


DISCUSSION ARTICLE

Discussion

Richard A. LEVINE

I have enjoyed the opportunity this discussion afforded to restudy the valuable contributions of Meng and van Dyk to data augmentation methods. The current article, as the next installment, and in some sense culmination, of their study of deterministic data augmentation, the EM algorithm, in Meng and van Dyk (1997, 1998) and continuing with stochastic data augmentation, efficient MCMC sampling schemes, in Meng and van Dyk (1999), equips us with more essential elements for the data augmentation toolkit, whether we are students of the theory, algorithmic designers, practical users, or patrons of the art.

"The Art of Data Augmentation" (the article) studies the stochastic algorithm side of data augmentation, most notably represented by MCMC methods. Besides suggesting many useful approaches to designing MCMC schemes through efficient data augmentation, van Dyk and Meng have brought to light three important themes in the art of data augmentation (as a craft).

1. Decision making: weighing statistical efficiency against computational complexity.
2. Positive recurrent subchains of null recurrent Markov chains.
3. Violations of the Markov property in Monte Carlo sampling.

The theme of this discussion is decision theory. I thus focus primarily on the first item: the decision problem of choosing an appropriate data augmentation scheme for the problem at hand. I will comment only briefly on the equally important, but less amenable to general rules and guidelines, latter two items.

Thematic exposition: The tremendous improvement in computational power over the past decade has led to the considerable breadth of application of MCMC methods and, consequently, an explosion in MCMC methodological developments and algorithms. The MCMC practitioner is thus left with the daunting task of choosing the best (or even a reasonable) MCMC sampler for the statistical analysis of interest. The decision theoretic problem of selecting from amongst a wide variety of MCMC algorithms for a given application may be approached from three angles: input, implementation, and output.

Richard A. Levine is Assistant Professor, Department of Statistics, 373 Kerr Hall, University of California, Davis, CA 95616 (E-mail: ralevine@ucdavis.edu).

Journal of Computational and Graphical Statistics, Volume 10, Number 1, Pages 51–58

- *Input*: Starting information or initial input for an algorithm consists of, for example, initial seeds or guesses, distributional assumptions such as curvature, peaks, or tail probabilities, and physical properties of the system of study. Depending on information available to the practitioner, some MCMC procedures may be preferable to others.
- *Implementation*: The implementation and modifications of an algorithm will greatly affect both the speed and the cost of simulating random variates. For example, alternative MCMC implementations may better take advantage of specific structures or characteristics of the distribution or model of interest than competing algorithms.
- *Output*: The adaptation of output may constrain the use of MCMC methods. Inferences drawn from the output of a given algorithm may be better at answering the questions of interest than competing inferential procedures evaluated from the same output or output of alternative algorithms.

As the artistic half of my brain lies in the musical arts, this discussion will play out in a three-movement concerto-style format. The opening movement, *allegro con brio*, develops an energetic exposition of the decision theory theme, providing an initial statement of the theme and illustrating its use in implementing random scan Gibbs samplers. The second movement carries the decision theory theme to efficient data augmentation. The *adagio* portion, as the term suggests, highlights the beauty of the work of van Dyk and Meng (abbreviated vDM from here onward) from a decision theoretic point of view. The movement ends with a faster, witty *andantino* section varying the theme to shed light on issues and questions about efficient augmentation. The third and final movement, *vivace*, plays excitedly, but briefly, on the Markov chain theory theme. This short movement suggests not an end to the music, but a beginning of more concertos, and perhaps symphonies, on the art of data augmentation.

1. *Allegro con brio*: RANDOM SCAN GIBBS SAMPLERS

In this section, I introduce decision-making in an MCMC context through some recent work in implementing random scan Gibbs samplers (for details see Levine and Casella 2000). Though the focus of this work is slightly different than efficient data augmentation, besides "tooting our own horn," it provides for a simple illustration of the decision theory framework. Furthermore, the common goal of our work and that of vDM highlights interesting questions and observations about efficient data augmentation which I discuss in Section 2.

Assume we are interested in constructing a Gibbs sampler to generate samples from the posterior distribution $\pi(\theta|Y_{\text{obs}})$ where $\theta = (\theta_1, \ldots, \theta_d)$ is a d-dimensional random vector. One question of interest then is how to choose the order in which to generate variates from each full conditional distribution such that we obtain samples from the distribution $\pi(\theta|Y_{\text{obs}})$? The seminal Gibbs sampling paper by Geman and Geman (1984) suggests a *random scan* strategy by which we randomly choose the coordinate(s) of θ to update. The random scan provides much flexibility in choice of sampling strategy. In particular,

if one coordinate of θ is more difficult to describe, say less precise or more variable, the random scan allows us to visit that coordinate more often (Liu, Wong, and Kong 1995). This observation is analogous to a comparison of data augmentation schemes using the Bayesian fraction of missing information as mentioned by vDM.

Standard application of the random scan Gibbs sampler updates each coordinate of θ with equal probability each iteration of the Gibbs sampler. Although this updating scheme may seem "fair," it is counter to our intuition of visiting more variable components more often; that is, using unequal visitation or *selection* probabilities. The question then is, how do we choose a set of probabilities from which we may decide which coordinate of θ to update each iteration of the random scan Gibbs sampler? And when does this set of probabilities differ markedly from equal selection probabilities of $1/d$ for each coordinate, each iteration of the Gibbs sampler? In general, how do we construct the optimal random scan Gibbs sampler in terms of these selection probabilities?

Levine and Casella (2000) address these questions from a decision theoretic framework. Let $\alpha = (\alpha_1, \ldots, \alpha_d)$ denote the probabilities of updating the components of θ in an iteration of the random scan Gibbs sampler. We require $0 < \alpha_i < 1$ for all i and $\sum \alpha_i = 1$ for an ergodic chain. One method for choosing an optimal set of selection probabilities, α, is to minimize the convergence rate of the induced Markov chain. However, as stated by vDM, this criterion is difficult to apply in practice. Statistical decision theory suggests an alternative, more practical, objective function.

Suppose interest lies in estimating $\mu = E_\pi(h(\theta))$ where $h \in L^2(\pi)$. The natural estimator for μ is the sample mean $\hat{\mu}$ based on the samples generated by the random scan Gibbs sampler. From a decision theoretic point of view, the best sampling strategy will generate variates $\theta^{(1)}, \ldots, \theta^{(t)}$ such that $\hat{\mu}$ is as close to $\mu = E_\pi(h(\theta))$ as possible on average. Under squared error loss as the measure of closeness, the best scan is chosen by minimizing the asymptotic risk

$$R(\alpha, h) = \text{var}_\pi\left(h(\theta)\right) + 2 \lim_{t\to\infty} \sum_{i=1}^{t-1} \left(\frac{t-i}{t}\right) \text{cov}\left(h(\theta^{(0)}), h(\theta^{(i)})\right). \tag{1.1}$$

Note that the estimator in the classic decision theoretic sense is the sampling scheme, say δ_α, which is characterized by α. The parameter of interest is in effect the function h. Therefore, the scan δ_α with smallest risk for every function $h \in L^2(\pi)$ is desired. Any action taken as a consequence of the risk function will apply to the variables α and h. Minimization of this risk function over α is equivalent to a limiting risk efficiency analysis as described by Lehmann and Casella (1998, chap. 6).

Unfortunately, a nontrivial "estimator" α which minimizes the risk uniformly in h does not exist (Lehmann and Casella 1998, sec. 1.1). However, we may protect against the worst possible function $h \in L^2(\pi)$ by minimizing the maximum risk over this function space. Mathematically, the solution to

$$\inf_{\alpha} \sup_{h\in L^2(\pi)} R(\alpha, h) \tag{1.2}$$

is sought. The value of α that minimizes the maximum risk corresponds to the *minimax* scan.

Of course, in practice, finding the minimax scan with respect to (1.2) is difficult and at the least computationally intensive. This is where the art of MCMC (data augmentation) enters the picture. Levine and Casella (2000) present a number of schemes for finding, at the least, an approximate minimax scan. The highlight of our constructions is an adaptive scan that estimates/approximates the risk function (1.1) as the chain proceeds. The consequence is a violation of the Markov property. We show that, however, under similar regularity conditions for the ergodic theory of standard random scan Gibbs samplers, the induced chain still has $\pi(\theta \mid Y_{\text{obs}})$ as the stationary distribution.

2. *Adagio/Andantino*: EFFICIENT DATA AUGMENTATION

The similarities between the decision theoretic framework of Section 1 and the autocorrelation criterion of vDM are apparent if we think of the selection probabilities α as the working parameter. Nonetheless, the more general description in Section 1 calls for an analogous development for efficient data augmentation schemes.

Let $\mathcal{P}$ denote the class of data augmentation schemes from which we will choose. In the context of vDM, we may think of this class as a set of prior functions indexed by the working parameter α. Our goal is to find the optimal data augmentation scheme. Following the development of Section 1, the "minimax data augmentation scheme" is the data augmentation scheme $p \in \mathcal{P}$ which is the solution to

$$\inf_{p \in \mathcal{P}} \sup_{h \in L^2(\pi)} R(p,\ h), \tag{2.1}$$

where the risk function $R(p,\ h)$ is as in (1.1).

The three efficient data augmentation schemes of vDM may be delineated as follows. As in vDM, let $\tilde{Y}_{\text{aug}} = \{Y_{\text{obs}},\ \tilde{Y}_{\text{mis}}\}$ with transformed missing data $Y_{\text{mis}} = \mathcal{D}_{\alpha,\theta}(\tilde{Y}_{\text{mis}})$.

- Conditional augmentation: $p \equiv \delta_{A(\theta)}$, the point-mass prior (delta function) at the parameterization $A(\theta) = \mathcal{D}_{\alpha,\theta}$. Optimize with respect to the conditional model $p(Y_{\text{mis}},\ \theta | Y_{\text{obs}}, \alpha)$.
- Marginal augmentation: $p \equiv p(\alpha|\theta)$ fixed (so no optimization required). Risk calculation, nonetheless, is computed with respect to the marginal model $p(Y_{\text{mis}}, \theta | Y_{\text{obs}}) = \int p(Y_{\text{mis}}, \theta | Y_{\text{obs}}, \alpha)\ p(d\alpha)$.
- Combined conditional/marginal augmentation: $p \equiv p(\alpha|\omega, \theta)$. Optimize over $\omega \in \Omega$ with respect to the conditional marginalized model $p(Y_{\text{mis}}, \theta | Y_{\text{obs}}, \omega) = \int p(Y_{\text{mis}}, \theta | Y_{\text{obs}}, \alpha) p(d\alpha|\omega)$.

Again, construction of the minimax scan as in (2.1), although beautifully elegant in theory, is infeasible for all practical purposes. The "art" of efficient data augmentation is obtaining the minimax scan in finite time (i.e., relatively small computational effort). vDM thus choose the working parameter or prior according to one of these three efficient augmentation schemes. Furthermore, they suggest parameterizations/transformations $\mathcal{D}_{\alpha,\theta}$ appropriate for the problem at hand and choose approximations of the objective function (1.1), namely the lag-1 autocorrelation function of linear functions h and the EM criteria

(under a normal approximation for the combined strategy). Note the similar choices by Liu and Wu (1999) as well.

The primary differences between the decision-making strategies of Levine and Casella (2000) and vDM are in the choice of the objective function and the argument over which we optimize. However, we consider the two methods complimentary: Levine and Casella (2000) and vDM consider two different implementation aspects of MCMC routines. An interesting observation too is that the search for efficient data augmentation schemes of vDM also addresses the input (design) component of the decision process.

The minimax strategy opens a number of curiosities about the vDM approach.

- beyond first order approximation (lag-1 autocorrelation) for the objective function;
- optimally efficient data augmentation schemes when functions h are not linear;
- adaptive strategies;
- computational complexity built into the loss function;
- the art of data augmentation: how do we choose the parameterization $\mathcal{D}_{\alpha,\theta}$?

We will detail our thoughts on these questions in the following paragraphs.

The choice of objective function in vDM is implicitly driven by an assumption that functions of interest h are linear in the parameters. This supposition simplifies the search for the optimal strategy and lends credence to the choice of maximum autocorrelation as an objective function. Note, however, that the decision theoretic motivation in Section 1 uses a more general risk function consisting of an infinite sum over all covariance lags. Granted, practical application will involve a finite sum approximation of this risk function. Levine and Casella (2000) suggest using the first two terms of the expansion. Nonetheless, it begs the question of whether consideration of only the lag-1 autocorrelation is sufficient and if not, how many terms are necessary?

Restriction to linear functions h of the parameters is also a reasonable simplifying assumption under which, at the least, nearly optimal augmentation schemes may be constructed. In many circumstances, however, we may be interested in particular functions, or a class of functions, say $\mathcal{H}$ of the parameters. In such settings, we may find the minimax strategy over this class $\mathcal{H}$ or minimize $R(p,\ h)$ over $\mathcal{P}$ for a specific function of interest h.

We are particularly intrigued by the distinction made by vDM between designers and users of the data augmentation algorithms. We agree that design effort may be substantial, but the bottom line is user time once the algorithm is constructed. A tradeoff of increased design time for decreased user time is preferable.

My current research efforts have focused on trying to bridge the gap between design and use of the algorithms; thus the construction of adaptive schemes as briefly mentioned in Section 1. The adaptive schemes allow, in a sense, the user to be part of the algorithm design (or at least the algorithm designs itself as it proceeds). The work of Levine and Casella (2000) may be applied to construction of efficient data augmentation schemes. In particular, in situations where the objective function (the EM criterion or the 1-lag autocorrelation) or the integrals in the marginalized models (as in the combined strategy) are difficult to compute analytically, we may estimate/approximate unknown pieces of these and related quantities using draws from previous generations of the Monte Carlo sample. The data augmentation scheme is thus refined numerically as we approach the stationary distribution.

This idea of adaptive efficient data augmentation is motivated by the comments in vDM Section 9.2 that "it is possible to use different data augmentation schemes for different conditional draws, which can lead to algorithms that converge faster." The adaptive strategy provides an automated routine for choosing the different data augmentation schemes (over the working parameter or working prior) for different conditional draws. In our work with random scan Gibbs sampling schemes, tuning of the selection probabilities adaptively as the algorithm runs does in fact improve the convergence rate of the induced (Markov) chain to the stationary distribution.

In these considerations, I cannot emphasize enough the strategy of vDM to search for algorithms that are simple and fast. Adaptive strategies run the risk of being computationally intensive (so we decrease the design time by increasing user time). The decision process in choosing an "optimal" algorithm must take into account the statistical properties (e.g., convergence, convergence rate, variance of estimators, Bayesian and/or frequentist optimality criteria) *and* the computational properties (e.g., flops, coding time, run time) of the routine.

I agree with vDM that choosing between computational implementability and statistical optimality is part of the "artistic aspect of the search for efficient data augmentation schemes" in that it is very much problem specific. However, I believe quantification of the notion of implementability is possible. Two ideas come to mind in this vein that are worthy of mention and more thinking.

First, the decision theoretic framework suggests an interesting avenue for incorporating computational complexity into the search for optimal designs. In particular, the loss function, upon which the risk (1.1) is based, may include a computational cost term. Though I cannot detail more at this point as I am just starting to look into such ideas from the complexity literature in computer science, it is analogous to adding a sampling cost term to the loss function in (sequential) sampling designs and decisions.

Second, we note that Casella (1996) also suggested a decision-theoretic approach in the search for appropriate MCMC algorithms in practice. Part of his exposition focuses on Rao-Blackwellization of output from MC algorithms. Briefly, conditioning on ancillary information (such as uniform random variate generations in accept-reject routines) improves the precision of the estimators. The choice of conditioning arguments, whether it be full conditional expectations or term-wise conditional expectations, leads to a tradeoff in computational complexity against statistical optimality. In fact, we can nest Rao-Blackwellized estimators in terms of the amount of ancillary information we integrate out (i.e., the number of variables on which we condition) and find the optimal estimator in terms of computational cost and precision. In this sense, the less ancillary information we integrate out, the less precise the estimator, but the less computationally intensive.

I am not sure how to apply these ideas toward the search for efficient data augmentation schemes. However, Rao-Blackwellization provides a statistical principle by which computational complexity may be quantitatively weighed against statistical optimality. The artistic aspect of MC output processing still exists. For example, on what variables do we condition? How do we construct the objective function for choosing the best estimator? What parameters lend themselves to feasible improvements in precision? Perhaps an analogous representation may present itself for quantifying implementability in efficient data augmentation, though the art of data augmentation will always play a role.

3. *Vivace*: MARKOV CHAIN THEORY

The decision process of choosing between Monte Carlo routines is close to my heart, thus the elaborate discussion on these issues in the previous two sections. Two very interesting aspects of vDM worthy of re-emphasis outside this theme are (1) the use of positive recurrent subchains of null or nonpositive recurrent Markov chains and (2) successful violations of the Markov property in Monte Carlo sampling.

vDM show three methods for ensuring a subchain embedded in a nonpositive recurrent Markov chain on a larger space is in fact ergodic: through (1) limiting sequences of "proper" transition kernels; (2) limiting sequences of posterior distributions; and (3) invariant measures over unimodular groups (developed by Liu and Wu 1999). These state-of-the-art methods for MCMC theory and application for complex models, along with the work of Hobert (in press), lead the way to promising work in this area.

The Markov property is not a prerequisite for successful application of Markovian-based Monte Carlo methods. Though the MCMC theory requires special attention, vDM illustrate a sampler in Section 6 that is not Markovian but has the correct stationary distribution. The theory of adaptive Gibbs samplers of Levine and Casella (2000) also induces non-Markovian chains that have the correct stationary distributions. We should thus not be afraid of deviations from the Markov property.

