure $H(d\theta)$ is also the left Haar measure, then we say G_{Θ} is *unimodular*. For example, in the toy model in Section 2, we have $M(Y_{mis}; \theta) = Y_{mis} - \theta$, and hence G_{Θ} is the additive group on the real line, with Lebesgue measure being its unimodular Haar measure.

Theorem 4. *For a given posterior distribution of interest, $p(\theta|Y_{obs}) \propto p(Y_{obs}|\theta)p_0(\theta)$, suppose we have an SA Y_{mis} and an AA $\tilde{Y}_{mis}$ linked by a one-to-one and continuously differentiable transformation $\tilde{Y}_{mis} = M(Y_{mis};\theta) \equiv M_\theta(Y_{mis})$ (for fixed θ). In addition, assume that*

C1. *Θ forms a group (as induced by the mappings M_θ) with a unimodular Haar measure;*

C2. *the model prior density $p_0(\theta)$ with respect to the Haar measure satisfies the condition $p_0(\theta \cdot \theta') \propto p_0(\theta)p_0(\theta')$, where "·" is the group operator.*

Then the ASIS algorithm is identical to the optimal PX-DA algorithm, that is, PX-DA with the Haar measure prior, for the expanded model $(\theta, \alpha, Y^{\alpha}_{mis}, Y_{obs})$, where α is the working parameter and $Y^{\alpha}_{mis} = M_\alpha(Y_{mis})$, that is, $M(Y_{mis};\theta)$ with θ replaced by α.

The proof of this theorem is given in Appendix B. The only truly restrictive condition in Theorem 4 is Condition C2, because it concerns the model prior. Currently we can neither explain intuitively why this condition is needed nor know whether it can be relaxed. The former is not completely unexpected, since in Section 2.4 we have shown that whether we can minimize $\mathcal{R}_{1,2}$ depends on the form of the prior. It is also not as restrictive as one might first believe, especially for Bayesian inference under noninformative priors. For example, it is always satisfied by the constant prior $p_0(\theta) \propto 1$, or equivalently the Haar measure itself (assuming the propriety of the resulting posterior, of course). This applies to, for example, all likelihood computation via DA. But the Haar measure is not the only measure that satisfies Condition C2. If M_θ is a scale group $M(Y_{mis};\theta) = Y_{mis}/\theta$ (the operator of the group is the usual multiplication), then priors such as $p_0(\theta) \propto \theta^\delta$ for any constant δ also satisfy Condition C2. This is indeed the class of "locally invariant scale priors" suggested by Berger (1985) as desirable noninformative priors (see also Meng and Zaslavsky 2002).

These connections cry out for a"missing link": why are invariant priors that are good for "objective" Bayesian inference also good for fast MCMC? We do not believe this is a coincidence, but rather a reflection of a deep connection between the inference model and the computational model, perhaps along the lines of fiducial inference or structural inference (see Liu and Wu 1999; Hannig, Iyer, and Patterson 2006; Hannig 2009). Additional open problems are discussed below.

6. LIMITATIONS AND OPEN PROBLEMS

Whenever one is already able to implement AA and SA, there is little reason not to try ASIS, global or component-wise. That one of AA and SA is particularly slow is not a reason for not using it; on the contrary, it is this very beauty-and-beast discordance that renders much of the power of ASIS. However, if such a discordance is still not enough to overcome the "stickiness" of the chain, then we may need more advanced interweaving strategies.

