Statistical Science
2010, Vol. 25, No. 4, 429–449
DOI: 10.1214/09-STS309

Cross-Fertilizing Strategies for Better EM Mountain Climbing and DA Field Exploration: A Graphical Guide Book

David A. van Dyk and Xiao-Li Meng

Abstract. In recent years, a variety of extensions and refinements have been developed for data augmentation based model fitting routines. These developments aim to extend the application, improve the speed and/or simplify the implementation of data augmentation methods, such as the deterministic EM algorithm for mode finding and stochastic Gibbs sampler and other auxiliary-variable based methods for posterior sampling. In this overview article we graphically illustrate and compare a number of these extensions, all of which aim to maintain the simplicity and computation stability of their predecessors. We particularly emphasize the usefulness of identifying similarities between the deterministic and stochastic counterparts as we seek more efficient computational strategies. We also demonstrate the applicability of data augmentation methods for handling complex models with highly hierarchical structure, using a high-energy high-resolution spectral imaging model for data from satellite telescopes, such as the *Chandra X-ray Observatory*.

Key words and phrases: AECM, blocking, collapsing, conditional augmentation, ECM, ECME, efficient augmentation, data augmentation, Gibbs Sampling, marginal augmentation, model reduction, NEM, nesting.

1. INTRODUCTION

Numerous statistical algorithms involving data augmentation have enjoyed remarkable popularity in the biological, medical, physical, social, engineering and other sciences. These algorithms include both deterministic versions such as the Expectation Maximization (EM) algorithm (Dempster, Laird and Rubin, 1977) and its many extensions and stochastic versions such as the Data Augmentation (DA) algorithm (Tanner and Wong, 1987), the method of auxiliary variables (Besag and Green, 1993) and other Markov chain Monte Carlo (MCMC) methods including the Gibbs sampler (Geman and Geman, 1984). The popularity of these algorithms rests in their suitability for fitting highly structured models (e.g., missing data models, latent variable models, hierarchical models, etc.) with high dimensional parameters. Such models are themselves growing ever more popular in modern statistical practice precisely because complex data generation mechanisms are often naturally defined in terms of unobserved quantities. This aides inference because the unobserved quantities often have a direct physical interpretation and are of scientific interest themselves. From a probabilistic point of view, complex correlation structures are much more easily described in terms of unobserved quantities and the conditional independence structures of hierarchical models. Thus, formulating multi-level models in terms of unobserved variables enables us to parse complex highly-structured data. A primary advantage of algorithms involving data augmentation is that even in these settings they are relatively easy to implement (as illustrated in the spectral model of Section 2) and enjoy stable convergence properties (e.g., EM-type algorithms exhibit monotone convergence in likelihood).

David A. van Dyk is Professor and Chair, Department of Statistics, University of California, Irvine, California 92697, USA (e-mail: dvd@ics.uci.edu). Xiao-Li Meng is Whipple V. N. Jones Professor of Statistics and Chair, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, USA (e-mail: meng@stat.harvard.edu).

In this paper we review, summarize and compare much of the recent work on algorithms involving data augmentation, with EM-like algorithms on the deterministic side and Gibbs-sampler-type MCMC samplers on the stochastic side. This work is primarily aimed at extending the applicability of the algorithms and improving their computational speed. We focus on methods that build on the statistical insight of the algorithms while maintaining their attractive properties (e.g., simplicity and stability), rather than numerical methods that can sacrifice these properties. We present basic ideas and concepts but gloss over much of the technical detail, which are documented in the cited references. To this end, we include a series of schematic graphic representations of the various algorithms that we hope can clarify and highlight their relationships, especially in visualizing the similarities between the deterministic algorithms and their stochastic counterparts. We begin with two overview schematics. Figure 1 describes the relationships among the various EM-type algorithms and Figure 2 describes the synergy between the deterministic and stochastic algorithms that we discuss in this article.

The paper is organized into seven additional sections. As a running example, Section 2 introduces a model for Poisson spectral imaging designed to analyze data from the *Chandra X-ray Observatory* and similar photon counting devices. Section 3 focuses on methods designed to simplify calculation in complex models, specifically data augmentation and model reduction in the context of both mode-finding and sampling algorithms. Section 4 reviews general strategies for improving convergence rates such as blocking and collapsing. These methods are illustrated in Sections 5 and 6 in the context of nesting, conditional augmentation, marginal augmentation, joint augmentation and partial collapsing. Finally, Section 7 applies some of these methods to the running example and Section 8 concludes with a brief discussion.

2. A POISSON SPECTRAL MODEL

This section briefly outlines a model for spectral analysis in astronomy that is designed to summarize high-resolution X-ray and γ-ray spectra. The treatment here is simplified for illustrational purposes. Details can be found in van Dyk et al. (2001), Protassov et al. (2002), Hans and van Dyk (2003), van Dyk and Kang (2004), van Dyk et al. (2006) and Park, van Dyk and Siemiginowska (2008). The spectral model is de-

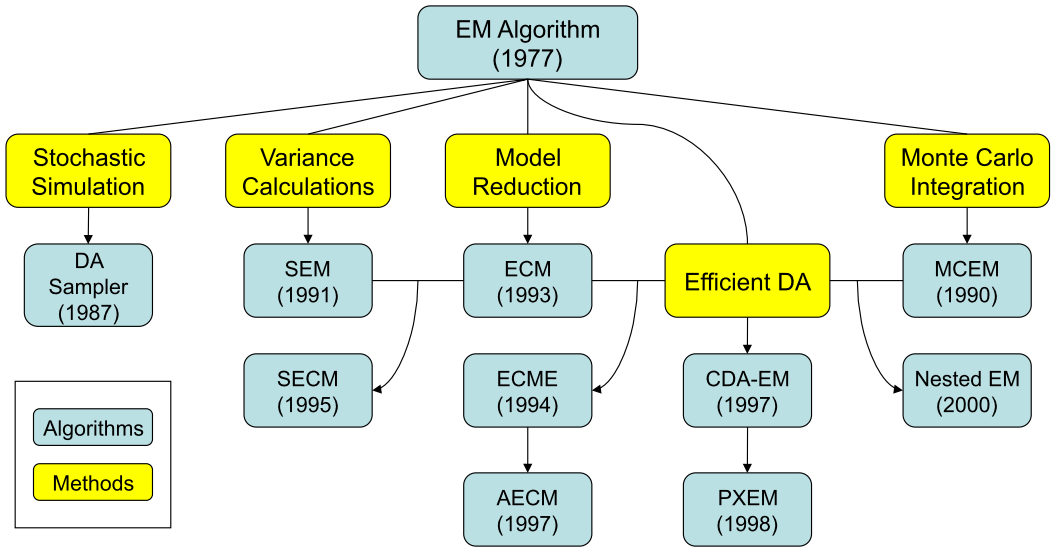

FIG. 1. *A family tree of algorithms inspired by EM. The tree illustrates how various techniques have been combined with the basic framework of EM to formulate new algorithms. It should be regarded as a description of the historical inspiration of the various algorithms rather than as a hierarchy of generalizations and special cases. The basic* stochastic simulation *EM-type algorithm, known as DA, is described in Section 3.1 and Figure 3.* Model reduction *and ECM are described in Section 3.3 and Figure 4.* Efficient data augmentation, *including CDA-EM and PXEM, is described in Section 5 and Figures 5 and 6. It is combined with model reduction to formulate ECME and AECM in Section 6 and Figure 11. The use of* Monte Carlo integration *with and without efficient data augmentation in MCEM and nested EM is discussed in Section 5.5 and illustrated in Figures 8–10. The* variance calculations *of SEM and SECM are developed in Meng and Rubin (1991) and van Dyk, Meng and Rubin (1995), respectively. The arrows illustrate the development and combination of techniques that inspired the generalizations of the EM algorithm.*

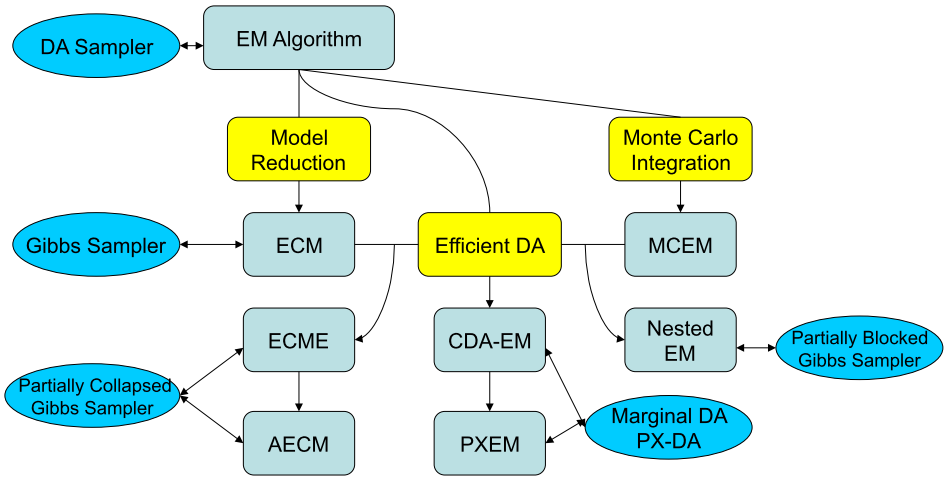

FIG. 2. *The synergy between EM-type algorithms and their stochastic counterparts. The figure shows the cross-fertilization of EM-type algorithms and DA-type samplers. The relationships between EM and DA and between ECM and the Gibbs sample are illustrated in Figures 3 and 4, respectively. Marginal data augmentation and PX-DA are described in Section 5.2. The partially blocked Gibbs sampler that inspired the nested EM algorithm is illustrated in Figure 8 and the partially collapsed Gibbs sampler and its connection with the ECME and AECM algorithms are discussed in Section 6.*

signed to summarize the relative frequency of the energy of photons (X-ray or γ-ray) arriving at a space-based detector. Because of the digital nature of the detector, energies are collected as counts in a number of energy bins (e.g., as many as 4096 on the detectors aboard the *Chandra X-ray Observatory*). These detectors have much higher resolution than their predecessors, and thus smaller expected counts per bin. Independent Poisson distributions are therefore more appropriate to model the counts than the commonly used Gaussian approximation.

Specifically, we model a spectrum as a mixture of a "continuum" term and an "emission line." The continuum characterizes the electromagnetic emission over a broad range of photon energies, while the emission line can be viewed as an aberration from the continuum in a narrow range of energies. A typical spectrum might be composed of multiple continua and multiple emission lines. For simplicity, we suppose there is only one of each in the model. In particular, we parameterize the intensity in bin $j \in \mathcal{J} = \{1, \ldots, J\}$ as

$$(1) \quad \lambda_j(\theta) = \delta_j f(\theta^C, E_j) + \nu p_j(\mu, \sigma^2), \quad j \in \mathcal{J},$$

where δ_j is the known width of bin j, $f(\theta^C, E_j)$ represents the continuum term and is a function of the continuum parameter, θ^C, E_j is the known mean energy in bin j, ν is the expected photon counts corresponding to the emission line, μ and σ are the center and scale (or rather "width") of the emission line, and $p_j(\mu, \sigma^2)$, which is a function of μ and σ^2, is the proportion of the emission line counts that are expected to fall in bin j. We typically quantify $p_j(\mu, \sigma^2)$ via a Gaussian distribution, a t distribution or, in the case of a very narrow line, a delta function. (These are all standard astronomical approximations to the distribution of the strictly positive photon energies of an emission line.) The collection of parameters, θ^C, (ν, μ, σ^2) and θ^A (defined below) are together represented by θ. Here we consider two simple forms of the continuum $f(\theta^C, E_j)$, (1) a log linear model, for example, the power law $\gamma E_j^{-\beta}$, and (2) a free (i.e., saturated) model, $f(\theta^C, E_j) = \theta_j^C$, typically including a smoothing prior distribution such as a Markov chain for θ_j^C, $j \in \mathcal{J}$, for example, $\theta_j^C | \theta_1^C, \ldots, \theta_{j-1}^C \sim \mathrm{N}(\theta_{j-1}^C, 1/\omega_j)$ for $j = 2, \ldots, J$, where $\omega = (\omega_2, \ldots, \omega_J)$ is a smoothing parameter and we assume a flat prior distribution for θ_1^C.

Unfortunately, the photon counts are degraded in the observed data. For example, *instrument response* is a characteristic of the detector that results in blurring of the photons, that is, a photon that arrives in bin j has probability M_{ij} of being detected in bin $i \in \mathcal{I} = \{1, \ldots, I\}$. The $I \times J$ matrix $\{M_{ij}\}$ is determined by on-going calibration of the detector and is presumed known. (Because calibration can be conducted at higher resolution than the binning of the detector, the instrument response matrix may not be square.) Another complication is *absorption*, a process by which a proportion of photons in a given energy bin are absorbed by matter between the astronomical source and the detector. This results in stochastic censoring, where

the censoring rate varies with energy. A similar process occurs in the telescope itself: the detector's *effective area* depends on the energy of the photons. Finally, the counts are contaminated by background events. Because of these degradations, we model the observed counts as independent Poisson variables with parameters

$$\xi_i(\theta) = \sum_{j=1}^{J} M_{ij}\lambda_j(\theta)d_j g(\theta^A, E_j) + \theta_i^B, \quad i \in \mathcal{I}, \tag{2}$$

where d_j is the (presumed) known effective area of the detector for energy bin j as a proportion of the total detector area, $g(\theta^A, E_j)$ is the probability that a photon of energy E_j is *not* absorbed by matter between the source and the detector and θ_i^B is the Poisson intensity of the background, which is generally estimated via real-time calibration in space. The absorption model, $g(\theta^A, E_j)$, may be a (constrained) log linear model with θ^A denoting the model parameter. Note that $\lambda_j(\theta)$ in (2) is given by (1).

How to construct simple, stable and efficient algorithms for fitting this model is the running example for the rest of this article.

