DISCUSSION ARTICLE

Rejoinder

David A. VAN DYK and Xiao-Li MENG

Responding to such a diverse collection of comments is both enjoyable and challenging. The discussions vary from insightful theoretical explorations, to impressive technical investigations, to imaginative potential applications. To address this range in a coherent fashion, our rejoinder responds to each discussion in turn. We highlight each of the comments with a thematic adjective, which we hope the authors would find accurate if not most complimentary. Since we obviously aim for an informative rejoinder, we cannot agree fully with every discussant. Nonetheless, we are thankful to all of the discussants for their valuable time and thoughts. Our heartfelt thanks also goes to the editor, Andreas Buja, for his lightning speed in handling our submission, and for his effort in organizing the discussion.

1. LEVINE: INSIGHTFUL!

We thoroughly enjoyed Levine's concerto even as we tried to be the most critical of critics. Early in the composition we identified a slight cacophony (*passing some of the effort of designing to the user*), only to discover later, pleasantly, that it was simply a proleptic contrasting note to dramatize the grand finale (*emphasis on trading the designer's effort for the user's time*). Bravo!

Few composers can conceal, privately or publicly, their excitement when a critic highlights every key note in their composition, especially the subtle ones (e.g., *violation of the Markovian property*). We are thus grateful to Levine for such an insightful discussion. We particularly appreciate Levine's attempt to establish a quantitative framework for analyzing intricate issues such as computational complexity and implementability versus statistical efficiency, and the tradeoff between the designer's and the user's time. For computational complexity and implementability, there is indeed a huge literature in computer science, numerical analysis, and several other related fields. The proceedings volumes of the International Conference on Monte Carlo and Quasi Monte Carlo methods (Niederreiter 1995, 1998, 1999, 2001) are good resources to sample some of the most recent development in these topics. Statisticians' contributions have been quite limited so far, but we generally

Journal of Computational and Graphical Statistics, Volume 10, Number 1, Pages 98–111

share Levine's optimism that more can and will be done statistically, especially regarding the issue of Monte Carlo efficiency.

For example, producing and/or analyzing simulated data poses an intriguing modeling issue that statisticians seem to be uniquely qualified to address. On the one hand, with simulated data, there is no issue of model uncertainty, since we know exactly how the data were generated (at least in principle). On the other hand, depending on the degree to which we "take advantage of specific structures or characteristics of the distribution or model of interest," as Levine recognized, we can construct more and more efficient Monte Carlo methods with *the same* simulation size or, equivalently, the same computational effort, *excluding* the computational effort needed for using these specific structures and characteristics.

Therefore, the issue of model uncertainty with simulated data has a different character than that of statistical design and inference with real data. The question is not which model is true or approximately true, since all models that can link the simulated data to our estimand are true and known. The question is rather which model represents the best compromise among computational complexity, human effort, and statistical efficiency. For DA algorithms, all augmented-data models, as defined by (1.1) or (1.2), are correct and known by construction. But as we demonstrated in the article, a good (not necessarily optimal) choice of the augmented-data model can lead to substantially more efficient algorithms. Meng and Schilling (in press) provide another striking example, in the context of analyzing simulated data (i.e., computing normalizing constants from MCMC output), where by constructing a better analysis model one can achieve orders of magnitude of improvement in Monte Carlo efficiency with relatively little increase in computational load. The construction there, however, like that in the current article, has an "artistic" aspect. Whereas we, like Levine, believe the "artistic" aspect can never be completely removed, we do hope that Levine's decision theoretic framework will provide a more coherent way of gaining insight into and guidelines for better MCMC designs and analysis in general.

A key in our methods, as well as in Levine's more general framework, is to find an *effective* objective function. By "effective" we mean an objective function that can be optimized or nearly optimized easily and that this optimizing value is a good approximation of the optimizer of the ideal objective function. Perhaps the most exciting message from Meng and van Dyk (1999) and our current article is that it is entirely possible to find such an effective objective function (e.g., our EM criterion (2.5)) even when the ideal objective function (e.g., the geometric rate (2.2)) cannot even be evaluated. This is very much in the same spirit of the EM algorithm, which can maximize an observed-data likelihood function that cannot be evaluated directly. The key is to *transfer* the desired optimizer to that of a simpler function, which itself may not even be a rough approximation to the ideal objective function. One can easily see this by picturing two functions that do not resemble each other at all except that their "deep valleys" occur in practically the same location. Incidentally, the use of this powerful "optimizer transfer" method for statistical applications was the topic of a discussion article by Lange, Hunter, and Yang (2000) that appeared a year ago in this journal. Our problem is somewhat easier than those studied by Lange, Hunter, and Yang or in the EM literature because in order to balance algorithmic optimality and human effort we seek only reasonable approximations to the ideal optimizer. Although the EM criterion is motivated by the maximum lag-1 autocorrelation over linear functionals, we use

it because we wish to limit complexity, not because we believe in the linearity or that (2.3) is a good approximation to (2.2). The effectiveness of this much simpler objective function is evident from the first two applications we presented, since in each case it led to Haar measure, which is the *theoretically* optimal prior under the specific group-theoretic setting of Liu and Wu (1999). In the third case, it led to right Haar measure, which is a natural generalization of the Haar measure when the transformation group is not unimodular (Liu and Sabatti 2000) even though currently there is no mathematical proof of the theoretical optimality of the right Haar measure.