A possibility is the following "nested ASIS." Consider a model with parameter θ, two layers of latent variables X and Y, and observed data Z such that $\theta \to X \to Y \to Z$ are

Markov dependent. If we treat (θ, X) as parameters, then Y is an SA for (θ, X); we may construct an AA, $\tilde{Y}$ for (θ, X), and implement an interweaving strategy. Part of this strategy will involve drawing (θ, X) given Y and drawing (θ, X) given $(\tilde{Y}, Z)$. If drawing (θ, X) jointly given Y is infeasible, then, as an alternative to CIS described earlier, we may consider a *nested interweaving strategy* for θ, with Y now playing the role of observed data. Noting that X is an SA for θ for this subproblem, we just need to construct an AA, $\tilde{X}$ for θ, and use one iteration of interweaving for this subproblem. Similarly, nested interweaving may be implemented for drawing (θ, X) given $(\tilde{Y}, Z)$. It can be shown that, for the multilayer extension of the toy model, that is,

$$\begin{aligned} &p(\theta) \propto 1, \\ &X|\theta \sim N(\theta, \sigma_1^2), \qquad Y|(\theta, X) \sim N(X, \sigma_2^2), \qquad Z|(\theta, X, Y) \sim N(Y, \sigma_3^2), \end{aligned}$$

a well-constructed nested interweaving strategy also leads to convergence in one step, that is, i.i.d. draws. The usefulness of nested ASIS for realistic models seems worth investigating.

On the theoretical side, an important open problem for GIS is whether an AA–SA pair $\{Y_{mis,A}, Y_{mis,S}\}$ minimizes $\mathcal{R}_{\mathcal{N}}(Y_{mis,1}, Y_{mis,2})$ among a suitable class of DA pairs $\{Y_{mis,1}, Y_{mis,2}\}$, where $\mathcal{N} = \sigma(Y_{mis,1}) \cap \sigma(Y_{mis,2})$. All evidence we have so far, empirical and theoretical (e.g., Theorem 4), supports this, but we surmise nontrivial conditions are needed to make it a mathematical theorem. A more challenging problem is to extend such results to CIS, for which we do not even have a result that parallels Theorem 4. Furthermore, for CIS, we yet need to determine if it is possible to replace the lower bound $\tilde{\mathcal{S}}_G$ in (5.11) by the actual Gibbs sampler speed $\mathcal{S}_G$.

Another theoretical problem is to either relax Condition C2 in Theorem 4, or to identify two schemes to be interwoven that could lead to an optimal algorithm without Condition C2. Despite the restrictiveness of Theorem 4, it provides a great starting point for this exploration, which ultimately may lead to a Bayesian counterpart of Basu's theorem. As discussed in Section 2.4, we may need to incorporate prior information into SA and AA in some way, in order to decrease the impact of the "non-factoring" part of (2.22). Even when this is not feasible and C2 fails, the ASIS sampler can still be quite effective, or at least faster than both of the original algorithms, as all our current evidence suggests. Moreover, Theorem 4 reveals an interesting relationship between reparameterization and parameter-expansion: the best parameter-expansion scheme sometimes corresponds to interweaving two special parameterizations: SA and AA. See Appendix C for an example that suggests possible extensions to Theorem 4.

Whereas much remains to be done, we hope we have unearthed a general strategy for efficient MCMC, guided by the classical concepts of sufficiency and ancillarity. Our strategy is analogous to boosting algorithms (e.g., Schapire 1990; Fruend 1995) in that interweaving a set of weaker learners/algorithms may lead to a much stronger one. We believe that there is a vast "efficiency via boosting" MCMC kingdom yet to be explored, and that our ASIS is only one of many "sesames" that can open its door.

SUPPLEMENTARY MATERIALS

Appendices: The file provides three appendices cited in the main article:

Appendix A: Auxiliary material for Section 3, which includes

A.1 Details of the MCMC Steps in the Poisson time series example.

A.2 Autocorrelations plots of the Monte Carlo draws (Figures A.1–A.3).

Appendix B: Auxiliary material for Section 5, which includes

B.1 A reducible chain as a result of combining two transition kernels (Figure B.1).

B.2 Proof of Lemma 1.

B.3–B.6 Proof of Theorems 1–4.

Appendix C: Auxiliary material for Section 6, which includes an example to illustrate both the relevance and the limitations of Theorem 4. (interweave_appendix.pdf)

ACKNOWLEDGMENTS

We thank NSF for partial financial support, and Kung-Sik Chan, Radu Craiu, David van Dyk, Vinay Kashyap, Donald Rubin, and eight referees for very helpful comments and suggestions on earlier versions. During revisions, comments from Gareth Roberts (who particularly suggested the bivariate example in Section 2.3) and his group, especially during the MCMC workshop at University of Warwick (March, 2009), have been very useful, to which we are grateful. We also thank Paul Baines for proofreading and for suggestions.