3. STATISTICAL CONCEPTS AND COMPUTATION

The EM algorithm is unique among common numerical optimization routines in that it is primarily formulated in statistical rather than mathematical terms. The missing data setup, the Expectation step and the complete-data computations of the Maximization step of EM stand in contrast, for example, to the derivatives and local linearization of the Newton–Raphson algorithm. Other EM-type optimizers and their related stochastic samplers extend this in that their motivation and implementation rely heavily on statistical concepts and insight. In this section we discuss two such concepts: data augmentation and model reduction. We show how their effective use of the divide-and-conquer strategy of reducing a complex problem into an iterated sequence of simpler ones has led to a rich class of statistical algorithms.

3.1 Data Augmentation

Computational methods based on data augmentation are generally applied to posterior distributions or likelihood functions. Here we generally take a Bayesian perspective, but are mindful of the fact that for computational purposes a likelihood function is equivalent to a posterior density under a constant prior distribution. Thus, the object of study can be written as

$$p(\theta|Y^{\text{obs}}) = \int p(\theta, \phi|Y^{\text{obs}})\mu(d\phi), \tag{3}$$

where Y^{obs} is the observed data, μ is a common measure such as a Lebesgue or counting measure, θ is the unobserved quantity of primary interest, and ϕ includes nuisance parameters, latent variables, missing data or any other unobserved quantity of secondary interest. Embedding $p(\theta|Y^{\text{obs}})$ into a model on a larger space such as $p(\theta, \phi|Y^{\text{obs}})$ in this way is called the method of data augmentation. This method can be used to either compute the mode of θ under the marginal distribution given in (3) or to obtain a sample from (3) which in turn can be used to approximate the posterior mean, variance, quantiles, etc., via Monte Carlo simulation.

In the spirit of the EM literature, we use a more inclusive notation Y^{aug} in place ϕ, where Y^{aug} is called the augmented data and represents the combination of Y^{obs} and any latent variables or missing data. The target posterior distribution can be expressed as

$$p(\theta|Y^{\text{obs}}) \propto p(Y^{\text{obs}}|\theta)p(\theta), \tag{4}$$

where $p(\theta)$ is a prior distribution and $p(Y^{\text{obs}}|\theta)$ yields a likelihood. In this way, data augmentation methods can be viewed as embedding (4) into a larger *augmented data* model, via

$$\int_{\mathcal{M}(Y^{\text{aug}})=Y^{\text{obs}}} p(Y^{\text{aug}}|\theta)\mu(dY^{\text{aug}}) = p(Y^{\text{obs}}|\theta), \tag{5}$$

where $\mathcal{M}$ is some many-to-one mapping from Y^{aug} to Y^{obs}. Using the factorization

$$p(Y^{\text{aug}}|\theta) = p(Y^{\text{aug}}|Y^{\text{obs}}, \theta)p(Y^{\text{obs}}|\theta), \tag{6}$$

we recognize that (5) can be maintained with any choice of $p(Y^{\text{aug}}|Y^{\text{obs}}, \theta)$, that is, as long as $p(Y^{\text{aug}}|\theta)$ yields the correct marginal distribution $p(Y^{\text{obs}}|\theta)$. In some cases we can use this flexibility to introduce artificial augmented data purely for computational reasons. Thus, we can choose $p(Y^{\text{aug}}|Y^{\text{obs}}, \theta)$ in order to optimize or improve computational performance rather for statistical modeling, as we shall discuss in Section 5.

Data augmentation can lead to useful algorithms if the conditional distributions, $p(Y^{\text{aug}}|Y^{\text{obs}}, \theta)$ and $p(\theta|Y^{\text{aug}})$, are easy to work with (e.g., to sample, maximize and/or compute expectations). Thus, a useful choice of an augmented data model specifies a division of a model into two simpler conditional models which are typically much easier to analyze.

The EM algorithm computes a posterior mode using the conditional distributions via the familiar two-step iteration, consisting of

E-step: Compute

$$Q(\theta|\theta^{(t)}) = \mathrm{E}\big[\log p(\theta|Y^{\mathrm{aug}})|Y^{\mathrm{obs}}, \theta^{(t)}\big],$$

M-step: Set $\theta^{(t+1)} = \operatorname{argmax}_\theta Q(\theta|\theta^{(t)})$,

where the parenthetical superscript t indexes the iteration. This iteration is known to increase $p(\theta|Y^{\mathrm{obs}})$ and converges to a stationary point of $p(\theta|Y^{\mathrm{obs}})$ that is generally, but not always, a (local) mode of $p(\theta|Y^{\mathrm{obs}})$ (Dempster, Laird and Rubin, 1977; Wu, 1983; Vaida, 2005). The two steps of this iteration give EM its name, that is, the Expectation or E-step and the Maximization or M-step.

The Data Augmentation (DA) algorithm of Tanner and Wong (1987) replaces the two steps of the EM algorithm with two sampling steps, each samples one of two full conditional distributions:

Step 1: $(Y^{\mathrm{aug}})^{(t+1)} \sim p(Y^{\mathrm{aug}}|Y^{\mathrm{obs}}, \theta^{(t)})$,
Step 2: $\theta^{(t+1)} \sim p(\theta|(Y^{\mathrm{aug}})^{(t+1)})$.

This iteration produces a Markov chain, $\{\theta^{(t)}, t = 1, 2, \ldots\}$, which under mild regularity conditions has the desired stationary distribution, $p(\theta|Y^{\mathrm{obs}})$ (see Roberts, 1996; Tierney, 1994, 1996, for convergence results). The EM and DA algorithms are compared in Figure 3. In all of the figures in this article, conditioning on Y^{obs} is suppressed, and hexagons, circles and squares (or their elongated versions) represent expectation steps, (conditional) maximization steps and random draws, respectively.

3.2 Data Augmentation in the Spectral Model

Table 1 lists a hierarchy of augmented data structures used to construct EM and DA algorithms for fitting the spectral model described in Section 2. In the notation of Table 1 more dots in the accent above a variable represent greater degrees of augmentation; variables with fewer dots are (sometimes stochastic) functions of those with more dots. The set $\mathcal{S}$ is the collection of photon sources, here simply $\mathcal{S} = \{C, L\}$, where C represents the continuum and L the emission line. The superscript on "Y" represents the photon source; a "+" in the superscript indicates a mixture of both sources.

Reading top-to-bottom in Table 1, the relationships among the variables are as follows. The vectors $\ddot{Y}^C$ and $\ddot{Y}^L$ contain the exact energies of photons attributed to the continuum and emission line, respectively. Because photon arrivals follow a Poisson process, the length of both of these vectors are Poisson variables; the length of $\ddot{Y}^L$ has expectation ν. These energies are binned and the resulting counts recorded as $\ddot{Y}^s = (\ddot{Y}^s_1, \ldots, \ddot{Y}^s_J)$, for $s \in \mathcal{S}$. Absorption and the varying effective area of the instrument cause an energy-varying proportion of these

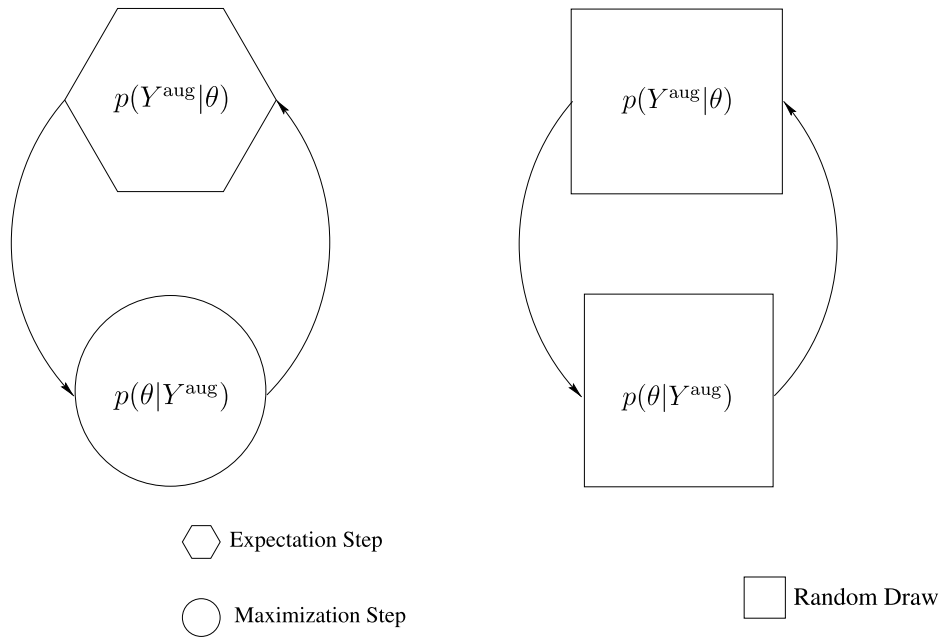

FIG. 3. *The EM (left panel) and DA (right panel) algorithms. In the maximization step of EM, we compute θ to maximize the conditional expectation of* $\log p(\theta|Y^{\mathrm{aug}})$, *with the expectation computed in the expectation step; see Section 3.1.*

TABLE 1
Data augmentation in the spectral model. For all variables, $j \in \mathcal{J}$, $i \in \mathcal{I}$, and $s \in \mathcal{S}$, where $\mathcal{J}$ indexes the ideal bins, $\mathcal{I}$ indexes the detector bins, and $\mathcal{S}$ indexes the sources

Level	Variable	Notation	Range
1.	The ideal data: no blurring, binning, background contamination, absorption[a] or mixing of sources	$\{\ddot{Y}^C, \ddot{Y}^L\}$	Positive, keV[b]
2.	The binned ideal counts	$\{\ddot{Y}_j^C, \ddot{Y}_j^L\}$	Counts
3.	The binned ideal counts after absorption	$\{\dot{Y}_j^C, \dot{Y}_j^L\}$	Counts
4.	The mixed and binned ideal counts after absorption	$\dot{Y}_j^+$	Counts
5.	The mixed, binned and blurred ideal counts after absorption	Y_i^+	Counts
6.	The mixed, binned and blurred ideal counts after absorption and background contamination, this is, the observed data	Y_i^{obs}	Counts

[a]In the statistical model the effective area of the instrument is handled in exactly the same way as absorption. Thus, in this table, absorption includes the effective area of the detector.
[b]The ideal data are the photon energies measured in kiloelectron volts (keV).

counts to be lost. In particular,

$$\dot{Y}_j^s | \ddot{Y}_j^s, \theta \sim \text{Binomial}(\ddot{Y}_j^s, d_j g(\theta^A, E_j)), \quad j \in \mathcal{J}, s \in \mathcal{S}. \tag{7}$$

For the observer, the continuum and emission line counts are combined, $\dot{Y}_j^+ = \dot{Y}_j^C + \dot{Y}_j^L$ for each j. Blurring, due to instrument response, shuffles photons among the bins and into the observed bin counts via

$$Y^+ | \dot{Y}^+, \theta \sim \sum_{j=1}^{J} \text{Multinomial}(\dot{Y}_j^+, M_j), \tag{8}$$

where $Y^+ = (Y_1^+, \ldots, Y_I^+)$, $\dot{Y}^+ = (\dot{Y}_1^+, \ldots, \dot{Y}_J^+)$, and M_j is the jth column of M, $j \in \mathcal{J}$. Because M may not be a square matrix, the lengths of Y^+ and $\dot{Y}^+$ may differ. Finally, background contamination leads to the observed bin counts,

$$Y_i^{\text{obs}} | Y_i^+, \theta \sim Y_i^+ + \text{Poisson}(\theta_i^B), \quad i \in \mathcal{I}. \tag{9}$$

This augmented-data construction leads to easy implementation for two reasons. First, each level of augmented data follows a standard distribution given θ and the data in the rows lower in Table 1. Reading Table 1 bottom-to-top, each conditional distribution can be derived using the Bayes theorem. For illustration, we report the details of just two:

$$Y_i^+ | Y_i^{\text{obs}}, \theta \sim \text{Binomial}\left(Y_i^{\text{obs}}, \frac{\xi_i(\theta) - \theta_i^B}{\xi_i(\theta)}\right), \quad i \in \mathcal{I}, \tag{10}$$

where $\xi_i(\theta)$ is defined in (2), and

$$\ddot{Y}_j^L | \dot{Y}_j^L, \theta \sim \dot{Y}_j^L + \text{Poisson}(\eta_j), \quad j \in \mathcal{J}, \tag{11}$$

where $\eta_j = \nu p_j(\mu, \sigma^2)(1 - d_j g(\theta^A, E_j))$. The other necessary conditional distributions can be found in Appendix B of van Dyk et al. (2001). Thus, the E-step of EM and the corresponding draw of DA are straightforward. Second, given the data in Table 1, the posterior distribution of θ is a set of independent standard distributions. For example, given $\ddot{Y}^L \overset{\text{i.i.d.}}{\sim} \text{N}(\mu, \sigma^2)$, it is easy to compute the posterior distribution of (ν, μ, σ^2), recalling that the length of $\ddot{Y}^L$ is a Poisson random variable with mean ν. The posterior distributions of the other components of θ are also standard and simple to derive. Thus, the M-step of EM and the corresponding draw of DA are again easy to implement. Incorporating proper prior information can be accomplished using the appropriate semi-conjugate prior distributions as described in van Dyk et al. (2001).