As a final note, we recommend that the text by Robert and Casella (1999) be added to the noteworthy references in Section 1 of vDM. The book provides a nice development of MCMC theory with many neat applications and an extensive reference list.

Coda: I again congratulate Meng and van Dyk for a stimulating and thought-provoking composition. The work is music to our ears. I hope my contribution adds some melodious harmonies and complimentary variations to the themes developed. I look forward to future artistic contributions to the craft of data augmentation. I thank the editor, Andreas Buja, for inviting me to join these discussions.

REFERENCES

Casella, G. (1996), "Statistical Inference and Monte Carlo Algorithms" (with discussion), *Test*, 5, 249–344.

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.

Hobert, J. P. (in press), "Stability Relationships Among the Gibbs Sampler and its Subchains," *Journal of Computational and Graphical Statistics*.

Lehmann, E. L., and Casella, G. (1998), *Theory of Point Estimation* (2nd ed.), New York: Wiley.

Levine, R. A., and Casella, G. (2000), "Optimizing Random Scan Gibbs Samplers," Technical Report, Department of Statistics, University of California, Davis.

Liu, J. S., 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 for Data Augmentation," *Journal of the American Statistical Association*, 94, 1264–1274.

Meng, X., and van Dyk, D. (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.

——— (1998), "Fast EM-type Implementations for Mixed Effects Models," *Journal of the Royal Statistical Society*, Ser. B, 60, 559–578.

Robert, C., and Casella, G. (1999), *Monte Carlo Statistical Methods*, New York: Springer.


Discussion

James P. HOBERT

I am grateful to Andreas Buja for giving me the opportunity to discuss Van Dyk and Meng's "The Art of Data Augmentation," which is hereafter referred to as vDM. I congratulate Van Dyk and Meng on an outstanding article! It is extremely well written, theoretically interesting, and the results are useful from a practical standpoint. From my perspective, the most interesting aspect of vDM is the realization that nonpositive recurrent Markov chains can play an important role in solving practical problems.

My discussion consists of two sections. The univariate version of vDM's Student's t example is examined in detail in Section 1 with the hope of accomplishing three things: (1) providing a slightly different explanation of exactly why it is possible to use a nonpositive recurrent Markov chain in MCMC; (2) setting notation that is required later in the discussion; and (3) pointing out connections to my own work. In Section 2, I demonstrate that it is not too difficult to establish *drift* and *minorization* conditions for vDM's Markov chains. Hence, it is possible to (1) prove that the chains converge at a geometric rate; (2) calculate exactly how much burn-in is necessary; and (3) take advantage of central limit theorems to assess Monte Carlo error through regenerative simulation methods.

1. THE STUDENT'S t EXAMPLE

Assume that $Y_1, \ldots, Y_n$ are iid $t(\nu, \mu, \lambda)$. (By $X \sim t(\nu, \mu, \lambda)$, I mean the density is proportional to $\left[\nu\lambda + (x-\mu)^2\right]^{-(\nu+1)/2}$.) Suppose that ν is known and let $\pi(\mu, \lambda)$ be a prior that yields a proper posterior given by

$$\pi(\mu, \lambda|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\mu, \lambda)\pi(\mu, \lambda)}{m(\boldsymbol{y})}, \tag{1.1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ and $m(\boldsymbol{y})$ is the marginal density of the data. The goal is to sample from this posterior.

J. P. Hobert is Associate Professor, Department of Statistics, University of Florida, Gainesville, FL 32611 (E-mail: jhobert@stat.ufl.edu).

Journal of Computational and Graphical Statistics, Volume 10, Number 1, Pages 59–68

vDM note that if it is assumed that

$$Y_1|q_1 \sim \mathrm{N}\left(\mu, \frac{\alpha\lambda}{q_1}\right) \quad \text{and} \quad q_1 \sim \mathrm{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2\alpha}\right),$$

where $\alpha > 0$, then the marginal distribution of Y_1 is $t(\nu, \mu, \lambda)$. In particular, α disappears when q_1 is integrated out. (By $X \sim \mathrm{Gamma}(a, b)$, I mean the density is proportional to $x^{a-1}e^{-xb}$.) Let (Y_j, q_j), $j = 1, \ldots, n$, be independent random pairs and consider the following hierarchical model

$$\begin{aligned} Y_j|q_j, \mu, \lambda, \alpha &\sim \mathrm{N}\left(\mu, \frac{\alpha\lambda}{q_j}\right) \\ q_j|\alpha &\sim \mathrm{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2\alpha}\right) \\ (\mu, \lambda, \alpha) &\sim \pi(\mu, \lambda)\pi(\alpha), \end{aligned} \tag{1.2}$$

where $\pi(\alpha)$ is an improper prior. The posterior density of $(\boldsymbol{q}, \mu, \lambda, \alpha)$ under this hierarchy is proportional to

$$\left[\prod_{j=1}^{n} f(y_j|q_j, \mu, \lambda, \alpha) f(q_j|\alpha)\right] \pi(\mu, \lambda)\pi(\alpha), \tag{1.3}$$

where $\boldsymbol{q} = (q_1, \ldots, q_n)^T$. Note that once $q_1, \ldots, q_n$ have been integrated out of (1.3), the only remaining α's are in $\pi(\alpha)$. More specifically,

$$\left[\prod_{j=1}^{n} \int f(y_j|q_j, \mu, \lambda, \alpha) f(q_j|\alpha)\, dq_j\right] \pi(\mu, \lambda)\pi(\alpha) = \pi(\alpha)\pi(\mu, \lambda|\boldsymbol{y})m(\boldsymbol{y}), \tag{1.4}$$

where $\pi(\mu, \lambda|\boldsymbol{y})$ and $m(\boldsymbol{y})$ are as in (1.1). Consequently, since $\pi(\alpha)$ is improper, so is the posterior corresponding to the hierarchical model (1.2). Denote the improper posterior density in (1.3) by $\pi^*(\boldsymbol{q}, \mu, \lambda, \alpha|\boldsymbol{y})$. It is important to keep in mind that integrating π^* with respect to $\boldsymbol{q}$ and α yields ∞ and *not* $\pi(\mu, \lambda|\boldsymbol{y})$. This is the reason for the "$*$" superscript.

Equation (1.4) shows that $\int \pi^*(\boldsymbol{q}, \mu, \lambda, \alpha|\boldsymbol{y})\, d\boldsymbol{q} < \infty$. Suppose it is also true that

$$\int \pi^*(\boldsymbol{q}, \mu, \lambda, \alpha|\boldsymbol{y})\, d\mu\, d\lambda\, d\alpha < \infty\,.$$

It is then possible to define two legitimate conditional densities as

$$\pi^*(\boldsymbol{q}|\mu, \lambda, \alpha, \boldsymbol{y}) = \frac{\pi^*(\boldsymbol{q}, \mu, \lambda, \alpha|\boldsymbol{y})}{\int \pi^*(\boldsymbol{q}, \mu, \lambda, \alpha|\boldsymbol{y})\, d\boldsymbol{q}}$$

and

$$\pi^*(\mu, \lambda, \alpha|\boldsymbol{q}, \boldsymbol{y}) = \frac{\pi^*(\boldsymbol{q}, \mu, \lambda, \alpha|\boldsymbol{y})}{\int \pi^*(\boldsymbol{q}, \mu, \lambda, \alpha|\boldsymbol{y})\, d\mu\, d\lambda\, d\alpha}\,.$$

By "legitimate conditional densities" I mean, for example, that for each fixed $(\boldsymbol{q}, \boldsymbol{y})$, the function $\pi^*(\mu, \lambda, \alpha|\boldsymbol{q}, \boldsymbol{y})$ is a legitimate density in the variable (μ, λ, α). Even though there

is no proper joint density that will yield these conditionals (Hobert and Casella 1998), one can still apply the DA algorithm (also known as the two-variable Gibbs sampler) to form a Markov chain. Write the resulting Markov chain as

$$\Phi = \left\{ \left(\boldsymbol{q}_i, (\mu_i, \lambda_i, \alpha_i)^T \right) : \; i = 0, 1, 2, \ldots \right\},$$

where $\boldsymbol{q}_i = (q_{i1}, \ldots, q_{in})^T$. Also define

$$\Phi_1 = \{ \boldsymbol{q}_i : \; i = 0, 1, 2, \ldots \}$$

and

$$\Phi_2 = \left\{ (\mu_i, \lambda_i, \alpha_i)^T : \; i = 0, 1, 2, \ldots \right\},$$

which are also Markov chains.

The improper posterior density, π^*, is invariant for Φ and there are two important consequences of this regarding the behavior of these Markov chains. (I am assuming irreducibility and aperiodicity throughout.) First, Φ, Φ_1, and Φ_2 are all *unstable*; that is, none of the three is *positive recurrent*. Second, it follows from (1.4) that

$$\pi(\alpha)\, \pi(\mu, \lambda | \boldsymbol{y})$$

is an invariant density for Φ_2. Recall that $\pi(\alpha)$ is improper, but $\pi(\mu, \lambda | \boldsymbol{y})$ is a proper posterior. This suggests that the instability of Φ_2 is due to the α component. Moreover, it also suggests that the limiting distribution of the (μ, λ) component could be $\pi(\mu, \lambda | \boldsymbol{y})$, which is precisely the posterior from which a sample is desired.

Let $\Phi_{\text{MA}} = \left\{ (\mu_i, \lambda_i)^T : i = 0, 1, 2, \ldots \right\}$, where MA stands for "marginal augmentation." One might conjecture that if Φ_{MA} were a Markov chain, it would immediately follow that Φ_{MA} is positive recurrent with invariant density $\pi(\mu, \lambda | \boldsymbol{y})$. However, the results in Section 3.4 of Meng and van Dyk (1999) show that this is not the case, and further regularity conditions are needed to reach the conclusion. vDM's Lemma 1 provides checkable sufficient conditions for Φ_{MA} to be a positive recurrent Markov chain with invariant density $\pi(\mu, \lambda | \boldsymbol{y})$. These conditions are satisfied when $\pi(\mu, \lambda) \propto \lambda^{-1}$ and $\pi(\alpha) \propto \alpha^{-1}$. The bottom line is that one can simulate the unstable Markov chain, Φ, and use the (μ, λ) components to explore the proper posterior $\pi(\mu, \lambda | \boldsymbol{y})$. Moreover, this chain mixes much faster than the standard DA algorithm.

More generally, let $g(\boldsymbol{u}, \boldsymbol{v})$ be any non-negative function such that

$$\int g(\boldsymbol{u}, \boldsymbol{v})\, d\boldsymbol{u} \quad \text{and} \quad \int g(\boldsymbol{u}, \boldsymbol{v})\, d\boldsymbol{v}$$

are both finite, but $\int \int g(\boldsymbol{u}, \boldsymbol{v})\, d\boldsymbol{u}\, d\boldsymbol{v} = \infty$. As above, define two conditional densities as

$$f_{U|V}(\boldsymbol{u}|\boldsymbol{v}) = \frac{g(\boldsymbol{u}, \boldsymbol{v})}{\int g(\boldsymbol{u}, \boldsymbol{v})\, d\boldsymbol{u}} \quad \text{and} \quad f_{V|U}(\boldsymbol{v}|\boldsymbol{u}) = \frac{g(\boldsymbol{u}, \boldsymbol{v})}{\int g(\boldsymbol{u}, \boldsymbol{v})\, d\boldsymbol{v}}\,.$$

Let $\psi = \{(\boldsymbol{u}_i, \boldsymbol{v}_i) : i = 0, 1, 2, \ldots\}$ denote the Markov chain arising from application of the DA algorithm with these two conditionals. As above, $\psi_1 = \{\boldsymbol{u}_i : i = 0, 1, 2, \ldots\}$ and $\psi_2 = \{\boldsymbol{v}_i : i = 0, 1, 2, \ldots\}$ are also Markov chains, and all three of these chains

are unstable. Hobert (in press) develops general results concerning stability relationships among these Markov chains and concerning ways in which positive recurrent chains can arise in such situations.

The Student's t example is considered further in the next section. Specifically, *drift* and *minorization* conditions are established for Φ_{MA}.

2. DRIFT AND MINORIZATION FOR Φ_{MA}

It is assumed throughout this section that $\pi(\mu, \lambda) \propto \lambda^{-1}$ and $\pi(\alpha) \propto \alpha^{-1}$. Let (μ', λ') and (μ, λ) denote the current and next states of Φ_{MA}, respectively. Suppressing dependence on the data, the Markov transition density of Φ_{MA} is given by

$$k(\mu, \lambda | \mu', \lambda') = \int \int \pi^*(\mu, \lambda, \alpha | \boldsymbol{q}, \boldsymbol{y}) \, \pi^*(\boldsymbol{q} | \mu', \lambda', \alpha', \boldsymbol{y}) \, d\boldsymbol{q} \, d\alpha \, .$$

(The fact that Φ_{MA} is a Markov chain implies that the right-hand side is not a function of α'.) Meng and van Dyk (1999) described how to simulate from k, and since I will make extensive use of their method, it is repeated here.

Draws from $\pi^*(\boldsymbol{q} | \mu, \lambda, \alpha, \boldsymbol{y})$ and $\pi^*(\mu, \lambda, \alpha | \boldsymbol{q}, \boldsymbol{y})$ can be used to simulate from k. Specifically, given (μ', λ'), draw $\boldsymbol{q}$ from $\pi^*(\boldsymbol{q} | \mu', \lambda', \alpha', \boldsymbol{y})$ where α' is arbitrary. Then draw (α, μ, λ) from $\pi^*(\mu, \lambda, \alpha | \boldsymbol{q}, \boldsymbol{y})$ and take (μ, λ). Sampling from $\pi^*(\boldsymbol{q} | \mu, \lambda, \alpha, \boldsymbol{y})$ and $\pi^*(\mu, \lambda, \alpha | \boldsymbol{q}, \boldsymbol{y})$ is straightforward. Indeed, conditional on (μ, λ, α), the components of $\boldsymbol{q}$ are independent with

$$q_i | \mu, \lambda, \alpha \sim \text{Gamma}\left(\frac{\nu + 1}{2}, \frac{1}{2\alpha}\left(\frac{(y_i - \mu)^2}{\lambda} + \nu\right)\right)$$

for $i = 1, \ldots, n$. Now define

$$e(\boldsymbol{q}, \boldsymbol{y}) = \frac{1}{q_\cdot} \sum_{i=1}^{n} y_i q_i \quad \text{and} \quad v(\boldsymbol{q}, \boldsymbol{y}) = \frac{1}{q_\cdot} \sum_{i=1}^{n} q_i \left[y_i - e(\boldsymbol{q}, \boldsymbol{y})\right]^2 ,$$

where $q_\cdot = \sum_{i=1}^{n} q_i$. Note that $e(\boldsymbol{q}, \boldsymbol{y})$ and $v(\boldsymbol{q}, \boldsymbol{y})$ are the expectation and variance of a discrete random variable taking the values $y_1, \ldots, y_n$ with probabilities $q_1/q_\cdot, \ldots, q_n/q_\cdot$. Simulation from the density of (μ, λ, α) given $\boldsymbol{q}$ can be done sequentially using the following

$$\alpha | \boldsymbol{q} \sim \text{IG}\left(\frac{n\nu}{2}, \frac{\nu q_\cdot}{2}\right),$$

$$\lambda | \alpha, \boldsymbol{q} \sim \text{IG}\left(\frac{n-1}{2}, \frac{q_\cdot v(\boldsymbol{q}, \boldsymbol{y})}{2\alpha}\right),$$

and

$$\mu | \lambda, \alpha, \boldsymbol{q} \sim \text{N}\left(e(\boldsymbol{q}, \boldsymbol{y}), \frac{\alpha\lambda}{q_\cdot}\right) .$$

I use $\text{IG}(a, b)$ to denote the distribution of $1/X$ when $X \sim \text{Gamma}(a, b)$.

Now, let A denote a measurable set in $\mathbb{R} \times \mathbb{R}_+$. For $t \in \{1, 2, 3, \ldots\}$, let $P^t\left[(\mu, \lambda), A\right]$ denote the t-step Markov transition kernel associated with Φ_{MA}; that is,

$$P^t\left[(\mu, \lambda), A\right] = \Pr\left[(\mu_{s+t}, \lambda_{s+t}) \in A \mid (\mu_s, \lambda_s) = (\mu, \lambda)\right]$$

for any $s \in \{1, 2, 3, \ldots\}$. Of course,

$$P^1\left[(\mu', \lambda'), A\right] = \int_A k(\mu, \lambda | \mu', \lambda')\, d(\mu, \lambda).$$

Note that $P^t\left[(\mu_0, \lambda_0), \cdot\right]$ is the probability measure of (μ_t, λ_t) conditional on starting the chain from the point (μ_0, λ_0). Abusing notation a little, let $\pi(\cdot|\boldsymbol{y})$ denote the probability measure corresponding to the posterior density $\pi(\mu, \lambda|\boldsymbol{y})$; for example, $\pi(A|\boldsymbol{y}) = \int_A \pi(\mu, \lambda|\boldsymbol{y})\, d(\mu, \lambda)$.