[Received January 2010. Revised July 2010.]

REFERENCES

Albert, J., and Chib, S. (1993), “Bayesian Analysis of Binary and Polychotomous Response Data,” *Journal of the American Statistical Association*, 88, 669–679. [551,552]

Amit, Y. (1991), “On Rates of Convergence of Stochastic Relaxation for Gaussian and Non-Gaussian Distributions,” *Journal of Multivariate Analysis*, 38, 82–99. [559]

Amit, Y., and Grenander, U. (1991), “Comparing Sweep Strategies for Stochastic Relaxation,” *Journal of Multivariate Analysis*, 37, 197–222. [542]

Basu, D. (1955), “On Statistics Independent of a Complete Sufficient Statistic,” *Sankhya, Ser. A*, 15, 377–380. [533,540]

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis* (2nd ed.), New York: Springer. [565]

Beskos, A., Papaspiliopoulos, O., Roberts, G. O., and Fearnhead, P. (2006), “Exact and Computationally Efficient Likelihood-Based Estimation for Discretely Observed Diffusion Processes,” *Journal of the Royal Statistical Society, Ser. B*, 68, 1–29. [536]

Chan, K. S., and Ledolter, J. (1995), “Monte Carlo EM Estimation for Time Series Models Involving Counts,” *Journal of the American Statistical Association*, 90, 242–252. [545]

Cox, D. R. (1981), “Statistical Analysis of Time Series: Some Recent Developments,” *Scandinavian Journal of Statistics*, 8, 93–115. [533,545]

Craiu, R., and Meng, X.-L. (2005), “Multi-Process Parallel Antithetic Coupling for Forward and Backward Markov Chain Monte Carlo,” *The Annals of Statistics*, 33, 661–697. [539]

Davis, R. A., and Rodriguez-Yam, G. (2005), "Estimation for State-Space Models Based on a Likelihood Approximation," *Statistica Sinica*, 15, 381–406. [545]

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood Estimation From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 39, 1–38. [532,534,543]

Deutsch, F. (2001), *Best Approximation in Inner Product Spaces*, New York: Springer-Verlag. [563]

Diaconis, P., Khare, K., and Saloff-Coste, L. (2007), "Stochastic Alternating Projections," technical report, Dept. of Statistics, Stanford University. [563]

Freund, Y. (1995), "Boosting a Weak Learning Algorithm by Majority," *Information and Computation*, 121, 256–285. [566]

Frühwirth-Schnatter, S., and Wagner, H. (2006), "Gibbs Sampling for Parameter-Driven Models of Time Series of Small Counts With Application to State Space Modelling," *Biometrika*, 93, 827–841. [545]

Gelfand, A. E., and Carlin, B. P. (1995), Comment on "Bayesian Computation and Stochastic Systems," by J. Besag, P. Green, D. Higdon, and K. Mengersen, *Statistical Science*, 10, 43–46. [532]

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409. [532]

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995), "Efficient Parameterisations for Normal Linear Mixed Models," *Biometrika*, 82, 479–488. [532,533,535,540]

——— (1996), "Efficient Parametrizations for Generalized Linear Mixed Models," in *Bayesian Statistics 5*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 165–180. [532,533,540]

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis* (2nd ed.), London: CRC Press. [532]

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741. [532]

Hamada, M., and Wu, C. F. J. (1995), "Analysis of Censored Data From Fractionated Experiments: A Bayesian Approach," *Journal of the American Statistical Association*, 90, 467–477. [554,555]

Hannig, J. (2009), "On Generalized Fiducial Inference," *Statistica Sinica*, 19, 491–544. [565]

