Springer
New York
Berlin
Heidelberg
Hong Kong
London
Milan
Paris
Tokyo

Eric D. Feigelson   G. Jogesh Babu
Editors

Statistical Challenges in Astronomy

With 104 Illustrations

Springer

3

Hierarchical Models, Data Augmentation, and Markov Chain Monte Carlo

David A. van Dyk[1]

ABSTRACT The ever increasing power and sophistication of today's high energy astronomical instruments is opening a new realm of high quality data that is quickly pushing beyond the capabilities of the "classical" data-analysis methods in common use. In this chapter we discuss the use of highly structured models that not only incorporate the scientific model (e.g., for a source spectrum) but also account for stochastic components of data collection and the instrument (e.g., background contamination and pile up). Such hierarchical models when used in conjunction with Bayesian or likelihood statistical methods offer a systematic solution to many challenging data analytic problems (e.g., low count rates and pile up). Hierarchical models are becoming increasingly popular in physical and other scientific disciplines largely because of the recent development of sophisticated methods for statistical computation. Thus, we discuss such methods as the EM algorithm, data augmentation, and Markov chain Monte Carlo in the context of high energy high resolution low count data.
This paper is followed by a commentary by astronomer Michael Strauss.

3.1 Introduction

Today's highly sophisticated astronomical instruments offer a new window into the complexities of the visible and invisible universe. As the state of instrumentation evolves to produce ever finer resolution in spectral, spatial, and temporal data ever more sophisticated statistical techniques are required to properly handle this data. For example, standard off-the-shelf methods such as χ^2 fitting and background subtraction are ill-equipped to handle the high resolution low count per bin data available from such instruments as the Chandra X-ray Observatory. See Siemiginowska et al. (1997) and van Dyk et al. (2001) for a general discussion of such issues. The Gaussian assumptions implicit in such methods are not justified with low counts and the resulting fits and error bars are therefore unreliable. Testing

[1]Department of Statistics, Harvard University

for model features such as spectral lines or a source above background is always a challenging task and standard methods such as the F-test, likelihood ratio test, and Cash statistic though commonly used in practice are inappropriate (Protassov et al. 2002). An even greater challenge is properly accounting for pile-up in X-ray detectors, a task that confounds standard techniques and thus demands more sophisticated statistical methods.

In this chapter, we outline a paradigm for data analysis that we believe is robust enough to systematically handle these and many other statistical challenges presented by modern astronomical instruments. It is important conceptually to break any data analysis scheme into (at least) three components, all of which are critical and must be done thoughtfully to ensure sound inference. These components are *model building*, *statistical inference*, and *statistical computation*.

The importance of careful model building is evident in the complexity and subtlety of the physical mechanisms giving rise to the observed data of modern instrumentation. The instrument response blurs the energy and sky coordinates of photons, counts are contaminated with background, the effective area of the instrument and the propensity of photons to be absorbed vary with energy, pile-up masks the energy and count of incoming photons, source spectral models are complex and may include emission and absorption features as well as a continuum. A statistical model should aim to describe all such components of data generation. Thus, by a *model* we mean much more than a parametric description of how the mean source flux varies with energy and/or sky coordinates. Models that include statistical descriptions of the processes that degrade the data can guide us in accounting for these degradations and eliminate the need for ad-hoc corrections, e.g., for pile-up and background. Because of the complexity of these models, we organize them into a hierarchical structure, which is formulated in terms of various unobserved quantities (e.g., counts without background contamination). Such unobserved quantities are often called *augmented data* and play an important role in the computational methods we suggest.

Once a model is formulated, statistical inference involves drawing inferences (e.g., point estimates and error bars) regarding unobserved quantities such as the model parameters describing the flux of the source. Important model-based modes of statistical inference include maximum likelihood and Bayesian inference. With large samples the asymptotic Gaussian behavior of the maximum likelihood estimate can be the basis for sound frequentist inference. Nevertheless, we generally take a Bayesian perspective for a number of practical reasons such as a ready mechanism for combining information from multiple sources, mathematical justification in small samples, and an obvious framework for handling nuisance parameters. Despite the placement of this chapter in a Bayesian section, we say very little about the relative merit of Bayesian and frequentist methods; our emphasis is on model building and statistical computation. Because of the aforementioned

practical advantages of Bayesian methods, they are often the only tractable methods available for fitting complex models—which is motivation enough for many practical minded statisticians to "be Bayesian." Here we give only enough details of Bayesian and likelihood methods to motivate the computational tools, giving somewhat more emphasis to Bayesian methods. For further reading on Bayesian methods, we recommend one of the several high-quality recent texts on the subject such as Gelman et al. (1995), Carlin & Louis (1996) and Gilks et al. (1996), as well as other chapters in this volume including those by Connors, Loredo & Chernoff, and Berger *et al.*.

