

# Rejoinder: Be All Our Insomnia Remembered ...

Yaming Yu

Department of Statistics, University of California, Irvine

Xiao-Li Meng

Department of Statistics, Harvard University

**August 2, 2011**

## **1 Dream on: From DA to GIS to CIS**

### **1.1 Parallel Dreams**

The evolutionary history from DA to GIS and more generally to CIS may well be cited by a future Stephen Stigler to advance a new Stigler’s Law: “No scientific idea is originated from a single team.” Putting aside the well-known connection between EM and DA (see Tanner and Wong, 2010 and van Dyk and Meng, 2010), we have witnessed how the idea of introducing a non-identifiable parameter into DA schemes—for the purpose of better algorithmic efficiency—was independently and simultaneously developed by two research teams (Liu and Wu, 1999 and Meng and van Dyk, 1999). Subsequently, the idea of utilizing or combining multiple DA schemes has been pursued by (at least) three teams, from seemingly different angles. Roberts and Papaspiliopoulos’s team has been investigating the partially non-centered parametrization (see, e.g., Papaspiliopoulos, Roberts and Sköld, 2007), whose power and versatility are nicely illustrated by the discussion by Papaspiliopoulos, Roberts, and Sermaidis (PRS). The idea of partially non-centering is to introduce a tuning parameter  $w$  into the non-centering scheme (i.e., a DA scheme) and then seek its optimal value for the fastest convergence. It is therefore equivalent to the conditional augmentation approach (Meng and van Dyk, 1999 and van Dyk and Meng, 2001), where  $w$  is known as a *working parameter* and is determined by the same optimality criterion; often the optimality calculations are approximate because exact optimality is hard to achieve in practice.

This *conditional augmentation* approach—meaning the algorithm is conditional upon a fixed value of the working parameter—contrasts with the *marginal augmentation* approach (Meng and van Dyk, 1999), where the working parameter is marginalized out after being assigned a *working prior*. The marginal augmentation approach, also known as the parameter-expanded DA (PX-DA; Liu and Wu, 1999), has resulted in some intriguing findings. For example, a great theoretical result established by Liu and Wu (1999) is that, under certain regularity conditions, the optimal working prior is the (typically improper) right Haar measure. Consequently, the resulting Markov chain is typically non-positive recurrent on the joint space of the desired target and the working parameter, since the latter is not identifiable given the observed-data model. Yet, this non-positive recurrent Markov chain contains a properly converging sub-sequence (to be more precise, sub-sub-sequence; see Hobert, 2001a and 2001b). Not only is its stationary distribution our target distribution, it also has the fastest convergence rate among a general class of DA algorithms as defined by Liu and Wu (1999) via an elegant group theory formulation.

Such theoretically fascinating and practically useful algorithms have caused much insomnia for those of us who want to understand them fully, in order to make them as generally applicable as possible. Whereas the number of sandwiches Hobert’s team indulged during their insomnia can only be speculated, the theoretical insight provided by their “sandwiched” unification (Hobert and Marchev, 2008) is unquestionable. Indeed, when we first learned about the sandwich formulation, we were struggling to understand another intriguing phenomenon. That is, why can the typically simple ASIS/GIS perform as well as marginal augmentation, whose construction requires more sophistication comparatively (e.g., most improper working priors will lead to incorrect algorithms; see Meng and van Dyk, 1999)? However, as succinctly formulated in Hobert and Roman’s (HR) discussion, the sandwich algorithms themselves are “sandwiched” between DA and GIS, providing the missing link we sought in the big picture.

Our own work on ASIS/GIS started with the thesis of Yu (2005), where the interweaving strategy was invented to deal with a Chandra X-ray data set, as detailed in our main article. Soon, however, we realized that it is not merely a trick for one particular problem, but rather a general strategy for addressing the much debated question: to center or not to center. This realization was exciting, but our earlier report (Yu and Meng, 2007) did not have enough theoretical muscles, at least not the type routinely found in the work of Hobert’s team or Roberts’ team. Despite our

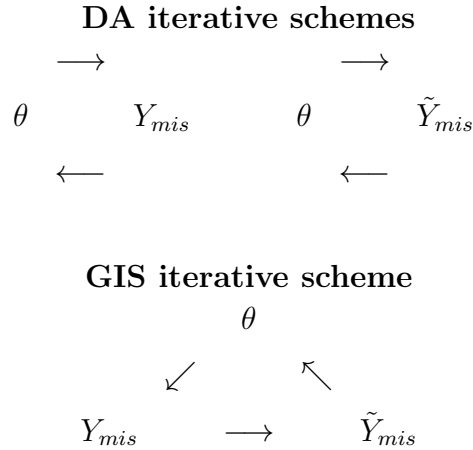


Figure 1: Comparing DA iterative schemes with a GIS iterative scheme.

suggestive empirical demonstrations, how can we be sure that they are not another demonstration of the wisdom that “simulations are doomed to be successful”?

## 1.2 Progressive Insights

The rejection of our initial paper (by another journal) led to our own insomnia — what is the correct theoretical explanation for the apparent, sometimes dramatic, empirical gains? After much searching, we had to laugh at ourselves because the answer was already in the proof of the old Theorem 1 of Yu and Meng (2007), where we conveniently bound the maximal correlation  $\mathcal{R}_{1,2}$  in the current Theorem 1 by its upper bound, that is, one. We thus “conveniently” deprived ourselves of the theoretical insights this maximal correlation can offer! In other words, the magic of interweaving comes from its built-in potential for “breaking the link” via conditional independence as measured by  $\mathcal{R}_{1,2}$  or its conditional variants (see Theorems 2 and 3 in our main article, as well as Romanovič, 1975 and Huang, 2010 for closely related definitions, which we came across after finishing our main article). This is perhaps best seen from Figure 1, which compares ordinary DA algorithms with GIS, using the generic notation of our main article.