Hannig, J., Iyer, H. K., and Patterson, P. (2006), "Fiducial Generalized Confidence Intervals," *Journal of the American Statistical Association*, 101, 254–269. [565]

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97–109. [532]

Hills, S. E., and Smith, A. F. M. (1992), "Parameterization Issues in Bayesian Inference," in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 227–246. [532]

Hobert, J. P. (2001a), Discussion of "The Art of Data Augmentation," by D. van Dyk and X.-L. Meng, *Journal of Computational and Graphical Statistics*, 10, 59–68. [564]

——— (2001b), "Stability Relationships Among the Gibbs Sampler and Its Subchains," *Journal of Computational and Graphical Statistics*, 10, 185–205. [564]

Hobert, J. P., and Marchev, D. (2008), "A Theoretical Comparison of the Data Augmentation, Marginal Augmentation and PX-DA Algorithms," *The Annals of Statistics*, 36, 532–554. [538,539,564]

Kypraios, T. (2007), "Efficient Bayesian Inference for Partially Observed Stochastic Epidemics and a New Class of Semi-Parametric Time Series Models," Ph.D. thesis, Dept. of Mathematics and Statistics, Lancaster University. [543]

Lavine, M. (2003), "A Marginal Ergodic Theorem," in *Bayesian Statistics 7*, eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, Oxford, U.K.: Oxford University Press, pp. 577–585. [564]

Liu, C. H. (2003), "Alternating Subspace-Spanning Resampling to Accelerate Markov Chain Monte Carlo Simulation," *Journal of the American Statistical Association*, 98, 110–117. [538,554,555]

Liu, C. H., and Rubin, D. B. (1996), “Markov-Normal Analysis of Iterative Simulations Before Their Convergence,” *Journal of Econometrics*, 75, 69–78. [557]

Liu, C. H., and Sun, D. X. (2000), “Analysis of Interval-Censored Data From Fractionated Experiments Using Covariance Adjustments,” *Technometrics*, 42, 353–365. [554]

Liu, C. H., Rubin, D. B., and Wu, Y. N. (1998), “Parameter Expansion to Accelerate EM—The PX-EM Algorithm,” *Biometrika*, 85, 755–770. [552]

Liu, J. S. (1994), “Fraction of Missing Information and Convergence Rate of Data Augmentation,” in *Computing Science and Statistics: Proceedings of the 26th Symposium on the Interface*, Fairfax Station, VA: Interface Foundation of North America, pp. 490–496. [534]

Liu, J. S., and Wu, Y. N. (1999), “Parameter Expansion for Data Augmentation,” *Journal of the American Statistical Association*, 94, 1264–1274. [533,534,543,552,564,565]

Liu, J. S., Wong, W. H., and Kong, A. (1994), “Covariance Structure of the Gibbs Sampler With Applications to Comparisons of Estimators and Augmentation Schemes,” *Biometrika*, 81, 27–40. [535,559,562]

——— (1995), “Correlation Structure and Convergence Rate of the Gibbs Sampler for Various Scans,” *Journal of the Royal Statistical Society, Ser. B*, 57, 157–169. [535,559,562]

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, 228–238. [564]

Meng, X.-L. (1994), “On the Rate of Convergence of the ECM Algorithm,” *The Annals of Statistics*, 22, 326–339. [534,561,563]

Meng, X.-L., and Rubin, D. B. (1991), “Using EM to Obtain Asymptotic Variance–Covariance Matrices: The SEM Algorithm,” *Journal of the American Statistical Association*, 86, 899–909. [534]

——— (1993), “Maximum Likelihood Estimation via the ECM Algorithm: A General Framework,” *Biometrika*, 80, 267–278. [544,557,561]

Meng, X.-L., and van Dyk, D. A. (1996), “Minimum Information Ratio and Relative Augmentation Function,” in *Proceedings of the Statistical Computing Section*, Alexandria, VA: American Statistical Association, pp. 73–78. [534]

——— (1997), “The EM Algorithm—An Old Folk Song Sung to a Fast New Tune” (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 59, 511–567.