Because of the highly-structured nature of the statistical models that we propose, sophisticated computational methods (e.g., the EM algorithm, the data augmentation algorithm, and Markov chain Monte Carlo) are often required. The methods we suggest are designed to be computationally stable and generally easy to implement. The details of the algorithm often follow directly from the hierarchical model specification via simple statistical calculations.

The remainder of this chapter is organized into five sections. In Section 3.2 we introduce a simple example, accounting for background contamination of counts. We use this example to motivate hierarchical modeling and the method of data augmentation, which are in turn generalized and more fully developed in Section 3.3. The computational methods are introduced and illustrated using the motivating example of background contamination in Section 3.4. In Section 3.5 we outline how these methods can be used to tackle the difficult problem of photon pile-up. Concluding remarks regarding the direction of modern statistical analysis appear in Section 3.6.

3.2 A Motivating Example

In this section we introduce a simple example that is used throughout the chapter to motivate ideas and methods. The example is simple so as not to distract attention from the statistical methods. As illustrated in Section 3.5, however, hierarchical models, data augmentation, and MCMC can tackle much more complicated problems.

Suppose we have observed counts, Y, contaminated with background in a (source) exposure and have observed a second exposure of pure background resulting in counts, Z. Throughout we assume the source exposure is τ_S seconds and the pure background exposure is τ_B seconds with both exposures using the same area of the detector. To model the source exposure, we assume Y follows a Poisson distribution[2] with intensity $\lambda_B + \lambda_S$, where

[2]Recall $Y \overset{d}{\sim}$ Poisson (λ) (read as Y is distributed as Poisson with intensity λ) indicates that Y follows the Poisson distribution with intensity and expectation λ, i.e.,

λ_B and λ_S represent the expected counts during the source exposure due to background and source respectively. Thus, the distribution function for Y given λ_B and λ_S is

$$p(Y|\lambda_B, \lambda_S) = e^{-(\lambda_B+\lambda_S)}(\lambda_B + \lambda_S)^Y/Y! \text{ for } Y = 0, 1, 2, \ldots. \quad (3.1)$$

We wish to estimate λ_S and treat λ_B as a *nuisance* parameter, a parameter that is of little interest, but must be included in the model. The expected counts during the background exposure are assumed to be the same as in the source exposure, but corrected for the exposure time, $\lambda_B\tau_B/\tau_S$. I.e.,

$$p(Z|\lambda_B, \lambda_S) = e^{-(\lambda_B\tau_B/\tau_S)}(\lambda_B\tau_B/\tau_S)^Z/Z! \text{ for } Z = 0, 1, 2, \ldots. \quad (3.2)$$

Maximum likelihood estimation involves estimating λ_B and λ_S by the values the maximize the likelihood function, i.e., the product of Equations 3.1 and 3.2. Under certain regularity conditions (e.g., $\lambda_B, \lambda_S > 0$), maximum likelihood estimates asymptotically follow a Gaussian distribution. This result leads to confidence intervals and error bars with (asymptotic) frequentist properties.

Bayesian inference is based on the *posterior* distribution,

$$p(\lambda_S, \lambda_S B|Y, Z) \propto p(Y|\lambda_B, \lambda_S)p(Z|\lambda_B, \lambda_S)p(\lambda_B, \lambda_S), \quad (3.3)$$

where $p(\lambda_B, \lambda_S)$ is the *prior* distribution which quantifies information regarding the values of the λ_S and λ_B available prior to observing the data. The posterior distribution combines such prior information with the information in the observed counts. The posterior distribution is a complete summary of our information, but if it is similar to Gaussian in shape, it is often summarized by its mean vector and variance matrix that can be used as point estimates and to compute error bars. The posterior distribution can also be used to compute a ζ-level credible region, R, such that $\int_R p(\lambda_S, \lambda_B|Y, Z)d\lambda_S d\lambda_B = \zeta$. Such probability statements should be regarded as summaries of the available information for the model parameters, in contrast to the frequentist interpretation of a confidence interval.