We conclude our reply to Levine by emphasizing that in the course of searching for optimal or near optimal methods/algorithms, one should not lose sight of the "robustness" of certain common methods. For example, while the random scan Gibbs sampler may not be optimal in most applications, it generally performs well compared to a badly estimated "optimal" order. Furthermore, it does not depend on the function of interest. In practice we are often interested in several functions and the optimal order for one function can be seriously suboptimal for another, as alluded to by Levine. These issues are quite similar to those involved with the optimal allocation of sample sizes in stratified sampling (e.g., Neyman 1934; Kish 1965). It is well-known that the "optimal" sample size allocation can provide a much worse estimator than simple *proportionate sampling* when the strata variances used for the optimal allocation are badly estimated or are based on a different estimator. Incidentally, the issues of step orderings within the Gibbs sampler and within its deterministic counterpart, the ECM algorithm (Meng and Rubin 1993), appear to be less related than one might expect given the well-known intrinsic connections between the Gibbs sampler and EM-type algorithms; see van Dyk and Meng (1997) and the rejoinder of Meng and van Dyk (1997). Within the framework of the random scan Gibbs sampler, there is much exibility that may be exploited for computational gain. In addition to Levine's strategy of more frequently visiting slow-mixing components (Liu, Wong, and Kong 1995) we suggest more frequently visiting components that require less complex computation (van Dyk 2000) or a combination of both strategies.

2. HOBERT: IMPRESSIVE!

Like Levine's composition, Hobert's *opéra seria* is a pleasure to read, at least for those who indulge in rigor, intricacy, and elegance. The only "complaint" we could offer is that we might have just lost a publication in *The Annals of Statistics* because Hobert's contribution covers some theoretical investigation we were planning to do!

Our first reaction to Hobert's impressive theoretical bound for the (limiting) marginal DA algorithm is that the corresponding bound for the standard DA algorithm must be much larger. Indeed, using the same technique and notation as Hobert we were able to obtain the following (best) bound for the standard DA algorithm when $\nu > 3$ and $n = 2$:

$$D(t|\theta_0) \equiv ||P^{(t)}[(\mu_0, \lambda_0), \cdot] - \pi(\cdot|\mathbf{y})|| \leq (.9944037)^t + (7.714286) \times (.9868403)^t; \qquad \text{(R.1)}$$

see the Appendix to this rejoinder for details. Consequently, we need $t \geq 824$ for the *bound* to be less than .01, in contrast to Hobert's $t \geq 335$ for the (limiting) marginal DA algorithm.

Although having a larger upper bound does not logically imply slower convergence, it is another strong indication that the (limiting) marginal DA algorithm is superior. Incidentally, we surmise that when the sample size n is larger than 2, the bounds, although more difficult to calculate analytically, should tend to be smaller because, for larger n, the joint posterior surface of the parameter and augmented-data sufficient statistics should tend to be smoother (e.g., exhibits fewer "bumps"), and thus the chain should mix more rapidly.

Our second reaction is that, although these bounds are practical, they must be very conservative, as suggested by the empirical results presented by Meng and van Dyk (1999) and in the current article. To check this we need to compute the actual total variation (TV) distance $D(t|\theta_0)$ as a function of t. For arbitrary t, the analytical calculation of the exact value of $D(t|\theta_0)$ is very difficult for either algorithm even with $n = 2$, but we can easily estimate the distance by replicating each algorithm with given t and θ_0.

2.1 A Subtlety of Marginal DA With an Improper Working Prior

Before we describe the Monte Carlo estimation of the TV distance, we point out a subtlety of the marginal DA algorithm with *improper* working prior that was revealed during our study of the TV distance. Take the univariate t model as an illustration. The standard DA (SDA) algorithm uses $\tilde{q} = \{\tilde{q}_1, \ldots, \tilde{q}_n\}$ as missing data, and thus has a joint stationary density $p(\theta, \tilde{q})$ (suppressing the conditioning on Y_{obs}; $\theta = (\mu, \lambda)$ under Hobert's notation). The marginal DA (MDA) algorithm uses $q = \alpha\tilde{q}$ as the missing data when the prior on α is *proper*, and it has a joint stationary distribution $p(\theta, \alpha, q)$. By the definition of $\tilde{q}$ $(= q/\alpha)$, the $(\theta, \tilde{q})$ margin of this joint stationary distribution is the same as the stationary distribution from SDA, $p(\theta, \tilde{q})$, even though the subchain $\{(\theta^{(t)}, \tilde{q}^{(t)} \equiv q^{(t)}/\alpha^{(t)}), t = 1, \ldots\}$ of MDA is not a Markov chain in general.

In the (optimal) limiting case when the hyper working parameter $\gamma = 0$ (and $\beta = 0$), MDA effectively gets rid of α because of the invariance property (2) of Lemma 1 (p. 10). Consequently, as is made clear by the explicit stochastic mapping given in Section 6, under Scheme 1 and with $\gamma = 0$, this limiting MDA (LMDA), which is the one underlying Hobert's theoretical investigation, can be implemented directly using the same $\tilde{q}$ as with SDA. Starting from $\theta^{(0)} = \theta_0$, LMDA first samples from $p(\tilde{q}^{(t+1)}|\theta^{(t)})$ exactly as with SDA, and then samples from some $p^*(\theta^{(t+1)}|\tilde{q}^{(t+1)})$. This $p^*(\theta|\tilde{q})$, however, has to be different from the $p(\theta|\tilde{q})$ of SDA; otherwise the two algorithms would be identical. Consequently, the joint chain from LMDA, $\{(\theta^{(t)}, \tilde{q}^{(t)}), t = 1, \ldots\}$ does *not* have the same joint stationary distribution as that of SDA, $p(\theta, \tilde{q}) = p(\theta|\tilde{q})p(\tilde{q})$. Rather, its joint stationary distribution is $p^*(\theta|\tilde{q})p(\tilde{q})$. The reason that LMDA maintains the same marginal stationary density $p(\tilde{q})$ as SDA is that both algorithms maintain the marginal stationary density $p(\theta)$, as we proved in this article. (This margin is often of primary interest.) Consequently, since both algorithms draw $\tilde{q}$ from $p(\tilde{q}|\theta)$, their $\tilde{q}$ marginal stationary densities must also be the same because $p(\tilde{q}) = \int p(\tilde{q}|\theta)p(\theta)d\theta$.