As seen from Figure 1, each iteration of GIS cycles through the parameter  $\theta$  and the two sets of augmented data by first drawing  $Y_{mis}$  given  $\theta$ , then  $\tilde{Y}_{mis}$  given  $Y_{mis}$ , and then  $\theta$  given  $\tilde{Y}_{mis}$ . (Henceforth we suppress the conditioning on the observed data  $Y_{obs}$  when there is no confusion.) Hence two consecutive draws of  $\theta$  would be independent if  $Y_{mis}$  and  $\tilde{Y}_{mis}$  are conditionally inde-

pendent given  $Y_{obs}$  (or in HR’s notation, the targeted chain on  $X(= \theta)$  would produce i.i.d. draws, if  $Y$  and  $\tilde{Y}$  are independent). In general, the weaker the dependence between  $Y_{mis}$  and  $\tilde{Y}_{mis}$ , the better the efficiency we expect from GIS. HR’s formulation of our Theorem 1, by explicitly introducing a Markov transition function  $Q(Y_{mis}, \tilde{Y}_{mis})$ , makes this point ever clearer. In particular, in the original sandwich algorithms,  $(\theta, Y_{mis})$  and  $(\theta, \tilde{Y}_{mis})$  have the same joint distribution. Consequently, as HR pointed out, the  $Q$  function becomes their reversible  $R$  transition function, which defines a Markov chain on the  $Y_{mis}$  space (the same as the  $\tilde{Y}_{mis}$  space in this case) with  $\mathcal{R}_{1,2}$  being its convergence rate. Therefore, the key inequality in our Theorem 1, as nicely re-expressed by HR’s inequality (8), demonstrates that the faster this  $Y_{mis}$  chain converges, the more we can expect the sandwiched chain or more generally the GIS chain to converge. Note that we only say that “the more we can expect” because  $\mathcal{R}_{1,2}$  enters as part of an upper bound, rather than an exact formula, of the actual convergence rate.

The diagram for GIS in Figure 1 also illustrates another theoretical insight of HR’s that we “conveniently” overlooked. HR pointed out that although our Theorem 1 assumed a three-way joint distribution for  $\{Y_{mis}, \tilde{Y}_{mis}, \theta\}$  via the conditional distribution of  $(Y_{mis}, \tilde{Y}_{mis})$  given both  $\theta$  and  $Y_{obs}$ , our proof never used this assumption. What actually is needed are only the three *two-way* joint distributions, namely  $p(\theta, Y_{mis})$ ,  $p(\theta, \tilde{Y}_{mis})$ , and  $p(Y_{mis}, \tilde{Y}_{mis})$ . But they may not be compatible with each other, that is, there may not exist a three-way joint distribution on  $\{Y_{mis}, \tilde{Y}_{mis}, \theta\}$  whose three two-way margins are given by them. The GIS diagram in Figure 1 clearly shows that all moves only require specifications of the two-way distributions.

For mathematicians, unnecessary assumptions are often a tell-tail sign of incompetency. But as statisticians we wondered about other reasons for us to have overlooked this issue. The answer came when we asked ourselves whether the restrictive GIS class of algorithms, as defined by requiring the existence of a three-way joint distribution, forms a “complete class” within the more general class of algorithms as recognized by HR. That is, given the original DA schemes  $p(\theta, Y_{mis})$  and  $p(\theta, \tilde{Y}_{mis})$ , whether any algorithm in the general class is matched by one in the restrictive class in terms of convergence rate. The answer is negative, because in the general class, we can always make  $Y_{mis}$  and  $\tilde{Y}_{mis}$  independent by imposing  $p(Y_{mis}, \tilde{Y}_{mis}) = p(Y_{mis})p(\tilde{Y}_{mis})$ , where  $p(Y_{mis})$  and  $p(\tilde{Y}_{mis})$  are respectively the  $Y$ -margins of the original DA specifications  $p(\theta, Y_{mis})$  and  $p(\theta, \tilde{Y}_{mis})$ . Thus, the resulting GIS will produce i.i.d. draws for reasons discussed previously. In the restrictive

class, however, we do not have the freedom to arbitrarily specify the dependence between  $Y_{mis}$  and  $\tilde{Y}_{mis}$  because the dependence is determined by the three-way joint distribution  $p(Y_{mis}, \tilde{Y}_{mis}, \theta)$ , and hence the resulting GIS does not produce i.i.d. draws except in special cases such as the toy model in our main article.

The above observation might excite a casual reader — does this mean that, by using the unrestrictive GIS class, we can routinely produce i.i.d. draws? If true, that of course would be a paradise for everyone (with the possible exception of those who have invested their lives in developing MCMC algorithms). However, in order to implement GIS, we need to be able to make draws from the conditional distribution  $p(\tilde{Y}_{mis}|Y_{mis})$  (or from  $p(Y_{mis}|\tilde{Y}_{mis})$  by reversing the cycle in Figure 1). Therefore, by making  $\tilde{Y}_{mis}$  and  $Y_{mis}$  independent, we effectively have committed ourselves to drawing directly from  $p(\tilde{Y}_{mis})$ . But if we were able to do so, then we would not need any MCMC, because given the drawn  $\tilde{Y}_{mis}$  from  $p(\tilde{Y}_{mis})$ , we can directly obtain a desired sample of  $\theta$  by drawing from  $p(\theta|\tilde{Y}_{mis})$ , a step that is already required by the original DA algorithm.