3.3 Model Reduction

Model reduction involves using a set of (typically complete) conditional distributions in a computation method designed to learn about the corresponding joint distribution. Reducing the augmented-data model significantly broadens the applicability of algorithms involved in data augmentation, while maintaining their stable convergence properties (e.g., Meng and Rubin, 1993). In particular, if we partition θ into P subvectors, $\theta = (\theta_1, \ldots, \theta_P)$, reducing the aug-

mented data model involves working with the set of conditional distributions $p(\theta_1|Y^{\text{aug}}, \theta_{-1}), \ldots, p(\theta_P|Y^{\text{aug}}, \theta_{-P})$ in place of directly working with $p(\theta|Y^{\text{aug}})$; here $\theta_{-p} = (\theta_1, \ldots, \theta_{p-1}, \theta_{p+1}, \ldots, \theta_P)$. For example, the ECM algorithm (Meng and Rubin, 1993) replaces the maximization in the M-step of EM with a sequence of P conditional maximizations or CM-steps of the form

CM-step p: Set $\theta^{(t+p/P)} = \text{argmax}_\theta\, Q(\theta|\theta^{(t)})$ subject to $\theta_{-p}^{(t+p/P)} = \theta_{-p}^{(t+(p-1)/P)}$.

The ECM algorithm is useful when the CM-steps exist in closed form but the M-step does not. ECM is illustrated with $P = 2$ in the left panel of Figure 4.

The same strategy can be applied to the DA sampler. By replacing the draw from $p(\theta|Y^{\text{aug}})$ with a sequence of draws from the corresponding full conditional distributions, the sampler becomes a $(P + 1)$-step Gibbs sampler. This sampler is illustrated in the right panel of Figure 4. In the context of sampling, we can also reduce $p(Y^{\text{aug}}|\theta)$ into a set of conditional distributions. Partitioning the expectation step, however, has proven much more illusive. One strategy involves using the law of iterated expectations in the computation of the E-step and results in the Nested EM algorithm; see Section 5.5.

Rather than using a partition of θ, a more general model reduction scheme updates θ by conditioning on a sequence of functions of θ. It is only required that the functions allow movement anywhere in the parameter space, that is, the functions are "space-filling" as described by Meng and Rubin (1993). Again, the same strategy can be used in sampling algorithms, such as the Bayesian IPF sampler used to fit constrained models on contingency tables (Schafer, 1997; Gelman et al., 2003). Recent work by Yu and Meng (2010) further explores the use of this strategy to improve MCMC algorithms by employing a sequence of sufficient and auxiliary data augmentation schemes that are space filling.

3.4 Model Reduction in the Spectral Model

To illustrate model reduction in an augmented data model, we consider the second form of the continuum model, namely, the free model $f(\theta^C, E_j) = \theta_j^C$ with a Markov-chain-type smoothing prior $\theta_j^C|\theta_1^C, \ldots, \theta_{j-1}^C \sim \text{N}(\theta_{j-1}^C, 1/\omega_j)$ for $j = 2, \ldots, J$, where $\omega = (\omega_2, \ldots, \omega_J)$ is a smoothing parameter and we assume a flat prior for θ_1^C. For simplicity, we assume there is no emission line and that $\delta_j = g(\theta^A, E_j) = 1$ for each j, that is, the bins are of the same size and that there is no absorption. In this case, we use only rows 4–6 of Table 1 in our data augmentation scheme to de-

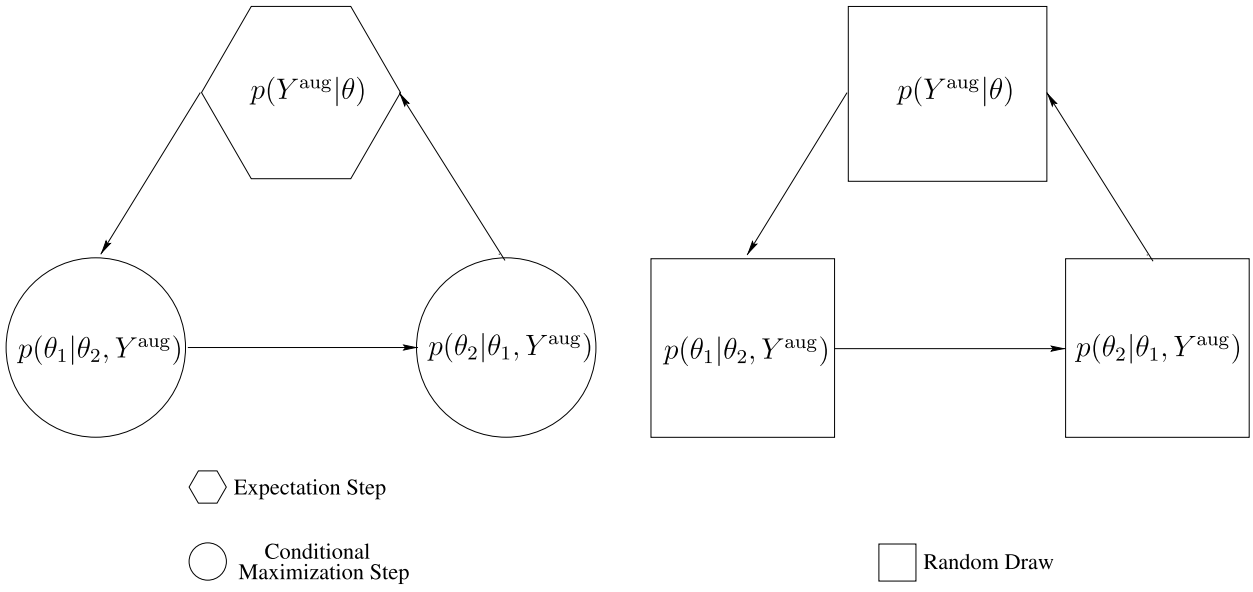

FIG. 4. *The ECM algorithm and the Gibbs sampler. The left panel shows a three-step ECM algorithm composed of an E-step and two CM-steps. The corresponding Gibbs sampler is illustrated in the right panel and is composed of three steps including a data augmentation step. In the conditional maximization steps of ECM, we compute the component of θ to maximize the conditional expectation of the log of the quantity in the ◯, with the expectation computed in the expectation step; see Section 3.3.*

rive

$$Q(\theta|\theta^{(t)}) = \sum_{j=1}^{J}[\mathrm{E}(\dot{Y}_j^C|Y^{\mathrm{obs}},\theta^{(t)})\log\theta_j^C - \theta_j^C] - \frac{1}{2}\sum_{j=2}^{J}\omega_j(\theta_j^C - \theta_{j-1}^C)^2. \tag{12}$$

Once we have computed the expectation in (12), we need only optimize $Q(\theta|\theta^{(t)})$ as a function of θ. Unfortunately, this optimization cannot be done analytically when some $\omega_j > 0$. However, the partial derivative of $Q(\theta|\theta^{(t)})$ with respect to θ_j^C is a quadratic function of θ_j^C if we fix θ_{-j}^C. Thus, as is discussed by Fessler and Hero (1995) and is improved in Section 7, we can construct an ECM algorithm with J CM-steps of the form

$$(\theta_j^C)^{(t+1)} = \max\left\{0, \frac{1}{A_j}\left(B_j + \sqrt{B_j^2 + A_j\mathrm{E}(\dot{Y}_j^C|Y^{\mathrm{obs}},\theta^{(t)})}\right)\right\},$$

where

$$A_j = \omega_j + \omega_{j+1} \quad \text{and}$$
$$B_j = -\left(1 - \omega_j(\theta_{j-1}^C)^{(t+1)} - \omega_{j+1}(\theta_{j+1}^C)^{(t)}\right)/2.$$

4. IMPROVING RATES OF CONVERGENCE

EM-type algorithms and their stochastic counterparts have seen many applications largely because of their computational stability and simple implementation. Nonetheless, these methods are legitimately criticized for their slow convergence in some settings. Strong posterior correlations among the components updated in each step lead to full conditional distributions that are far less variable than the corresponding marginal distributions. This in turn leads to smaller step sizes and slower progress toward the mode or toward the stationary distribution. Much work has been focused on developing algorithms with improved rates of convergence that continue to enjoy the simplicity and stability that makes data augmentation so useful in practice. As we shall see with both data augmentation and model reduction, *less is better* if one hopes for speed, while more is often better if one hopes for simplicity. In this section we discuss the sometimes conflicting strategies for improving the computational performance of methods based on data augmentation.

4.1 The EM and DA Rates of Convergence

Before we can develop criteria for speeding up data augmentation methods, we need mathematical measures of their rates of convergence. For EM, such a measure is given by ρ_{EM}, the spectral radius of the so-called matrix fraction of missing information (Dempster, Laird and Rubin, 1977),

$$I - I^{\mathrm{obs}}[I^{\mathrm{aug}}(Y^{\mathrm{aug}})]^{-1}, \tag{13}$$

where I is an identity matrix, I^{obs} is the observed Fisher information matrix, and $I^{\mathrm{aug}}(Y^{\mathrm{aug}}) = -\partial^2 Q(\theta|\theta^*)/(\partial\theta\,\partial\theta)|_{\theta=\theta^*}$ with θ^* the posterior mode; our notation for I^{aug} emphasizes that both $Q(\theta|\theta')$ and the augmented-data information matrix depend on the choice of augmented data model. Here we use the traditional terms (e.g., Fisher information) of the EM literature, which primarily focus on likelihood calculation, even though we are dealing with the more general posterior computation. In particular, I^{obs} is the negative of the second derivative of the log posterior density evaluated at the posterior mode.

We call ρ_{EM} the global rate of convergence and $I - I^{\mathrm{obs}}(I^{\mathrm{aug}})^{-1}$ the matrix rate of convergence of the EM algorithm. More general formulations of the rate of convergence for ECM and other EM-type algorithms are given by Meng and Rubin (1993, 1994), Meng (1994), Meng and van Dyk (1997) and van Dyk (2000b). For the EM algorithm, our goal is to minimize ρ_{EM} as a function of the data augmentation scheme.

For the DA algorithm, the geometric rate of convergence (Amit, 1991) is

$$1 - \inf_{h:\mathrm{Var}(h(\theta)|Y^{\mathrm{obs}})=1} \mathrm{E}[\mathrm{Var}(h(\theta)|Y^{\mathrm{aug}})|Y^{\mathrm{obs}}]. \tag{14}$$

Although this quantity and the maximum lag one autocorrelation (Liu, 1994) are valuable for theoretical calculations, they are generally difficult to work with analytically in particular models. The EM-approximation of van Dyk and Meng (2001) is essentially based on a Gaussian approximation to the posterior distribution and simply replaces these quantities by ρ_{EM}. Van Dyk and Meng (2001) illustrate that this approximate EM criterion can lead to substantial improvements in DA samplers. Thus, one of our basic strategies is to focus on methods that reduce ρ_{EM} with an understanding that such methods are useful in formulating efficient data augmentation schemes for both deterministic and stochastic algorithms.

4.2 Blocking and Collapsing

As the formulations of the matrix rates of convergence for more complex EM-type algorithms in the above cited articles illustrate, analysis of convergence is significantly more complex with multi-step algorithms. In the analysis of DA and Gibbs samplers, the spectral radius and the norm of the forward operator are useful measures of the convergence behavior of a Markov chain (Liu, Wong and Kong, 1994; Liu, 2001). Based on these measures, Liu, Wong and Kong (1994) introduced two strategies that have emerged as important general techniques for improving the behavior of Gibbs-type samplers.

To illustrate these techniques, consider a P-step sampler that simulates each component of $\theta = (\theta_1, \ldots, \theta_P)$ in turn conditioning on the most recently sampled values of the other $P-1$ components of θ. The first strategy, known as *blocking*, involves combining two or more draws into a single draw. For example, the last two steps could be combined into a single draw of (θ_{P-1}, θ_P) given the other $P-2$ components of θ. *Collapsing*, on the other hand, involves the construction of a sampler on a subspace of the original sampler. For example, we might compute the marginal distribution of θ_{-P} by integrating out θ_P and construct a $(P-1)$-step sampler using the full conditional distributions of the first $P-1$ components of the original partition of θ. Each of these components is updated conditioning on the most recently sampled values of the other $P-2$ components of θ_{-P} to construct a Markov chain with stationary distribution equal to the marginal distribution of θ_{-P}.

Liu (2001) shows that both of these strategies are expected to improve the convergence behavior of the original P-step sampler in that they reduce the norm of its forward operator. (For Gibbs samplers with more than two steps, the norm may not be equal to the rate of convergence of the Markov chain.) He also showed that collapsing reduces the norm by at least as much as blocking. Thus, good general advice is to collapse whenever possible, and to block if you can when collapsing is not possible. Liu's technical results apply only when blocking is applied to the last steps of each iteration of a Gibbs sampler and/or when the subparameter sampled in the last step is collapsed out of the sampler, as we discussed for illustration in the previous paragraph. Nonetheless, experience shows that both strategies are more generally useful and should be implemented whenever feasible.

Analogous advice applies to EM-type algorithms. In the comparison of the EM and ECM algorithms, blocking suggests that fewer CM-steps should be preferred and that the ECM algorithm is expected to converge more slowly than the corresponding EM algorithm. While this is good general advice, it does not always hold mathematically; Meng (1994) gives a simple example in which ECM outperforms EM. We emphasize that the motivation of ECM, however, is not faster convergence but easier implementation. We generally consider ECM when the M-step of EM is not tractable and, thus, the EM algorithm itself is not feasible.

Collapsing is also a useful strategy in the context of EM-type algorithms. The next section is devoted to methods that aim to reduce the information in Y^{aug} and thus effectively collapse a portion of Y^{aug} out of the iteration. Section 6 describes intermediate strategies that allow partial collapse when full collapse is not possible, as in the ECME and AECM algorithms.