——— (1998), “Fast EM Implementations for Mixed-Effects Models,” *Journal of the Royal Statistical Society, Ser. B*, 60, 559–578. [532,533]

——— (1999), “Seeking Efficient Data Augmentation Schemes via Conditional and Marginal Augmentation,” *Biometrika*, 86, 301–320. [532,543,552,564]

Meng, X.-L., and Zaslavsky, A. M. (2002), “Single Observation Unbiased Priors,” *The Annals of Statistics*, 30, 1345–1375. [565]

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), “Equation of State Calculations by Fast Computing Machines,” *Journal of Chemical Physics*, 21, 1087–1092. [532]

Meyn, S. P., and Tweedie, R. L. (1993), *Markov Chains and Stochastic Stability*, London: Springer-Verlag. [535, 564]

Michalak, S. (2001), “Multilevel Bernoulli Models for Evaluating Medical Departments in VA Hospitals,” Ph.D. thesis, Dept. of Statistics, Harvard University. [545]

Papaspiliopoulos, O., and Roberts, G. O. (2008), “Stability of the Gibbs Sampler for Bayesian Hierarchical Models,” *The Annals of Statistics*, 36, 95–117. [533,535,538]

Papaspiliopoulos, O., Roberts, G. O., and Skold, M. (2003), “Non-Centered Parameterisations for Hierarchical Models and Data Augmentation” in *Bayesian Statistics 7*, eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, Oxford, U.K.: Oxford University Press, pp. 307–326. [532,533,535,536,540]

——— (2007), “A General Framework for the Parametrization of Hierarchical Models,” *Statistical Science*, 22, 59–73. [532,533,535,536,540,543,544,560]

Roberts, G. O., and Sahu, S. K. (1997), “Updating Schemes, Correlation Structure, Blocking and Parameterisation for the Gibbs Sampler,” *Journal of the Royal Statistical Society, Ser. B*, 59, 291–317. [532,533]

Roberts, G. O., and Tweedie, R. L. (2001), "Geometric L^2 and L^1 Convergence Are Equivalent for Reversible Markov Chains," *Journal of Applied Probability*, 38A (Probability, Statistics and Seismology), 37–41. [535]

Rubin, D. B. (1976), "Inference and Missing Data" (with discussion), *Biometrika*, 63, 581–592. [554]

Schapire, R. E. (1990), "The Strength of Weak Learnability," *Machine Learning*, 5, 197–227. [566]

Smith, A. F. M., and Roberts, G. O. (1993), "Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society, Ser. B*, 55, 3–23. [532]

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–540. [532]

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions," *The Annals of Statistics*, 22, 1701–1727. [537]

van Dyk, D. A., and Meng, X.-L. (1997), "On the Orderings and Groupings of Conditional Maximizations Within ECM-Type Algorithms," *Journal of Computational and Graphical Statistics*, 6, 202–223. [542]

——— (2001), "The Art of Data Augmentation" (with discussion), *Journal of Computational and Graphical Statistics*, 10, 1–111. [532,534,543,552,564]

——— (2010), "Cross-Fertilizing Strategies for Better EM Mountain Climbing and DA Field Exploration: A Graphical Guide Book," *Statistical Science*, 25, 429–449. [534]

Wu, C. F. J. (1983), "On the Convergence Properties of the EM Algorithm," *The Annals of Statistics*, 11, 95–103. [532]

Yu, Y. (2005), "Three Contributions to Statistical Computing," Ph.D. thesis, Dept. of Statistics, Harvard University. [533,543]

——— (2008), "On the Maximal Correlation Coefficient," *Statistics & Probability Letters*, 78, 1072–1075. [560]

Yu, Y., and Meng, X.-L. (2007), "Espousing Classical Statistics With Modern Computation: Sufficiency, Ancillarity and an Interweaving Generation of MCMC," technical report, Dept. of Statistics, University of California, Irvine. [545]

Zeger, S. L. (1988), "A Regression Model for Time Series of Counts," *Biometrika*, 75, 621–629. [545]