We emphasize that LMDA is not an MDA in its usual sense, since (3.2) no longer defines a *proper* marginal augmentation model when $p(\alpha)$ is improper. Consequently, the fact that $\{(\theta^{(t)}, \tilde{q}^{(t)}), t = 1, \ldots\}$ has the same *joint* stationary density under any *proper* working prior on α, yet has a different joint stationary density when the working prior

becomes improper as $\gamma \to 0$ is not paradoxical. As $\gamma \to 0$, there is no joint limiting distribution for (θ, α, q). Therefore, without further conditions, we do not even know if the "marginal distribution" of $(\theta, \tilde{q} = q/\alpha)$ is proper. This again shows the caution one needs to exercise when dealing with nonpositive recurrent chains, as one is certainly tempted to guess the joint stationary density for $(\theta, \tilde{q})$ would remain unchanged as $\gamma \to 0$, especially given the fact that both marginal stationary densities do remain unchanged.

For a user, this subtlety is not of any concern when the purpose of using LMDA is to obtain draws from $p(\theta)$, and thus $\tilde{Y}_{\text{mis}}$ (e.g., $\tilde{q}$) and Y_{mis} (e.g., q) are *auxiliary variables* introduced purely for constructing the algorithm. When the joint distribution of $(\theta, \tilde{Y}_{\text{mis}})$ is of interest because $\tilde{Y}_{\text{mis}}$ corresponds to "real" missing data or latent variables, as possible in the mixed effect model where $\tilde{Y}_{\text{mis}}$ represents the random effects, a user can easily obtain MCMC draws from the desired joint target density $p(\theta, \tilde{Y}_{\text{mis}})$ by using the *one-step shifted* joint draws $\{(\theta^{(t-1)}, \tilde{Y}_{\text{mis}}^{(t)}), t = 1, \ldots\}$ from LMDA. This is because $\theta^{(t-1)}$ has the desired limiting marginal distribution $p(\theta)$ and $\tilde{Y}_{\text{mis}}^{(t)}$ was drawn from the desired conditional distribution $p(\tilde{Y}_{\text{mis}}|\theta = \theta^{(t-1)})$. Apparently, this slight complication with LMDA is a consequence of its fast mixing. This suggests that it can be beneficial, in terms of mixing, to consider generalizations of the Gibbs sampler which use sampling distributions that are not the full conditionals from the target distribution but preserve the desired *marginal* stationary distribution(s). In the current setting this means $p^*(\theta|\tilde{Y}_{\text{mis}})$ must satisfy $\int p^*(\theta|\tilde{Y}_{\text{mis}})p(\tilde{Y}_{\text{mis}})\mu(dY_{\text{mis}}) = p(\theta)$; details will be given in Meng and van Dyk (2001).

2.2 Estimating Marginal and Joint Total Variation Distance

Having generalized the standard DA setting to include LMDA, we now describe our methods for estimating the TV distance. Specifically, suppose we have a DA chain on the joint space (u, v) which is equipped with a dominating product measure $\lambda \times \mu$. Starting from $u = u_0$, each iteration of DA first samples from $p(v|u)$ and then samples from $p^*(u|v)$, where $p^*(u|v)$ may or may not be the same as $p(u|v)$, as explained above. The *marginal total variation* (MTV) distance between the target *marginal* density $p(u)$ and $p^{(t)}(u)$, the density of $u^{(t)}$ given the initial value u_0, can be expressed as

$$\begin{aligned} D_{\text{MTV}}(t) &= \frac{1}{2}\sup_{|f|\le 1}\int f(u)[p^{(t)}(u) - p(u)]\lambda(du) = \frac{1}{2}\int |p^{(t)}(u) - p(u)|\lambda(du) \\ &= \frac{1}{2}\int \left|1 - \frac{p(u)}{p^{(t)}(u)}\right| p^{(t)}(u)\lambda(du) \\ &= \frac{1}{2}\int \left|1 - \frac{p(u)}{\int p^*(u|v)p^{(t)}(v)\mu(dv)}\right| p^{(t)}(u)\lambda(du). \end{aligned}$$

(The factor $1/2$ is needed for consistency with the TV distance used by Hobert.) Consequently, we can estimate $D_{\text{MTV}}(t)$ by

$$\hat{D}_{\text{MTV}}(t) = \frac{1}{2K}\sum_{i=1}^{K}\left|1 - \frac{p(u_i^{(t)})}{\frac{1}{K}\sum_{j=1}^{K} p^*(u_i^{(t)}|v_j^{(t)})}\right| = \frac{1}{2}\sum_{i=1}^{K}\left|\frac{1}{K} - \frac{p(u_i^{(t)})}{\sum_{j=1}^{k} p^*(u_i^{(t)}|v_j^{(t)})}\right|,$$

where $\{(u_k^{(t)}, v_k^{(t)}), k = 1, \ldots, N\}$ are iid realizations of $(u^{(t)}, v^{(t)})$ obtained by running the DA chain *independently* K times with the same starting value, u_0.