Since Φ_{MA} satisfies the usual regularity conditions, for any starting value, (μ_0, λ_0),

$$\|P^t\left[(\mu_0, \lambda_0), \cdot\right] - \pi(\cdot|\boldsymbol{y})\| \to 0 \quad \text{as} \quad t \to \infty,$$

where $||\cdot||$ denotes the *total variation* distance. Moreover, the left-hand side is non-increasing in t. The plots in Meng and van Dyk (1999) and in vDM suggest that this convergence occurs quickly. More formally, one would like to be able to say that the convergence is *geometric*; that is, that there exists a constant $0 < \phi < 1$ and a function $M : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}_+$ such that for any (μ_0, λ_0),

$$\|P^t\left[(\mu_0, \lambda_0), \cdot\right] - \pi(\cdot|\boldsymbol{y})\| \leq M(\mu_0, \lambda_0)\phi^t \tag{2.1}$$

for all t.

Establishing that Φ_{MA} is *geometrically ergodic* (i.e. establishing the existence of M and t in (2.1)) provides one with "peace of mind" about the mixing rate of Φ_{MA}, but it is also important from a practical standpoint. Indeed, if one has (or can bound) M and ϕ, then one can make precise statements about how long the chain should be burned-in. Furthermore, (2.1) implies that there are central limit theorems available that can be used to assess the Monte Carlo error of estimates of posterior quantities of interest (Chan and Geyer 1994). See Jones and Hobert (2000a) for an introduction to these ideas.

One can prove that Φ_{MA} is *geometrically ergodic* and, simultaneously, get an upper bound on $M(\mu_0, \lambda_0)\phi^t$ by establishing *drift* and *minorization* conditions. There are several different ways to do this (Meyn and Tweedie 1994; Rosenthal 1995; Roberts and Tweedie 1999). The version I describe here is based on the work of Rosenthal (1995). A drift condition holds if for some function $V : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}_+$, some $0 < \rho < 1$, and some $L < \infty$

$$\mathrm{E}[V(\mu, \lambda)|(\mu', \lambda')] \leq \rho V(\mu', \lambda') + L\,. \tag{2.2}$$

As the notation suggests, this expectation is with respect to $k(\mu, \lambda|\mu', \lambda')$, the Markov transition density. An associated minorization condition holds if for some density function $h(\mu, \lambda)$ (with support $\mathbb{R} \times \mathbb{R}_+$) and some $\varepsilon > 0$

$$k(\mu, \lambda|\mu', \lambda') \geq \varepsilon h(\mu, \lambda) \quad \forall \quad (\mu', \lambda') \in S \tag{2.3}$$

where $S = \{(\mu, \lambda) : V(\mu, \lambda) \leq d\}$. Here d is any number larger than $2L/(1 - \rho)$.

Despite the fact that the Markov transition density of Φ_{MA} is not available in closed form, it is still possible to establish drift and minorization for Φ_{MA} as is now demonstrated. I use the drift function

$$V(\mu, \lambda) = \frac{1}{\lambda} \sum_{i=1}^{n} (y_i - \mu)^2 ,$$

and I begin with the minorization condition. The goal is to find a density, $h(\mu, \lambda)$, and an $\varepsilon > 0$ such that

$$k(\mu, \lambda \mid \mu', \lambda') \geq \varepsilon h(\mu, \lambda) \tag{2.4}$$

for all $(\mu', \lambda') \in S$ where

$$S = \{(\mu, \lambda) : V(\mu, \lambda) \leq d\} = \left\{ (\mu, \lambda) : \frac{1}{\lambda} \sum_{i=1}^{n} (y_i - \mu)^2 \leq d \right\} .$$

(Until the drift condition has been established, d is just considered to be a fixed positive constant.) This is done by constructing a density, $g(\boldsymbol{q})$, with support $\mathbb{R}^n_+$ and an $\varepsilon > 0$ such that

$$\pi^*(\boldsymbol{q} | \mu', \lambda', \alpha', \boldsymbol{y}) \geq \varepsilon g(\boldsymbol{q}) \tag{2.5}$$

for all $(\mu', \lambda') \in S$. (For convenience, α' is taken to be 1.) Then for all $(\mu', \lambda') \in S$,

$$\begin{aligned} k(\mu, \lambda \mid \mu', \lambda') &= \int\int \pi^*(\mu, \lambda, \alpha | \boldsymbol{q}, \boldsymbol{y})\, \pi^*(\boldsymbol{q} | \mu', \lambda', \alpha', \boldsymbol{y})\, d\boldsymbol{q}\, d\alpha \\ &\geq \varepsilon \int\int \pi^*(\mu, \lambda, \alpha | \boldsymbol{q}, \boldsymbol{y})\, g(\boldsymbol{q})\, d\boldsymbol{q}\, d\alpha \\ &= \varepsilon h(\mu, \lambda) . \end{aligned}$$

and (2.4) is established. I now construct $g(\boldsymbol{q})$ and ε of (2.5) with the help of the following lemma, which is proven in Jones and Hobert (2000b) (see also Rosenthal 1996).

Lemma 1. *Let Gamma$(\delta, \beta; x)$ denote a Gamma(δ, β) density evaluated at $x > 0$. If $\delta > 1$, $b > 0$, and $c > 0$ are fixed then, as a function of x,*

$$\inf_{0<\beta<c} \mathit{Gamma}(\delta, b + \beta/2; x) = \begin{cases} \mathit{Gamma}(\delta, b; x) & \text{if } x < x^* \\ \mathit{Gamma}(\delta, b + c/2; x) & \text{if } x > x^*, \end{cases}$$

where

$$x^* = \frac{2\delta}{c} \log\left(1 + \frac{c}{2b}\right) .$$

Here is a formal statement of the minorization condition:

Proposition 1. **(Minorization)** *The Markov transition density $k(\mu, \lambda | \mu', \lambda')$ satisfies the following minorization condition*

$$k(\mu, \lambda | \mu', \lambda') \geq \varepsilon h(\mu, \lambda) \quad \forall\, (\mu', \lambda') \in S$$

where $h(\mu, \lambda)$ *is a density on* $\mathbb{R} \times \mathbb{R}_+$ *given by*

$$h(\mu, \lambda) = \int \int \pi^*(\mu, \lambda, \alpha | \boldsymbol{q}, \boldsymbol{y}) \left[\prod_{i=1}^{n} \frac{g(q_i)}{\int_0^\infty g(q)\, dq} \right] d\boldsymbol{q}\, d\alpha$$

and $\varepsilon = \left(\int_0^\infty g(q)\, dq \right)^n$. *The function* $g(\cdot)$ *is given by*

$$g(q) = \begin{cases} \textit{Gamma}\left(\frac{\nu+1}{2}, \frac{\nu}{2}; q\right) & \textit{if} \quad 0 < q < q^* \\ \textit{Gamma}\left(\frac{\nu+1}{2}, \frac{\nu+d}{2}; q\right) & \textit{if} \quad q > q^* \end{cases}$$

where

$$q^* = \frac{\nu + 1}{d} \log\left(1 + \frac{d}{\nu}\right) .$$

(Note that ε *can be calculated with two calls to the incomplete gamma function.)*

Proof: It is enough to show that for all $(\mu', \lambda') \in S$ we have

$$\pi^*(\boldsymbol{q} | \mu', \lambda', \alpha', \boldsymbol{y}) \geq \prod_{i=1}^{n} g(q_i) .$$

For $i = 1, \ldots, n$, define

$$S_i = \left\{ (\mu, \lambda) : \frac{(y_i - \mu)^2}{\lambda} \leq d \right\} .$$

Clearly, $S \subset S_i$ for $i = 1, \ldots, n$. Recall that $\pi^*(\boldsymbol{q} | \mu', \lambda', \alpha', \boldsymbol{y})$ is a simply product of n Gamma densities. Hence,

$$\begin{aligned} \inf_{(\mu', \lambda') \in S} \pi^*(\boldsymbol{q} | \mu', \lambda', \alpha', \boldsymbol{y}) &= \inf_{(\mu', \lambda') \in S} \left[\prod_{i=1}^{n} \text{Gamma}\left(\frac{\nu+1}{2}, \frac{(y_i - \mu')^2}{2\lambda'} + \frac{\nu}{2}; q_i \right) \right] \\ &\geq \prod_{i=1}^{n} \left[\inf_{(\mu', \lambda') \in S} \text{Gamma}\left(\frac{\nu+1}{2}, \frac{(y_i - \mu')^2}{2\lambda'} + \frac{\nu}{2}; q_i \right) \right] \\ &\geq \prod_{i=1}^{n} \left[\inf_{(\mu', \lambda') \in S_i} \text{Gamma}\left(\frac{\nu+1}{2}, \frac{(y_i - \mu')^2}{2\lambda'} + \frac{\nu}{2}; q_i \right) \right] \end{aligned}$$

and an application of Lemma 1 yields the result.

□

Here is a formal statement of the drift condition:

Proposition 2. **(Drift)** *Let* $V(\mu, \lambda) = \frac{1}{\lambda} \sum_{i=1}^{n} (y_i - \mu)^2$. *If* $n = 2$ *and* $\nu > 1$ *then*

$$\mathrm{E}\left[V(\mu, \lambda) \mid \mu', \lambda' \right] \leq \left[\frac{\nu + 1}{2(\nu - 1)^2} \right] V(\mu', \lambda') + \frac{2\nu^2}{(\nu - 1)^2} .$$

Note that $(\nu + 1)/[2(\nu - 1)^2] < 1$ *as long as* $\nu > (5 + \sqrt{17})/4 \approx 2.28$.

Before proving Proposition 2, a couple of remarks are in order regarding its lack of generality. The sole reason for the restriction to the $n = 2$ case is to keep the notation

somewhat under control. It is definitely possible to generalize the result to the $n > 2$ case (Jones 2001). The restrictions on ν seem a bit tougher to get around, and removing these might require coming up with a better drift function.

Proof: The required expectations are calculated using several layers of iterated expectation. Specifically,

$$\begin{aligned} \mathrm{E}(\cdot \mid \mu', \lambda') &= \mathrm{E}\left[\mathrm{E}(\cdot \mid \boldsymbol{q}) \mid \mu', \lambda'\right] \\ &= \mathrm{E}\left\{\mathrm{E}\left[\mathrm{E}(\cdot \mid \alpha, \boldsymbol{q}) \mid \boldsymbol{q}\right] \mid \mu', \lambda'\right\} \\ &= \mathrm{E}\left(\mathrm{E}\left\{\mathrm{E}\left[\mathrm{E}(\cdot \mid \lambda, \alpha, \boldsymbol{q}) \mid \alpha, \boldsymbol{q}\right] \mid \boldsymbol{q}\right\} \mid \mu', \lambda'\right). \end{aligned}$$

I start off with a general n, and then later specialize to $n = 2$. First,

$$\begin{aligned} \mathrm{E}\left[\sum_{i=1}^{n}(y_i - \mu)^2 \mid \lambda, \alpha, \boldsymbol{q}\right] &= \sum_{i=1}^{n}\mathrm{E}\left[(y_i - \mu)^2 \mid \lambda, \alpha, \boldsymbol{q}\right] \\ &= \sum_{i=1}^{n}\mathrm{var}\left[(y_i - \mu) \mid \lambda, \alpha, \boldsymbol{q}\right] \\ &\quad + \sum_{i=1}^{n}\left\{\mathrm{E}\left[(y_i - \mu) \mid \lambda, \alpha, \boldsymbol{q}\right]\right\}^2 \\ &= \sum_{i=1}^{n}\mathrm{var}\left[\mu \mid \lambda, \alpha, \boldsymbol{q}\right] + \sum_{i=1}^{n}\left\{y_i - \mathrm{E}\left[\mu \mid \lambda, \alpha, \boldsymbol{q}\right]\right\}^2 \\ &= \frac{n\alpha\lambda}{q.} + \sum_{i=1}^{n}\left[y_i - e(\boldsymbol{q}, \boldsymbol{y})\right]^2. \end{aligned}$$

Thus,

$$\begin{aligned} \mathrm{E}\left[\frac{1}{\lambda}\sum_{i=1}^{n}(y_i - \mu)^2 \mid \alpha, \boldsymbol{q}\right] &= \frac{n\alpha}{q.} + \mathrm{E}\left[\lambda^{-1} \mid \alpha, \boldsymbol{q}\right]\sum_{i=1}^{n}\left[y_i - e(\boldsymbol{q}, \boldsymbol{y})\right]^2 \\ &= \frac{n\alpha}{q.} + \frac{\alpha(n-1)}{q.v(\boldsymbol{q}, \boldsymbol{y})}\sum_{i=1}^{n}\left[y_i - e(\boldsymbol{q}, \boldsymbol{y})\right]^2. \end{aligned}$$

Hence, if $n\nu > 2$, then

$$\begin{aligned} \mathrm{E}\left[\frac{1}{\lambda}\sum_{i=1}^{n}(y_i - \mu)^2 \mid \boldsymbol{q}\right] &= \frac{\mathrm{E}(\alpha \mid \boldsymbol{q})}{q.}\left[n + \frac{(n-1)}{v(\boldsymbol{q}, \boldsymbol{y})}\sum_{i=1}^{n}\left[y_i - e(\boldsymbol{q}, \boldsymbol{y})\right]^2\right] \\ &= \frac{\nu}{n\nu - 2}\left[n + \frac{(n-1)}{v(\boldsymbol{q}, \boldsymbol{y})}\sum_{i=1}^{n}\left[y_i - e(\boldsymbol{q}, \boldsymbol{y})\right]^2\right] \\ &= \frac{n\nu}{n\nu - 2} + \frac{\nu(n-1)}{n\nu - 2}\frac{\sum_{i=1}^{n}\left[y_i - e(\boldsymbol{q}, \boldsymbol{y})\right]^2}{v(\boldsymbol{q}, \boldsymbol{y})}. \end{aligned} \tag{2.6}$$

Thus, I need to evaluate (or bound)

$$\mathrm{E}\left[\frac{\sum_{i=1}^{n}\left[y_i - e(\boldsymbol{q}, \boldsymbol{y})\right]^2}{v(\boldsymbol{q}, \boldsymbol{y})} \mid \mu', \lambda'\right]. \tag{2.7}$$

This is where the notation gets out of control in the $n > 2$ case, and hence this is where I specialize to $n = 2$. When $n = 2$, it is easy to show that

$$\frac{\sum_{i=1}^{n} \left[y_i - e(\boldsymbol{q}, \boldsymbol{y})\right]^2}{v(\boldsymbol{q}, \boldsymbol{y})} = \frac{q_1}{q_2} + \frac{q_2}{q_1} .$$

Thus, using the conditional independence of the q_i's, it follows that

$$\begin{aligned}
\mathrm{E}\left[\frac{q_1}{q_2} + \frac{q_2}{q_1} \mid \mu', \lambda'\right] &= \mathrm{E}(q_1 \mid \mu', \lambda')\,\mathrm{E}(q_2^{-1} \mid \mu', \lambda') + \mathrm{E}(q_2 \mid \mu', \lambda')\,\mathrm{E}(q_1^{-1} \mid \mu', \lambda') \\
&= \frac{(\nu+1)}{(\nu-1)}\left[\frac{(y_2-\mu')^2 + \lambda'\nu}{(y_1-\mu')^2 + \lambda'\nu} + \frac{(y_1-\mu')^2 + \lambda'\nu}{(y_2-\mu')^2 + \lambda'\nu}\right] \\
&\leq \frac{(\nu+1)}{(\nu-1)}\left[\frac{(y_1-\mu')^2}{\lambda'\nu} + \frac{(y_2-\mu')^2}{\lambda'\nu} + 2\right] \\
&= \frac{2(\nu+1)}{(\nu-1)} + \frac{(\nu+1)}{\nu(\nu-1)}\, V(\mu', \lambda')
\end{aligned}$$

Combining this with (2.6) yields the result.

□

Together, Propositions 1 and 2 imply that Φ_{MA} is geometrically ergodic when $n = 2$ and $\nu > 2.3$. As mentioned above, Proposition 2 can be extended to the $n > 2$ case, but this requires some tedious analysis of (2.7). It would be interesting to see if similar results hold in the multivariate case.

Here's an example of one of the benefits of establishing (2.2) and (2.3). Suppose that $n = 2$, $\nu = 10$, and that the data are (y_1, y_2). Suppose further that the starting value for Φ_{MA} will be

$$(\mu_0, \lambda_0) = \left(\frac{y_1 + y_2}{2}, \frac{(y_1 - y_2)^2}{2}\right) .$$

Using Propositions 1 and 2 in conjunction with Theorem 12 of Rosenthal (1995) leads to:

$$\|P^t\left[(\mu_0, \lambda_0), \cdot\right] - \pi(\cdot|\boldsymbol{y})\| \leq (.98554)^t + (4.649)(.97704)^t .$$

Hence, $\|P^{335}\left[(\mu_0, \lambda_0), \cdot\right] - \pi(\cdot|\boldsymbol{y})\| < .01$. Thus, the distribution of $(\mu_{335}, \lambda_{335})$ is within .01 of the true posterior in total variation! Therefore, a burn-in of 335 is adequate and, best of all, this obviates *ad hoc* convergence diagnostics! (Just in case the reader wishes to reproduce these results: in Rosenthal's notation, I used $d = 5.94801$ and $r = .034$.)

Other benefits of establishing (2.2) and (2.3) are the availability of central limit theorems and the ability to use regenerative simulation methods for assessing Monte Carlo error (Mykland, Tierney, and Yu 1995; Robert 1995). See Jones and Hobert (2000a) for more explanation and simple examples.

ACKNOWLEDGMENTS

I am grateful to Galin Jones for checking some of my calculations and for helpful suggestions. This research was partially supported by NSF Grant DMS-00-72827.