In the context of EM, we can sometimes also collapse θ via a profile loglikelihood. Suppose that $\theta = (\theta_1, \theta_2)$ and that we are able to compute the profile likelihood $\tilde{\ell}(\theta_1; Y^{\text{obs}}) = \ell(\theta_1, \hat{\theta}_2(\theta_1, Y^{\text{obs}})|Y^{\text{obs}})$, where $\hat{\theta}_2(\theta_1, Y^{\text{obs}})$ is the maximizer of $\ell(\theta_1, \theta_2|Y^{\text{obs}})$ when θ_1 is fixed. There are two ways to construct an EM algorithm in this situation. The first way is to construct a data augmentation, Y^{aug}, to implement EM for the full parameter $\theta = (\theta_1, \theta_2)$ via the full augmented data loglikelihood, $\ell(\theta_1, \theta_2|Y^{\text{aug}})$. That is, we do not take advantage of the potential computational gain of using the profile likelihood. The second way is to construct a data augmentation, $\tilde{Y}^{\text{aug}}$, to augment the profile likelihood $\tilde{\ell}(\theta_1; Y^{\text{obs}})$ and then implement the EM algorithm for the subparameter θ_1 only. Note that here we use the notation $\tilde{\ell}(\theta_1; Y^{\text{obs}})$ rather than $\tilde{\ell}(\theta_1|Y^{\text{obs}})$ to emphasize that $\tilde{\ell}(\theta_1; Y^{\text{obs}})$ may not necessarily be a proper loglikelihood in the sense of being derived from a log density or probability of Y^{obs}. We can nonetheless use EM, because it is possible to construct an EM algorithm for maximizing any objective function $D(\theta; Y^{\text{obs}})$ as long as we can find an augmented objective function $D(\theta; Y^{\text{aug}})$ such that $\exp\{D(\theta; Y^{\text{aug}}) - D(\theta; Y^{\text{obs}})\}$ is a proper conditional density function of Y^{aug} given θ and Y^{obs}; see the rejoinder of Meng and van Dyk (1997) for more discussion on this flexibility of EM. Therefore, it is possible to use EM for the profile likelihood by treating $\tilde{\ell}(\theta_1; Y^{\text{obs}})$ as an objective function. This collapsing through profiling has not been generally recognized, but can significantly improve the speed, when compared to the first way of directly applying the EM algorithm to the full likelihood. See Meng (1997) for

more discussion and an example involving a zero inflated Poisson model.

5. EFFICIENT DATA AUGMENTATION

Inherent in the definition of the augmented data model is a choice: There are infinitely many augmented data models satisfying (5). In this section we discuss various criteria for this choice that result in efficient algorithms. By "efficient data augmentation" we mean using augmentation schemes that improve *speed*, while maintaining *stability and simplicity*. Here we discuss techniques that are able to achieve all three criterion: They reduce the augmented data in the construction of the algorithm to improve speed while maintaining stability and simplicity.

The basic idea is similar to collapsing in the Gibbs sampler. Suppose that an EM algorithm or a data augmentation sampler can be constructed with a baseline data augmentation scheme that we denote $\tilde{Y}^{\text{aug}}$. Further suppose that $\tilde{Y}^{\text{aug}} = Y_1^{\text{aug}} \cup Y_2^{\text{aug}}$, where both Y_1^{aug} and Y_2^{aug} are legitimate data augmentation schemes in that they both contain Y^{obs}. It is easy to show that $I^{\text{aug}}(\tilde{Y}^{\text{aug}}) \geq I^{\text{aug}}(Y_1^{\text{aug}})$ [i.e., that $I^{\text{aug}}(\tilde{Y}^{\text{aug}}) - I^{\text{aug}}(Y_1^{\text{aug}})$ is semi-positive definite] and that $\mathrm{E}[\mathrm{Var}(h(\theta)|\tilde{Y}^{\text{aug}})|Y^{\text{obs}}] \leq \mathrm{E}[\mathrm{Var}(h(\theta)|Y_1^{\text{aug}})|Y^{\text{obs}}]$, where $h(\cdot)$ is any real-valued function, the first expression being an asymptotic variant of the second (Meng and van Dyk, 1999). Thus, by (13) and (14), construction of an alternate algorithm using only Y_1^{aug} as the augmented data results in faster convergence. This strategy effectively collapses $\tilde{Y}^{\text{aug}} \setminus Y_1^{\text{aug}}$ out of the algorithm. We will discuss direct applications of this idea when we discuss the nesting strategy in Section 5.5. Less direct applications are the topic of Sections 5.1–5.3. The methods described in these sections do not directly decompose $\tilde{Y}^{\text{aug}}$ into two components but still aim to either reduce $I^{\text{aug}}(\tilde{Y}^{\text{aug}})$ or to increase $\mathrm{E}[\mathrm{Var}(h(\theta)|\tilde{Y}^{\text{aug}})|Y^{\text{obs}}]$.

5.1 Conditional Augmentation

The methods of conditional, marginal and joint augmentation all take advantage of the flexibility in (5) to introduce less informative augmented data in order to construct a more efficient algorithm. To search for a good augmented data model using any of the three methods, we begin by parameterizing the augmented data model using a *working parameter*. We define a working parameter to be a parameter in the augmented data model that is not identifiable under the observed data model, $p(Y^{\text{obs}}|\theta)$. In particular, we generalize (5) via

$$\int_{\mathcal{M}(Y^{\text{aug}})=Y^{\text{obs}}} p(Y^{\text{aug}}|\theta,\alpha)\mu(dY^{\text{aug}}) = p(Y^{\text{obs}}|\theta) \tag{15}$$

for all α in some class $\mathcal{A}$. Notice that the right-hand side of (15) does not depend on the working parameter. An effective method of introducing α is to let $Y^{\text{aug}} = \mathcal{D}_{\alpha,\theta}(\tilde{Y}^{\text{aug}})$, where $\mathcal{D}_{\alpha,\theta}$ is a one-to-one mapping for any θ and $\alpha \in \mathcal{A}$ and $\tilde{Y}^{\text{aug}}$ is the baseline augmented data. Typically $\tilde{Y}^{\text{aug}}$ is the standard augmented data used to construct EM-type algorithms or samplers for fitting a particular model. In the context of the EM algorithm, we can compute the scalar rate of convergence, $\rho_{\text{EM}}(\alpha)$, for each α. Conditional augmentation simply optimizes $\rho_{\text{EM}}(\alpha)$ as function of α and then conditions on the optimal value of α throughout the iteration. Meng and van Dyk (1997) call an EM algorithm constructed with the resulting optimal data augmentation scheme an *efficient data augmentation EM algorithm*. For clarity, we refer to it here as a *conditional data augmentation EM algorithm* or CDA-EM. Although this choice of augmented data model is based on the EM rate of convergence, the same model can be used to construct data augmentation samplers. This is an example of the approximate EM criterion discussed in Section 4.1.

It is worth noting that the optimization required by conditional augmentation occurs as part of the derivation of the algorithm. The value α is fixed when we run the algorithm; see Figure 5. The methods of marginal and joint augmentation, on the other hand, avoid this initial optimization problem by averaging over or fitting α on the fly, and, more importantly, they can lead to better algorithms.

5.2 Marginal Augmentation

Marginal augmentation also begins with (15), but, in addition to a working parameter, introduces a *working prior distribution*, $p(\alpha)$. The working prior distribution is typically chosen so that α and θ are independent, so that

$$\int_{\mathcal{M}(Y^{\text{aug}})=Y^{\text{obs}}} \left[\int p(Y^{\text{aug}}|\theta,\alpha)p(d\alpha)\right]\mu(dY^{\text{aug}}) = p(Y^{\text{obs}}|\theta). \tag{16}$$

Note that if we define the resulting augmented data model as $p(Y^{\text{aug}}|\theta) = \int p(Y^{\text{aug}}|\theta,\alpha)p(d\alpha)$, we obtain $\int p(Y^{\text{aug}}|\theta)\mu(dY^{\text{aug}}) = p(Y^{\text{obs}}|\theta)$. Thus, (16) results

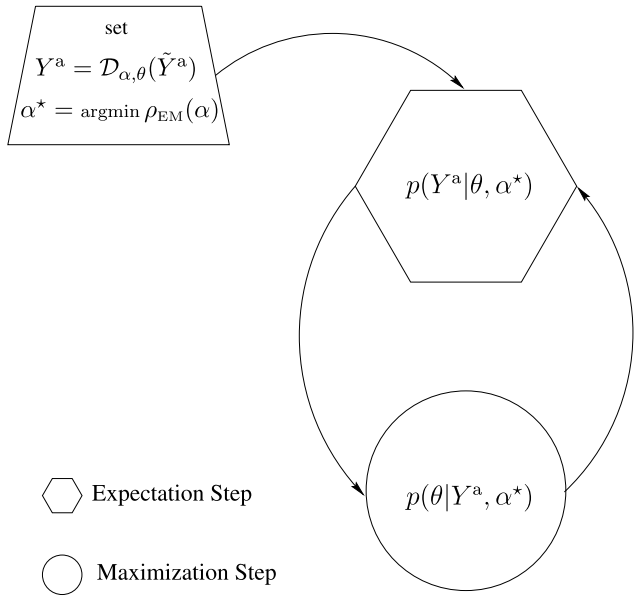

FIG. 5. *The EM algorithm constructed with conditional data augmentation.* [*In the maximization step we compute* θ *to maximize the conditional expectation of* $\log p(\theta|Y^{\mathrm{aug}}, \alpha^\star)$, *with the expectation computed in the expectation step. Here we use the superscript "a" as an abbreviation for "aug" or "augmented."*]

in a legitimate data augmentation scheme. [Marginal augmentation was introduced by Meng and van Dyk (1999) and is very closely related to the PX-DA sampler of Liu and Wu (1999).]

This strategy is motivated by a desire to reduce the information in Y^{aug} for θ. Since conditioning tends to increase information, marginalization may be advantageous. In particular, for any function $h(\cdot)$, we have

$$\begin{aligned}&\mathrm{E}[\mathrm{Var}(h(\theta)|Y^{\mathrm{aug}})|Y^{\mathrm{obs}}]\\ &\quad = \mathrm{E}[\mathrm{E}[\mathrm{Var}(h(\theta)|Y^{\mathrm{aug}}, \alpha)|Y^{\mathrm{obs}}, \alpha]|Y^{\mathrm{obs}}]\\ &\qquad + \mathrm{E}[\mathrm{Var}[\mathrm{E}(h(\theta)|Y^{\mathrm{aug}}, \alpha)|Y^{\mathrm{aug}}]|Y^{\mathrm{obs}}].\end{aligned} \tag{17}$$

If $p(Y^{\mathrm{aug}}|\theta, \alpha)$ is generated by $Y^{\mathrm{aug}} = \mathcal{D}_\alpha(\tilde{Y}^{\mathrm{aug}})$ using the baseline augmentation, $\tilde{Y}^{\mathrm{aug}}$, then $\mathrm{E}[\mathrm{Var}(h(\theta)|Y^{\mathrm{aug}}, \alpha)|Y^{\mathrm{obs}}, \alpha]$ does not depend on α and (17) implies

$$\begin{aligned}&\mathrm{E}[\mathrm{Var}(h(\theta)|Y^{\mathrm{aug}})|Y^{\mathrm{obs}}]\\ &\quad \geq \mathrm{E}[\mathrm{Var}(h(\theta)|Y^{\mathrm{aug}}, \alpha)|Y^{\mathrm{obs}}, \alpha]\end{aligned}$$

for any α, and, thus, in terms of the geometric rate, marginal augmentation is superior to conditional augmentation (Meng and van Dyk, 1999). This result, however, depends on the working parameter being introduced via $Y^{\mathrm{aug}} = \mathcal{D}_\alpha(\tilde{Y}^{\mathrm{aug}})$, a transformation depending only on α. When the transformation depends on the model parameters as well, conditional augmentation can be superior. See Meng and van Dyk (1999) or Liu and Wu (1999) for details.

Although there is no need to choose α when using marginal augmentation, we are left with the choice of working prior distributions. One strategy for choosing $p(\alpha)$ (van Dyk and Meng, 2001) suggests parameterizing the working prior, $p(\alpha|\psi)$, and chooses ψ as a level-two working parameter via a conditional augmentation criterion. Liu and Wu (1999) show that, under certain conditions, the Haar measure leads to an optimal algorithm with the correct stationary distribution. In general, however, using an improper working prior distribution may not even lead to the correct stationary distribution, let alone optimality; see Meng and van Dyk (1999), van Dyk and Meng (2001) and van Dyk (2009). When it exists, the use of the Haar measure typically leads to a joint chain on the enlarged space $(\alpha, \theta, Y^{\mathrm{aug}})$ that is nonpositive recurrent, but the marginal chain on the original space θ converges properly to the desired posterior distribution $p(\theta|Y^{\mathrm{obs}})$; see Hobert (2001), Marchev and Hobert (2004) and Hobert and Marchev (2008) for additional discussion.

5.3 Joint Augmentation

There is no known easy way to implement EM-type algorithms that use marginal augmentation. A similar strategy, however, uses the augmentation scheme (15), but rather than optimizing ρ_{EM} as a function of α before running the algorithm or marginalizing α out as in (16), this method fits α jointly with θ in the M-step. In particular, Liu, Rubin and Wu (1998) presents the PXEM algorithm as a fast adaptation of conditional augmentation in the context of the EM algorithm in the case when $p(\theta) \propto 1$, for example, in maximum likelihood estimation. Van Dyk (2000a) slightly extended the framework to the Bayesian case, by defining

$$\begin{aligned}&Q_{\mathrm{px}}(\theta, \alpha|\theta', \alpha_0)\\ &\quad = \int \log[p(Y^{\mathrm{aug}}|\theta, \alpha)p(\theta)]\\ &\qquad \cdot p(Y^{\mathrm{aug}}|Y^{\mathrm{obs}}, \theta', \alpha_0)\,dY^{\mathrm{aug}}.\end{aligned}$$

As illustrated in Figure 6, the PXEM iteration sets $(\theta^{(t+1)}, \alpha^{(t+1)})$ equal to the maximizer of $Q_{\mathrm{px}}(\theta, \alpha|\theta^{(t)}, \alpha_0)$, where α_0 is some fixed value.[1] The *par-*

[1] We need not condition on $\alpha = \alpha^{(t)}$ in Q_{px} because $Q_{\mathrm{px}}(\theta, \alpha|\theta', \alpha') \geq Q_{\mathrm{px}}(\theta', \alpha'|\theta', \alpha')$ implies $p(\theta|Y^{\mathrm{obs}}) \geq p(\theta'|Y^{\mathrm{obs}})$ for any values of θ' and α'. In particular, $Q_{\mathrm{px}}(\theta^{(t+1)}, \alpha^{(t+1)}|\theta^{(t)}, \alpha_0) \geq Q_{\mathrm{px}}(\theta^{(t)}, \alpha_0|\theta^{(t)}, \alpha_0)$ implies $p(\theta^{(t+1)}|Y^{\mathrm{obs}}) \geq p(\theta^{(t)}|Y^{\mathrm{obs}})$; see Liu, Rubin and Wu (1998) and van Dyk (2000a).