Implicitly, the counts from the source exposure, Y, are made up of two components, $Y = Y_S + Y_B$, where Y_S are counts from the source exposure due to the source and Y_B are the counts due to background. Since neither Y_S nor Y_B are observed, we call these counts *missing data.* We note that if Y_S and Y_B had been observed, our statistical analysis would be greatly simplified since we could confine attention to $Y_S \overset{d}{\sim} \text{Poisson}(\lambda_S)$. Of course, it is impossible to observe Y_S and Y_B. Nonetheless, this "thought experiment" offers insight into computational methods that are useful both for Bayesian and likelihood-based inference. In particular, the method of data augmentation is an elegant computational construct allowing us to take

$p(Y = y) = e^{-\lambda}y^{\lambda}/y!$.

advantage of the fact that if it were possible to collect additional data, statistical analysis could be greatly simplified. This is true regardless of why the so-called "missing-data" are not observed. There is a large class of powerful statistical methods designed for "missing-data" problems. These methods have broad application in astrophysics (and in the physical sciences generally) once we note that quantities observed with measurement error can be regarded as "missing-data".

To illustrate the method of data augmentation, we begin by reformulating our model in terms of Y_S and Y_B. In particular, consider the multi-level or *hierarchical model*

Level 1: $Y|Y_B, \lambda_S \overset{d}{\sim} \text{Poisson}(\lambda_S) + Y_B$,

Level 2: $Y_B|\lambda_B \overset{d}{\sim} \text{Poisson}(\lambda_B)$ and $Z|\lambda_B \overset{d}{\sim} \text{Poisson}(\lambda_B\tau_B/\tau_S)$,

Level 3 (optional): specify a prior distribution for λ_B and λ_S.

Notice that in each level of the model, we specify the distribution of random quantities conditioning on unobserved quantities whose distribution is specified in lower levels of the model. For example, in Level 1, we condition on Y_B, the distribution of which is specified in Level 2. *The power of such a hierarchical model is that it separates a complex model into a number of easy to handle smaller parts.*

If Y_S and Y_B were observed, Level 1 specifies the form of the likelihood for λ_S, i.e.,

$$L(\lambda_S|Y_S) = e^{-\lambda_S}\lambda_S^{Y_S}, \quad (3.4)$$

and Level 2 specified the form of the likelihood for λ_B, i.e.,

$$L(\lambda_B|Y_B, Z) = e^{-\lambda_B k}\lambda_B^{Y_B+Z}, \quad (3.5)$$

where $k = (\tau_S + \tau_B)/\tau_S$. Notice that Equations 3.1 and 3.2 are relatively complex functions of λ_S and λ_B and are harder to, for example, maximize than are Equations 3.4 and 3.5.

It is also easy to estimate the "missing data" in this hierarchical model. In particular, if λ_B and λ_S were known, the conditional distribution of Y_B given Y can be computed using Bayes Theorem,

$$p(Y_B|Y, \lambda_S, \lambda_B) = \frac{p(Y|Y_B, \lambda_S, \lambda_B)\, p(Y_B|\lambda_S, \lambda_B)}{p(Y|\lambda_S, \lambda_B)} \quad (3.6)$$

$$= \binom{Y}{Y_B}\left(\frac{\lambda_B}{\lambda_S + \lambda_B}\right)^{Y_B}\left(\frac{\lambda_S}{\lambda_S + \lambda_B}\right)^{Y-Y_B}. \quad (3.7)$$

That is,

$$Y_B|Y, \lambda_S, \lambda_B \overset{d}{\sim} \text{Binomial}^3\left[Y, \lambda_B/(\lambda_S + \lambda_B)\right]. \quad (3.8)$$

[3]Recall $Y \overset{d}{\sim} \text{Binomial}(n, P)$ indicates that Y follows a binomial distribution with n

Thus, given the model parameters, we can predict the "missing data" (e.g., by its conditional expectation with error bars based on its conditional standard deviation). Likewise, given the "missing data" we can estimate the model parameters (e.g., using maximum likelihood or a Bayesian estimate). This leads to an iterative strategy that updates the "missing data" given the model parameters and then the model parameters given the "missing data." Such computational methods include the EM algorithm and the Data Augmentation (DA) algorithm and are referred to generally as the method of data augmentation.