REFERENCES

Chan, K. S., and Geyer, C. J., (1994), Comment on "Markov Chains for Exploring Posterior Distributions", *The Annals of Statistics*, 22, 1747–1758.

Hobert, J. P. (in press), "Stability Relationships Among the Gibbs Sampler and its Subchains," *Journal of Computational and Graphical Statistics*, 10.

Hobert, J. P., and Casella, G. (1998), "Functional Compatibility, Markov Chains, and Gibbs Sampling with Improper Posteriors," *Journal of Computational and Graphical Statistics*, 7, 42–60.

Jones, G. L. (2001) "Convergence Rates and Monte Carlo Standard Errors for Markov Chain Monte Carlo Algorithms," unpublished Ph.D. dissertation, University of Florida.

Jones, G. L., and Hobert, J. P. (2000a), "Honest Exploration of Intractable Probability Distributions via Markov Chain Monte Carlo," under revision for *Statistical Science*.

——— (2000b), "Upper Bounds on the Distance to Stationarity for the Block Gibbs Sampler for a Hierarchical Random Effects Model," Technical Report, University of Florida.

Meng, X.-L., and van Dyk, D. A. (1999), Seeking Efficient Data Augmentation Schemes Via Conditional and Marginal Augmentation, *Biometrika*, 86, 301–320.

Meyn, S. P., and Tweedie, R. L. (1994), "Computable Bounds for Geometric Convergence Rates of Markov Chains," *The Annals of Applied Probability*, 4, 981–1011.

Mykland, P., and Tierney, L., and Yu, B. (1995), "Regeneration in Markov Chain Samplers," *Journal of the American Statistical Association*, 90, 233–241.

Robert, C. P. (1995), "Convergence Control Methods for Markov Chain Monte Carlo Algorithms," *Statistical Science*, 10, 231–253.

Roberts, G. O., and Tweedie, R. L. (1999), "Bounds on Regeneration Times and Convergence Rates for Markov Chains," *Stochastic Processes and their Applications*, 80, 211–229.

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

——— (1996), "Analysis of the Gibbs Sampler for a Model Related to James–Stein Estimators," *Statistics and Computing*, 6, 269–275.


DISCUSSION ARTICLE

Discussion

Dave HIGDON

Thanks to professors vanDyk and Meng for a very interesting article. I found their explanation of the link between EM augmentation approaches and auxiliary variable methods quite helpful, as were their criteria for selecting a promising marginal augmentation scheme from a family of possibilities. In the following, I make some comments on marginal augmentation and updating in MCMC and talk briefly how auxiliary variables and MCMC play a role in some spatial modeling applications I am currently involved with. Before getting to that, I briefly mention well-known MCMC approaches for unaugmented posteriors. I realize that some of the examples from the article are purely to illustrate its ideas. I mention these straightforward approaches only to remind practitioners that simple Hastings and Metropolis updates often prove fruitful, alleviating the need for augmentation at all.

1. UNAUGMENTED SAMPLERS

Since the thrust of this article is augmentation approaches in conjunction with Gibbs updates (i.e., sampling directly from the appropriate conditional distributions), I want to remind potential MCMC practitioners that using simple Metropolis or Hastings updates for the unaugmented posterior distribution can be quite effective. For example, the univariate probit regression has a posterior density $p(\beta_0, \beta_1|Y)$ that can be updated using univariate single-site Metropolis updates with proposals centered at the current value of each β_k. Plots corresponding to Figure 4 are shown in Figure 1 for this simple Metropolis sampler. Here univariate updates with uniform proposals centered at the current value of the parameter. The width was tuned so that the proposals were accepted around 50% of the time. Note that a Metropolis-based approach to MCMC in binomial models is used in the software program MLwiN (Rasbash et.al. 2000) as described by Browne and Draper (2000), while BUGS (Spiegelhalter, Thomas, Best, and Gilks 1996) also samples from the unaugmented posterior using a Gibbs update with adaptive rejection sampling (Gilks and Wild 1992).

Dave Higdon is Assistant Professor, Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708-0251 (E-mail: higdon@stat.duke.edu).

Journal of Computational and Graphical Statistics, Volume 10, Number 1, Pages 69–74

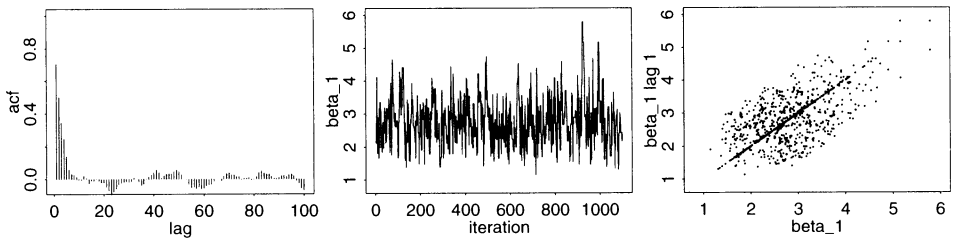

Figure 1. Convergence the Simple Metropolis Sampling on the Simple Probit Regression Example.

For the random effects model in Section 8, one can run a sampler on the unaugmented posterior $p(\beta, \xi, \sigma^2, T)$ by using Gibbs updates for $(\beta, \xi|\sigma^2, T)$ and Hastings updates for σ^2 and T. Drawing the candidate for $1/\sigma^2$ from a $\Gamma(b/\sigma^2, b)$ so that the mean of the proposal is the current value of $1/\sigma^2$ with b a tuning parameter, often works well in practice. Similarly, a $\text{Wishart}_{mb}(mbT^{-1})$ for T^{-1} draw could be used as the Hastings proposal for T^{-1}. A drawback of using a Hastings-based approach on the difficult parameters is that one must tune the proposal distributions so that the chain mixes sufficiently well and it requires evaluations of the posterior density—which can be costly if the covariance matrix Σ is difficult to work with. This is often the case in spatial problems where one must evaluate the determinant of a rather large covariance matrix. Note that the marginal augmentation approaches in this article are based on Gibbs updates which require no tuning, even in the fairly extreme examples of Section 8. This is a very attractive feature.

2. MARGINAL UPDATING

Quite often, data augmentation leads to improved mixing in MCMC because the augmented sampling scheme allows for marginal, or joint, updates where only single site updating was feasible under the original, unaugmented chain. For a simple example, consider the unaugmented parameters $x = (x_1, x_2)^T$ with density $f(x)$. A straightforward MCMC scheme might update x a single component at a time. Whereas augmenting the parameters with auxiliary variable u and conditional distribution $f(u|x)$ can lead to a sampler over the augmented distribution $f(x, u) = f(x)f(u|x)$ which allows for marginal updating of $x|u$.

standard sampler	*augmented sampler*
1. $x_1\|x_2$	1. $(x_1, x_2)\|u$
2. $x_2\|x_1$	2. $u\|(x_1, x_2)$

The marginal update for x can make the augmented sampler far more efficient than the standard sampler, even though the Markov chain is exploring a higher dimensional space.

One key to the success of the sampling approaches used in the applications of Sections 6–8 is that the data augmentation schemes use marginal updating on some of the parameters. For example, the updating scheme for the probit regression example of Section 7.1 steps

through the updates:

1. $\xi|\sigma^2, \beta$
2. $\sigma^2|\xi$
3. $\beta|\sigma^2, \xi$

with Step 2 being the marginal update and Steps 2 and 3 together giving the update of (σ^2, β) given ξ. Since update Step 2 uses a partial conditional for σ^2 (i.e., conditional only on ξ, not (ξ, β)) it is necessary that β be updated directly after σ^2 is updated. For example, an MCMC scheme that cycles through the above updates in 2-1-3 order would not maintain the right stationary distribution. Also it is required that update 3 be a Gibbs update, rather than a Metropolis or Hastings update, since the previous update in Step 2 has marginalized over the current value of β. Furthermore, if simulating directly from the partial conditional for σ^2 in Step 2 were difficult, a Metropolis or Hastings update would be suitable as long as a Gibbs update was used in Step 3. Some ideas from partial conditioning are explored in Appendix 2 of Besag, Green, Higdon, and Mengersen (1995).

Of course, a main aim of vanilla data augmentation is also to allow for a joint update of $\theta|\tilde{Y}_{\text{mis}}, Y_{\text{obs}}$ which was difficult or impossible without augmentation. The marginal augmentation approach described here goes farther by stochastically splitting the missing data $\tilde{Y}_{\text{mis}}$ into two components (α, Y_{mis}) where $Y_{\text{mis}} = \mathcal{D}_\alpha(\tilde{Y}_{\text{mis}})$ and α has prior "density" $p(\alpha|\omega)$. The MCMC updates for the vanilla and marginal augmentation samplers then become

vanilla augmentation	*marginal augmentation*
1. $\tilde{Y}_{\text{mis}}\|\theta$	1. $\tilde{Y}_{\text{mis}}\|\theta$
2. $\theta\|\tilde{Y}_{\text{mis}}$	2. set $Y_{\text{mis}} = \mathcal{D}_\alpha(\tilde{Y}_{\text{mis}})$
	3. $(\alpha, \theta)\|Y_{\text{mis}}$

so that the marginal augmentation scheme is very close to the vanilla scheme with Y_{mis} serving as the missing data. This suggests choosing $p(\alpha|\omega)$ so that Y_{mis} gives minimal information about θ beyond what is given in Y_{obs}. This is very similar in spirit to the suggestion given by vanDyk and Meng of choosing $p(\alpha|\omega)$ so that the dependence between θ and α is maximal since both seek to split $\tilde{Y}_{\text{mis}}$ into two pieces—α, which is informative about θ and Y_{mis}, which is not.

3. CONNECTIONS TO SPATIAL MODELING

Markov chain Monte Carlo and data augmentation come up frequently in fairly large-scale spatial modeling applications that I have been involved with lately. A slightly simplified version of a typical model gives the observed data Y as a function of a spatial process Z plus white noise ϵ

$$Y = f(Z) + \epsilon, \tag{3.1}$$

where $f(\cdot)$ describes how the observations Y are related to the spatial process Z. Often, $f(Z)$ is simply a restriction of Z to n spatial locations. Here (3.1) is a special case of the

mixed model of Section 8. I am curious if the authors have any suggestions for applying marginal sampling in this case. For small or moderate n, one can maginalize over Z and use Metropolis updates for the parameters controlling Z. If posterior realizations of Z are required, they can be obtained post hoc. In large n situations, marginalization is not typically feasible. I suspect marginal augmentation can play a role in such cases.

When n is large, it is often useful to construct a basis representation for Z on a lattice—see Higdon (1998), Nychka, Wikle, and Royle (1999), for example—so that

$$Z = BX,$$

where the columns of the matrix B are the basis functions and the vector X holds the basis coefficients. A simple, often independent, model is specified for X. Again, this leads to a mixed model. However, the dimension of X is typically quite large so that the marginal augmentation scheme of Section 8 may not be feasible. An alternative marginal augmentation scheme may prove fruitful, however. With such a problem, MCMC can be carried out using straightforward single-site updating for the components of X. Alternatively, the components of X can be updated simultaneously using a conjugate gradient-based approach (Wikle, Milliff, Nychka, and Berliner 2001; Nychka, Wikle, and Royle 1999).

In other applications $f(Z)$ may represent a very complicated physical process so that the resulting observations Y will be very indirect and $f(Z)$ cannot be written down, but only computed via simulation code. Such *inverse* problems are common in tomography, hydrology, and geology applications, just to name a few. Here Z cannot be marginalized out so that one must update Z as well as hyperparameters λ_z that control Z in any MCMC based approach. In cases when the data are fairly uninformative about Z, the sampler can be quite slow mixing. This happens because Z is very high dimensional so that the full conditional for λ_z may be very peaked even though the marginal posterior for λ_z is quite spread out. Thus, the high dependence between Z and its hyperparameters λ_z can result in a slow mixing sampler.

At the Center for Multiscale Modeling and Distributed Computing here at Duke, I, along with other researchers, are working on inverse problems in hydrology that require complicated flow simulation code to evaluate $f(Z)$ where Z is a spatial description of the hydrolic conductivity of the aquifer being studied. A straightforward MCMC-based approach uses Hastings updates for subsets of components of Z which require posterior evaluations for the acceptance probabilities (Lee, Higdon, and Bi 2000). Accepting or rejecting each Hastings candidate Z' means that the forward simulator must be run to evaluate the posterior. The computational burden of this MCMC approach can make such a scheme infeasible for large problems.

To combat this problem, we are developing a multiscale augmentation approach that takes advantage of parallel processing. To actually parallelize the flow simulator is a daunting task. However, the simulator can be modified to run at a coarser resolution, taking in a coarser version of the inputs Z_c. This coarse simulator $f_c(Z_c)$ will not attain the accuracy of the original simulator $f(Z)$ for two main reasons. First, information is lost because the simulator now runs at a coarsened spatial resolution. Second, information is lost because fine scale detail is smoothed out in the coarsened input Z_c. However,

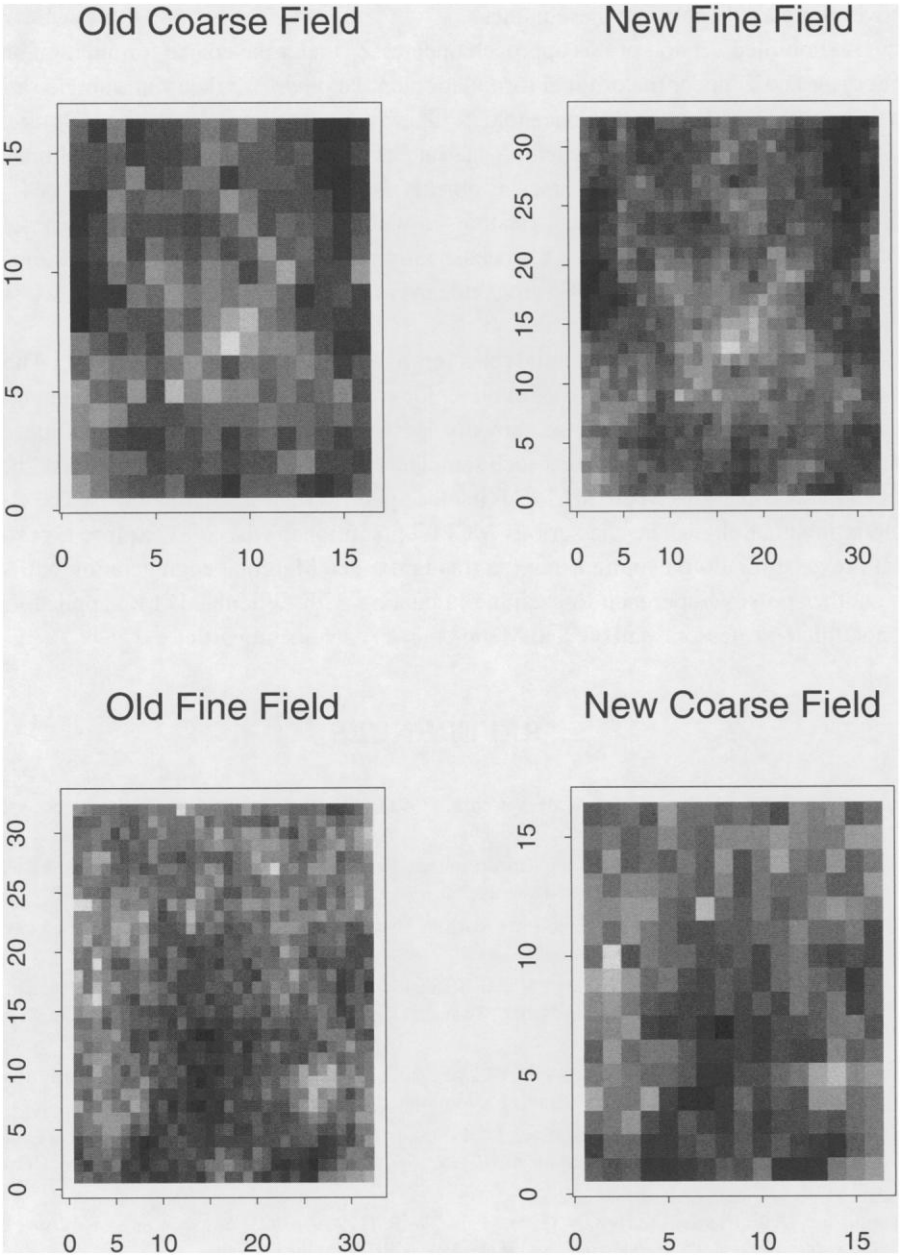

Figure 2. An accepted Hastings step from (z_c, z) on the left to (z_c', z') on the right where information from the coarse formulation is captured in the simulation run at the fine scale.

larger scale properties of the simulator and the inputs are fairly well preserved in the coarser system. Besides the computational savings, a sampler running on the coarsened formulation is typically better mixing since the corresponding posterior has fewer local modes and chain moves more easily between these modes.

A simplified version of this approach updates Z_c under the coarse formulation on one processor and Z under the original formulation on the other. We then run a single sampler over the augmented parameter space for (Z_c, Z) with joint posterior density $p_c(Z_c) \times p(Z)$. In a two-dimensional problem where Z_c has half as many elements as Z in each dimension, $f_c(Z_c)$ is computed about four times as quickly as $f(Z)$. So after four updates of Z_c and a single update of Z we propose a Hastings move from (Z_c, Z) to (Z_c', Z') where Z_c' is a coarsened version of Z and Z' is a stochastically refined version of Z_c. This candidate is accepted according to the usual Hastings rule over the augmented posterior. Figure 2 shows an accepted swap between two scales.