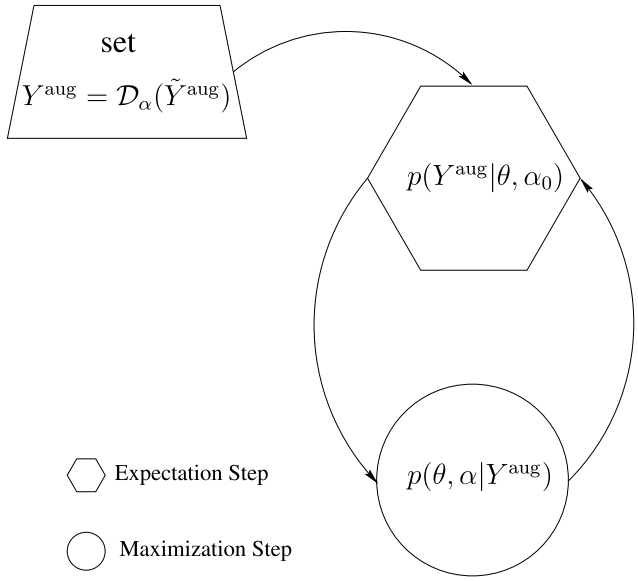

FIG. 6. *The PXEM Algorithm.* [*In the maximization step we compute* θ *and* α *to maximize the conditional expectation of* $\log p(\theta, \alpha|Y^{\text{aug}})$, *with the expectation computed in the expectation step.*]

ticular value of α_0 is generally irrelevant for a PXEM iteration and is simply set to some convenient value throughout the iteration (e.g., $\alpha_0 = 1$ for scale working parameters and $\alpha_0 = 0$ for location working parameters). In this regard, the PXEM iteration could be rewritten to avoid the dependence on α_0, but it is generally deemed easier to simply set α_0 at one arbitrary value and avoid potentially complex algebraic manipulations. The situation is similar when using marginal augmentation with an improper working prior distribution. In that case the posterior distribution of α is improper leading to the technical concerns discussed in Section 5.2. With PXEM the observed data likelihood does not depend on α which can lead to numerical problems if the updated value of α is carried forward in the iteration.

We expect PXEM to perform at least as well as an algorithm that fixes α (i.e., CDA-EM) in terms of the global rate of convergence because it essentially removes the conditioning on α in the data-augmentation scheme. Removing this conditioning reduces I^{aug} (in a positive semidefinite ordering sense) and thus improves the rate of convergence of EM (see Meng and van Dyk, 1997, and Liu, Rubin and Wu, 1998, for details). It is in this regard that PXEM is an example of efficient data augmentation: it effectively reduces the augmented data information in order to improve the rate of convergence without sacrificing simplicity or stability. This does not mean that PXEM generally dominates a CDA-EM algorithm because different augmentation schemes are used in the context of the two strategies. In particular, like marginal data augmentation, PXEM is generally implemented with a transformation, $Y^{\text{aug}} = \mathcal{D}_\alpha(\tilde{Y}^{\text{aug}})$. However, unlike that of conditional data augmentation, this transformation does not depend on θ; see Figure 6. Liu, Rubin and Wu (1998) give an alternative explanation for the efficient performance of PXEM, that by fitting α, we are performing a covariance adjustment to capitalize on information in the data-augmentation scheme. They also illustrate the substantial computational advantage PXEM can offer over other EM-type algorithms for ML estimation. In the context of Bayesian calculations, van Dyk and Tang (2003) show how one-step-late methods (Green, 1990) can be used to accomplish the required optimizations of the PXEM M-step.

5.4 A Graphical Comparison of CDA-EM and PXEM

To illustrate the differences between the CDA-EM and PXEM algorithms, we consider a simple Gaussian model. Suppose

$$X_i \sim \mathrm{N}(\theta, 1/2) \quad \text{for } i = 1, \ldots, n \tag{18}$$

and

$$Y_i \sim \mathrm{N}(\theta, 1/2) \quad \text{for } i = 1, \ldots, m, \tag{19}$$

where the $X = (X_1, \ldots, X_n)$ is observed and $Y = (Y_1, \ldots, Y_m)$ is completely missing. Obviously, the maximum likelihood estimate of θ is $\bar{X}$ and the missing Y is not relevant. Nonetheless, for illustration, we can construct an EM algorithm that treats Y as missing data. In particular, with (X, Y) being the augmented data, we have

$$\begin{aligned} Q\big(\theta|\theta^{(t)}\big) &= 2\theta\left[n\bar{X} + \sum_{i=1}^{m} \mathrm{E}\big(Y_i|\theta^{(t)}\big)\right] - (n+m)\theta^2 \\ &= 2\theta\big(n\bar{X} + m\theta^{(t)}\big) - (n+m)\theta^2, \end{aligned}$$

which can be compared to the observed data loglikelihood, $\ell(\theta)$, as in the first panel of Figure 7, where $n = 1$, $m = 5$, $\bar{X} = 0$, and $\theta^{(t)} = 5$. The panel illustrates that $\ell(\theta)$ and $Q(\theta|\theta^{(t)})$ have the same derivative at $\theta^{(t)}$ and that their optimizers are the maximum likelihood estimate, $\theta^\star$, and $\theta_{\text{EM}}^{(t+1)}$, respectively. (For diagrams illustrating EM's iteration and rate of converge, see Navidi, 1997.)

To use CDA-EM and PXEM, we introduce a working parameter α, via the transformation, $Z_i = Y_i - \alpha\theta \sim \mathrm{N}[(1-\alpha)\theta, 1/2]$ for $i = 1, \ldots, m$, and treat $Z = (Z_1, \ldots, Z_m)$ as the missing data. Since α is not iden-

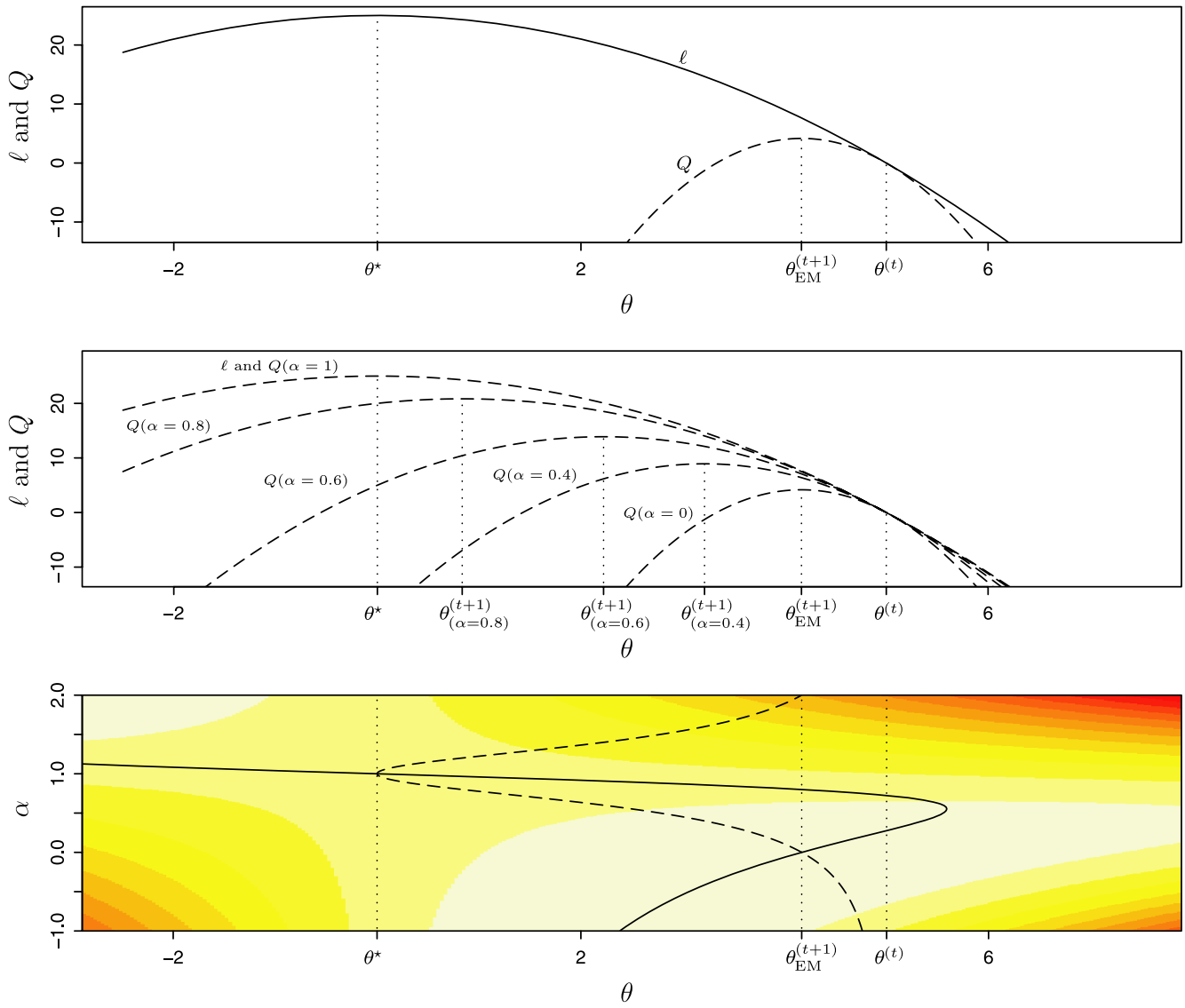

FIG. 7. *Comparing the CDA-EM and the PXEM algorithms. The first panel compares* $\ell(\theta)$ *and* $Q(\theta|\theta')$ *for an EM algorithm applied to a simple Gaussian problem. The functions are normalized to be tangent at* $\theta^{(t)}$. *The second panel compares* $\ell(\theta)$ *with* $Q(\theta, \alpha|\theta^{(t)}, \alpha)$ *for several values of* α *and shows how the missing data becomes less informative and* $Q(\theta, \alpha|\theta^{(t)}, \alpha)$ *becomes a better approximation to* $\ell(\theta)$ *as* α *get closer to the optimal value,* $\alpha = 1$. *The final plot is a heat map of* $Q(\theta, \alpha|\theta^{(t)}, \alpha' = 0)$ *that is optimized in the PXEM algorithm. Lighter colors correspond to higher functional values. The function has two critical points, one at* $(\theta = 0, \alpha = 1)$ *and one at* $(\theta = 0, \alpha = -\infty)$. *The solid and dashed curves give the optimal value of* θ *as a function of* α *by maximizing* $Q(\theta, \alpha|\theta^{(t)}, \alpha)$ *and* $Q(\theta, \alpha|\theta^{(t)}, \alpha' = 0)$, *respectively. The CDA-EM update is a saddle point of* $Q(\theta, \alpha|\theta^{(t)}, \alpha' = 0)$ *and the PXEM update occurs in the limit as* $\alpha \to -\infty$. *Nonetheless, both algorithms return the maximum likelihood estimate in one iteration,* $\theta^{(t+1)} = 0$.

tifiable given X, it is a valid working parameter. In this case,

$$\begin{aligned} Q(\theta, \alpha|\theta^{(t)}, \alpha') &= 2\theta\left[n\bar{X} + (1-\alpha)\sum_{i=1}^{m} \mathrm{E}(Z_i|\theta^{(t)}, \alpha')\right] \\ &\quad - [n + m(1-\alpha)^2]\theta^2 \\ &= 2\theta\left[n\bar{X} + m(1-\alpha)(1-\alpha')\theta^{(t)}\right] \\ &\quad - [n + m(1-\alpha)^2]\theta^2. \end{aligned}$$

The method of conditional data augmentation requires $I^{\mathrm{aug}}(\alpha) = 2[n + m(1-\alpha)^2]$ be computed by differentiating $Q(\theta, \alpha|\theta^{(t)}, \alpha')$ twice with respect to θ and minimized it as a function of α. The optimal value occurs when $\alpha = 1$, in which case the distribution of the missing data does not depend on θ. The second panel of Figure 7 compares $Q_\alpha(\theta|\theta^{(t)}) \equiv Q(\theta, \alpha|\theta^{(t)}, \alpha)$ computed with several values of α with $\ell(\theta)$. As α grows closer to one, $\theta^{(t+1)}$ grows closer to θ_{MLE}. With the op-

timal value of α in this example, $Q_\alpha(\theta|\theta^{(t)})$ and $\ell(\theta)$ coincide, and CDA-EM converges to $\theta^\star$ in one iteration. In general, the algorithm does not converge in one step, but the underlying strategy of choosing a working parameter so that $Q_\alpha(\theta|\theta^{(t)})$ is closer to $\ell(\theta)$ is always the goal.