In the next two sections we abstract and generalize the important features of this example to construct robust tools for analysis of the high resolution high quality data available with today's sophisticated instruments. In Section 3.4 we show how data augmentation can be used to compute maximum likelihood estimates, Bayesian posterior modes and means, as well as error bars. Generally these methods involve maximizing, simulating, and computing expectations of standard distribution functions. Such simple distributions often arise naturally from a hierarchical model expressed in terms of the "missing data," e.g., Equations 3.4, 3.5, and 3.8. Details of the computation stability as well as examples which illustrate the computational simplicity appear in the following sections.

3.3 Data Augmentation and Hierarchical Models

The term "data augmentation" originated with computational methods designed to handle missing data, but as illustrated in Section 3.2, the method is really quite general and often useful when there is no missing data per se. In particular, for Monte Carlo integration in Bayesian data analysis we aim to obtain a sample from the posterior distribution, $p(\boldsymbol{\theta}|\mathrm{Y})$. In some cases, we can *augment* the model to $p(\boldsymbol{\theta}, \mathrm{X}|\mathrm{Y})$, where X may be missing data or any other unobserved quantity (e.g., counts due to background). With judicious choice of X, it may be much easier to obtain a sample from $p(\boldsymbol{\theta}, \mathrm{X}|\mathrm{Y})$ than directly from $p(\boldsymbol{\theta}|\mathrm{Y})$. Once we have a sample from $p(\boldsymbol{\theta}, \mathrm{X}|\mathrm{Y})$, we simply discard the sample of X to obtain a sample from $p(\boldsymbol{\theta}|\mathrm{Y})$. The notation here is more general, but the idea is exactly that of Section 3.2; we use statistical insight to construct $p(\boldsymbol{\theta}, \mathrm{X}|\mathrm{Y})$ so that both $p(\boldsymbol{\theta}|\mathrm{X}, \mathrm{Y})$ and $p(\mathrm{X}|\boldsymbol{\theta}, \mathrm{Y})$ are simple or at least standard distributions.

Absorption Lines. Absorption can be accounted for by supposing the expected counts in energy bin i are $\mathcal{F}_i\pi_i$, where $\mathcal{F}_i$ would be the expected counts if there were no absorption and π_i is the expected proportion of counts in energy bin i that are not absorbed. (We might, for example, parameterize $\mathcal{F}_i$ as a power law.) In particular, we might model the counts in energy bin i as $Y_i \overset{\text{d}}{\sim}$ Poisson $(\mathcal{F}_i\pi_i)$. To formulate this model using data augmentation, we let Y_i^+ be the unobserved counts that the detector would have detected if no photons were absorbed. We can then formulate the hierarchical model,

LEVEL 1: $Y_i|Y_i^+, \mathcal{F}_i, \pi_i \overset{\text{d}}{\sim}$ Binomial (Y_i^+, π_i),

LEVEL 2: $Y_i^+|\mathcal{F}_i \overset{\text{d}}{\sim}$ Poisson $(\mathcal{F}_i)$,

LEVEL 3 (optional): specify prior distributions for $\mathcal{F}_i$ and π_i.

Again, the power of the data augmentation is the ability to partition the model complexity into simpler pieces, in this case a binomial absorption model and a Poisson spectral model with no absorption.

Many standard absorption models (including absorption lines) and continuum spectral models (e.g., power laws and bremsstrahlung emission) can be formulated using simple transformations of π_i and $\mathcal{F}_i$ that are linear functions of energy. In this case, given the "missing" absorbed photon counts both LEVEL 1 and LEVEL 2 specify Generalized Linear Models that are well studied and generally easy to fit. Likewise, given the model parameters and the observed data, the absorbed photons follow a simple model, $Y_i^+ \overset{\text{d}}{\sim}$ Poisson $[(1-\pi_i)\mathcal{F}_i] + Y_i$.

Emission Lines. Spectral models often include emission lines,

$$\mathcal{F}_i = c(E_i) + \sum_{k=1}^{K} \delta_{ik}$$

where $c(E_i)$ is the expected continuum counts in energy bin i and δ_{ik} is the expected counts from emission like k in energy bin i. For each photon, we postulate a variable that specifies whether the photon is due to the continuum or a particular emission line. This unobserved specification variable is treated as "missing data." Given this variable we can fit the continuum using the counts due to the continuum without the complication of emission lines. Likewise we can fit each of the emission lines (e.g., parameters specifying a Gaussian or Lorenzian distribution) using the counts attributed to that line. Conversely, given the parameter of continuum and the emission lines, the specification variable for each photon follows a simple multinomial distribution.