In conclusion, I see two potential roles for marginal augmentation in MCMC. The first role is in standard, commonly used models for which some tried-and-true augmentation schemes have been developed and shown to perform well in a variety of situations. The examples in this article suggest that such samplers may be useful candidates for workhorses in automated situations where MCMC is being applied to such standard models. The second role is in situations where the serious MCMC practitioner wishes to explore a posterior that proves difficult to explore using the standard tools. Marginal augmentation will serve as another possible approach to examine in dealing with difficult MCMC applications. I congratulate Professors vanDyk and Meng on a very interesting article.

REFERENCES

Besag, J., Green, P. J., Higdon, D. M., and Mengersen, K. (1995), "Bayesian Computation and Stochastic Systems" (with discussion), *Statistical Science*, 10, 3–66.

Browne, W. J., and Draper, D. (2000), "Implementation and Performance Issues in the Bayesian and Likelihood Fitting of Multilevel Models," *Computational Statistics*, 15, 391–420.

Higdon, D. (1998), "A Process-Convolution Approach to Modeling Temperatures in the North Atlantic Ocean (disc P191-192)," *Environmental and Ecological Statistics*, 5, 173–190.

Lee, H., Higdon, D. M., and Bi, Z. (2000), "Markov Random Field Models for High-Dimensional Parameters in Simulations of Fluid Flow in Porous Media," Technical Report, Institute of Statistics and Decision Sciences, Duke University.

Nychka, D., Wikle, C., and Royle, J. A. (1999), "Large Spatial Prediction Problems and Nonstationary Random Fields," Technical report, National Center for Atmospheric Research.

Rasbash, J., Browne, W. J., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I., and Lewis, T. (2000), *A User's Guide to MLwiN, Version 2.1*. London: Institute of Education, University of London.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. R. (1996), *BUGS: Bayesian inference Using Gibbs Sampling*, Version 0.5, (version ii), Cambridge, MA: MRC Biostatistics Unit.

Wikle, C. K., Milliff, R. F., Nychka, D., and Berliner, L. M. (2001), "Spatio-Temporal Hierarchical Bayesian Modeling: Tropical Ocean Surface," *Journal of the American Statistical Association*, 97, to appear.

Wild, P., and Gilks, W. R. (1993), "Algorithm AS 287: Adaptive Rejection Sampling From Log-Concave Density Functions," *Applied Statistics*, 42, 701–708.

DISCUSSION ARTICLE

Discussion

Chuanhai LIU

I thank the editor for inviting me to discuss such an excellent overview of the state of the art in data augmentation (DA)-based computation and its applications, some of which my coauthors and I investigated previously (without using the conditional DA approach). For example, the multivariate student-t (Liu 1999); the univariate probit model (Chen and Liu 1999); and a robust extension of linear mixed-effects models using the multivariate t distribution (Pinheiro, Liu, and Wu in press). My discussion here contributes one more example, the multivariate probit regression model. This example shows that CA/PX-DA cannot only make DA/Gibbs faster, but also simplify DA/Gibbs implementations. This is a useful point that is missing in the article by van Dyk and Meng. Along with this example, which is useful in its own right, I also provide my view of the state of the art of DA and its possible impact on future development of statistical methods in terms of its role in statistical computing and in extending the multivariate probit model.

1. THE MULTIVARIATE PROBIT REGRESSION MODEL

The multivariate extension of the probit regression model (Bliss 1935) is known as the Ashford-Sowden model (Ashford and Sowden 1970; Amemiya 1974). Let $y_i = (y_{i,1}, \ldots, y_{i,q})'$ denote the vector of the q binary responses and let $x_i = (x_{i,1}, \ldots, x_{i,p})'$ denote the p-dimensional covariate vector of the ith observation for $i = 1, \ldots, n$. Let $z_i = (z_{i,1}, \ldots, z_{i,q})'$ be the vector of latent variables; that is, the vector of real or hypothetical unobservable quantitative responses that result in observed binary responses y_i. The multivariate probit linear regression model is specified as follows.

1. The latent variables $z_{i,1}, \ldots,$ and $z_{i,q}$ follow the multivariate normal distribution with the location vector $\alpha' x_i$ and variance covariance matrix $\Psi = (\phi_{j,k})$ for $i = 1, \ldots, n$; that is,

$$z_i|\theta \overset{\text{iid}}{\sim} \mathbf{N}_q(\alpha' x_i, \Psi) \qquad (i = 1, \ldots, n), \tag{1.1}$$

   where

$$\theta \in \Theta \equiv \{(\alpha, \Psi) : \alpha \in \mathcal{R}^{pq}, \Psi > 0\}.$$

Chuanhai Liu is a Member of Technical Staff, Bell Laboratories, Lucent Technologies, 600 Mountain Avenue, Room 2C-262, Murray Hill, NJ 07974 (E-mail: liu@research.bell-labs.com).

Journal of Computational and Graphical Statistics, Volume 10, Number 1, Pages 75–81

2. Given $z = (z_1, \ldots, z_n)$, $y_{i,j}$ is assigned to 1 or 0 according to the sign of $z_{i,j}$:

$$y_{i,j}|\{\theta, z\} = I_{\{z_{i,j}>0\}} \qquad (j = 1, \ldots, q;\ i = 1, \ldots, n).$$

As noticed in the literature (e.g., Chib and Greenberg 1998), the parameters θ are not identifiable from the observed-data model. For example, (α, Ψ) and $(\alpha D, D\Psi D)$ are unidentifiable for any positive definite $(q \times q)$ diagonal matrix D. It is common to restrict Ψ to be a correlation matrix; that is, $\Psi = R = (r_{j,k})$ with $r_{j,j} = 1$ for $j = 1, \ldots, q$, and thereby restrict the parameter space to

$$\Theta^{(r)} \equiv \left\{(\alpha^{(r)}, \Psi^{(r)}) : \alpha^{(r)} \in \mathcal{R}^{pq}, \Psi^{(r)} > 0, \Psi^{(r)}_{1,1} = \cdots = \Psi^{(r)}_{q,q} = 1, \nu^{(r)} > 0\right\}.$$

For notational convenience, let

$$(\alpha, R) \equiv \left(\alpha^{(r)}, \Psi^{(r)}\right) \in \Theta^{(r)}.$$

For Bayesian inference, we use the prior distribution specified by

$$\text{pr}(\vec{\alpha'}|R) = \mathbf{N}_{pq}(0, \Delta^{-1} \otimes R), \tag{1.2}$$

and

$$\text{pr}(R) \propto |R|^{-(q+1)/2}, \tag{1.3}$$

where $\vec{\alpha'}$ is the pq-dimensional vector obtained by stacking the p row vectors of the $(p \times q)$ regression coefficient matrix α, and $\Delta = \text{diag}(\delta_1, \ldots, \delta_p) \geq 0$ is a known $(p \times p)$ diagonal matrix. The prior distribution $\text{pr}(R)$ is obtained from Jefferys' prior distribution (e.g., Box and Tiao 1973) $\text{pr}(\Psi) \propto |\Psi|^{-(q+1)/2}$ on an unrestricted covariance matrix Ψ, which implies that $(\Psi_{1,1}, \ldots, \Psi_{q,q})$ and $R = \Psi^{(r)}$ are independent a priori and $\text{pr}(R) \propto |R|^{-(q+1)/2}$. This prior for the regression coefficients and residual variance-covariance matrix has nice properties as discussed in Liu and Sun (2000) and the references therein.

2. A STANDARD DA/GIBBS SAMPLER

Denote by X the $(n \times p)$ design matrix whose ith row is x_i', by Z the $(n \times q)$ latent response matrix whose ith row is z_i, by Y_{com} the complete data $\{y_i, z_i : i = 1, \ldots, n\}$, by Y_{obs} the observed data $\{y_{i,j} : i = 1, \ldots, n; j = 1, \ldots, q\}$, and by Y_{mis} the missing data $\{z_{i,j} : j = 1, \ldots, q; i = 1, \ldots, n\}$. Given Y_{obs} and $\theta^{(r)}$, (y_i, z_i) are independent for all $i = 1, \ldots, n$. Let $z_{i,-j}$ be the subvector of z_i containing all the components of z_i but $z_{i,j}$. A standard Gibbs scheme can be described as follows.

I-step: Impute missing values $Y_{\text{mis}} = \{z_{i,j} : j = 1, \ldots, q; i = 1, \ldots, n\}$ by cycling through the conditional distributions $\text{pr}(z_{i,j}|z_{i,-j}, y_i, \theta^{(r)})$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, q$, each of which is a univariate truncated normal distribution.

P-step: First draw α from $\text{pr}(\alpha|Y_{\text{com}}, R)$ and then R from $\text{pr}(R|Y_{\text{com}}, \alpha)$.

The P-step can be modified to draw (α, R) from $\text{pr}(\alpha, R|Y_{\text{com}})$.

As is commonly known, it is difficult to take draws from posterior distributions of unknown correlation matrices (Chib and Greenberg 1998; Barnard, McCulloch, and Meng

in press). But in what follows, we show that the PX-DA (Liu and Wu 1999) and CA-DA (Liu 1999) computation techniques lead to a simple way of generating the unknown correlation matrix. Moreover, these algorithms converge faster than the standard Gibbs sampler.

3. A SIMPLE AND EFFICIENT DA/GIBBS SAMPLER

Although $\theta \in \Theta$ is not identifiable from the observed data, the parameter $\theta \in \Theta$ is identifiable from the (partially augmented unobservable) complete data $\{(x_i, y_i, z_i) : i = 1, \ldots, n\}$. More specifically, Θ can be viewed as an expanded parameter space with the associated reduction function

$$R(\theta) = \left(\alpha \text{diag}\left(\Psi_{1,1}^{-1/2}, \ldots, \Psi_{q,q}^{-1/2}\right), \right.$$
$$\left. \text{diag}\left(\Psi_{1,1}^{-1/2}, \ldots, \Psi_{q,q}^{-1/2}\right) \Psi \text{diag}\left(\Psi_{1,1}^{-1/2}, \ldots, \Psi_{q,q}^{-1/2}\right)\right).$$

The corresponding expanded model preserves the observed-data model (Liu, Rubin, and Wu 1998; Liu and Wu 1999). Perhaps surprisingly, expanding the parameter space in this way is helpful at least for computational simplicity and efficiency. For example, the technique of PX-DA (Liu and Wu 1999) is readily to apply with making explicit adjustments to the current draws of the latent variables after each P-step because of the I-step involves a cycle of Gibbs steps, as discussed by (Liu 1999).

The intuition behind expanding the parameter space is shown in the derivation of the detailed CA-DA implementation. Drawing R is difficult because the scale parameters—that is, the diagonal elements of the covariance matrix—are fixed. Intuitively, the difficulty can be overcome by making R an unconstrained covariance matrix by borrowing the scales from the latent variables. In other words, the constraints on the covariance matrix in one coordinate system may disappear in another coordinate system. The above intuition leads to considering the following one-to-one mapping from $\{z_{i,j}, \alpha, R\}$ to $\{e_{i,j}, \beta, \Psi\}$ for creating draws of the correlation matrix:

$$\left.\begin{array}{lll} \alpha_{k,j} = \phi_{j,j}^{-1/2}\beta_{k,j} & & (k = 1, \ldots, p; j = 1, \ldots, q), \\ z_{i,j} = \phi_{j,j}^{-1/2} e_{i,j} + \phi_{j,j}^{-1/2}(\beta_{1,j}, \ldots, \beta_{p,j}) x_i & & (i = 1, \ldots, n; j = 1, \ldots, q), \\ r_{j,l} = \phi_{j,j}^{-1/2}\phi_{j,l}\phi_{l,l}^{-1/2} & & (j = 2, \ldots, q, l = 1, \ldots, j-1), \end{array}\right\}, \tag{3.1}$$

where $\Psi = (\phi_{j,l})$ is a $(q \times q)$ positive definite matrix, and

$$\sum_{i=1}^{n} e_{i,j}^2 = 1 \qquad (j = 1, \ldots, q).$$

The steps that draw $z_{i,j}$ and α implicitly draw $\{e_{i,j} : i = 1, \ldots, n; j = 1, \ldots, q\}$, $\beta = (\beta_{k,j})$, and $(\phi_{1,1}, \ldots, \phi_{q,q})$ because

$$\sum_{i=1}^{n} \left(z_{i,j} - (\alpha_{1,j}, \ldots, \alpha_{p,j}) x_i\right)^2 = \phi_{j,j}^{-1} \sum_{i=1}^{n} e_{i,j}^2 = \phi_{j,j}^{-1} \qquad (j = 1, \ldots, q).$$

Thus, drawing R and $(\phi_{1,1}, \ldots, \phi_{q,q})$ is equivalent to drawing Ψ from $\text{pr}(\Psi | \{e_{i,j}\}, \beta)$. From the likelihood function (1.1), the prior distribution ((1.2) and (1.3)), and the Jacobean

of the transformation (3.1), which is given in the Appendix, we have

$$\text{pr}(\Psi|\{e_{i,j}\},\beta) \propto |\Psi|^{-(n+p+q+1)/2} \exp\left\{-\frac{1}{2}\text{tr}\Psi^{-1}\text{S}\right\},$$

where

$$S = \sum_{i=1}^{n} e_i e_i' + \sum_{i=1}^{p} \delta_i \beta_i \beta_i' \quad \text{and} \quad \beta_i = (\beta_{i,1}, \ldots, \beta_{i,q});$$

that is, $\text{pr}(\Psi|\{e_{i,j}\},\beta)$ is the inverse Wishart distribution (e.g., Box and Tiao 1973). Thus, it is simple to draw Ψ, and thereby R. The draws of $z_{i,j}$ and α are then adjusted based on the current draws of $\phi_{1,1}, \ldots,$ and $\phi_{q,q}$ using the transformation (3.1).

Note that for all $i = 1, \ldots, n$ and $j = 1, \ldots, q$ the adjusted values of $z_{i,j}$ are consistent with Y_{obs} because the adjusted value of $z_{i,j}$ has the same sign as its unadjusted value. By *re*-drawing the scale parameters $\phi_{1,1}, \ldots,$ and $\phi_{q,q}$, this MCMC sampling scheme contains a way to adjust the current draws of the missing data and the regression coefficient, and therefore has a faster rate of convergence than the standard Gibbs sampler. But, what is most interesting is that this scheme avoids drawing the correlation matrix directly, which ultimately makes it much simpler.

4. COMPUTING

Chambers (1999, 2000) provided an excellent overview of the current state of the art in statistical computing. Currently, the Omega project (Chambers and Temple Lang 1999; Temple Lang 2000; http://www.omegahat.org) has developed software for statistical computing based on distributed computing, component-based software, and the object-orientation and other features of the Java language. Using a similar Java-based software design, the simplicity of DA schemes can be demonstrated as follows (the details of a complete implementation are omitted).

First, we define DataAugmentation abstractly (as a Java interface), so that development of common tools such as a convergence monitor and posterior estimation can be independent of the implementation of the DataAugmentation algorithm. This DA interface can be simply specified as the Java interface

```
public interface DataAugmentation  {
       public void imputationStep();
       public void posteriorStep();
}
```

along with some possible utility methods/functions to access the data and parameters. Software that uses the DA method is declared to refer to this interface, making it independent of specific Java classes that implement the interface.

Second, we can proceed to implement DataAugmentation with the method/function `posteriorStep()` left unimplemented:

```
public abstract class MultiProbit implements DataAugmentation  {
       public void imputationStep() {
                //detailed implementation of the I-step goes here
       }
       public void posteriorStep();// not implemented here\\
}
```

The method of Chib and Greenberg (1998) can be used to implement `posterior Step()` in order to have a complete implementation of DataAugmentation. Thus, in Java we may have the following.

```
public class MultiProbitChibGreenberg extends MultiProbit  {
     public void posteriorStep() {
               // detailed implementation of the P-step goes here
     }
}
```

When a data augmentation scheme has a different P-step, we can implement it by simply extending MultiProbit or MultiProbitChibGreenberg.

```
public class MultiProbitEfficient extends MultiProbit  {
      public void posteriorStep() {
               // detailed implementation of the P-step goes here
       }
}
```

More efficient DA schemes, when available, can be implemented in the same fashion.

The implication of this example is that new data schemes that modify only the P-step are in general simpler to implement than those that modify the I-step. From a similar analogy made by Rubin (1997) in the context of statistics; that is, the I-step corresponds to data collection and the P-step corresponds to data analysis, the data analysis tool (the P-step) is likely to be changed accordingly when the design for collecting data (the I-step) is changed. Of course, one would like to implement a simpler and more efficient algorithm than the current algorithm even if both the I-step and P-step have to be re-implemented. Thus, it is interesting to look for new DA algorithms for the multivariate probit model based on new ideas, such as conditional DA.

5. EXTENSIONS

The current advances in statistical computing, especially MCMC methods for Bayesian methods, allow new statistical models to be used as data analysis tools. For example, replacing the normal distribution with a student-t distribution with a small number of degrees of freedom leads to a robust probability model (Liu 2000), which can also be viewed as an approximation to logistic regression model (Mudholkar and George 1978; Albert and Chib 1993). Alternatively, the logistic model can be regarded as an approximation to the probit model (Grizzle 1971) or a special case of its extension. Accordingly, the extension of the multivariate probit model can be obtained by replacing the multivariate normal distri-

bution with the multivariate student-t distribution. Using the associated MCMC sampling schemes is also straightforward. Further extensions can be made to handle multivariate ordinal responses (work in progress with Andrew Gelman and Rostislav Protassov). The implication is that the study on simple and efficient DA algorithms is expected to provide new statistical models for data analysis.