As shown in our main article, for our restrictive GIS class, the sampling from  $p(\tilde{Y}_{mis}|Y_{mis})$  is accomplished by first sampling from  $p(\theta|Y_{mis})$ , which again is already required by the original DA algorithm. We then sample from  $p(\tilde{Y}_{mis}|\theta, Y_{mis})$ , which is often a trivial step (e.g., a deterministic evaluation), but nevertheless it requires a legitimate three-way specification  $p(\theta, Y_{mis}, \tilde{Y}_{mis})$ . In other words, the theoretically unnecessary three-way compatibility assumption for our Theorem 1 was actually a practical necessity for realizing the gain in efficiency of GIS in all the examples in our article, which perhaps explains (as an afterthought of course) why it did not occur to us to abandon the three-way distribution requirement in Theorem 1. HR’s insight on expanding our GIS class therefore suggests a new task: seeking a GIS class that does not require three-way compatibility, yet is still implementable. Part of the difficulty, of course, lies in an old thorny problem: how do we quantify implementability, which is problem-dependent?

## 2 Statistical Ambien for MCMC Insomnia

### 2.1 Two Ways to Break the Cycle

As discussed above, partially non-centering is equivalent to conditional augmentation, whereas the interweaving strategy, especially ASIS, has close ties with marginal augmentation. It has

long been noticed that, in some simple cases, both approaches can achieve i.i.d. algorithms, as demonstrated by PRS's and HR's normal examples, and yet they do not dominate each other in either the EM or MCMC context (see Meng and van Dyk 1997, 1999 and van Dyk and Meng, 2010). Part of our past insomnia stemmed from a desire to understand the connection between the two approaches, and to formulate practical guidelines. We are happy to report that the GIS diagram in Figure 1 as applied to PRS's/HR's example finally provides the Ambien we have been looking for.

Specifically, in both examples the authors started with a centered/sufficient augmentation,  $(\theta, Y_{mis})$ . They then created a partially non-centered augmentation by letting  $\tilde{Y}_{mis} = Y_{mis} - w\theta$ , to use PRS's notation (corresponding to HR's  $c = -w$ ). The question is how to choose  $w$ . Let us apply the GIS diagram in Figure 1 to the current setting. Given the original augmentation as represented by the arrow from  $\theta$  to  $Y_{mis}$ , we have two choices of  $w$  to break the cycle. The first is to make  $\tilde{Y}_{mis}$  independent of  $\theta$  and hence to break the  $\tilde{Y}_{mis} \rightarrow \theta$  link, which is what the partially non-centering or conditional augmentation approach aims to achieve. The second is to choose  $w$  such that  $\tilde{Y}_{mis}$  is independent of  $Y_{mis}$ , thereby breaking the  $Y_{mis} \rightarrow \tilde{Y}_{mis}$  link, which is what the marginal augmentation and ASIS (and more generally GIS/CIS) approaches try to accomplish. This is also nicely depicted by PRS's Figure 2, which conveys more geometric insight than the Figure 1 in our main article. The two classes of methods are therefore not directly comparable, since they are designed to break different links, although they share the ultimate goal.

This explains why in HR's example (in their Section 2.1) there are two choices of  $w = -c$  that lead to i.i.d. draws. If our aim is to break the  $\tilde{Y}_{mis} \rightarrow \theta$  link, that is, to reduce as much as possible the dependence between  $\tilde{Y}_{mis}$  and  $\theta$ , then an intuitive approach is to set  $\tilde{Y}_{mis}$  equal to the part of  $Y_{mis}$  not explained by  $\theta$ . That is, we should make  $\tilde{Y}_{mis} = Y_{mis} - w\theta$  (a function of) *the residual from regressing  $Y_{mis}$  on  $\theta$* , that is,

$$w = \frac{\text{Cov}(\theta, Y_{mis}|Y_{obs})}{\text{V}(\theta|Y_{obs})} = \frac{\text{Cov}(\theta, E[Y_{mis}|\theta, Y_{obs}]|Y_{obs})}{\text{V}(\theta|Y_{obs})}. \quad (2.1)$$

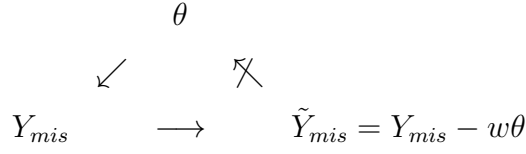
For HR's example (as well as for PRS's example), noticing

$$E[Y_{mis}|\theta, Y_{obs}] = \frac{1}{1+V}\theta + \frac{V}{1+V}Y_{obs},$$

we have from (2.1)

$$w = \frac{1}{1+V} \frac{\text{Cov}(\theta, \theta|Y_{obs})}{\text{V}(\theta|Y_{obs})} = \frac{1}{1+V}.$$

## Conditional Augmentation/Partially Non-centering



## Marginal Augmentation/GIS/ASIS

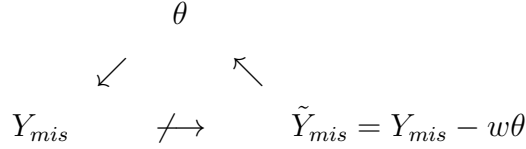


Figure 2: Two ways to reduce dependence. The  $\not\rightarrow$  operation indicates the potential for breaking the corresponding link. Note that conditional augmentation alternates between  $\theta$  and  $\tilde{Y}_{mis}$  only.