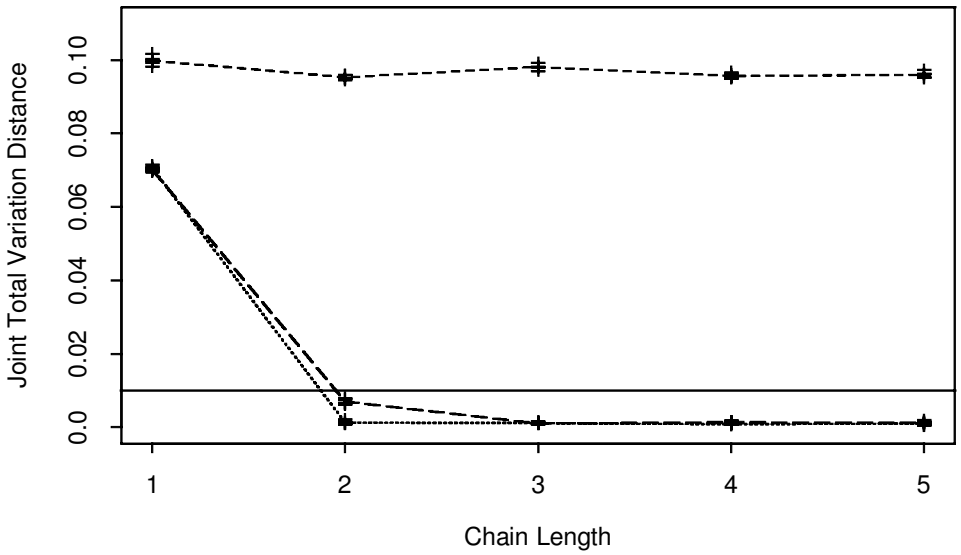

Figure A. The JTV Distance Between the Sampling and Stationary Distributions. We use Hobert's model specification and plot $\hat{D}_{\text{JTV}}(t)$ for $t = 1,\ldots, 5$ for the SDA (long dashed line) and LMDA (dotted line) algorithms; the solid line represents Hobert's threshold of .01 and the + signs represent the Monte Carlo replications. The dashed line near .10 is a Monte Carlo estimate of the JTV distance between $p^{(t)}(\theta, \tilde{q})$ of the LMDA algorithm and the joint stationary distribution of SDA.

A potential difficulty with $\hat{D}_{\text{MTV}}(t)$ is that it requires the evaluation of $p^*(u|v)$, which can be somewhat complicated when it is induced by LMDA (e.g., with the t model). One way to avoid this problem is to estimate the *joint total variation* (JTV) distance between $p^{(t)}(u, v)$ and the joint stationary density $\tilde{p}(u, v) = p^*(u|v)p(v)$, where $p(v) = \int p(v|u)p(u)\lambda(du)$. We can then use JTV as an *upper bound* for MTV because $D_{\text{JTV}}(t) \geq D_{\text{MTV}}(t)$ for all t. Since $p^{(t)}(u, v) = p^*(u|v) \int p(v|u')p^{(t-1)}(u')\lambda(du')$, we have $\tilde{p}(u, v)/p^{(t)}(u, v) = p(v)/\int p(v|u')p^{(t-1)}(u')\lambda(du')$, and thus

$$\begin{aligned} D_{\text{JTV}}(t) &= \frac{1}{2}\int\int |p^{(t)}(u, v) - \tilde{p}(u, v)|\lambda(du)\mu(dv) \\ &= \frac{1}{2}\int \left|1 - \frac{p(v)}{\int p(v|u)p^{(t-1)}(u)\lambda(du)}\right| p^{(t)}(v)\mu(dv). \end{aligned}$$

We thus can estimate $D_{\text{JTV}}(t)$ by

$$\hat{D}_{\text{JTV}}(t) = \frac{1}{2}\sum_{i=1}^{K}\left|\frac{1}{K} - \frac{p(v_i^{(t)})}{\sum_{j=1}^{k} p(v_i^{(t)}|u_j^{(t-1)})}\right|,$$

where $\{(u_k^{(t-1)}, v_k^{(t)}), k = 1, \ldots, N\}$ are iid realizations of $(u^{(t-1)}, v^{(t)})$, again obtained by running K independent DA chains with the same starting value, u_0.