APPENDIX

Let $e_i = (e_{i,1}, \ldots, e_{i,q})'$ for $i = 1, \ldots, n$, and let $\alpha_j = (\alpha_{1,j}, \ldots, \alpha_{p,j})'$ and $\beta_j = (\beta_{1,j}, \ldots, \beta_{p,j})'$ for $j = 1, \ldots, q$. It is straightforward to show that

$$\frac{\partial(\alpha_1, \ldots, \alpha_q, z_1, \ldots, z_n)'}{\partial(\phi_{2,1}, \ldots, \phi_{q,1}, \ldots, \phi_{q,q-1})} = \mathbf{0},$$

$$\left| \frac{\partial(r_{2,1}, \ldots, r_{q,1}, \ldots, r_{q,q-1})'}{\partial(\phi_{2,1}, \ldots, \phi_{q,1}, \ldots, \phi_{q,q-1})'} \right|_+ = \prod_{j=1}^{q} \phi_{j,j}^{-(q-1)/2},$$

$$\frac{\partial(\phi_{1,1}^{-1/2}, \ldots, \phi_{q,q}^{-1/2})'}{\partial(\phi_{1,1}, \ldots, \phi_{q,q})} = 2^{-q} \prod_{j=1}^{q} \phi_{j,j}^{-3/2},$$

$$\frac{\partial(\alpha_1, \ldots, \alpha_q, z_1, \ldots, z_n)'}{\partial(\beta_1, \ldots, \beta_q)} = P_1 \text{Diag}(\phi_{1,1}^{-1/2} I_p, \ldots, \phi_{q,q}^{-1/2} I_p),$$

$$\frac{\partial(\alpha_1, \ldots, \alpha_q, z_1, \ldots, z_n)'}{\partial(e_1, \ldots, e_{n-1})} = P_2 \left(I_{n-1} \otimes \text{Diag}(\phi_{1,1}^{-1/2}, \ldots, \phi_{q,q}^{-1/2}) \right),$$

and

$$\frac{\partial(\alpha_1, \ldots, \alpha_q, z_1, \ldots, z_n)'}{\partial(\phi_{1,1}^{-1/2}, \ldots, \phi_{q,q}^{-1/2})} = P_3,$$

where I_p is the p-dimensional identity matrix and all the matrices P_1, P_2, and P_3 are independent of Ψ. Thus, we have for the Jacobean of the transformation (3.1), as a function of Ψ,

$$\begin{aligned}
&\left| \frac{\partial(\alpha, z_1, \ldots, z_n, R)}{\partial(\beta, e_1, \ldots, e_{n-1}, \Psi)} \right|_+ \\
&\quad = 2^{-q} \prod_{j=1}^{q} \phi_{j,j}^{-(q+2)/2} \left| \frac{\partial(\alpha_1, \ldots, \alpha_q, z_1, \ldots, z_{n-1}, z_n)'}{\partial(\beta_1, \ldots, \beta_q, e_1, \ldots, e_{n-1}, \phi_{1,1}^{-1/2}, \ldots, \phi_{q,q}^{-1/2})} \right|_+ \\
&\quad = 2^{-q} \prod_{j=1}^{q} \phi_{j,j}^{-(q+2)/2} \prod_{j=1}^{q} \phi_{j,j}^{-p/2} \prod_{j=1}^{q} \phi_{j,j}^{-(n-1)/2} \left| P_1 \vdots P_2 \vdots P_3 \right|_+ \\
&\quad \propto \prod_{j=1}^{q} \phi_{j,j}^{-(q+2)/2} \prod_{j=1}^{q} \phi_{j,j}^{-p/2} \prod_{j=1}^{q} \phi_{j,j}^{-(n-1)/2} \\
&\quad = \prod_{j=1}^{q} \phi_{j,j}^{-(n+q+p+1)/2}.
\end{aligned}$$

ACKNOWLEDGMENTS

I thank John Chambers, Ming-Hui Chen, Diane Lambert, Rostislav Protassov, and Duncan Temple Lang for helpful discussions and comments.

REFERENCES

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

Amemiya, T. (1974), "Bivariate Probit Analysis: Minimum Chi-Square Methods," *Journal of the American Statistical Association*, 69, 940–944.

Ashford, J. R., and Sowden, R. R. (1970), "Multi-Variate Probit Analysis," *Biometrics*, 26, 535–546.

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

Bliss, C. J. (1935), "The Calculation of the Dosage-Mortality Curve," *Annals of Applied Biology*, 22, 307–330.

Box, G. E. P., and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, New York: Wiley.

Chambers, J. (1999), "Computing With Data: Concepts and Challenges," *The American Statistician*, 53, 73–84.

——— (2000). "Users, Programmers, and Statistical Software," *Journal of Computational and Graphical Statistics*, 9, 404–422.

Chambers, J., and Temple Lang, D. (1999), "Omegahat—A Component-Based Statistical Computing Environment," in *Proceedings of the 52nd Session of the ISI*, Bulletin of the International Statistical Institute, Tome LVIII, Book 2, Helsinki, Finland, pp. 125–128.

Chen, M. H., and Liu, C. (1999), Comments on "Simulated Sintering: Markov Chain Monte Carlo With Spaces of Varying Dimensions" by J. S. Liu and C. Sabatti, in *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, New York: Oxford University Press.

Chib, S., and Greenberg, E. (1998), "Analysis of Multivariate Probit Models," *Biometrika*, 85, 347–361.

Grizzle, J. E. (1971), "Multivariate Logit Analysis," *Biometrics*, 27, 1057–1062.

Liu, C. (1999), "Covariance Adjustment for Markov Chain Monte Carlo—A General Framework," Technical Report, Bell Labs. under revision.

——— (2000), "Robit Regression: A Simple Robust Alternative to Logistic and Probit Regression," Technical Report, Bell Labs.

Liu, C., Rubin, D. B., and Wu, Y. (1998), "Parameter Expansion to Accelerate EM: The PX-EM Algorithm," *Biometrika*, 85, 755–770.

Liu, C., and Sun D. X. (2000), "Analysis of Interval Censored Data From Fractionated Experiments Using Covariance Adjustments," *Technometrics*, 42, 353–365.

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

Mudholkar, G. S., and George E. O. (1978), "A Remark on the Shape of the Logistic Distribution," *Biometrika*, 65, 667–668.

Pinheiro J. C., Liu, C., and Wu, Y. N. (in press), "Efficient Algorithms for Robust Estimation in Linear Mixed-Effects Models Using the Multivariate t-Distribution," *Journal of Computational and Graphical Statistics.*

Rubin, D. B. (1997), Comments on "The EM Algorithm—An Old Folk Song Sung to Fast New Tune" by X. L. Meng and van Dyk, *Journal of the Royal Statistical Society*, Ser. B, 59, 541–542.

Temple Lang, D. (2000), "The Omega Project: New Possibilities for Statistical Software," *Journal of Computational and Graphical Statistics*, 9, 423–451.

van Dyk, D., and Meng, X. L. (2001), "The Art of Data Augmentation," *Journal of Computational and Graphical Statistics*, 10, 1–50.

DISCUSSION ARTICLE

Discussion

Gabriel HUERTA, Wenxin JIANG, and Martin A. TANNER

MIXTURES OF TIME SERIES MODELS

This insightful article by van Dyk and Meng (vDM) makes the important point that advances in data augmentation algorithms offer a wide variety of tools for statistical inference. Time series methods are no exception and mixture modeling within this context may help to improve forecasting and to detect changes in structure across time.

The time series modeling approach that we adopt is based on the idea of mixing models through the neural network paradigm known as hierarchical mixtures-of-experts (HME)—see Jordan and Jacobs (1994). The HME approach easily allows for model comparison and permits one to represent the mixture weights as a function of time or other covariables. With the additional hierarchy, it is possible to localize the comparisons to specific *regions* or *regimes*. Furthermore, the defining elements of the mixture do not have to be restricted to a particular class of models permitting very general comparisons. In this comment, parameters are estimated via maximum likelihood using the EM-algorithm—extensions to a full Bayesian approach using MCMC may follow one or more of the many lines outlined by vDM. We see this comment as a call to the Chagalls of this world to use their artist abilities to develop quick mixing stochastic algorithms for this important, yet complex class of HME models.

Let $\{y_t\}_0^n$ be a time series of endogenous or response variables, and $\{\mathbf{x}_t\}_0^n$ be a time series of exogenous variables or covariates. Suppose the main interest is to draw inference on $\{y_t\}_0^n$ conditional on $\{\mathbf{x}_t\}_0^n$. Let the conditional probability density function (pdf) of y_t be $f_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}; \theta)$, where θ is a parameter vector; $\mathcal{X}$ is the σ-field generated by $\{\mathbf{x}_t\}_0^n$, representing the external information; and for each t, $\mathcal{F}_{t-1}$ is the σ-field generated by $\{y_s\}_0^{t-1}$ representing "the previous history" at time $t-1$. Typically, the conditional pdf f_t

Gabriel Huerta is Assistant Professor, Department of Probability and Statistics, Centro de Investigación en Matemáticas, Apartado Postal 402, Guanajuato, Gto. 36000, México (E-mail: ghuerta@cimat.mx). Wenxin Jiang is Assistant Professor, and Martin A. Tanner is Professor, Department of Statistics, Northwestern University, 2006 Sheridan Road, Evanston, IL 60208-4070 (E-mail: wjiang@nwu.edu; mat132@neyman.stats.nwu.edu).

Journal of Computational and Graphical Statistics, Volume 10, Number 1, Pages 82–89

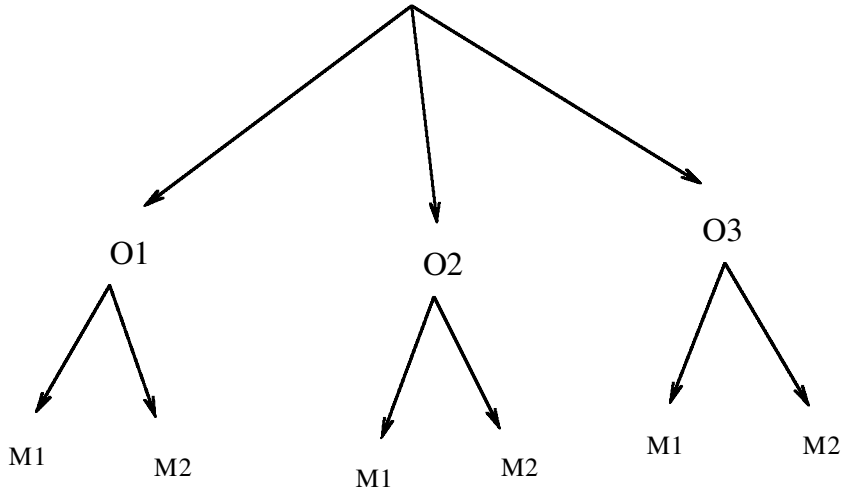

Figure 1. A Graphical Representation of a Two-Layer HME.

is assumed to depend on $\mathcal{X}$ through $\mathbf{x}_t$ only. In HME, the pdf f_t of the response variable is assumed to be a conditional mixture of the pdfs from simpler, well established models. In a time series context, this mixture can be represented by the finite sum

$$f_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}; \theta) = \sum_J \mathrm{g}_t(J|\mathcal{F}_{t-1}, \mathcal{X}; \gamma)\pi_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}, J; \eta), \tag{1.1}$$

where the functions $\mathrm{g}_t(\cdot|\cdot, \cdot; \gamma)$ are the mixture weights; $\pi_t(\cdot|\cdot, \cdot, J; \eta)$ are the conditional pdfs from simpler models each defined by a label J; and γ and η are vectors of sub-parameters from θ.

The simpler models in HME are often referred to as the "experts." In a time series context, one "expert" could be an AR(1) model, another "expert" could be a GARCH(1,1) model or an EGARCH(1,1) model. For example, in a situation where it is not clear whether to use a stochastic or a deterministic trend, one expert could be a *trend-stationary process*, another a *difference-stationary process*. A somewhat simpler situation occurs when all the experts propose a model of the same type; for example, linear autoregressive, but perhaps with different values for the coefficients or for the model order.

Furthermore, the HME models considered have an additional layer designed with the purpose of local time series modeling. The HME partitions the covariate space, which could include time, into O overlapping regions called "overlays." In each overlay, M models are to compete with each other, in the hope that the model most suitable to the specific region is favored by a high weight (see Figure 1). By having multiple overlays, the hierarchical mixture model allows for modeling multiple switching across regions.

Therefore, the expert index J can be expressed as $J = (o, m)$, where the overlay index o takes a value from $\{1, \ldots, O\}$, and the model-type index m from $\{1, \ldots, M\}$. We allow the same type of model m to assume different versions or more specifically different parameter values, at each possible overlay.

The mixing weights are often referred to as "gating functions." They can depend on the previous history, exogenous information (see McCulloch and Tsay 1993), or can exclusively

depend on t. The gating functions may have the form

$$\mathrm{g}_t(o, m|\mathcal{F}_{t-1}, \mathcal{X}; \gamma) = \left\{ \frac{e^{v_o + \mathbf{u}_o^T \mathbf{w}_t}}{\sum_{s=1}^{O} e^{v_s + \mathbf{u}_s^T \mathbf{w}_t}} \right\} \left\{ \frac{e^{v_{m|o} + \mathbf{u}_{m|o}^T \mathbf{w}_t}}{\sum_{l=1}^{M} e^{v_{l|o} + \mathbf{u}_{l|o}^T \mathbf{w}_t}} \right\}, \tag{1.2}$$

where the v's and $\mathbf{u}$'s are parameter components of γ; and $\mathbf{w}_t$ is an "input" at time t, which is measurable with respect to the σ-field induced by $\mathcal{F}_{t-1}$ { $\mathcal{X}$. For example, the input $\mathbf{w}_t$ could be the covariate $\mathbf{x}_t$, the "two-lag" history $(y_{t-1}, y_{t-2})^T$, or exclusively depend on time t.

In the context where one is interested in how the weighting for individual models is assigned across different time periods, $\mathbf{w}_t$ can be taken as (t/n). Therefore, one can adopt the following parametric form for the gating functions:

$$\mathrm{g}_t(o, m|\mathcal{F}_{t-1}, \mathcal{X}; \gamma) = g_{om}(t; \gamma) \equiv \left\{ \frac{e^{v_o + u_o(t/n)}}{\sum_{s=1}^{O} e^{v_s + u_s(t/n)}} \right\} \left\{ \frac{e^{v_{m|o} + u_{m|o}(t/n)}}{\sum_{l=1}^{M} e^{v_{l|o} + u_{l|o}(t/n)}} \right\}. \tag{1.3}$$

Here γ includes all the following components: $v_1, u_1, \ldots, v_{O-1}, u_{O-1}, v_{1|1}, u_{1|1}, \ldots, v_{M-1|1}, u_{M-1|1}, \ldots, v_{M-1|O}, u_{M-1|O}$. For identifiability, we set $v_O = u_O = v_{M|o} = u_{M|o} = 0$ for all $o = 1, \ldots, O$. The free vector of parameters γ in the gating functions automatically determines the location and the "softness" of the splitting of the regions.

Note that this framework defines the two-layer HME architecture of Jordan and Jacobs (1994), where the first layer of gating functions hypothesizes O overlays on the entire time axis, and the second layer of gating functions defines weights for each of the M model types within each overlay. When the input space for the gating functions is time, the hierarchical mixture model can identify the region over which a model or a set of models is (are) dominant in a data-adaptive manner. Thus, the present approach allows for modeling multiple regime switching. Further details of this approach, as well as related asymptotic theory were presented by Huerta, Jiang, and Tanner (2000).

Inference on the parameter θ can be based on the log-likelihood function, conditional on y_0, $\mathcal{X}$ and "averaged" in time, which is

$$\mathcal{L}_n(\cdot) = n^{-1} \sum_{t=1}^{n} \log f_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}; \cdot). \tag{1.4}$$

We denote the maximum likelihood estimate (MLE) of θ as $\hat{\theta} = \arg\max \mathcal{L}_n(\cdot)$. To obtain the MLE, the EM algorithm starts with an initial estimate of the parameters θ^0. Then a sequence $\{\theta^i\}$ is obtained by iterating between the following two steps:

For $i = 0, 1, 2, \ldots,$

E-step: Construct

$$Q^i(\theta) = \sum_{t=1}^{n} \sum_{o,m} h_{om}(t; \theta^i) \log\{\pi_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}, o, m; \eta) \mathrm{g}_t(o, m|\mathcal{F}_{t-1}, \mathcal{X}; \gamma)\}, \tag{1.5}$$

where $\theta = (\gamma, \eta)$, $\theta^i = (\gamma^i, \eta^i)$, $h_{om}(t; \theta^i) = h_{om}(t; \theta)|_{\theta=\theta^i}$, and

$$h_{om}(t; \theta) = \frac{g_{om}(t; \gamma) \pi_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}, o, m; \eta)}{\sum_{s=1}^{O} \sum_{l=1}^{M} g_{sl}(t; \gamma) \pi_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}, s, l; \eta)} \tag{1.6}$$

is the "posterior probability" of choosing the expert (o, m) at time t.

M-step: Find $\theta^{i+1} = \arg\max_\theta Q^i(\theta)$.