This result provides a theoretical insight into why the optimal partially non-centering scheme is given by  $w = (V + 1)^{-1}$  for PRS's example, and why  $c = -w = -(V + 1)^{-1}$  is the first value in HR's example to render i.i.d. draws. In the latter case, it is not because the interweaving strategy is effective (note that HR interweave  $\tilde{Y}_{mis}$  with  $Y_{mis}$  regardless of whether  $\tilde{Y}_{mis}$  is already optimal by itself); rather, one of the two DA schemes being interwoven,  $(\theta, \tilde{Y}_{mis})$ , already provides i.i.d. draws. Recall our Theorem 1 says, in HR's notation, that  $\|K_{YM}\| = 0$  whenever  $\|\tilde{K}_{DA}\| = 0$ .

On the other hand, if our aim is to break the  $Y_{mis} \rightarrow \tilde{Y}_{mis}$  link, then we need to make  $\tilde{Y}_{mis}$  (a function of) *the residual from regressing  $\theta$  on  $Y_{mis}$* . This can be achieved by noting that  $\tilde{Y}_{mis} = -w(\theta - w^{-1}Y_{mis})$ , and hence  $w^{-1}$  should be the regression coefficient from regressing  $\theta$  on  $Y_{mis}$ , that is,

$$w = \frac{V(Y_{mis}|Y_{obs})}{Cov(\theta, Y_{mis}|Y_{obs})} = \frac{V(Y_{mis}|Y_{obs})}{Cov(E[\theta|Y_{mis}, Y_{obs}], Y_{mis}|Y_{obs})}. \quad (2.2)$$

For HR's example,

$$E[\theta|Y_{mis}, Y_{obs}] = E[\theta|Y_{mis}] = \frac{A}{A+V}Y_{mis}, \quad (2.3)$$

where the first equality holds because  $Y_{mis}$  is a sufficient augmentation. We have from (2.2) that

$$w = \frac{V(Y_{mis}|Y_{obs})}{\frac{A}{A+V}Cov(Y_{mis}, Y_{mis}|Y_{obs})} = 1 + \frac{V}{A},$$

which yields the second  $c = -w = -V/A - 1$  reported by HR that leads to i.i.d. draws.

## 2.2 Residual Augmentation

Although the exact results in the last section are harder to obtain without joint normality, they suggest some general guidelines. The most important insight is that the choice of  $\tilde{Y}_{mis}$  should form a *residual* in the posterior space  $p(Y_{mis}, \theta | Y_{obs})$  (for non-Bayesian computations, we follow HR's more general notation  $f(X, Y)$ ). What type of residuals to use depends on whether we plan to use partially non-centering (i.e., conditional augmentation) or the interweaving strategy. In the former case we should consider

$$\tilde{Y}_{mis} = Y_{mis} - E[Y_{mis} | \theta, Y_{obs}], \quad (2.4)$$

whereas in the latter case we should aim for

$$\tilde{Y}_{mis} = \theta - E[\theta | Y_{mis}, Y_{obs}]. \quad (2.5)$$

A couple of remarks are in order. First, both (2.4) and (2.5) only involve conditional mean calculations with respect to the two original DA full conditionals, namely  $p(Y_{mis} | \theta, Y_{obs})$  and  $p(\theta | Y_{mis}, Y_{obs})$ . Therefore, at least in principle, both can be calculated or approximated analytically. This, however, is not enough because in order to implement the resulting algorithm, we need to be able to sample from  $p(\theta | \tilde{Y}_{mis}, Y_{obs})$  and from  $p(\tilde{Y}_{mis} | \theta, Y_{obs})$  effectively. For (2.4), a possible general strategy is to find a linear approximation of the form

$$E[Y_{mis} | \theta, Y_{obs}] \approx \alpha + \beta\theta, \quad (2.6)$$

where  $\alpha$  and  $\beta$  are known quantities but they can depend on  $Y_{obs}$ . We then let

$$\tilde{Y}_{mis} = Y_{mis} - \beta\theta, \quad (2.7)$$

which approximately eliminates the posterior correlation between  $\theta$  and  $\tilde{Y}_{mis}$ . In other words,  $w = \beta$  is an approximately optimal working parameter. The specification of the joint distribution of  $(\theta, \tilde{Y}_{mis})$  (given  $Y_{obs}$ ) then follows a simple linear transformation of  $(\theta, Y_{mis})$ . Again, this does not automatically imply that sampling from  $p(\theta | \tilde{Y}_{mis}, Y_{obs})$  is as easy as sampling from the original  $p(\theta | Y_{mis}, Y_{obs})$ , but the empirical evidence so far suggests that it is often manageable. When this simple approximation is inadequate and we cannot effectively implement the DA algorithm based on (2.4), it should serve as a warning sign that reducing the dependence between  $\theta$  and  $\tilde{Y}_{mis}$  is not

fruitful for the problem at hand. Hence it may help to consider the alternative strategy, namely, reducing the dependence between  $Y_{mis}$  and  $\tilde{Y}_{mis}$  via (2.5).