For PXEM, α' is fixed at the identity value of the transformation from Y to Z (i.e., $\alpha' = 0$) and θ and α are updated at each iteration by jointly optimizing $Q(\theta, \alpha|\theta^{(t)}, \alpha' = 0)$. The third panel of Figure 7 plots this function using a heat map, where brighter colors represent higher values and darker colors represent lower values. The solid line superimposed on the plot is the optimal value of θ as a function of α and is given by

$$\frac{\sum_{i=1}^n X_i + m(1-\alpha)\theta^{(t)}}{n + m(1-\alpha)^2}. \tag{20}$$

For example, with $\alpha = 0$ the curve gives $\theta_{\mathrm{EM}}^{(t+1)}$. The dashed line gives the optimal value of θ as a function of α under CDA-EM. This curve corresponds to the modes of the dashed curves in the second panel. The solid and dashed curves in the third panel differ because CDA-EM and PXEM differ in how they treat α' in $Q(\theta, \alpha|\theta^{(t)}, \alpha')$. PXEM fixes α' at the identity value under the transformation from Y to Z (i.e., PXEM fixes $\alpha' = 0$), whereas CDA-EM does not update α in the iteration and sets $\alpha' = \alpha$ throughout. The function $Q(\theta, \alpha|\theta^{(t)}, \alpha' = 0)$ plotted in panel 3 increases along the solid curve as α goes to $-\infty$ and the solid curve asymptotes to $\theta = 0$, the maximum likelihood estimate. Thus, both CDA-EM run with $\alpha = 1$ and PXEM converge to the maximum likelihood estimate in one iteration.

One might be tempted to think that PXEM is superior to CDA-EM because it optimizes $Q(\theta, \alpha|\theta^{(t)}, \alpha' = 0)$ over both θ and α at each iteration, whereas CDA-EM optimizes $Q(\theta, \alpha|\theta^{(t)}, \alpha' = \alpha)$ over only θ under a constraint that fixes α at a prespecified value. That is, one might expect PXEM to increase ℓ more because it increases Q more. This reasoning, however, not only blurs the difference in how the two algorithms treat α', but also oversimplifies the rates of convergence of EM-type algorithms. An algorithm that increases Q more at every iteration does not necessary converge faster. This can be seen clearly in the first panel of Figure 7. The optimal update is $\theta^\star$, but $\theta^\star$ is far from the maximizer of Q. Our goal is not to increase Q more, but to make Q a better approximation of the log likelihood. As another example, the EM algorithm by definition increases Q by at least as much in its M-step as ECM can in a sequence of CM-steps. Nonetheless, Meng (1994) shows that ECM can converge faster than EM. In the present example, CDA-EM sets $\alpha = 1$ and updates θ to $\theta^{(t+1)} = 0$ which is a saddle point of $Q(\theta, \alpha|\theta^{(t)}, \alpha' = 0)$. Even though $Q(\theta, \alpha|\theta^{(t)}, \alpha' = 0)$ evaluated at the CDA-EM update is less than when it is evaluated at the PXEM update, both updates have $\theta^{(t+1)} = 0$ and thus give the same value of the observed data log likelihood. The rate of convergence is more directly determined by (13) than by the relative increase in Q. It is this rate that CDA-EM aims to optimize and that PXEM improves by eliminating the conditioning on α; see Section 5.3.

5.5 Nesting

Nested EM and DA-type algorithms involve iteratively using a data augmentation method to accomplish one of the steps of a larger algorithm also involving data augmentation. Figures 8–10 illustrate three different ways this might be done. To motivate the nesting strategy, we begin with the partially-blocked Gibbs sampler illustrated in Figure 8 (van Dyk, 2000b). Although we consider a sampler composed using three full conditional distributions, the ideas apply immediately to samplers with arbitrarily many conditional distributions. In particular, suppose we wish to sample from $p(\theta|Y^{\mathrm{obs}})$, where $\theta = (\theta_1, \theta_2, \theta_3)$ by using a Gibbs sampler which samples

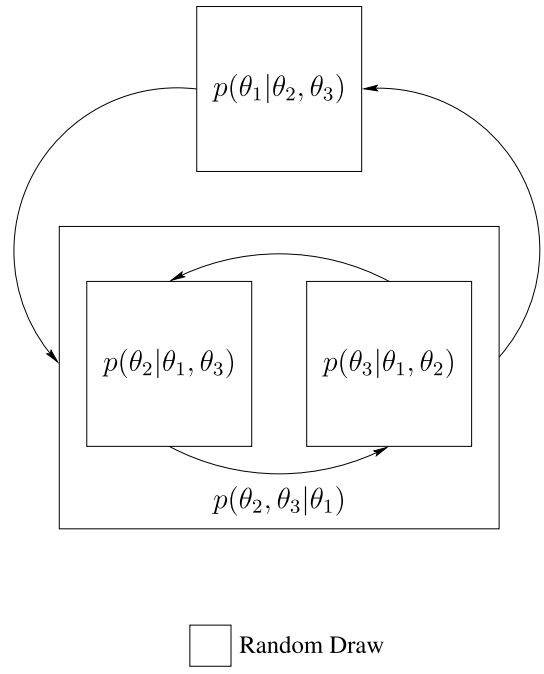

FIG. 8. *The partially blocked Gibbs sampler. The inner loop is iterated N times.*

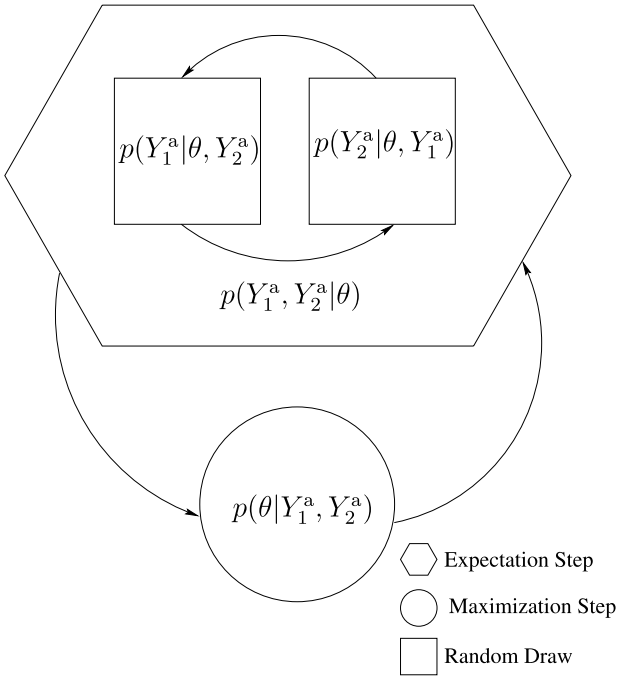

FIG. 9. *The MCEM algorithm. The inner loop is iterated several times.* [*In the maximization step we compute* θ *to maximize the conditional expectation of* $\log p(\theta|Y_1^{\text{aug}}, Y_2^{\text{aug}})$, *with the expectation computed in the Monte Carlo expectation step. Here we use the superscript "a" as an abbreviation for "aug" or "augmented."*]

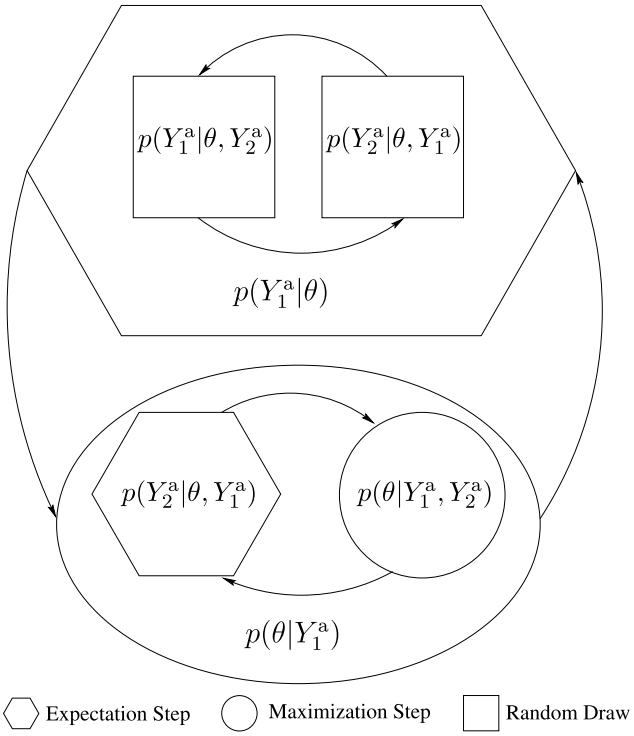

FIG. 10. *The nested EM algorithm with Monte-Carlo E-step implemented with a two-step Gibbs sampler. The inner loops are both iterated several times.* [*In the maximization step we compute* θ *to maximize the conditional expectation of* $\log p(\theta|Y_1^{\text{aug}}, Y_2^{\text{aug}})$, *with the expectation computed in the expectation steps, see Section 5.5. Here we use the superscript "a" as an abbreviation for "aug" or "augmented."*]

from each of $p(\theta_1|\theta_2, \theta_3, Y^{\text{obs}})$, $p(\theta_2|\theta_1, \theta_3, Y^{\text{obs}})$, and $p(\theta_3|\theta_1, \theta_2, Y^{\text{obs}})$ in turn. If sampling from $p(\theta_1|\theta_2, \theta_3, Y^{\text{obs}})$ is expensive relative to sampling from the other two conditional distributions, it may be beneficial to sample once from $p(\theta_1|\theta_2, \theta_3, Y^{\text{obs}})$ and then to sample from $p(\theta_2|\theta_1, \theta_3, Y^{\text{obs}})$ and $p(\theta_3|\theta_1, \theta_2, Y^{\text{obs}})$ N times each in turn. If N is large, the internal Gibbs sampler delivers an approximate draw from the joint distribution $p(\theta_2, \theta_3|\theta_1)$. If this approximation is good, we are essentially running a blocked Gibbs sampler with conditional distributions $p(\theta_1|\theta_2, \theta_3, Y^{\text{obs}})$ and $p(\theta_2, \theta_3|\theta_1, Y^{\text{obs}})$. The partially blocked Gibbs sampler is useful when the advantage of blocking outweighs the cost of sampling from $p(\theta_2, \theta_3|\theta_1, Y^{\text{obs}})$ via a nested Gibbs sampler. This strategy may be helpful when θ_2 and θ_3 exhibit significant correlation given θ_1 and/or $p(\theta_1|\theta_2, \theta_3, Y^{\text{obs}})$ is particularly difficult to sample (e.g., van Dyk et al., 2001). Notice there is a subtle tradeoff here. If θ_2 and θ_3 are (nearly) conditionally independent given θ_1, then there is no need to run the inner iteration. If, on the other hand, they are highly correlated, then the inner iteration may need to be run many times in order to deliver a good draw. The key to success with this strategy is repeating the expensive draw of $p(\theta_1|\theta_2, \theta_3)$ as seldom as possible.

In the context of the EM algorithm, we can implement a similar strategy when the augmented data naturally divide into two or more parts. This strategy takes advantage of the fact that an EM algorithm that treats only part of Y^{aug} as missing and collapses over the rest is faster in terms of ρ_{EM} (Meng and van Dyk, 1997). Thus, we aim to construct an EM algorithm using only part of Y^{aug}. Although this algorithm typically does not have a closed form M-step, the maximization can be accomplished by a second, typically closed-form, EM algorithm that treats the remainder of Y^{aug} as missing data. The resulting nested EM algorithm (van Dyk, 2000b) has an improved rate of convergence but, because of the nesting, each iteration requires more time to compute. If the computational complexity of the E-step is relegated to the outer loop, this trade-off can go in favor of the nesting strategy when considering the actual computing time required. This advantage can be pronounced when the outer E-step requires a Gibbs sampler to compute the necessary conditional expectations. This is possible with the Monte Carol EM

(MCEM) algorithm (Wei and Tanner, 1990), as is illustrated by van Dyk (2000b). The MCEM algorithm is compared with the nested EM algorithm in Figures 9 and 10.

6. PARTIAL COLLAPSING AS A UNIFIED APPROACH

While the partially-blocked nature of the sampler in Figure 8 is clear, the nested EM algorithm in Figure 10 partially removes $\tilde{Y}^{\text{aug}} \setminus Y_1^{\text{aug}} \subset Y_2^{\text{aug}}$ from the data augmentation scheme in the spirit of conditional augmentation. In this regard, the nested EM algorithm is a type of "partially collapsed" EM algorithm. In this section we discuss a different strategy for partially collapsing quantities out of an EM or DA algorithm. In particular, in algorithms that involve model reduction, we can collapse quantities in some but not all of the CM-steps or conditional draws. It is in this sense that we use the term "partially collapsed."

Collapsing involves constructing an algorithm on a marginal distribution of the target space of the original algorithm. That is, we construct an algorithm that works on a *collapsed parameter space* of the *original parameter space*. (Here the parameter space includes all unknowns including latent variables and missing data.) Although this strategy is computational efficient it can be practically difficult if some or all of the full conditional distributions on the collapsed parameter space are complex or nonstandard distributions. Given that the augmented data are introduced specifically to simplify the full conditional distributions, it is not surprising that reducing that augmented data can sacrifice this simplicity. Partially collapsed methods aim to reap some of the gains of collapsing in this situation. In particular, when some of the conditional distributions on the collapsed parameter space are simple or at least no more complicated that the corresponding conditional distribution of the original parameter space, partially collapsed methods mix conditional distributions from the two (or perhaps more) parameter spaces in the construction of EM-type algorithms and DA-type samplers. For example, if a conditional maximization or draw given the augmented data are not easier than the corresponding maximization or draw given the observed data, then we may as well use the version that does not involve data augmentation, that is the collapsed version. As we shall discuss, this strategy has lead to a number of useful algorithms.

6.1 The ECME and AECM Algorithms

In order to improve the rate of convergence of the ECM algorithm, Liu and Rubin (1995) formulated the Expectation Conditional Maximization Either or ECME algorithm in which they suggest replacing one or more of the CM-steps of the ECM algorithm with

Direct CM-step p: Set $\theta^{(t+p/P)} = \text{argmax}_\theta \log p(\theta|Y^{\text{obs}})$ subject to $\theta_{-p}^{(t+p/P)} = \theta_{-p}^{(t+(p-1)/P)}$.