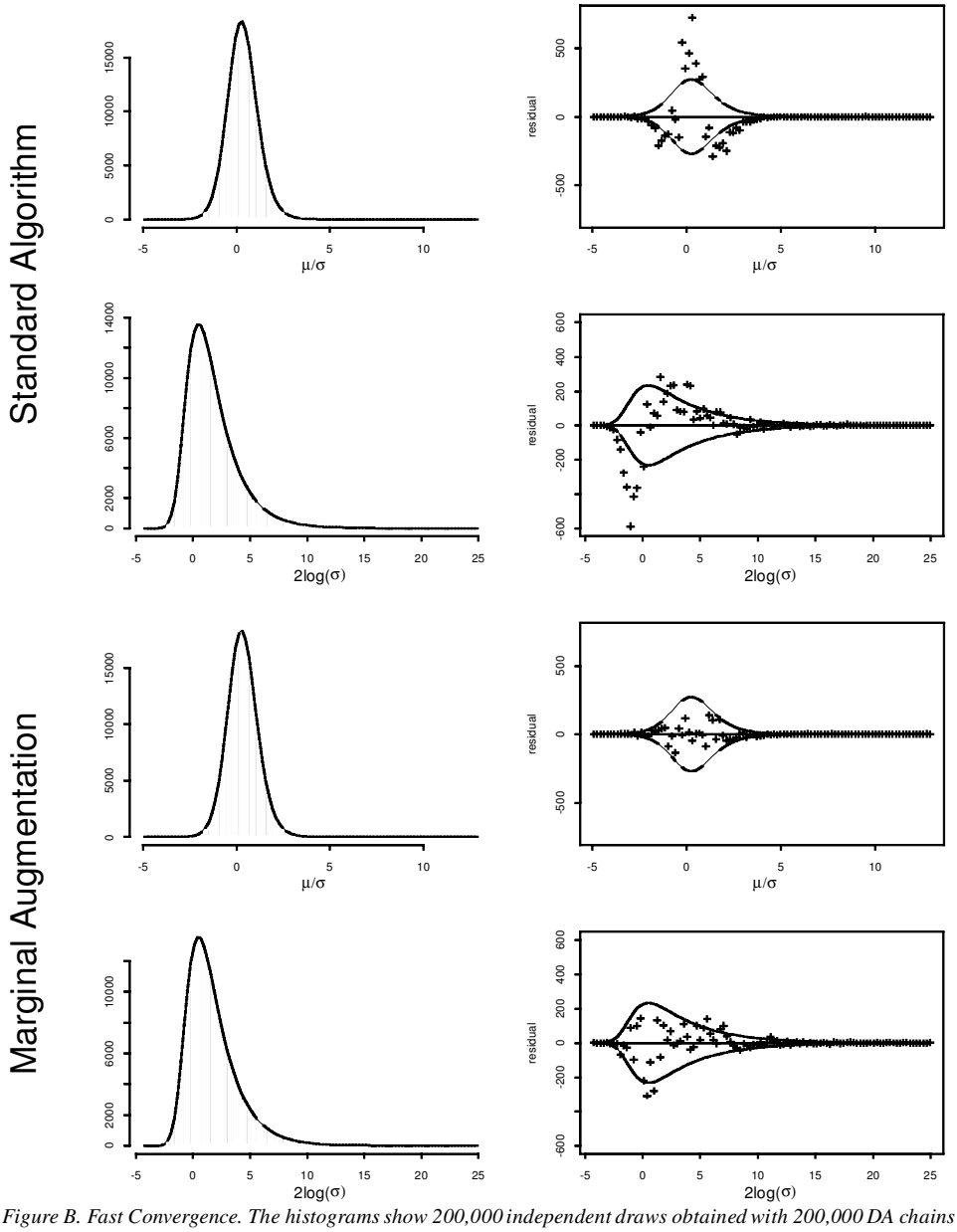

Figure B. Fast Convergence. The histograms show 200,000 independent draws obtained with 200,000 DA chains of length one, all using Hobert's starting values and model specification. The solid curves are the exact posterior obtained by numerical integration. The plots in the second column compare residuals (observed count less expected count in each bin of the histograms under the target distribution) with twice the standard error of the observed counts under the target distribution. Although both algorithm perform very well in this setting, LMDA is faster.

We can use $\hat{D}_{\text{JTV}}(t)$ to check the conservativeness of the theoretical bounds for the LMDA algorithm and for the SDA algorithm. Figure A shows the results for both algorithms; LMDA is represented by the dotted line and SDA by the long dashed line. It is seen that both SDA and LMDA obtain a JTV distance of less than .01 in only two iterations; the solid line corresponds to a JTV distance of .01. At the second iteration, the JTV distance corresponding to SDA is about five times as large as that of LMDA. The theoretical bounds of 335 and 824 iterations, respectively, for LMDA and SDA are clearly exceedingly conservative! Here, each evaluation of $\hat{D}_{\text{JTV}}(t)$ is based on $K = 20{,}000$ Monte Carlo draws and we replicated five times for each algorithm and each chain length. The replications are represented by plus signs and indicate minimal Monte Carlo variation. The short dashed line near .10 shows the TV distance between $p^{(t)}(\theta, \tilde{q})$ from LMDA and the target joint posterior distribution of SDA, which confirms that this quantity does not converge to zero. Figure B further illustrates the quick convergence of both algorithms for this problem, by comparing 200,000 independent draws of the model parameters based on chains of length one. The first column compares the draws with the target posterior distribution and the second column shows the residual (from the expected count) for each bin in the histograms. The residuals can be compared with twice the standard error of the bin counts, under the assumption that the draws are actually from the posterior distribution, as indicated by the solid lines. This shows again that LMDA is superior to SDA, albeit both algorithms perform very well in this case.

3. HIGDON: INSPIRING . . .

Higdon's comparison of the DA algorithm and the "unaugmented" Metropolis–Hastings (MH) algorithm inspired us to make a comparison of the two methods at an abstract level. The DA algorithm is obviously a special case of MH, since it is a special case of the Gibbs sampler, which in turn is a particular application of MH that always accepts the proposed move (e.g., Gelman 1992). Perhaps less realized, however, is that MH itself is an application of the auxiliary variable method, and hence is itself a form of data augmentation. At the $(t+1)$st iteration, the MH algorithm first samples from a proposal density: $Y^{(t+1)} \sim q(Y|X^{(t)})$. We then let $X^{(t+1)} = Y^{(t+1)}$ or $X^{(t+1)} = X^{(t)}$ according to the MH acceptance ratio criterion, which depends only on $X^{(t)}$, $Y^{(t+1)}$, and a random number. Clearly this defines a joint Markov chain $\{(X^{(t)}, Y^{(t)}), t = 1, \ldots\}$, with the subchain $\{X^{(t)}, t = 1, \ldots\}$ being the MH chain. In other words, choosing a proposal density $q(y|x)$ is equivalent to choosing an auxiliary variable Y, which in turn is equivalent to choosing an augmented-data model $p(x, y)$ whose X margin is our target distribution $p(x)$. The difference between this and Tanner and Wong's DA algorithm is that MH does not alternate between $X|Y$ and $Y|X$ in general; indeed, $q(y|x)$ can even be free of x, as with the so-called independent MH sampler (e.g., Tierney 1994).