Multiple Model Components. So far, we have divided the unobserved quantities into two groups, the model parameters and the "missing data." More generally, we may partition $\boldsymbol{\theta}$ into $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)$, where some component of θ are model parameters of scientific interest, others may be

independent trials each with probability p, i.e., $\Pr(Y=y) = \binom{n}{y}p^y(1-p)^{n-y}$. As an example, Y may be the random number of heads in n independent flips of a (possibly unfair) coin that has probability p of coming up heads on each flip.

nuisance parameters, and still others may be "missing data" or other unobserved quantities. The key is that we select the unobserved quantities and the partition of $\boldsymbol{\theta}$ so that $p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{k-1},\boldsymbol{\theta}_{k+1},\ldots,\boldsymbol{\theta}_p,\mathrm{Y})$ is a standard distribution for each k. In this way we partition a complex problem into a sequence of simpler standard problems which we handle iteratively and one at a time. Thus, we can easily account for absorption, emission lines, instrument response, and background, all in the context of a Poisson model without sacrificing numerical stability, computational simplicity, or sound statistical inference. Details of such a model appear in van Dyk et al. (2001); see also van Dyk's discussion of Strauss (this volume).

3.4 Model Fitting

In Sections 3.2 and 3.3 we emphasize repeatedly that judicious choice of the "missing data," X, can lead to simple conditional models, $p(\boldsymbol{\theta}|\mathrm{X},\mathrm{Y})$ and $p(\mathrm{X}|\boldsymbol{\theta},\mathrm{Y})$, even when $p(\boldsymbol{\theta}|\mathrm{Y})$ is much more complex. In this section we show how these simple conditional models can be used to construct computation tools for likelihood-based and Bayesian model fitting. In recent years, these tools have become popular throughout the social, physical biological and engineering sciences primarily because of their computational stability and simplicity.

3.4.1 *The EM Algorithm*

Dempster et al. (1977) formulated the expectation maximization (EM) algorithm to compute a maximum likelihood estimate, that is

$$\hat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}|\mathrm{Y}), \tag{3.9}$$

where Y is the observed data, $\boldsymbol{\theta}$ is a model parameter, $L(\boldsymbol{\theta}|\mathrm{Y})$ is the likelihood function, and $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate. (More generally, we can replace $L(\boldsymbol{\theta}|\mathrm{Y})$ with a posterior distribution in Equation 3.9 and use EM to compute the posterior mode, $\hat{\theta}$.) In particular, Dempster et al. (1977) considered maximum likelihood estimation in the presence of incomplete data or problems that can be formulated as such (e.g., spectral imaging with background or degraded counts). In this context, the EM algorithm builds on the intuitive idea that (i) if there were no "missing data," maximum likelihood estimation would be easy, and (ii) if the model parameters were known, the "missing data" could easily be imputed (i.e., predicted) by its (conditional) expectation.

These two steps take on a simple form in the context of the background example described in Section 3.2. In particular, if Y_S had been observed, we could estimate λ_S with Y_S. Likewise, if λ_S and λ_B were known, Y_S could be estimated as the proportion of the observed counts, Y, implied by λ_S and λ_B, i.e., the conditional expectation of Y_S, $Y\lambda_S/(\lambda_B+\lambda_S)$. This leads naturally to a two-step iteration which converges to the maximum likelihood estimate. It should be noted that this procedure necessarily leads to a non-negative estimate of λ_S, whereas the common estimate resulting from "subtracting background," $Y - Z\tau_S/\tau_B$, may be negative.

The two steps in this simple iteration correspond to the M-step (i.e., maximization step) and the E-step (i.e., expectation step) of EM respectively, with the proviso that not the missing data, but rather the so-called augmented-data log likelihood should be imputed by its conditional expectation. In general, we begin by defining the *missing data*, X, and the corresponding loglikelihood, $L(\boldsymbol{\theta}|\mathrm{Y},\mathrm{X})$. EM starts with an initial value[4] $\boldsymbol{\theta}^{(0)}\in\Theta$ and iterates the following two steps for $t=0,1,\ldots$

E-step: Compute the conditional expectation of the loglikelihood corresponding to the augmented data (Y, X), given the observed data and the current parameter value,