Second, the residual augmentation given by (2.4) does not depend on the prior for  $\theta$ , which explains why the same  $c = -(V + 1)^{-1}$  leads to i.i.d. draws regardless of the value of the prior variance  $A$  in HR’s example. This is an advantage of conditional augmentation (i.e., partially non-centering), for once we identify a good partially non-centering scheme, we can expect its performance to be robust to the prior specification of  $\theta$ . (Note we only say “robust to,” not “invariant to” because the invariance result in HR’s example relies on the normality assumption; in general the lack of correlation between  $\theta$  and  $\tilde{Y}_{mis}$  does not imply independence.) In contrast, the augmentation (2.5) in general depends on the prior for  $\theta$ . This explains why ASIS does not produce i.i.d. draws in HR’s example when an informative prior is used, because both the sufficient and ancillary schemes are defined without involving the prior on  $\theta$ .

### 2.3 The Roles of Sufficiency and Ancillarity

A reader then may naturally ask: what is the role of sufficiency and ancillarity in breaking the  $Y_{mis} \rightarrow \tilde{Y}_{mis}$  link? This was indeed a cause of much of our insomnia. The beauty of ASIS, as noted by several discussants, is its conceptual simplicity, rooted in the familiar classical notions, and its ease in implementation, mostly repeating existing steps in a certain order. Kelly’s statement “*These qualities of ASIS are attractive to me as an astrophysicist, because I spend most of my time doing astrophysics research, not designing MCMC samplers.*” is a great reminder to all MCMC designers of the importance of keeping things simple. However, just as to teach intuitively requires one to understand deeply, for us as the designer we want to fully understand why and when certain algorithms work. Our initial intuition, as documented in our main article, came from the classical Basu’s theorem. But Basu’s theorem is about the independence between sufficient and ancillary statistics in the sampling model, not in the posterior model as we need; this mismatch has indeed given us some midnight sweats.

In Section 2.4 of our main article, we have examined this issue assuming that  $\theta$  and  $Y_{mis}$  are one-to-one given  $\tilde{Y}_{mis}$ . Under this assumption, the sufficiency of  $Y_{mis}$  and the ancillarity of  $\tilde{Y}_{mis}$  allow us to write

$$p(\tilde{Y}_{mis}, Y_{mis} | Y_{obs}) \propto p(Y_{obs} | Y_{mis}) p(\tilde{Y}_{mis}) \Delta(\tilde{Y}_{mis}, Y_{mis}), \quad (2.8)$$

with

$$\Delta(\tilde{Y}_{mis}, Y_{mis}) = J(\tilde{Y}_{mis}, Y_{mis})p\left(\theta(\tilde{Y}_{mis}, Y_{mis})\right), \quad (2.9)$$

where  $\theta(\tilde{Y}_{mis}, Y_{mis})$  is the map from  $Y_{mis}$  to  $\theta$  (with  $\tilde{Y}_{mis}$  fixed) as determined by  $\tilde{Y}_{mis} = M(Y_{mis}; \theta)$ , and  $J(\tilde{Y}_{mis}, Y_{mis})$  is the corresponding Jacobian given by (2.23) in our main article.

Therefore, what ASIS achieves is to separate  $\tilde{Y}_{mis}$  and  $Y_{mis}$  in the first two terms on the right-hand side of (2.8). If the  $\Delta$  function is also separable in the sense of taking a product form  $\Delta(\tilde{Y}_{mis}, Y_{mis}) = \Delta_1(\tilde{Y}_{mis})\Delta_2(Y_{mis})$ , then  $\tilde{Y}_{mis}$  and  $Y_{mis}$  will be independent conditional on  $Y_{obs}$ , thereby breaking the  $Y_{mis} \rightarrow \tilde{Y}_{mis}$  link.

Of course, in general  $\Delta(\tilde{Y}_{mis}, Y_{mis})$  is not perfectly separable, but it tends to be so when the prior is weak, as demonstrated in Section 2.4 of our main article. We note that  $\Delta(\tilde{Y}_{mis}, Y_{mis})$  is determined by only two factors: the map  $\tilde{Y}_{mis} = M(Y_{mis}; \theta)$  and the prior  $p(\theta)$ , neither of which involves  $Y_{obs}$ . When the relationship between  $\tilde{Y}_{mis}$  and  $Y_{mis}$  is strong given  $\theta$ , (2.9) explains why ASIS may not do too well when the prior is also strong (but see below), because then  $\Delta(\tilde{Y}_{mis}, Y_{mis})$  would be far from being separable, resulting in strong posterior dependence between  $\tilde{Y}_{mis}$  and  $Y_{mis}$ .

In contrast, the residual augmentation (2.5) eliminates at least the posterior (Pearson) correlation by regressing out the impact of the prior via choosing  $\tilde{Y}_{mis} = \theta - E(\theta|Y_{mis}, Y_{obs})$ , which reduces to  $\tilde{Y}_{mis} = \theta - E(\theta|Y_{mis})$  when  $Y_{mis}$  is sufficient. For a clear comparison, let us denote  $M_R(Y_{mis}; \theta) = \theta - E(\theta|Y_{mis})$  and define  $M_A(Y_{mis}; \theta)$  as the aforementioned map for the ancillary augmentation. That is, the difference between residual augmentation and ASIS is that the former uses  $\tilde{Y}_{mis} = M_R(Y_{mis}; \theta)$  whereas the latter uses  $\tilde{Y}_{mis} = M_A(Y_{mis}; \theta)$ .