Higdon's spatial modeling application also inspired us to consider the more difficult problem where the missing data/latent variable is a stochastic process Z. Although we do not have enough details of Higdon's specific application to assess whether our suggestion is sensible, it seems to us that one can try a positive working variable α to form a *randomly scaled* process αZ, and use the Haar working prior $p(\alpha) \propto \alpha^{-1}$. This should improve mixing

according to Liu and Wu's (1999) theoretical optimality result, which does not depend on the underlying target distribution as long as αZ is an admissible stochastic process for the underlying problem for any $\alpha > 0$. If one has to restrict the range of α in order for αZ to be admissible, then the Haar measure for the scale group may no longer be optimal and one may need to invoke the approximate methods in this article to seek a good choice of the (possibly improper) working prior. More complicated transformations such as a random rotation of a spatial process could also be tried as long as the transformed process is admissible for any transformation of a given transformation group. We fully agree with Higdon, however, that when the dimension of the underlying variable is large, this may require too much additional computation to be useful. Incidentally, the multiresolution approach Higdon used can be viewed as a form of data augmentation with the resolution level acting as the auxiliary variable; see Goodman and Sokal (1989) and Liu and Sabatti (2000).

4. LIU: INTRIGUING!

Liu made the intriguing observation that the marginal DA algorithm not only is fast, but also can be *simpler* than the standard DA algorithm. By first drawing a covariance matrix and then obtaining the implied correlation matrix, one avoids the problem of directly dealing with the unit-diagonal restriction of the correlation matrix. By moving in the space of *covariance* matrices, we can also move faster from one correlation matrix to another than is possible when we restrict ourselves to the space of *correlation* matrices. This is very much like simplifying movement from one point in a curved two-dimensional space to another point in the same space by allowing ourselves to leave the two-dimensional space along the way. Not only can the move be faster, it can also be much simpler—we can move along a straight line in the three-dimensional space that links the two points. In the curved two-dimensional space the shortest path could be much more complicated.

Incidentally, in his discussion of Meng and van Dyk (1997), Liu made another intriguing suggestion using covariance matrices. This suggestion is worth mentioning here because it is closely related to Levine's discussion on quantifying implementability. In the context of normal regression with a patterned covariance matrix, Liu suggested that one can mathematically formulate simplicity by requiring the augmented-data model to have a covariance matrix that allows closed-form (conditional) maximization. This requirement led to an explicit form of the augmented-data covariance matrix, which then defines a class of augmented-data models. We can then search this class for the augmented-data model that minimizes the augmented Fisher information (i.e., the EM criterion). These suggestions are valuable not only because they provide some theoretical insight, but also because they have direct practical implication in the difficult and important problem of modeling covariance matrices; see Barnard, McCulloch, and Meng (2000) and the references therein.

Liu's idea of using component-based software to exibly combine different parts of DA algorithms is also intriguing. We certainly look forward to its full development.

5. HUERTA, JIANG, AND TANNER: INVITING . . .

Huerta, Jiang, and Tanner's discussion is an article in its own right, focusing on a study of the so-called hierarchical mixtures-of-experts (HME) model in the context of time series analysis. As with many other data-fitting methods initially advanced by nonstatisticians, HME has a catchy name that statistical experts may find more inviting than the method itself. If one is not concerned with the sense in which an AR(1), GARCH(1,1), or EGARCH(1,1) is an "expert" or what the compelling reasons to mix these models might be, then it is not difficult to accept Huerta, Jiang, and Tanner's invitation to consider more efficient deterministic or stochastic algorithms for fitting the HME model. In particular, since HME is a finite mixture model with covariate-dependent weights, the data augmentation scheme developed by Pilla and Lindsay (in press) for efficient fitting of finite mixture models might suggest a strategy in this setting. Pilla and Lindsay's method uses an alternating data-augmentation scheme [i.e., an AECM algorithm in the terminology of Meng and van Dyk (1997)] to significantly improve the standard EM implementation for fitting the weights of a finite mixture model where the components themselves are assumed known.

6. WU AND ZHU: IMAGINATIVE . . .

It may take a reader some imagination to see the link between Wu and Zhu's discussion and our article. But imagination is exactly what Wu and Zhu aim for, both figuratively and literally. Modeling vision is a fascinating and challenging subject; we share Wu and Zhu's vision that the missing data framework can play an important role, both conceptually and methodologically. Indeed, Wu and Zhu considered DA as a modeling rather than a computational framework. This view is helpful for dealing with complex missing data problems or problems that can be formulated as such (e.g., latent variable modeling). In fact, an important reason for the popularity of data-augmentation based methods such as the EM algorithm and multiple imputation (Rubin 1987) is that these methods allow a data analyst to separate the task of dealing with missing data from that of modeling the complete data.

Incidentally, we view Wu and Zhu's "bottom-up" and "top-down" modeling strategies as another example of the difference in emphasis between the frequentist approach and the Bayesian method. The "bottom-up" strategy is consistent with frequentist thinking, which focuses on *sampling statistics* (e.g., Wu and Zhu's *features*). In contrast, the "top-down" framework focuses on the estimand (e.g., Wu and Zhu's *Summary*), a standard Bayesian philosophy. For complex problems such as modeling vision, both strategies are useful at the methodological level, but we agree with Wu and Zhu that at the conceptual level the Bayesian way of thinking can be far more fruitful.