When an iterative method is required to accomplish one or more of the CM-step of ECM, it is often no more difficult to maximize the conditional log posterior directly without recourse to data augmentation. In this case Liu and Rubin (1995) argue that the direct CM-step is expected to improve convergence without complicating implementation. We recognize this as a partially collapsed algorithm. If all of the ECM CM-steps were replaced by direct CM-steps the augmented data would be completely removed from the iteration. This would collapse ECM into a Gauss–Seidel optimizer, which is generally expected to be faster than ECM. Of course, if some of the CM-steps of ECM are simple closed-form optimizations while those of ECME require numerical optimization, the computational tradeoff can easily favor ECM over Gauss–Seidel.

Meng and van Dyk (1997) set up a more general framework by allowing different levels of augmented data in each CM-step. The resulting algorithm is called the Alternating Expectation Conditional Maximization or AECM algorithm and generalizes both the ECME and the SAGE (Fessler and Hero, 1994) algorithms. In particular, Meng and van Dyk suggest replacing the CM-step of ECM with

CM-step p: Set $\theta^{(t+p/P)} = \text{argmax}_\theta \text{E}[\log p(\theta|g_p(Y^{\text{aug}}))|\theta^{(t+(p-1)/P)}]$

subject to $\theta_{-p}^{(t+p/P)} = \theta_{-p}^{(t+(p-1)/P)}$. Here we have expanded $Q(\theta|\theta^{(t)})$ according to its original definition with two important changes. First, Y^{aug} is replaced by some function g_p of Y^{aug}. This allows us to reduce the data augmentation by differing amounts in each of the P CM-steps. Here we assume $g_p(Y^{\text{aug}})$ is a legitimate data augmentation scheme for each p. In particular, Y^{obs} is part of each $g_p(Y^{\text{aug}})$. Second, because the data augmentation varies among the CM-steps, we must compute and E-step each time the data augmentation changes, see Figure 11. Thus, in the expectation of each AECM CM-step we condition on the value of θ produced by the most recent CM-step, not

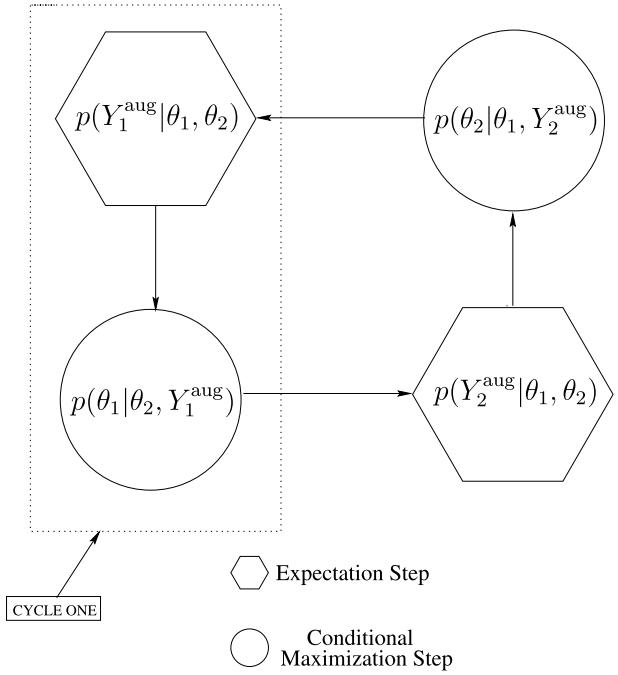

FIG. 11. *A two-cycle AECM algorithm.* [*In the conditional maximization steps, we compute the component of* $\theta = (\theta_1, \theta_2)$ *to maximize the conditional expectation of the log of the quantity in the* ◯, *with the expectation computed in the most recent expectation step, see Section* 6.1.]

the value produced at the end of the previous iteration. If the data augmentation is the same for several consecutive CM-steps (i.e., if g_p is the same) we need only recompute the E-step at the beginning of this sequence. The same requirement holds for ECME in that the steps must be appropriately ordered relative to the E-step. The CM-steps that involve data augmentation must all follow the E-step and be performed before any of the CM-steps that do not involve data augmentation, unless the E-step is repeated. These step-ordering requirements are necessary to ensure monotone convergence of the ECME and AECM algorithms (Meng and van Dyk, 1997). As we discuss next, similar step-ordering requirements apply to the partially collapsed Gibbs sampler.

6.2 The Partially Collapsed Gibbs Sampler

Consider the two-step data augmentation sampler described in Section 3.1. To clarify ideas, we rewrite this sampler with Y^{aug} replaced by ψ and with the conditioning on Y^{obs} suppressed:

Step 1: $\psi^{(t+1)} \sim p(\psi|\theta^{(t)})$,
Step 2: $\theta^{(t+1)} \sim p(\theta|\psi^{(t+1)})$.

Under the standard regularity conditions, we expect that after sufficient burn-in this sampler will effectively return correlated draws from its stationary distribution, $p(\psi, \theta)$. In order to speed up convergence to stationarity and reduce the correlation of the draws, we might take a cue from ECME and AECM and attempt to partially collapse the sampler. In particular, suppose we want to reduce the conditioning in Step 2. A reasonable and optimal strategy might seem to be the following:

Step 1: $\psi^{(t+1)} \sim p(\psi|\theta^{(t)})$,
Step 2: $\theta^{(t+1)} \sim p(\theta)$.

Clearly, $\psi^{(t+1)}$ and $\theta^{(t+1)}$ are independent and the stationary distribution of this sampler is $p(\psi)p(\theta)$ which is generally different than the target distribution, $p(\psi, \theta)$. In this simple example, we need only change the order of the two steps to regain a chain with the target distribution as its stationary distribution. Nonetheless, three important cautionary facts regarding partially collapsed Gibbs samplers are illustrated by this simple example.

First, the "full conditional distributions" of the partially collapsed sampler may not be compatible with *any* joint distribution. In the simple example, this is illustrated by the fact that one cannot find a joint distribution of (ψ, θ) such that ψ depends on θ but θ is independent of ψ. This incompatibility means that we have left the standard Gibbs sampler framework and that standard results as well as our intuition may fail. Second, as with ECME and AECM, the *order* of the steps may matter. Even in this simple case, the stationary distribution of the chain depends on the order of the steps.

Finally, the steps can *sometimes* be blocked to form a standard sampler. If we first draw θ from its marginal distribution and then ψ from its conditional distribution given θ, we are directly sampling from the joint distribution, and have thus blocked the two steps. In fact, blocking is a special case of partially collapsing. It is easy, however, to construct cases where partially collapsed samplers do not correspond to any blocked version of the ordinal sampler (van Dyk and Park, 2008; Park and van Dyk, 2009).

Given these cautionary facts, it is clear that care must be taken when partially collapsing a Gibbs sampler. Van Dyk and Park (2008) give a prescriptive method for construction such samplers that are guaranteed to maintain the target stationary distribution. They also argue that like blocking, partial collapsing improves the convergence characteristics of the chain, but not as much as complete collapsing. This, along with the fact that blocking is a special case of complete collapsing,

unifies the blocking and collapsing strategies. Generally, blocking is not as efficient as collapsing because blocking is only partial collapsing.

7. REFINED ALGORITHMS FOR THE SPECTRAL MODEL

By far the most computationally intensive aspects of the EM and DA algorithms for the spectral model described in Sections 3.2 and 3.4 are the removal of the background counts and the deblurring of the source counts, that is, computing the conditional expectation of or sampling Y_i^+ and $\dot{Y}_j^+$ for $i \in \mathcal{I}$ and $j \in \mathcal{J}$. These tasks involve looking up values in the typically large matrix, M, a time-consuming task even when sophisticated sparse-matrix techniques are implemented. Given the computation cost of these steps and the hierarchical structure of the data augmentation, nesting is an obvious strategy. As an illustration, we implement a nested EM algorithm. In this algorithm we start by setting Y_1^{aug} equal to $\dot{Y}_j^+$ for $j \in \mathcal{J}$. Because this augmentation is smaller than the complete data-augmentation scheme outlined in Table 1, fewer iterations of the EM algorithm are required. Because there is less augmented data, however, the M-step is not in closed form. Thus, we implement an inner EM algorithm to accomplish the M-step of the outer EM algorithm. This strategy is similar to the algorithm illustrated in Figure 10, except the outer E-step does not require a Gibbs sampler but is nonetheless computationally demanding. The inner EM iteration fixes Y_1^{aug} and updates only the first three rows of Table 1 in the inner E-step and θ in the M-step. If this inner EM converges slowly (e.g., there are many and/or weak emission lines), a relatively large number of inner iterations (e.g., 10) may substantially improve the speed of the algorithm. The outer E-step updates all of Y^{aug}.

The advantage of nesting is illustrated using a spectrum of the high redshift quasar S5 0014+81 collected with the *Chandra X-ray Observatory* as described by Elvis et al. (1994). The spectrum is modeled using a power law continuum, $f(\theta^C, E_j) = \gamma E_j^{-\beta}$, exponential absorption, $g(\theta^A, E_j) = e^{\xi/E_j}$, and a single Gaussian emission line with location, width, and intensity parameters[2] for a total of six free parameters. The first two panels of Figure 12 show the convergence of ν, the expected counts attributed to the line,

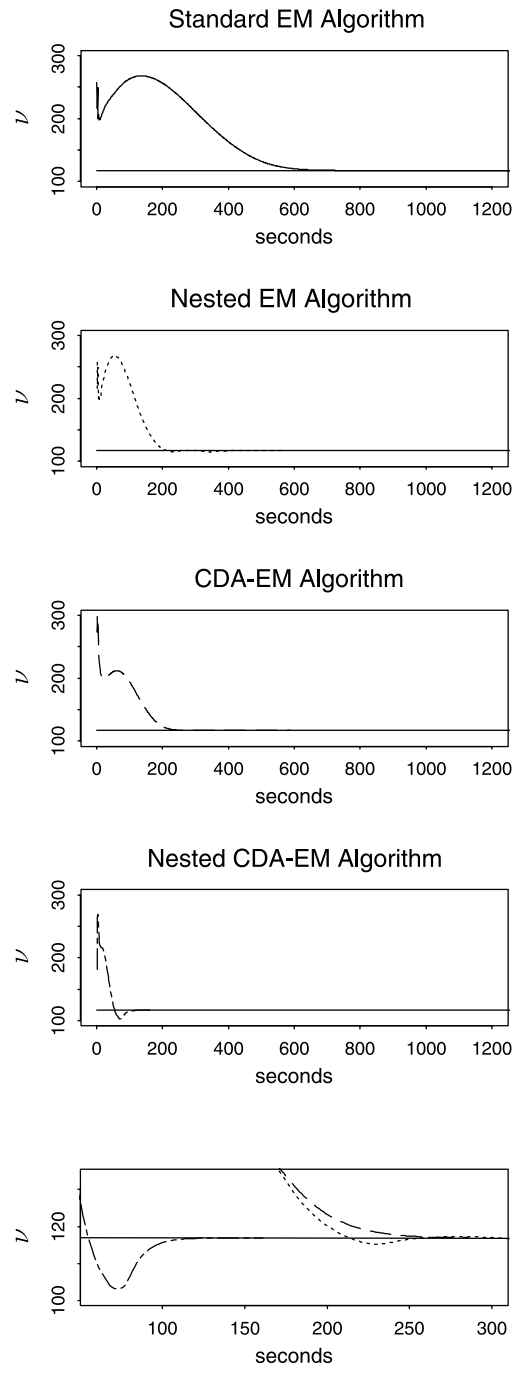

FIG. 12. *Various EM-type algorithms for fitting the spectral model. The figure illustrates the computational advantage of nesting and conditional augmentation. All five plots show the convergence of the parameter ν, the expected line count, as a function of C.P.U. time in seconds. The five plots correspond to the standard EM algorithm based on the data-augmentation scheme outlined in Table 1 (solid line); the nested EM algorithm (dotted line); the CDA-EM algorithm (dashed line); an algorithm that combines nesting and CDA-EM (dotted–dashed line); and a close up of the first 300 seconds comparing all but the standard EM algorithm. The solid horizontal line in each plot is the MLE of ν. The nested and CDA-EM used here are described in Section 7.*

[2] A Gaussian emission line is parameterized as $\frac{\nu}{\sigma}\phi(\frac{E-\mu}{\sigma})$, where ϕ is the standard normal probability density function, μ is the line location, σ is the line width, and ν is the line intensity.

for the EM and nested EM algorithms, respectively. The nested EM algorithm (run with 4 inner iterations) converges in about a third of the time required by the standard EM algorithm. The remaining panels in Figure 12 will be described shortly.

To further improve the convergence of the algorithms, we can reduce the augmented information for θ using the method of conditional augmentation. In particular, we reduce the counts attributed to the absorbed photons in the emission line, $\ddot{Y}_j^L - \dot{Y}_j^L$. Recall that absorption does not occur uniformly across the range of energies of an emission line, and the energies of the observed photons are biased towards areas of low absorption, complicating parameter estimation. Our typical strategy, as described in Table 1, is to treat the absorbed photons as missing data. Thus, in the augmented data, there is no absorption. It is important to note, however, that we need not account for (i.e., augment) all of the absorbed photons, rather we only need the absorption rate to be *constant* across the support energies of the emission line. Thus, a better strategy is to augmented fewer absorbed photons, just enough so that the absorption rates are equal across the range of energies of an emission line. In particular, suppose $a_{\min}$ is the lowest absorption rate, $1 - d_j g(\theta^A, E_j)$, where j varies over the support of the emission line. To reduce the volume of the augmented data, we can compute $\ddot{Y}_j^k$ acting as if the absorption rate were $1 - d_j g(\theta^A, E_j) - a_{\min}$. Here $a_{\min}$ is the optimal value of a working parameter, and we condition on it throughout. In this way, we add fewer counts to each bin. As an extreme example, consider a delta function emission line that is contained entirely within a single energy bin. In this case, the support of the emission line is one bin, $a_{\min} = 1 - d_j g(\theta^A, E_j)$ with j the index of the bin containing the line, $1 - d_j g(\theta^A, E_j) - a_{\min}$ is zero, and we need not impute any missing counts to account for absorption in the line. We emphasize that this does not change the model being fit, it only improves the efficiency of the computation. This strategy is used in the CDA-EM algorithm and is combined with nesting in the nested CDA-EM algorithm; both algorithms are illustrated in Figure 12. The nested EM algorithm and the CDA-EM algorithm (coincidentally) require similar computation time, combining the two strategies, however, is twice as fast as either alone. The final panel in Figure 12 is a more detailed comparison of the three improved algorithms. These algorithms are discussed and further illustrated in van Dyk and Kang (2004).