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathrm{E}\left[\log L(\boldsymbol{\theta}|\mathrm{Y},\mathrm{X})\middle|\mathrm{Y},\boldsymbol{\theta}^{(t)}\right] \; ; \tag{3.10}$$

M-step: Determine $\boldsymbol{\theta}^{(t+1)}$ by maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, that is, find $\boldsymbol{\theta}^{(t+1)}$ so that $Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ for all $\boldsymbol{\theta}\in\Theta$;

until convergence. The usefulness of the EM algorithm is apparent when both of these steps can be accomplished with minimal analytic and computation effort but the direct maximization in Equation 3.9 is difficult. In many common models (e.g., multivariate Gaussian, Poisson, binomial, exponential, etc.) $\log L(\boldsymbol{\theta}|\mathrm{Y},\mathrm{X})$ is linear in a set of simple "augmented-data sufficient statistics." Thus, as will be illustrated below, computing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ involves routine calculations. The M-step then only requires computing the maximum likelihood estimates as if there were no "missing data," by using the predicted augmented-data sufficient statistics from the E-step as data.

To illustrate these ideas, we return to the example of Section 3.2. We set $\mathrm{X}=\{Y_S,Y_B\}$, $\mathrm{Y}=\{Y,Y_B\}$, and $\boldsymbol{\theta}=(\lambda_B,\lambda_S)$. In this case, $\log L(\boldsymbol{\theta}|\mathrm{Y},\mathrm{X}) = \log L(\lambda_S|Y_S)+\log(\lambda_B|Y_B,Z)$; see Equations 3.4 and 3.5. Thus, $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) =$

$$-\lambda_S + \mathrm{E}\left(Y_S|Y,\boldsymbol{\theta}^{(t)}\right)\log\lambda_S - k\lambda_B + \left[Z+\mathrm{E}\left(Y_B|Y,\boldsymbol{\theta}^{(t)}\right)\right]\log\lambda_B . \tag{3.11}$$

Elementary calculations show the expectations in Equation 3.11 are given by $Y\lambda_S/(\lambda_B+\lambda_S)$ and $Y\lambda_B/(\lambda_B+\lambda_S)$, which is the E-step, and $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is maximized by $\lambda_S^{(t+1)} = \mathrm{E}(Y_S|Y,\lambda_S^{(t)})$ and $\lambda_B^{(t+1)} = [Z+\mathrm{E}(Y_B|Y,\lambda_S^{(t)})]/k$, which is the M-step.

[4]Parenthetic superscripts indicate iteration number.

3.4.2 The Data Augmentation Algorithm

In the context of Bayesian data analysis, numerical summaries of the posterior distributions are often computed via numerical integration. Because of the high dimension of the parameter space in most practical problems, Monte Carlo integration is really the only useful method. If we can obtain a sample from the posterior distribution, $\{\boldsymbol{\theta}^{(t)}, t = 1, \ldots, T\}$, Monte Carlo integration approximates the posterior mean of any function, g, of the parameter with

$$\mathrm{E}[g(\boldsymbol{\theta})|\mathrm{Y}] = \int g(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathrm{Y})d\boldsymbol{\theta} \approx \frac{1}{T}\sum_{t=1}^{T} g(\boldsymbol{\theta}^{(t)}), \tag{3.12}$$

where we assume $\mathrm{E}[g(\boldsymbol{\theta})|\mathrm{Y}]$ exists. For example, $g(\boldsymbol{\theta}) = \boldsymbol{\theta}$ and $g(\boldsymbol{\theta}) = (\boldsymbol{\theta} - \mathrm{E}(\boldsymbol{\theta}|\mathrm{Y}))(\boldsymbol{\theta} - \mathrm{E}(\boldsymbol{\theta}|\mathrm{Y}))'$ lead to the posterior mean and variance respectively. Probabilities, such as $\zeta = \Pr(\boldsymbol{\theta} \in R)$ can be computed using $g(\boldsymbol{\theta}) = I\{\boldsymbol{\theta} \in R\}$, where the function I takes on value 1 if the condition in curly brackets holds and zero otherwise. Likewise, quantiles of the distribution can be approximated by the corresponding quantiles of the posterior sample. In short, a robust data analysis requires only a sample from the posterior distribution.