7. MIRA AND GREEN: INTRACTABLE?

Some readers might find parts of Mira and Green's discussion "intractable" in the sense that they may be overwhelmed by the attempted "structure analysis" without knowing where the argument is heading. Indeed, we got dizzy reading Mira and Green's recasting

of our conditional and marginal DA methods! Of course, the real intractability is not in our methods, but with the goal implicitly underlying Mira and Green's attempt. Namely, to unify the MCMC "culinary art" with a single or a few standard recipes. It is indeed tempting to seek unification given the apparent "general ability" and "commonality" of methods like the MH algorithm, and given the great simplification unification might bring. On the other hand, the fact that so much effort has been devoted in recent years by statisticians—Mira, Green, and us included—to develop *efficient* MCMC algorithms for various real data analysis problems is a strong indication that such a unification is simply beyond reach. As we have shown in the article and in the response to Higdon, choosing an efficient auxiliary variable, *including* the MH proposal variable, is generally as difficult a task as choosing an efficient data augmentation scheme—there is as little indication in the MH algorithm itself of how to choose the proposal variable as there is in (2.1) or (3.1) of how to choose the augmented-data model $p(Y_{\text{aug}}|\theta, \alpha)$! Consequently, the sophistication required for the latter is not in any sense "contrasting" the generality of the former, but rather is an honest reection of the difficulty inherent in the subject. Standard recipes do exist, as with Chinese cooking one can always follow the stir-fry-with-soy-sauce or stir-fry-with-soy-sauce-and-MSG recipes no matter what one is cooking, but of course such a "simplified Chinese culinary art" could hardly thrill any chef or diner.

Mira and Green's information identity is a re-expression of the well-known "missing information principle": *complete information = observed information + missing information* (e.g., Orchard and Woodbury 1972; Dempster, Laird, and Rubin 1977; Meng and Rubin 1991). The "missing information" is the *expected* Fisher information for θ contained in the conditional model $p(Y_{\text{aug}}|Y_{\text{obs}}; \theta, \alpha)$ (evaluated at the observed-data posterior mode of θ and with α given), and it is well-known that the expected Fisher information can also be written as the variance of the score function under the same model. This yields Mira and Green's identity because the difference between the score function from the conditional model $p(Y_{\text{aug}}|Y_{\text{obs}}; \theta, \alpha)$ for θ and the first derivative with respect to θ of the log of the augmented-data posterior density does not depend on the unobserved part of Y_{aug}. In practice, it is somewhat easier to deal with $I_{\text{aug}}(\alpha)$ of (2.5) than the variance expression of Mira and Green because the former allows one to directly take advantage of the *linearity* of the complete-data *observed* Fisher information in a set of sufficient statistics (when they exist). Although Mira and Green's variance expression is mathematically equivalent, in specific problems it is typically evaluated by converting to (2.5) to simplify algebra.

We conclude our reply to Mira and Green, as well as our rejoinder, by emphasizing that we use the word "art" in its engineering sense. Successful engineering is a combination of scientific principles, good intuition, some experiences, a bit of sweat, and a touch of luck. Having worked on statistical algorithms for a number of years, we find that the construction of *efficient statistical algorithms* requires a similar combination. By *statistical algorithms* we mean not only those algorithms that are useful for statistical computation, but more importantly those algorithms that are motivated from statistical principles, such as the EM algorithm and the DA algorithm. The adjective *efficient* describes the "three S" requirement: simplicity, stability, and speed. It is our desire to achieve all three S's simultaneously that brings out the artistic aspect of the construction process. We hope that our article, together

with the discussions and rejoinder, makes it clear that being "artistic" is fundamentally important for achieving the state-of-the-art in MCMC cookery.

APPENDIX TO REJOINDER

A.1 THE MINORIZATION AND DRIFT CONDITIONS FOR SDA UNDER THE t MODEL

Here we adapt Hobert's minorization and drift conditions for SDA for fitting a univariate t model. Many of the required calculations follow directly from Hobert's discussion; the differences are outlined below.

We use the same drift function, $V(\mu, \sigma^2)$, and begin with the minorization condition. (We use σ^2 in place of Hobert's λ to be consistent with Meng and van Dyk (1999) and Section 6.) By Hobert's argument with $\pi^*(\mu, \sigma^2, \alpha|q, y)$ and $\pi^*(q|\mu', (\sigma^2)', \alpha', y)$ replaced with $p(\mu, \sigma^2|\tilde{q}, y)$, and $p(\tilde{q}|\mu', (\sigma^2)', y)$ respectively, it is sufficient to show Hobert's (2.5) holds. Because the draw of $\tilde{q}$ is identical for both algorithms (Hobert's q is the same as $\tilde{q}$ as he took $\alpha' = 1$), Hobert's (2.5) can be established for SDA almost exactly as argued by Hobert for LMDA using the same notational substitution. In particular, we have the following proposition.

Proposition A.1. *The Markov transition density $k(\mu, \sigma^2|\mu', (\sigma^2)')$ for SDA under the univariate t model satisfies the following condition*

$$k(\mu, \sigma^2|\mu', (\sigma^2)') \geq \epsilon h(\mu, \sigma^2) \quad \textit{for all} \quad (\mu', (\sigma^2)') \in S,$$

where $h(\mu, \sigma^2)$ is a density on $R \times R^+$ given by

$$h(\mu, \sigma^2) = \int p(\mu, \sigma^2|q, y) \left[\prod_{i=1}^{n} \frac{g(q_i)}{\int_0^\infty g(x)dx} \right] dq,$$

and $\epsilon = \left(\int_0^\infty g(x)dx\right)^n$. The function $g(\cdot)$ is as given in Proposition 1 of Hobert.