If statistical efficiency is the only consideration, then it would be sensible to only use  $M_R(Y_{mis}; \theta)$ . However, an astute reader may have noticed that whereas  $\tilde{Y}_{mis}$  of (2.4) is of the same dimension as the original  $Y_{mis}$ , the  $\tilde{Y}_{mis}$  of (2.5) is of the same dimension as  $\theta$ , which can be very different from that of  $Y_{mis}$ . Hence the derivation of  $p(\theta|\tilde{Y}_{mis})$  with  $\tilde{Y}_{mis} = M_R(Y_{mis}; \theta)$  can be complicated because  $\tilde{Y}_{mis}$  is typically not a one-to-one transformation of  $Y_{mis}$  given  $\theta$ . In contrast, the ancillary augmentation  $M_A(Y_{mis}; \theta)$  is often a one-to-one map of  $Y_{mis}$  given  $\theta$ , making it significantly easier to derive the conditional distribution of  $\theta$  given the ancillary augmentation. The price we pay for this ease of implementation is reflected by the posterior dependence of  $Y_{mis}$  and  $\tilde{Y}_{mis}$ , which in the simple case of (2.9) is captured by the  $\Delta$  function.

However, when the prior is not strong, the following heuristic argument suggests that  $M_A(Y_{mis}; \theta)$

may serve as a reasonable approximation to  $M_R(Y_{mis}; \theta)$ . The theory of Meng and Zaslavsky (2002) says that for many commonly used likelihoods with SOUP (single observation unbiased prior), a type of non-informative prior (e.g., a constant prior for a location parameter), the posterior mean  $\hat{\theta}(Y_{mis}) = E(\theta|Y_{mis})$  is unbiased for  $\theta$ . Consequently,

$$E[M_R(Y_{mis}; \theta)|\theta] = \theta - E[\hat{\theta}(Y_{mis})|\theta] = \theta - \theta = 0, \quad \text{for all } \theta \in \Theta. \quad (2.10)$$

This of course does not imply  $\tilde{Y}_{mis} = M_R(Y_{mis}; \theta)$  is ancillary, but it does rule out any  $\tilde{Y}_{mis}$  that is not first-order ancillary, that is, whose mean depends on  $\theta$ .

In general, we do not expect (2.10) to hold exactly, but the above derivation suggests that the use of ancillary augmentation may be viewed as a compromise between ensuring easy implementation and our desire to use (2.5)—or any of its variations such as (2.15) discussed below—that aims to break the  $Y_{mis} \rightarrow \tilde{Y}_{mis}$  link. For instance, in HR’s example, because of (2.3), the optimal residual augmentation for GIS is

$$M_R(Y_{mis}; \theta) = \theta - \frac{A}{A+V}Y_{mis},$$

which is increasingly better approximated by the ancillary augmentation  $M_A(Y_{mis}; \theta) = \theta - Y_{mis}$  as  $A/V \rightarrow \infty$ , that is, as the prior becomes weak.

Even when  $M_A(Y_{mis}; \theta)$  is not a good approximation to  $M_R(Y_{mis}; \theta)$ , our ASIS algorithm may still converge relatively fast because of the built-in robustness of ASIS/GIS as theoretically established in our main article. In HR’s example mentioned above, the joint normality of  $\{\theta, Y_{mis}, Y_{obs}\}$  makes it easy to obtain the convergence rates for sufficient augmentation (SA), ancillary augmentation (AA), and ASIS, which are respectively

$$r_{SA} = \frac{A}{(V+A)(V+1)}, \quad r_{AA} = \frac{AV}{(V+1)(A+1)}, \quad r_{ASIS} = \frac{AV}{(V+1)(A+1)(V+A)}. \quad (2.11)$$

These expressions reveal that both  $r_{SA}$  and  $r_{AA}$  can get arbitrarily close to one on their own, yet  $r_{ASIS} \leq 1/8$  for any combination of  $A$  and  $V$ , with equality if and only if  $A = V = 1$ . That is, the worst rate of convergence of ASIS is 0.125 in this example. Furthermore, because

$$\mathcal{R}_{1,2} = \text{Corr}(\tilde{Y}_{mis}, Y_{mis}|Y_{obs}) = \sqrt{\frac{V}{(A+1)(V+A)}}, \quad (2.12)$$

we have  $\mathcal{R}_{1,2} \rightarrow 1$  as  $A \rightarrow 0$ , i.e., the choice of  $\tilde{Y}_{mis} = \theta - Y_{mis}$  fails to break the  $Y_{mis} \rightarrow \tilde{Y}_{mis}$  link. Nevertheless, ASIS does not fail. On contrary, we have  $r_{ASIS} \rightarrow 0$  as  $A \rightarrow 0$  because  $r_{SA} \rightarrow 0$

and  $r_{AA} \rightarrow 0$  (in fact one of them suffices to ensure  $r_{ASIS} \rightarrow 0$ ). This again shows the power of ASIS/GIS — in order for it to converge fast, we only need *one* of  $\mathcal{R}_{1,2}$ ,  $r_1$  and  $r_2$  to be small, where  $r_1$  and  $r_2$  denote the convergence rates of the two DA schemes being interwoven.