Inference is greatly facilitated by the introduction of augmented data, resulting in the fact that the objective function Q^i has the form of a double sum of logarithms, instead of a "sum log sum" typical for the log-likelihood function $\mathcal{L}_n$. For this reason, the maximization of the objective function can be decomposed into a number of smaller maximization problems which involve fewer parameters and usually define "known" maximizations of widely used models. For example, suppose the expert pdf has the form

$$\pi_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}, o, m; \eta) = p_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}, m; \eta_{om}), \tag{1.7}$$

where η is decomposed into a collection of sub-parameter η_{om}, each of which only appears in the pdf of one expert. The parameter η_{om} carries an index o in addition to m to allow one type of model to take different versions (parameters) in different overlays. In such a situation, in the M-step, the maximization over the η_{om}'s and γ can be performed separately. For example, for each o, m, i,

$$\eta_{om}^{i+1} = \arg\max_{\eta_{om}} \sum_{t=1}^{n} h_{om}(t; \theta^i) \log p_t(y_t|\mathcal{F}_{t-1}, \mathcal{X}, m; \eta_{om}), \tag{1.8}$$

which become the "standard" (albeit weighted by the h's) maximum likelihood problem for model type-m.

When the MLE $\hat{\theta}$ is obtained, we are interested in evaluating the relative weighting of each of the M model types at time t. Two estimates are of interest in this regard. One is an empirical Bayes estimate of the posterior probability/weight of model type-m at time t. This is the *conditional* probability regarding the history up to time t and defined by:

$$\hat{P}_t(m|y_t, \mathcal{F}_{t-1}, \mathcal{X}) \equiv \hat{h}_m(t) \equiv \sum_{o=1}^{O} h_{om}(t; \hat{\theta}), \tag{1.9}$$

where $\hat{\theta}$ is the MLE. Another approach for weighting is to consider an empirical Bayes-type estimate of the *unconditional* probability/weight of model m at time t (unconditional on the current process y_t):

$$\hat{P}_t(m|\mathcal{F}_{t-1}, \mathcal{X}) \equiv \hat{g}_m(t) \equiv \sum_{o=1}^{O} g_{om}(t; \hat{\theta}). \tag{1.10}$$

As we shall see in an example, (1.9) can vary point-wise over time due to the conditioning on the specific history of the observations. The second weighting scheme (1.10) is smoother when describing a regional change of preference for model m. The term $\hat{h}_m(t)$ is an estimated "posterior" probability of model m, and $\hat{g}_m(t)$ is the corresponding estimate of the "prior" probability in the sense that the prior probability is not conditional y_t and the posterior probability is conditional on y_t. These estimates are not formal Bayesian priors and posterior probabilities for model m, since we have not assigned any prior $p(\theta)$ to the parameters θ, but instead are estimating the conditional or "unconditional" mixing weights at the MLE.

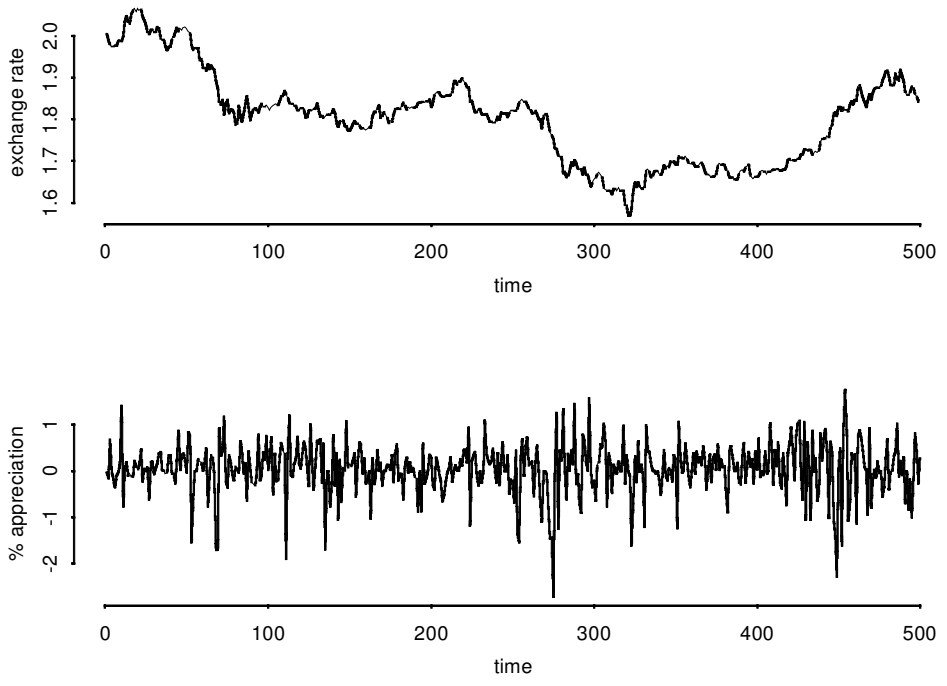

Figure 2. Exchange Rate Data. Top: 500 observations of spot rates between the German mark and U.S. dollar beginning in October 1986. Bottom: Logarithm of the first difference of the spot rates between the German mark and the U.S. dollar.

We consider a financial time series to illustrate this hierarchical mixture defined with GARCH and EGARCH models—see Bollerslev (1986) and Nelson (1991). The series consist of 500 daily observations of exchange rates between the German-mark and the U.S. dollar starting from October 9, 1986. In fact, Figure 2 presents a time plot of the 500 daily *spot rates* characterized by a nonstationary random walk behavior.

Also in the same figure, we present the logarithm of the first difference of spot rates as a function of time, a transformation which is widely used to induce covariance-stationarity and propose parametric models. Assuming that Y_t represents the log of the first difference in spot rates, as discussed in Andersen and Bollerslev (1998), three candidate models are used to obtain inferences on *volatilities* or innovation variances at each time t:

- An AR(1) model simply defined by:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \epsilon_t,$$

  where $\epsilon_t \sim N(0, \sigma^2)$.
- An AR(1)-GARCH(1,1) model represented by

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \epsilon_t,$$

  but now $\epsilon_t \sim N(0, \sigma_t^2)$. That is, the innovation variance can change in time and according to the evolution equation

$$\sigma_t^2 = \theta_0 + \theta_1 \epsilon_{t-1}^2 + \theta_2 \sigma_{t-1}^2,$$

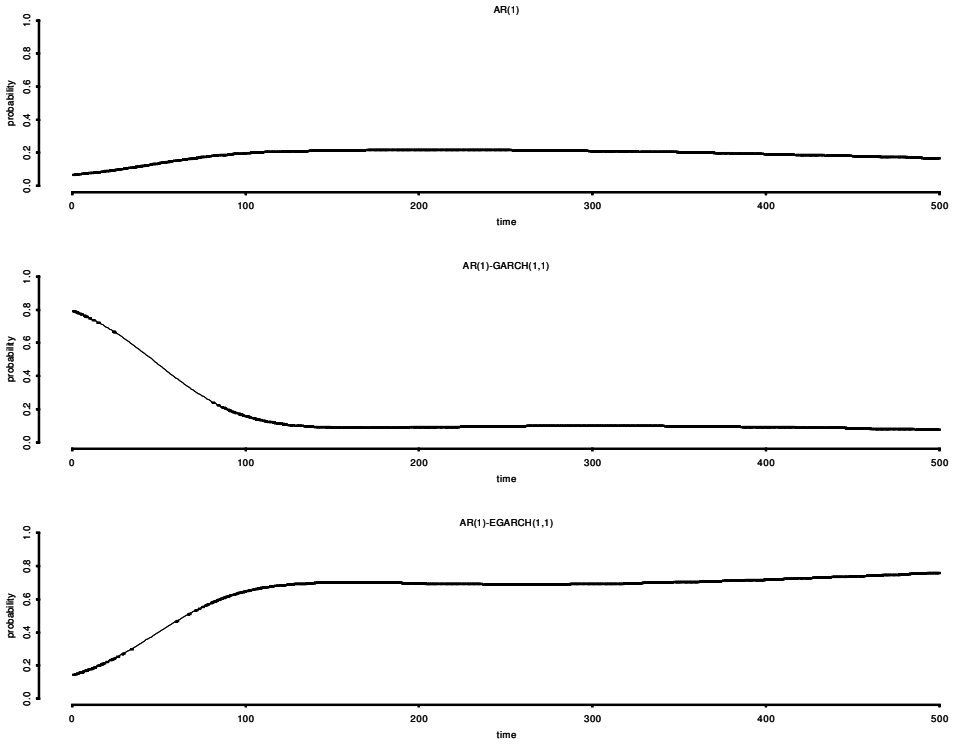

Figure 3. Exchange Rate Data Example. Maximum likelihood estimates of $g_m(t)$ for the three models considered: AR(1), AR(1)-GARCH(1,1), and AR(1)-EGARCH(1,1).

with non-negative parameters θ_0, θ_1 and θ_2.

- Finally, an AR(1)-EGARCH(1,1) is considered, where again

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \epsilon_t$$

but now the evolution in variance is defined in terms of the natural logarithm and the standardized innovations $z_t = \epsilon_t/\sigma_t$ through the expression

$$\log(\sigma_t^2) = \beta_0 + \beta_1 z_{t-1} + \beta_2(|z_{t-1}| - \sqrt{2/\pi}) + \beta_3 \log(\sigma_{t-1}^2),$$

with no restrictions on the parameters β_0, β_1, β_2, and β_3.

Within an HME framework taking $M = 3$: $\pi_t|_{o,m=1}$ denotes the pdf of Y_t given the history based on an AR(1); $\pi_t|_{o,m=2}$ the same pdf with an AR(1)-GARCH(1,1); and $\pi_t|_{o,m=3}$ denotes the pdf with an AR(1)-EGARCH(1,1). As before, the index o is added to model parameters and our initial exploration is based on a value of $O = 2$; that is, allowing for two overlays. We ran the EM algorithm with 20 different starting points. Parameters for the pdfs $\pi_t|_{o,m}$ were initialized at the individual model MLE's and initial parameters for the gating functions were generated from uniform distributions. Each EM was run for 500 iterations and solutions were ranked using the log-likelihood function $\mathcal{L}_n(\cdot)$.

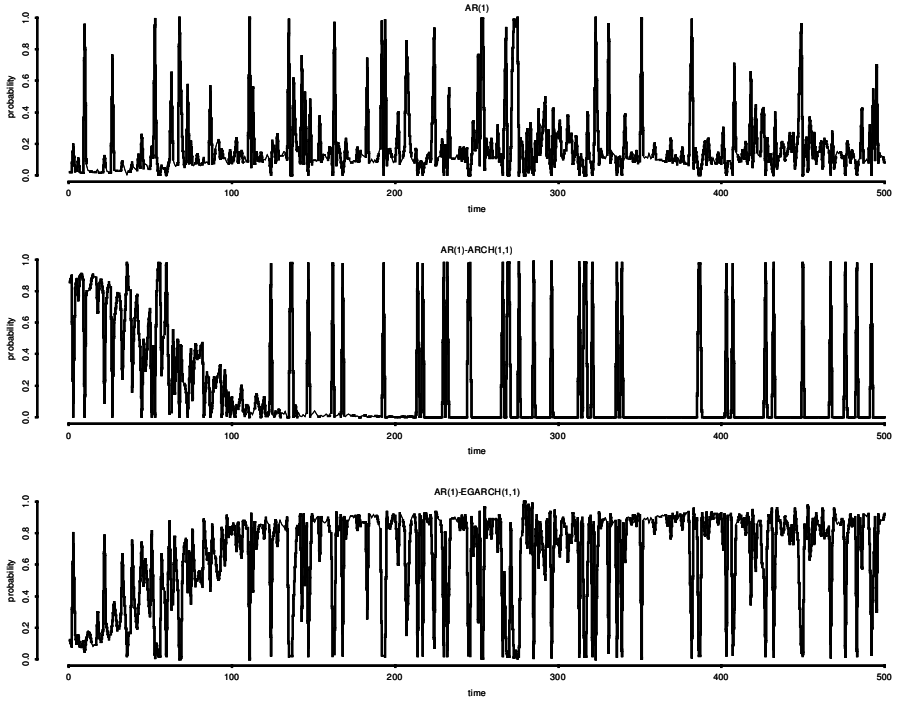

Figure 4. Exchange Rate Data Example. Maximum likelihood estimates of $h_m(t)$ for the 3 models considered: AR(1), AR(1)-GARCH(1,1), and AR(1)-EGARCH(1,1).

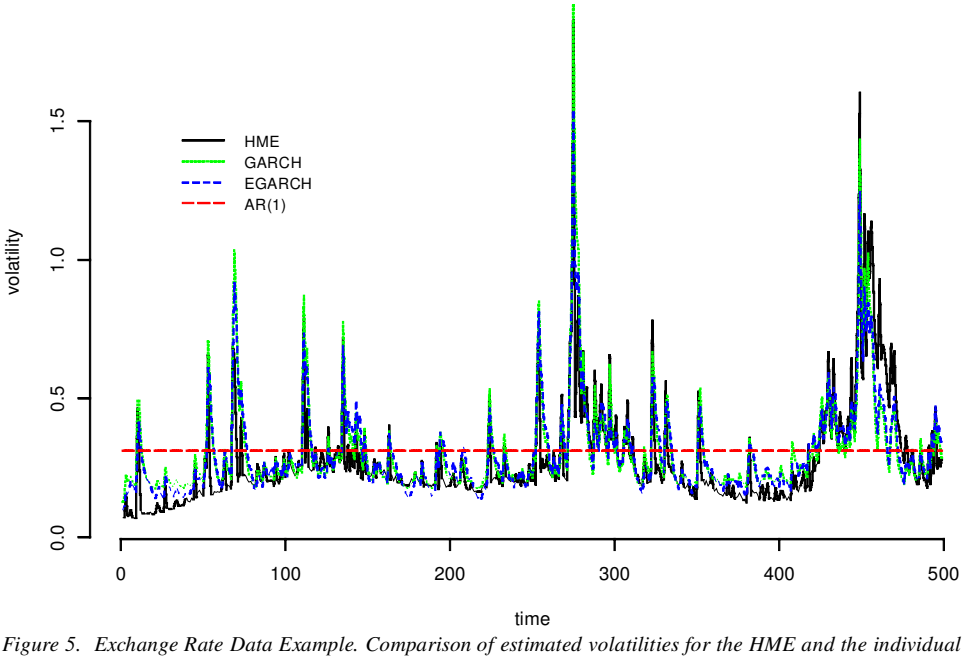

Figure 5. Exchange Rate Data Example. Comparison of estimated volatilities for the HME and the individual models AR(1), AR(1)-GARCH(1,1), and AR(1)-EGARCH(1,1).

Figure 3 presents the estimates $\hat{g}_m(t)$ for $m = 1, 2, 3$. In general, the model assigns a very low weight to the AR(1), with competing weights for AR(1)-GARCH(1,1) and AR(1)-EGARCH(1,1), for approximately the first 100 observations of the series. For the remaining segment of the time series, the preferred model is the AR(1)-EGARCH(1,1).

Figure 4 shows the estimates of $\hat{h}_m(t)$ for $m = 1, 2, 3$. Although the general "smoothed" pattern is similar to that exhibited by Figure 3, the model posterior probabilities have jumps of high probability for the three competing alternatives. This example reects how $h_m(t)$ can be highly impacted by single observations. The "ups and downs" in volatility experienced by exchange rate data across time lead to these model switches. Periods of almost constant variance can be well represented by an AR(1) model but when the data present periods of nonconstant variance, the GARCH or EGARCH structure dominates producing the large jumps in the functions $h_m(t)$.

Figure 5 presents a comparison of the estimated volatilities of the HME with those based on the individual models. The MLE, computed with a numerical optimizer, was used to obtain the volatilities for AR(1), GARCH(1,1), and EGARCH(1,1). For the present hierarchical mixture model, the volatilities were estimated using the EM-solution, recognizing that for this mixture model, the variance of Y_t given the history is the expectation with respect to the mixing weights $g_m(t)$ of individual model-variances plus the variance of the expectation functions for each defining model. We note that the volatility for the HME smoothes some of the high volatility peaks induced by other models but recognizes the overall pattern suggested by the AR(1)-GARCH(1,1) and the AR(1)-EGARCH(1,1) processes.

ACKNOWLEDGMENTS

M. Tanner was supported by NIH grant CA35464.

REFERENCES

Andersen, T. G., and Bollerslev, T. (1998), "ARCH and GARCH Models," *Encyclopedia of Statistical Sciences, Update Vol.*, 2, eds. S. Kotz, C. B. Read, and D. L. Banks, New York: Wiley, pp. 6–16.

Bollerslev, T. (1986), "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*, 31, 307–327.

Huerta, G., Jiang, W., and Tanner, M. A. (2000), "Time Series Modeling via Hierarchical Mixtures," Technical Report, Northwestern University.

Jordan, M. I., and Jacobs, R. A. (1994), "Hierarchical Mixtures of Experts and the EM Algorithm," *Neural Comp.*, 6, 181–214.

McCulloch, R. E., and Tsay, R. S. (1993), "Bayesian Inference and Prediction for Mean and Variance Shifts in Autoregressive Time Series," *Journal of the American Statistical Association*, 88, 968–978.

Nelson, D. B. (1991), "Conditional Heteroskedasticity in Asset Returns: A New Approach," *Econometrica*, 59, 347–370.

DISCUSSION ARTICLE

Discussion

Ying Nian WU and Song Chun ZHU

VISION AND THE ART OF DATA AUGMENTATION

We have learned a lot from studying the sequence of artful works on EM/data augmentation authored by van Dyk and Meng. In this note, we discuss some of our thoughts (or rather speculations) on the problem of vision from the perspective of missing data modeling and data augmentation.