The drift condition differs more significantly between SDA and LMDA. For SDA, the drift condition is given by

Proposition A.2. *Let $V(\mu, \sigma^2) = \sum_{i=1}^n (Y_i - \mu)^2/\sigma^2$. If $n = 2$ and $\nu > 1$, then*

$$\mathrm{E}\left[V(\mu, \sigma^2)|\mu', (\sigma^2)'\right] = \frac{2}{\nu - 1} V(\mu', (\sigma^2)') + \frac{4\nu}{\nu - 1}.$$

Note that $0 < 2/(\nu - 1) < 1$ if $\nu > 3$.

Proof: As in Hobert's calculation, the required expectation is calculated using iterative expectation; we suppress conditioning on the observed data throughout. Also following Hobert, we begin with general n. Namely,

$$\mathrm{E}\left[\sum_{i=1}^{n}(y_i - \mu)^2 \,\middle|\, \sigma^2, q, \mu', (\sigma^2)'\right] = \frac{n\sigma^2}{\sum_{i=1}^n q_i} + \sum_{i=1}^{n}(y_i - \hat{\mu})^2,$$

and

$$\mathrm{E}\left[\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 \;\Big|\; q,\mu',(\sigma^2)'\right] = \frac{n}{\sum_{i=1}^n q_i} + (n-1)\frac{\sum_{i=1}^n (y_i-\hat{\mu})^2}{\sum_{i=1}^n q_i\,(y_i-\hat{\mu})^2}.$$

Thus, we need to evaluate

$$\mathrm{E}\left[\frac{n}{\sum_{i=1}^n q_i} + (n-1)\frac{\sum_{i=1}^n (y_i-\hat{\mu})^2}{\sum_{i=1}^n q_i\,(y_i-\hat{\mu})^2} \;\Big|\; \mu',(\sigma^2)'\right],$$

which simplifies to

$$\mathrm{E}\left[\frac{1}{q_1}+\frac{1}{q_2} \;\Big|\; \mu',(\sigma^2)'\right] = \frac{2}{\nu-1}V(\mu',(\sigma^2)') + \frac{4\nu}{\nu-1}$$

when $n = 2$.

□

Inequality (R.1) follows from Propositions A.1 and A.2 along with Theorem 12 of Rosenthal (1995) using $d = 16.45$ and $r = .0775$.

REFERENCES

Barnard, J., McCulloch, R. E., and Meng, X. L. (2000), "Modeling Covariance Matrices in Terms of Standard Deviations and Correlations, With Application to Shrinkage," *Statistica Sinica*, 10, 1281–1311.

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.

Gelman, A. (1992), "Iterative and Non-iterative Simulation Algorithms," *Computing Science and Statistics*, 24, 433–438.

Goodman, J., and Sokal, A. D. (1989), "Multigrid Monte Carlo Method. Conceptual Foundations," *Physical Review D*, 40, 2035–2071.

Kish, L. (1965), *Survey Sampling*, New York: Wiley.

Lange, K., Hunter, D. R., and Yang, I. (2000), "Optimization Transfer Using Surrogate Objective Functions" (with discussion), *The Journal of Computational and Graphical Statistics*, 9, 1–59.

Liu, J. S., and Sabatti, C. (2000), "Generalized Gibbs Sampler and Multigrid Monte Carlo for Bayesian Computation," *Biometrika*, 87, 353–369.

Liu. J. S., and Wong, W. H., and Kong, A. (1995), "Covariance Structure and Convergence Rate of the Gibbs Sampler With Various Scans," *Journal of the Royal Statistical Society*, Ser. B, 57, 157–169.

Liu, J. S., and Wu, Y. N. (1999), "Parameter Expansion Scheme for Data Augmentation," *Journal of the American Statistical Association*, 94, 1264–1274.

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.

——— (1993), "Maximum Likelihood Estimation via the ECM Algorithm: A General Framework," *Biometrika*, 80, 267–278.

Meng, X. L., and Schilling, S. (in press), "Warp Bridge Sampling," revised for *The Journal of Computational and Graphical Statistics*.

Meng, X.-L., and van Dyk, D. A. (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.

——— (1999), "Seeking Efficient Data Augmentation Schemes via Conditional and Marginal Augmentation," *Biometrika*, 86, 301–320.

——— (2001) "Incompatible Gibbs Samplers, Stochastic Stability, and Estimating Total Variation Distance," *in preparation.*

Neyman, J. (1934), "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection," *Journal of the Royal Statistical Society*, Ser. A, 97, 558–606; Discussion, 607–625.

Niederreiter, H. (ed) (1995), *Monte Carlo and Quasi Monte Carlo*, New York: Springer-Verlag.

——— (1998), *Monte Carlo and Quasi Monte Carlo Method*, New York: Springer-Verlag.

——— (1999), *Monte-Carlo and Quasi-Monte Carlo Methods*, New York: Springer-Verlag.

——— (2001), *Monte Carlo and Quasi-Monte Carlo Methods*, New York: Springer-Verlag.

Orchard, T., and Woodbury, M. A. (1972), "A Missing Information Principle: Theory and Application," in *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, 1, pp. 697–715.

Pilla, R. S., and Lindsay, B. G. (in press), "Alternative EM Methods in High Dimensional Finite Mixtures," *Biometrika*, to appear.

Rosenthal, J. S. (1995), "Minorization Conditions and Convergence Rates of Markov Chain Monte Carlo," *Journal of the American Statistical Association*, 90, 558–566.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions" (with discussion), *The Annals of Statistics*, 22, 1701–1762.

van Dyk, D. A. (2000), "Nesting EM Algorithms for Computational Efficiency," *Statistica Sinica*, 10, 203–226

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.