In the highly structured models we described in Section 3.3 we must use sophisticated algorithms to obtain a posterior sample. Here we introduce the powerful *Data Augmentation (DA) algorithm* (Tanner & Wong 1987). A description of the more general *Gibbs sampler* (Metropolis et al. 1953) and *Metropolis-Hastings algorithms* (Hastings 1970) with applications in astronomy can be found in (van Dyk et al. 2001). All of these algorithms construct a Markov chain with stationary distribution equal to the posterior distribution (e.g., Gelfand & Smith 1990); i.e., once the chain has reached stationarity, it generates samples which are identically (but not independently) distributed according to the posterior distribution. These samples can then be used for Monte Carlo integration; hence these algorithms are known as Markov chain Monte Carlo or MCMC methods. (See Tierney [1996] for regularity conditions for using Equation 3.12 with MCMC draws [11].) From the onset then, it is clear that three important concerns when using MCMC in practice are (1) selecting starting values for the Markov chain, (2) detecting convergence of the Markov chain to stationarity, and (3) the effect of the lack of independence in the posterior draws. Space does not allow us to address all of these practical issues. Instead we direct interested readers to van Dyk et al. (2001) and the references therein.

In order to obtain a sample from $p(\boldsymbol{\theta}, \mathrm{X}|\mathrm{Y})$, the DA algorithm uses an iterative sampling scheme that samples first X conditional on $\boldsymbol{\theta}$ and Y and second samples $\boldsymbol{\theta}$ given (X, Y). Clearly, the DA algorithm is most useful when both of these conditional distributions are easily sampled from. The iterative character of the resulting chain naturally leads to a Markov chain,

which we initialize at some starting value, $\boldsymbol{\theta}^{(0)}$. For $t = 1, \ldots, T$, where T is dynamically chosen, we repeat the following two steps:

Step 1: Draw $\mathrm{X}^{(t)}$ from $p(\mathrm{X}|\mathrm{Y}, \boldsymbol{\theta}^{(t-1)})$,

Step 2: Draw $\boldsymbol{\theta}^{(t)}$ from $p(\boldsymbol{\theta}|\mathrm{Y}, \mathrm{X}^{(t)})$.

Since the stationary distribution of the resulting Markov chain is the desired posterior distribution, for large t, $\boldsymbol{\theta}^{(t)}$ approximately follows $p(\boldsymbol{\theta}|\mathrm{Y})$.

To illustrate the utility of the algorithm, we return to the background contamination model introduced in Section 3.2. Given some starting value, $\boldsymbol{\theta}^{(0)} = (\lambda_B^{(0)}, \lambda_S^{(0)})$ the two steps of the algorithm at iteration t become

Step 1: Draw $Y_B^{(t)}$ using the binomial distribution given in Equation 3.8 and set $Y_S^{(t)} = Y - Y_B^{(t)}$.

Step 2: Draw $\lambda_B^{(t)}$ and $\lambda_S^{(t)}$ from independent γ distributions[5]

$$\lambda_B^{(t)}\Big|Y_B^{(t)} \overset{\mathrm{d}}{\sim} \gamma\left(\alpha_B + Y_B + Z, \beta_B + k\right) \text{ and } \lambda_S^{(t)}\Big|Y_S^{(t)} \overset{\mathrm{d}}{\sim} \gamma\left(\alpha_S + Y_S, \beta_S + 1\right). \tag{3.13}$$

Here $\alpha_B, \beta_B, \alpha_S$, and β_S are hyperparameters which quantify prior information via a prior γ distribution on λ_S and λ_B; see van Dyk et al. (2001) for guidance in selecting these parameters. In the first step, we stochastically divide the source count into source counts and background counts based on the current values of λ_B and λ_S. In the second step we use this division to update λ_B and λ_S. Markov chain theory tells us the iteration converges to the desired draws from the posterior distribution.

3.5 Accounting for Pile Up

Pile-up occurs in X-ray detectors when two or more photons arrive in a single spatial cell during the same time frame (i.e., the discrete time units). Such coincident events are counted as a single event with energy equal to the sum of the energies of each of the individual photons. Thus, for bright sources pile-up can seriously distort both the count rate and the energy spectrum. Accounting for pile-up is perhaps the most important outstanding data-analytic challenge for Chandra. Conceptually, however, there is no difficulty in addressing pile-up in a hierarchical Bayesian framework using MCMC; we must stochastically separate a subset of the observed

[5]The $\gamma\ (\alpha, \beta)$ distribution is a continuous distribution on the positive real line with probability density function $p(Y) = \beta^\alpha Y^{\alpha-1}e^{-\beta Y}/\Gamma(\alpha)$, expected value α/β, and variance α/β^2 for positive α and β.