1. VISUAL COMPLEXITY AND THE MISSING DATA FRAMEWORK

When looking at our visual environment, we are not only aware of the rich details of the 3-D visual scene, but we also summarize the details into a simple description of "what is where," which provides the crucial information of a visual scene. Therefore, we can formulate the problem of vision in terms of the following three variables. (1) *Image*: a 2-D matrix (or a pair of matrix sequences). (2) *Details*: a representation of the 3-D scene in full detail. (3) *Summary*: the abstract description of "what is where." Of course, summary and details are relative concepts, and the two variables Details and Summary should be understood as the bottom and the top of a pyramid of increasingly abstract layers of visual concepts. We call this pyramid the "scene description." For instance, for a scenery image, the Summary may consist of abstract concepts such as river, trees, and their overall shapes, and Details may consist of concepts like water ripples and waves, tree leaves, and branches, and their shapes and locations.

The meaning of the Summary can be defined in terms of how the details look and how they are composed together. Mathematically, this amounts to a generative model $P(\text{Details} \mid \text{Summary})$, which decomposes the complexity in Details into deterministic redundancy

Ying Nian Wu is Assistant Professor, Department of Statistics, 8130 Math Sciences Building, University of California–Los Angeles, Los Angeles, CA 90095-1554 (E-mail: ywu@stat.ucla.edu). Song Chun Zhu is Assistant Professor, Department of Computer and Information Sciences, The Ohio State University, 2015 Neil Avenue, Columbus, OH 43210 (E-mail: szhu@cis.ohio-state.edu).

Journal of Computational and Graphical Statistics, Volume 10, Number 1, Pages 90–93

Figure 1. A Dog Walking on Snow. Purely bottom-up computation cannot find the contour of the dog.

and irrelevant randomness. Human vision perpetually summarizes complex details into simple patterns. With the detailed knowledge of the 3-D scene, the image can be rendered via Image = Graphics(Details). A more general form for this part of the model is P(Image | Details). Our prior knowledge on Summary can be represented by a distribution P(Summary). With this top-down generative model Summary → Details → Image, visual perception can be considered a process of computing the conditional distribution P(Summary, Details | Image).

This formulation clearly fits into the missing data framework, with Details being considered as the missing data. Then an EM/data augmentation algorithm can be derived as iterating the following two steps. (1) Scene reconstruction: imputing Details $\sim P$(Details | Image, Summary). (2) Scene understanding: abstracting Summary $\sim P$(Summary | Details).

Previous thinking on visual perception was often along the direction of bottom-up computation: Image → Details → Summary. This can be inadequate for visual perception because we sometimes need high-level knowledge to resolve uncertainties in perceiving low-level details. For example, in Figure 1, no matter how good our edge detector is, we cannot isolate the detailed contour of the dog without the help of the high-level knowledge. This point is reected mathematically in the scene reconstruction step where we need to impute Details conditioning on both Image (bottom-up information) and Summary (top-down knowledge).

2. MENTAL OPTICS AND THE ART OF DATA AUGMENTATION

The scene description (Summary, Details) has both geometrical and photometrical aspects. The geometrical aspect includes the shapes, poses, and relative positions of the objects above a certain scale. It can be considered the "sketching" part of the scene description. The photometrical aspect includes lighting condition, reectance properties of visible surfaces, as well as small-scale structures not describable in explicit geometrical terms. It can be considered the "painting" part of the scene description. So another way to look at the problem of vision is based on the three variables (Geometry, Photometry, Image). Estimating the Geometry from Image is crucial for our survival, and the estimated Geometry can be readily checked with the physical reality. Compared to the Geometry, the Photometry is only of secondary importance, and the introduction of Photometry may be viewed as an art of data augmentation, the purpose of which is mainly to assist the recovery of Geometry. For this reason, we should make Photometry and the augmented model P(Photometry) P(Image | Geometry, Photometry) as simple as possible, as long as the marginal P(Image | Geometry) leads to sufficiently accurate estimation of the Geometry. We call the mathematical representation of Photometry and the augmented model P(Photometry) P(Image | Geometry, Photometry) the "mental optics." There is no need for "mental optics" to go as deep as physical optics, because otherwise the modeling and computing can be made unnecessarily complicated without much gain. It is still largely a mystery how human brains perform this art of data augmentation. We need physics, psychology, and statistics to solve this puzzle.

Although the overall geometry provides the most important information of a visual scene, it is the complexity of the details and the photometrical aspect that defines perceptually realistic pictures. Therefore, understanding visual complexity and mental optics is crucial for visual perception and learning in computer vision and for realistic texturing and lighting in computer graphics.

3. CONCEPTUALIZATION AS DATA AUGMENTATION

For modeling images, one may argue that there are two major types of modeling strategies. One type consists of "exponential family models," which is based on the statistics of features; for example, responses from linear filters or edge detectors, which are computed deterministically from the observed image. The Markov random fields are models of this type (see, e.g., Wu, Zhu, and Liu 2000, and Zhu, Liu, and Wu 2000), and they are consistent with the bottom-up thinking in the research of visual perception. The other type consists of "data augmentation models," which introduce *hidden variables*; for example, linear basis, edges, bars, blobs, and so on as the causes for the observed image intensities. These hidden variables are to be imputed or inferred from the observed image. The models we discussed earlier are of this type and they are consistent with top-down thinking in the research of visual conception. In exponential family models, the data explain themselves (e.g., the Markov property of the Markov random fields), whereas in data augmentation models, the observed dependencies among the data are attributed to the sharing of common latent causes, and these latent causes become new concepts in our knowledge of data. For the purpose of conceptualization, the hidden causes should be independent so that they do not need further

explanation, and at the same time, the image given the hidden causes should follow a simple model, so that the hidden causes provide a simple explanation for the dependencies among the data. If there are still remaining dependencies among the augmented latent variables, then we can further augment more abstract concepts; for example, lines, curves, ows, organizations, templates, and so on. This art of data augmentation or conceptualization may lead to a representational (instead of operational) theory of low-level vision, and may shed new light on Julesz's textons and Marr's primal sketches (see, e.g., Zhu and Guo 2000).

In some sense, our conceptualization of the world is an art of data augmentation. The data we continuously observe over time include images, sounds, touches, pleasure, pain, and our actions, and we want to make sense of the data—that is, to build a model, P(data), for our survival. For this purpose, our brains perform a data augmentation by introducing an extra variable, world, to simplify the modeling of the complicated dependencies among the sensory data. So we have an augmented model P(world) P(data | world). In physics, people collect more data and find deeper laws, so the P(world) in physics becomes more profound, to the extent that the world and P(world) in quantum mechanics is so removed from the world and P(world) in our brains that we simply cannot imagine or conceive the quantum mechanical P(world) using our intuitive P(world).

ACKNOWLEDGMENTS

We thank the editor for kindly inviting us to contribute this discussion. The research is supported by NSF IIS-9877127 and NSF DMS-0072538.

REFERENCES

Wu, Y., Zhu, S. C., and Liu, X. (2000), "Equivalence of Julesz Ensembles and FRAME Models," *International Journal of Computer Vision*, 38, 245–261.

Zhu, S. C., and Guo, C. E. (2000), "Mathematical Modeling of Clutter: Descriptive vs. Generative Models," in *Proceedings of Spie Aerosense Conference On Automatic Target Recognition*, Orlando, FL.

Zhu, S. C., Liu, X., and Wu, Y. (2000), "Exploring Texture Ensembles by Efficient Markov Chain Monte Carlo—Towards a `Trichromacy' Theory of Texture," *IEEE Pattern Analysis and Machine Intelligence*, 22, 554–569.

DISCUSSION ARTICLE

Discussion

Antonietta MIRA and Peter J. GREEN

1. INTRODUCTION

As long-time enthusiasts for auxiliary variables methods in Bayesian MCMC, we are glad to have the opportunity to discuss this interesting article.

The authors have a reputation for picturesque and metaphorical titles, and this article is no exception. We were intrigued by the "artistic" aspirations here, but remain in some doubt as to whether this should be read in the sense of "nonscientific" or "aesthetically pleasing"! One of the objectives in preparing our discussion was to see whether trying to relate the ideas here more closely to existing auxiliary variables methods might both diminish the former characteristic and enhance the latter. This was only partially successful: is our intention doomed, and can the authors do better?

2. GRAPHICAL MODELS

Existing MCMC methods are set in the context of a full probability model for a (usually structured) collection of variables; in Bayesian MCMC we are interested in the distribution of some of these (the unknowns) given the remainder (the data). Auxiliary variables methods augment this collection of variables and elaborate the corresponding probability model, so that the original model is obtained either as the *marginal* for the original variables, or as their *conditional*, given the values of the auxiliaries. In either case, a simulation of the augmented model can be used in obvious ways to extract information about the original one.

In fact, the term "auxiliary variables" is commonly reserved for the marginal case, while the conditional one is usually called "simulated tempering." Both originate in statistical physics: see Edwards and Sokal (1988) and Marinari and Parisi (1992), respectively. Note that these are not treatments of special problems but quite general formulations.

Antonietta Mira is Associate Professor, Department of Economics, University of Insubria, Via Ravasi 2, 21100, Varese, Italy (E-mail: anto@aim.unipv.it). Peter J. Green is Professor, Department of Mathematics, University of Bristol, Bristol BS8 1TW, UK (E-mail: P.J.Green@bristol.ac.uk).

Journal of Computational and Graphical Statistics, Volume 10, Number 1, Pages 94–97

In our perspective, the "data" are simply the (original) variables that are to be conditioned upon, and are not otherwise logically distinguished from the rest of the variables. Thus, the term "data augmentation" (DA) seems something of an historical accident: the additional variables need have no direct interpretation as any kind of "data".

With this in mind, we examined the conditional and marginal forms of DA discussed in this article, especially in the light of the examples, hoping that the conditional independence structure of the augmented models would prove revealing about the intrinsic ideas here.

We find it helpful to regard the observed data Y as a subvector of what the authors call augmented data, rather than as a deterministic function $\mathcal{M}$ of it. The authors themselves do this in their Section 3. Thus we write Y instead of Y_{obs} and (Y, Z) instead of Y_{aug}; they would call Z "missing data". Our target is the posterior $p(\theta|Y)$ derived from a given joint model $p(\theta, Y)$: what is the structure of the augmented models $p^\star(\theta, \alpha, Z, Y)$ that the authors are studying?

This seems elusive. The basis for conditional augmentation, Equation (2.1), says simply that $p^\star(Y|\theta, \alpha) = p(Y|\theta)$. Superficially, this looks like the definition of simulated tempering, but note that this must hold for *all* α. It in fact asserts that Y and α are conditionally independent given θ, as well as requiring that the augmented model agrees with the original for the distribution of Y given θ. In marginal augmentation, (3.1) tells us only that the augmented model should again satisfy $p^\star(Y|\theta) = p(Y|\theta)$.

Thus, both forms of DA discussed here are examples of auxiliary variables, and no particular structure relating (θ, α, Z, Y) is assumed other than that just mentioned.

This lack of structure is reinforced by examination of the examples, where the joint distribution $p^\star(\theta, \alpha, Z, Y)$ variously factors as $p^\star(\alpha)p(\theta)p^\star(Z|\alpha)p^\star(Y|f(Z, \alpha), \theta)$ (Section 6), $p^\star(\alpha)p(\theta)p^\star(Z|\alpha, \theta)p(Y|\theta)$ (Section 7) and $p^\star(\alpha)p(\theta)p^\star(Z|\alpha, \theta)p^\star(Y|Z, g(\alpha, \theta))$ (Section 8), where f and g are particular functions.

Our concern here about lack of structure is not because cherished prejudices about the universal utility of conditional independence ideas are being threatened, but because it does seem to highlight what we see as a major problem with the ideas advocated in the article. Although the authors' results do assist us in *tuning* a given augmentation scheme to give optimal performance, they do not seem to say much about how to devise the scheme in the first place. The lack of commonality among the examples does seem to leave us with little on which to build a scheme for an entirely new model. It is an extreme contrast with the general ability of MCMC to provide usually workable algorithms on the basis of standard recipes, not requiring subtle analysis of the model.

3. SPECTRUM ANALYSIS

As noted by Besag and Green (1993), auxiliary variable methods in the sense of Edwards and Sokal (1988)—thus including DA algorithms and slice samplers—are two-block Gibbs samplers. Liu (1991) pointed out that the marginal chains of two-block Gibbs samplers have the "interleaving Markov" property. He proved that this property implies that lag one autocorrelations of the marginal chains are non-negative.

Furthermore it also implies reversibility which, in turn, implies that the even-lag autocovariances are non-negative. Using induction we obtain that the lag-n autocovariances

of the marginal chains are non-negative and monotone decreasing with n (Liu, Wong, and Kong 1995).

The following theorem appears in Liu (1991, p. 20):

Theorem 1. *Suppose P is the operator of a general Markov chain $\{X_n\}_{n=1}^{\infty}$. A necessary and sufficient condition for $\gamma_n = cov[f(X_0), f(X_n)]$ to be non-negative and monotone decreasing with n for all functions f square integrable and with zero mean with respect to π, is that P is a positive and self-adjoint operator.*

It follows that (for all such auxiliary variable algorithms) the operators of the two marginal Markov chains are non-negative.

The fact that we work with positive operators makes the two goals of reducing the variance and speeding up the convergence to stationarity no longer conicting, as they can be for general Markov chains. This is consistent with the simulation results in the article where DA algorithms, with an optimal or quasi-optimal choice of the working parameter, have better performance than the standard DA in both senses.

In this setting we can thus unambiguously talk about "good mixing" properties of the Markov chain while, in the MCMC literature, the term is sometimes used to mean fast convergence, and sometimes small variance of the resulting estimates.

The *drawback* of the fact that we are dealing with positive operators is that we will never be able to induce negative correlation along the realised chain, so that the variance of the resulting MCMC estimate cannot be less than the variance under iid sampling.

Using the representation of the lag-n autocovariance γ_n in terms of the spectral measure, we can argue, following Geyer (1992) and Mira (1998) that, in the DA and slice sampler setting, the lag-n autocovariances are strictly positive unless the chain produces independent samples. Furthermore, γ_n is strictly decreasing as n increases and is strictly convex in n if the Markov chain is irreducible. These properties of the autocovariances of a chain generated using a self-adjoint positive operator can be used to construct adaptive window estimators for the asymptotic variance along the line of what has been proposed in Geyer (1992).

4. COMPARISON OF CRITERIA FOR CONDITIONAL DA

The second criterion of Section 2 aims to maximize

$$E\left[\text{var}(\theta|Y_{\text{aug}}, \alpha)|Y_{\text{obs}}, \alpha\right].$$

This is often quite difficult to achieve and, as pointed out by the authors themselves, much more difficult than minimizing

$$I_{\text{aug}}(\alpha) = E\left[-\frac{\partial^2}{\partial\theta^2}\log p(\theta|Y_{\text{aug}}, \alpha)|Y_{\text{obs}}, \theta, \alpha\right]\Big|_{\theta=\theta^*}$$

which corresponds to the third criterion.

When is the easier-to-implement second criterion a good approximation to the third criterion?

Using the identity

$$\frac{\partial^2}{\partial\theta^2}\log p(y|\theta) = E\left(\frac{\partial^2}{\partial\theta^2}\log p(y, z|\theta)\Big| y, \theta\right) + \text{var}\left(\frac{\partial}{\partial\theta}\log p(y, z|\theta)\Big| y, \theta\right),$$

which is readily proved under regularity conditions by the standard device of interchanging integrals over z and derivatives with respect to θ, we find that this equality is true

$$I_{\text{aug}}(\alpha) = I_{\text{obs}} + \text{var}\left(\frac{\partial}{\partial\theta}\log p(\theta|Y_{\text{aug}},\alpha)\,\middle|\,Y_{\text{obs}},\theta,\alpha\right)\Bigg|_{\theta=\theta^*}.$$

Thus, the third criterion is equivalent to minimizing

$$\text{var}\left(\frac{\partial}{\partial\theta}\log p(\theta|Y_{\text{aug}},\alpha)\,\middle|\,Y_{\text{obs}},\theta,\alpha\right)\Bigg|_{\theta=\theta^*}.$$

We do not think the above identity has been brought to the attention of DA experts before and believe that it might help in evaluating the quality of the deterministic approximation method proposed by the authors, theoretically rather than simply empirically.

Notice that the argument we have developed for $I_{\text{aug}}(\alpha)$ can be similarly carried over for $I_{\text{aug}}(w)$ and $\tilde{I}_{\text{aug}}(w)$.

REFERENCES

Besag, J., and Green, P. J. (1993), "Spatial Statistics and Bayesian Computation" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 55, 25–37.

Edwards, R. G., and Sokal, A. D. (1988), "Generalization of the Fortuin-Kastelyn-Swendsen-Wang Representation and Monte Carlo Algorithm," *Physical Review*, D, 38, 2009–2012.

Geyer, C. J. (1992), "Practical Markov Chain Monte Carlo," *Statistical Science*, 7, 473–511.

Liu, J. S. (1991), "Correlation Structure and Convergence Rate of the Gibbs Sampler," unpublished PhD Thesis, University of Chicago.

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

Marinari, E., and Parisi, G. (1992), "Simulated Tempering: A New Monte Carlo Scheme," *Europhysics Letters*, 19, 451–458.

Mira, A. (1998), "Ordering, Slicing and Splitting Monte Carlo Markov Chains," unpublished PhD thesis, University of Minnesota.