## 2.4 Flexibility in Constructing Residual Augmentations

Various strategies adopted in our main article for forming ancillary augmentations turn out to be also suggestive for forming residual augmentations. For example, the use of the standard residual  $Y_{mis} - w\theta$  is implicitly driven by the usual least-squared formulation with additive residuals. For a multiplicative model with positive  $Y_{mis}$  and positive  $\theta$  (let us assume that both are scalars for simplicity), we can form the multiplicative counterpart of (2.4) as

$$\tilde{Y}_{mis} = \frac{Y_{mis}}{E[Y_{mis}|\theta, Y_{obs}]}. \quad (2.13)$$

For any  $b(\theta)$  such that the relevant expectations are finite, we have

$$E \left[ \frac{Y_{mis}b(\theta)}{E[Y_{mis}|\theta, Y_{obs}]} \middle| Y_{obs} \right] = E \left[ \frac{E[Y_{mis}|\theta, Y_{obs}]b(\theta)}{E[Y_{mis}|\theta, Y_{obs}]} \middle| Y_{obs} \right] = E[b(\theta)|Y_{obs}]. \quad (2.14)$$

By taking  $b_1(\theta) = \theta$  and then  $b_2(\theta) = 1$  in (2.14), we can easily verify that

$$\text{Cov}(\tilde{Y}_{mis}, \theta|Y_{obs}) = E[b_1(\theta)|Y_{obs}] - E[b_2(\theta)|Y_{obs}] E[\theta|Y_{obs}] = 0.$$

That is, like the additive residual (2.4), the multiplicative residual (2.13) is also uncorrelated with  $\theta$ . Similarly, we can establish that the multiplicative counterpart of (2.5), that is,

$$\tilde{Y}_{mis} = \frac{\theta}{E[\theta|Y_{mis}, Y_{obs}]}, \quad (2.15)$$

is uncorrelated with  $Y_{mis}$ . (Note again that when  $Y_{mis}$  is a sufficient augmentation, the above becomes  $\tilde{Y}_{mis} = \theta/E[\theta|Y_{mis}]$ .) In our main article, we have shown how ancillary augmentation schemes are often formed by re-centering and/or re-scaling (including rotating) a sufficient augmentation  $Y_{mis}$ . The additive and multiplicative residual formulations here further explain why such strategies have been successful.

Going beyond the additive and the multiplicative, we can also consider various transformations of  $\{\theta, Y_{mis}\}$  before forming the residual augmentation  $\tilde{Y}_{mis}$ . Suitably chosen transformations are likely to produce practically important gains when the joint posterior  $p(\theta, Y_{mis}|Y_{obs})$  is far from

normal. As discussed above, our goal is to reduce the *dependence* between  $\theta$  and  $\tilde{Y}_{mis}$  (for conditional augmentation) or between  $Y_{mis}$  and  $\tilde{Y}_{mis}$  (for interweaving). Using the additive residual forms in (2.4) or in (2.5) achieves this when joint normality holds (approximately) for  $(\theta, Y_{mis})$ . This suggests that for arbitrary  $(\theta, Y_{mis})$ , it may be worthwhile to consider one-to-one transformations in the form of  $\theta^* = g(\theta)$  and  $Y_{mis}^* = h(Y_{mis})$  to bring the resulting joint distribution  $(\theta^*, Y_{mis}^*)$  closer to normality before constructing the residual in (2.4)

$$\tilde{Y}_{mis}^* = Y_{mis}^* - E[Y_{mis}^* | \theta^*, Y_{obs}] = h(Y_{mis}) - E[h(Y_{mis}) | \theta, Y_{obs}], \quad (2.16)$$

or in (2.5)

$$\tilde{Y}_{mis}^* = \theta^* - E[\theta^* | Y_{mis}^*, Y_{obs}] = g(\theta) - E[g(\theta) | Y_{mis}, Y_{obs}]. \quad (2.17)$$

These transformations do not alter the original DA algorithm since they are one-to-one and are applied separately, though a price we pay for this simplicity is that the degree of joint normality we hope to achieve may never be reachable with separate transformations on the margins. Moreover, for discrete  $Y_{mis}$ , such as the restricted Boltzmann machine discussed by Wu, one-to-one transformation to normality is plainly impossible. Fortunately, Roberts' team has come up with a number of innovations to deal with discrete augmentations, as discussed in Section 2.6 of our main article and further by PRS, their Poisson example being particularly instructive.

### 3 More Insomnia and Needing Inception

While many theoretical questions remain, more insomnia is likely caused by practical issues. One important issue, as highlighted by both our main article and by PRS's discussion, is the balance between computing time per iteration and the convergence rate, that is, computational efficiency versus statistical efficiency. Convergence rate may be subject to theoretical analysis; computing time per iteration, however, depends on the model, the particular data set, the programming, and the computer being used. PRS suggest the *adjusted effective sample size*, a concept similar to the precision per CPU measure used in Kong et al. (2003). A nice illustration is provided by PRS in their Section 6 in the context of Bayesian inference for discretely observed diffusions. In that example, the centered (respectively, non-centered) scheme performs better when  $n$ , the number of data points, is large (respectively, small). The interweaving strategy initially performs better than both the centered and non-centered schemes, but as  $n$  becomes very large, it lags behind the

centered scheme, presumably because the improvement in convergence rate can no longer offset the increase in computing time per iteration.

A related issue is the use of Metropolis-within-Gibbs, which is common as illustrated by both PRS and Kelly. How should such Metropolis-Hastings (MH) steps be tuned, i.e., what is the ideal acceptance rate? Recall that our theoretical investigations focus on the case when each conditional draw of the interweaving strategy is available in closed form. The intuition for ASIS relies on this implicitly. Therefore it seems that the acceptance rates at such MH steps should be made to be reasonably large. But higher MH acceptance rates entail more computing time. Therefore we face the same problem of computational versus statistical efficiency. This trade-off lies at the heart of any sensible MCMC methodology, yet currently it is handled almost exclusively in an ad hoc fashion. Such grand challenges call for *Inception*—to borrow a catchy phrase from Hollywood—rather than mere insomnia.