Other strategies described in this article lead to additional improvements. The posterior distribution or likelihood of the location of a narrow emission line, for example, is typically highly multimodal. The Poisson nature of the data leads to small energy ranges with more counts than expected. These correspond to possible locations of a narrow emission line and may be relatively large modes of the likelihood if the actual line is weak. The standard EM and DA algorithms described here are not able to jump between these modes because line location is updated while conditioning on which photons are attributed to that line. Thus, the line location will be among the energies of these photons and only photons in this energy range will be attributed to the line in the next step. To get around this, van Dyk and Park (2004) and Park and van Dyk (2009) suggest EM-type and DA-type samplers that remove the conditioning on all or part of the augmented data *while updating the line locations*. The result is ECME and AECM algorithms for mode finding and partially collapsed Gibbs samplers for posterior exploration, all of which are much more efficient than the standard EM and DA algorithms.

8. CONCLUDING REMARKS

The highly flexible nature of multilevel modeling inhibits an off-the-shelf algorithmic approach to model fitting. However, the flexibility of a dynamic combination of data augmentation and model reduction give us tools to tackle these models. As illustrated in the spectral model, the many recent extensions and refinements of data augmentation methods can substantially improve computational speed while maintaining simplicity and stable convergence, thus greatly extending the applicability and power of data augmentation .

The data-augmentation and model-reduction strategies outlined in this article have been used either explicitly or implicitly to derive numerous efficient EM-type and DA-type algorithms with applications to a wide range of models including longitudinal data analysis for binary response and robust methods, robust regression, binary and grey-level Ising models, dynamic linear models, finite mixture models, Poisson image analysis, probit regression, multinomial probit models, switching-state space models, factor analysis, spectral analysis, etc. A small subset of examples can be found in Liu and Rubin (1994, 1995), Gelfand, Sahu and Carlin (1995), Meng and van Dyk (1997, 1998, 1999), van Dyk and Tang (2003), van Dyk and Park (2004), Higdon (1998), Pilla and Lindsay (2001), Liu, Rubin and Wu (1998), van Dyk (2000a, 2000b), Liu and Wu (1999), van Dyk and Meng (2001), Foulley and van Dyk (2000), van Dyk and Kang (2004), Imai and

van Dyk (2005a, 2005b), Gelman et al. (2008), Pope and Wong (2005) and Ghosh and Dunson (2009). We hope that this overview paper will help to both further stimulate methodological research and promote efficient implementation of EM-type and DA-type algorithms in practice. In other words, to paraphrase the title, we hope practitioners will have an easier time to climb likelihood surfaces using EM-type algorithms and to explore posterior landscape using DA-type samplers.

ACKNOWLEDGMENTS

David A. van Dyk is supported in part by NSF Grants DMS-04-06085, SES-05-50980 and DMS-09-07522. Xiao-Li Meng is supported in part by NSF Grants DMS-04-05953, DMS-05-05595, DMS-06-52743 and DMS-09-07185.

REFERENCES

AMIT, Y. (1991). On rates of convergence of scholastic relaxation for Gaussian and non-Gaussian distributions. *J. Multiple Anal.* **38** 82–89. MR1128938

BESAG, J. and GREEN, P. J. (1993). Spatial statistics and Bayesian computation. *J. Roy. Statist. Soc. Ser. B* **55** 25–37. MR1210422

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–37. MR0501537

ELVIS, M., MATSUOKA, M., SIEMIGINOWSKA, A., FIORE, F., MIHARA, T. and BRINKMANN, W. (1994). An ASCA GIS spectrum of S5 0014 + 813 at $z = 3.384$. *The Astrophysical Journal* **436** L55–L58.

FESSLER, J. A. and HERO, A. O. (1994). Space-alternating generalized expectation-maximization algorithm. *IEEE Trans. Signal Process.* **42** 2664–2677.

FESSLER, J. A. and HERO, A. O. (1995). Penalized maximum-likelihood image reconstruction using space-alternating generalized EM algorithm. *IEEE Trans. Image Process.* **4** 1417–1438.

FOULLEY, J.-L. and VAN DYK, D. A. (2000). The PX-EM algorithm for fast stable fitting of Henderson's mixed model. *Genetics Selective Evolution* **32** 143–163.

GELFAND, A. E., SAHU, S. K. and CARLIN, B. P. (1995). Efficient parameterization for normal linear mixed models. *Biometrika* **82** 479–488. MR1366275

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2003). *Bayesian Data Analysis*, 2nd ed. Chapman & Hall, London. MR2027492

GELMAN, A., VAN DYK, D. A., HUANG, Z. and BOSCARDIN, W. J. (2008). Transformation and parameter-expanded Gibbs samplers for multilevel and generalized linear models. *J. Comput. Graph. Statist.* **17** 95–122. MR2424797

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence* **6** 721–741.

GHOSH, J. and DUNSON, D. (2009). Default priors and efficient posterior computation in Bayesian factor analysis. *J. Comput. Graph. Statist.* **18** 306–320.

GREEN, P. J. (1990). On use of the EM algorithm for penalized likelihood estimation. *J. Roy. Statist. Soc. Ser. B* **52** 443–452. MR1086796

HANS, C. M. and VAN DYK, D. A. (2003). Accounting for absorption lines in high energy spectra. In *Statistical Challenges in Modern Astronomy III* (E. Feigelson and G. Babu, eds.) 429–430. Springer, New York.

HIGDON, D. M. (1998). Auxiliary variable methods for Markov chain Monte Carlo with applications. *J. Amer. Statist. Assoc.* **93** 585–595.

HOBERT, J. P. (2001). Discussion of "The art of data augmentation," by D. A. van Dyk and X. L. Meng. *J. Comput. Graph. Statist.* **10** 59–68. MR1936358

HOBERT, J. P. and MARCHEV, D. (2008). A theoretical comparison of the data augmentation, marginal augmentation and PX-DA algorithms. *Ann. Statist.* **36** 532–554. MR2396806

IMAI, K. and VAN DYK, D. A. (2005a). A Bayesian analysis of the multinomial probit model using marginal data augmentation. *J. Econometrics* **124** 311–334. MR2125369

IMAI, K. and VAN DYK, D. A. (2005b). MNP: R package for fitting multinomial the probit model. *J. Statist. Software* **14**.

LIU, C. and RUBIN, D. B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika* **81** 633–648. MR1326414

LIU, C. and RUBIN, D. B. (1995). ML estimation of the t distribution using EM and its extensions, ECM and ECME. *Statist. Sinica* **5** 19–39. MR1329287

LIU, C., RUBIN, D. B. and WU, Y. N. (1998). Parameter expansion for EM acceleration—the PXEM algorithm. *Biometrika* **75** 755–770. MR1666758

LIU, J. S. (1994). The fraction of missing information and convergence rate for data augmentation. In *Computing Science and Statistics. Computationally Intensive Statistical Methods. Proceedings of the 26th Symposium on the Interface* 490–497. Interface Foundation of North America, Fairfax Station, VA.

LIU, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer, New York. MR1842342

LIU, J. S., WONG, W. H. and KONG, A. (1994). Covariance structure of the Gibbs sampler with applications to comparisons of estimators and augmentation schemes. *Biometrika* **81** 27–40. MR1279653

LIU, J. S. and WU, Y. N. (1999). Parameter expansion for data augmentation. *J. Amer. Statist. Assoc.* **94** 1264–1274. MR1731488

MARCHEV, D. and HOBERT, J. P. (2004). Geometric ergodicity of van Dyk and Meng's algorithm for the multivariate student's t model. *J. Amer. Statist. Assoc.* **99** 228–238. MR2054301

MENG, X.-L. (1994). On the rate of convergence of the ECM algorithm. *Ann. Statist.* **22** 326–339. MR1272086

MENG, X.-L. (1997). The EM algorithm and medical studies: A historical link. *Stat. Methods Med. Res.* **6** 3–23.

MENG, X.-L. and RUBIN, D. B. (1991). Using EM to obtain asymptotic variance–covariance matrices: The SEM algorithm. *J. Amer. Statist. Assoc.* **86** 899–909.

MENG, X.-L. and RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80** 267–278. MR1243503

MENG, X.-L. and RUBIN, D. B. (1994). On the global and componentwise rates of convergence of the EM algorithm. *Linear Algebra Appl.* **199** 413–425. MR1274429

MENG, X.-L. and VAN DYK, D. A. (1997). The EM algorithm—an old folk song sung to a fast new tune (with discussion). *J. Roy. Statist. Soc. Ser. B* **59** 511–567. MR1452025

MENG, X.-L. and VAN DYK, D. A. (1998). Fast EM implementations for mixed-effects models. *J. Roy. Statist. Soc. Ser. B* **60** 559–578. MR1625942

MENG, X.-L. and VAN DYK, D. A. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika* **86** 301–320. MR1705351

NAVIDI, W. (1997). A graphical illustration of the EM algorithm. *Amer. Statist.* **51** 29–31.

PARK, T. and VAN DYK, D. A. (2009). Partially collapsed Gibbs samplers: Illustrations and applications. *J. Comput. Graph. Statist.* **18** 283–305.

PARK, T., VAN DYK, D. A. and SIEMIGINOWSKA, A. (2008). Searching for narrow emission lines in X-ray spectra: Computation and methods. *The Astrophysical Journal* **688** 807–825.

PILLA, R. S. and LINDSAY, B. G. (2001). Alternative EM methods for nonparametric finite mixture models. *Biometrika* **88** 535–550. MR1844850

POPE, C. A. and WONG, Y. (2005). Nested Monte Carlo EM algorithm for switching state-space models. *IEEE Trans. Knowledge Data Engineering* **17** 1653–1663.

PROTASSOV, R., VAN DYK, D. A., CONNORS, A., KASHYAP, V. and SIEMIGINOWSKA, A. (2002). Statistics: Handle with care—detecting multiple model components with the likelihood ratio test. *The Astrophysical Journal* **571** 545–559.

ROBERTS, G. O. (1996). Markov chain concepts related to sampling algorithms. In *Markov Chain Monte Carlo in Practice* (W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds.) 45–57. Chapman & Hall, London. MR1397967

SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London. MR1692799

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82** 528–550. MR0898357

TIERNEY, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22** 1701–1762. MR1329166

TIERNEY, L. (1996). Introduction to general state-space Markov chain theory. In *Markov Chain Monte Carlo in Practice* (W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds.) 59–74. Chapman & Hall, London. MR1397968

VAIDA, F. (2005). Convergence of the EM and MM algorithms. *Statist. Sinica* **15** 831–840. MR2233916

VAN DYK, D. and PARK, T. (2004). Efficient EM-type algorithms for fitting spectral lines in high-energy astrophysics. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: Contributions by Donald Rubin's Statistical Family* (A. Gelman and X.-L. Meng, eds.) 285–296. Wiley, New York. MR2138264

VAN DYK, D. and PARK, T. (2008). Partially collapsed Gibbs samplers: Theory and methods. *J. Amer. Statist. Assoc.* **103** 790–796.

VAN DYK, D. A. (2000a). Fitting mixed-effects models using efficient EM-type algorithms. *J. Comput. Graph. Statist.* **9** 78–98. MR1826277

VAN DYK, D. A. (2000b). Nesting EM algorithms for computational efficiency. *Statist. Sinica* **10** 203–225. MR1742109

VAN DYK, D. A. (2009). Marginal MCMC Methods. *Statist. Sinica*. To appear.

VAN DYK, D. A., CONNORS, A., ESCH, D. N., FREEMAN, P., KANG, H., KAROVSKA, M., KASHYAP, V., SIEMIGINOWSKA, A. and ZEZAS, A. (2006). Deconvolution in high-energy astrophysics: Science, instrumentation, and methods. *Bayesian Anal.* **1** 189–236. MR2221261

VAN DYK, D. A., CONNORS, A., KASHYAP, V. and SIEMIGINOWSKA, A. (2001). Analysis of energy spectra with low photon counts via Bayesian posterior simulation. *The Astrophysical Journal* **548** 224–243.

VAN DYK, D. A. and KANG, H. (2004). Highly structured models for spectral analysis in high-energy astrophysics. *Statist. Sci.* **19** 275–293. MR2140542

VAN DYK, D. A. and MENG, X.-L. (2001). The art of data augmentation (with discussion). *J. Comput. Graph. Statist.* **10** 1–111. MR1936358

VAN DYK, D. A., MENG, X.-L. and RUBIN, D. B. (1995). Maximum likelihood estimation via the ECM algorithm: Computing the asymptotic variance. *Statist. Sinica* **5** 55–75. MR1329289

VAN DYK, D. A. and TANG, R. (2003). The one-step-late PXEM algorithm. *Stat. Comput.* **13** 137–152. MR1963330

WEI, G. and TANNER, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *J. Amer. Statist. Assoc.* **85** 699–704.

WU, C. F. J. (1983). On the convergence properties of the EM algorithms. *Ann. Statist.* **11** 95–103. MR0684867

YU, Y. and MENG, X.-L. (2010). To center or not to center: That is not the question—An ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency (with discussion). *J. Comput. Graph. Statist.* To appear.