For partially non-centering (i.e., conditional augmentation), the choice of an “optimal” working parameter is the key. Here again one must balance implementability, computing time, and statistical efficiency. A crude approximation of the optimal scheme, such as the linear approximation (2.6), may be preferred if a finer approximation requires too much analysis, is too difficult to implement, or takes too much time per iteration. An important difference in implementation between partially non-centering and ASIS is that the optimal partially non-centering scheme is expected to be data-dependent (see, e.g., (2.4)), whereas constructing sufficient and ancillary augmentations is model-dependent but not data-dependent. The absence of this tuning parameter (the working parameter) is often a practical advantage of ASIS (with its trade-offs as discussed in Section 2.3). Of course, as long as the implementation is easy, one can also interweave a centered augmentation and a partially non-centered one derived from it, as in HR’s normal example.

With the ever-increasing complexity of statistical models, the need for general and effective MCMC is also ever increasing (and so is the insomnia of the designers of such algorithms). Kelly’s discussion both illustrates the potential for interweaving and the need for more research in this area. After constructing a sufficient augmentation  $\tilde{\delta}$ , he samples the parameters  $\gamma$  along a suitable direction that leaves  $\tilde{\delta}$  invariant. Like our example of normal regression under censoring, this demonstrates the flexibility of the interweaving strategy. One issue noted by Kelly is that sometimes it is difficult to obtain useful sufficient augmentations for certain parameters, which

of course limits the applicability of the method. There is much room for both theoretical and empirical investigations, and again the key is to have innovative ideas.

As an example of brain-storming, consider a highly complex model that carries many nuisance parameters. If the goal of MCMC is to obtain the marginal posterior of the few parameters of interest, to what extent can we ignore the poor convergence of the nuisance parameters? Obviously all components must converge in order for a joint chain to converge. However, as we mentioned in Section 1.1, by now there are a number of practical examples where the joint chain is not even positive recurrent, yet the sub-chain corresponding to the parameters of interest converges not only properly but also rapidly. If, in certain problems, it is indeed possible to ignore the poor convergence of nuisance parameters, then we can focus on improving the convergence for a few targeted parameters of interest. Our component-wise interweaving strategy (CIS) seems especially relevant in such situations.

Last but not least, we cannot have sweet dreams without thanking the discussants for their inspirations, especially HR's theoretical insight, PRS's methodological advance, Kelly's practical implementation, and Wu's unsupervised learning. Of course the editor, Professor Richard Levine, deserves all our morning gratefulness for his confidence in and help with our paper, and for his sharing our belief that we statisticians can have our own midsummer night's dream.

## Acknowledgements

We thank Nathan Stein for helpful comments and NSF for partial financial support.

## References

- [1] Hobert, J. P. (2001a). Discussion of paper by van Dyk and Meng. *Journal of Computational and Graphical Statistics* **10**, 59–68.
- [2] Hobert, J. P. (2001b). Stability relationships among the Gibbs sampler and its subchains. *Journal of Computational and Graphical Statistics* **10**, 185–205.
- [3] Hobert, J. P. and Marchev, D. (2008). A theoretical comparison of the data augmentation, marginal augmentation and PX-DA algorithms. *Ann. Statist.* **36**, 532–554.

- [4] Huang, T.-M. (2010). Testing conditional independence using maximal nonlinear conditional correlation. *Ann. Statist.* **38**, 2047–2091.
- [5] Kong, A., McCullagh, P., Meng, X.-L., Nicolae, D. and Tan, Z. (2003). A theory of statistical models for Monte Carlo integration (with discussion). *J. Roy. Statist. Soc. B.* **65**, 585–618.
- [6] Liu, J. S. and Wu, Y. N. (1999). Parameter expansion for data augmentation. *J. Amer. Statist. Assoc.* **94**, 1264–1274.
- [7] Meng, X.-L. and van Dyk, D. A. (1997). The EM algorithm — an old folk song sung to a fast new tune (with discussion). *J. Roy. Statist. Soc. B* **59**, 511–567.
- [8] Meng, X.-L. and van Dyk, D. A. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika* **86**, 301–320.
- [9] Meng, X.-L. and Zaslavsky, A. (2002). Single observation unbiased priors. *Ann. Statist.* **30**, 1345–1375.
- [10] Papaspiliopoulos, O., Roberts, G. O. and Sköld, M. (2007). A general framework for the parametrization of hierarchical models. *Statist. Sci.* **22**, 59–73.
- [11] Romanovič, V. A. (1975). The maximal partial correlation coefficient of two  $\sigma$ -algebras relative to a third  $\sigma$ -algebra. *Izv. Vysš. Učebn. Zaved. Matematika* **10**, 94-96.
- [12] Tanner, M. A. and Wong, W. H. (2010). From EM to data augmentation: the emergence of MCMC Bayesian computation in the 1980s. *Statist. Sci.* **25**, 506–516.
- [13] van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation (with discussion). *Journal of Computational and Graphical Statistics* **10**, 1–111.
- [14] van Dyk, D.A. and Meng, X.-L. (2010). Cross-fertilizing strategies for better EM mountain climbing and DA field exploration: a graphical guide book. *Statistical Science* **25**, 429–449.
- [15] Yu, Y. (2005). *Three Contributions to Statistical Computing*. Ph.D. Thesis, Department of Statistics, Harvard University.

- [16] Yu, Y. and Meng, X.-L. (2007). Espousing classical statistics with modern computation: sufficiency, ancillarity and an interweaving generation of MCMC. *Technical Report*, Department of Statistics, University of California, Irvine.