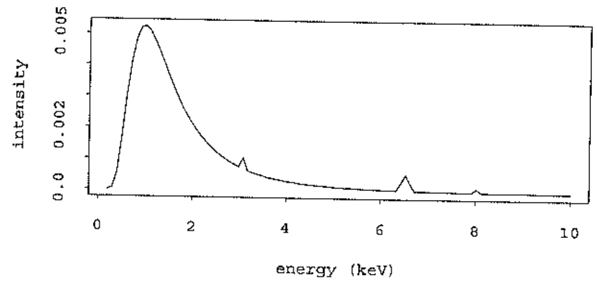

FIGURE 3.1. A Typical Energy Spectrum. We plot the expected photon count per bin per time frame as a function of energy and illustrate the smooth continuum with three small emission lines. This spectrum is plotted at low resolution (100 energy bins) to reduce the computational burden required for handling pile-up; see Figures 3.2.

counts into multiple counts of lower energy while conditioning on the current iteration of the model being fit. The attraction of hierarchical models in this setting is that they allow us to handle pile up ignoring all other model components. That is, when we separate counts into multiple counts of lower energy, the spectral model is completely specified and all the other degradations of the data (e.g., instrument response and background contamination) are accounted for by conditioning on the appropriate "missing data." Thus, we can attack pile up as an isolated problem.

Unfortunately, even in isolation handling pile up is challenging. The difficulty lies in computation. Simply enumerating the set of photons that could result in a particular observed event, let alone their relative probabilities, is an enormous task. Nonetheless, we believe there is great promise in Monte Carlo techniques which if carefully designed, can automatically exclude numerous possibilities that have minute probability. As an illustration, Figure 3.2 plots the conditional distribution of the energy of one of two photons with energy summing to 10 keV, assuming the energy spectra is as in Figure 3.1 and the point spread function is uniform across some area of the detector. The symmetry of the distribution in Figure 3.2 reflects the exchangeability of the component photon energies and the modes arises from the spectral emission lines in Figure 3.1. In practice, an observed energy can be the sum of more than two actual photon energies; in this case there is an 8% chance that there are three photons (and a 61% chance of only one photon, 29% chance of two photons, and 1% chance of four photons).

Care must be taken to efficiently sample from such complex distributions.

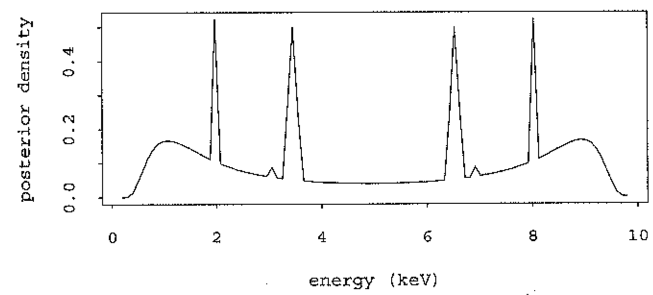

FIGURE 3.2. Un-piling Two Photons. The plot illustrates the conditional distribution of the energy of one of two photons with energy summing to 10 keV, assuming the energy spectra is as in Figure 3.1 and a uniform point spread function. Sophisticated Monte Carlo methods are required to simulate such a highly multi-modal distribution.

Development of Monte Carlo samplers for this task is an area of current research. Nonetheless, even with substantial simplifying assumptions (e.g., at most two photons can pile) preliminary results from our hierarchical model fit via MCMC show great promise. An example is given in the contributed paper by Kang et al. (this volume).

3.6 The Future of Data Analysis

The highly structured models described in this chapter reflect a new trend in applied statistics—it is becoming ever more feasible to build application specific models which are designed to account for the hierarchical and latent structures inherent in any particular data generation mechanism. Such multi-level models have long been advocated on theoretical grounds, but recently the development of new computational tools such as those described here has begun to bring such model fitting into routine practice. Although these methods offer great promise, they are by no means statistical black boxes that will automatically solve any problem. The flexibility of such models and computational methods require users to be statistically savvy. We, however, believe the benefits of superior scientific modeling far outweigh these costs. Indeed the future of data analysis lies in sophisticated application-specific modeling and methods.

Acknowledgments: The author gratefully acknowledges funding for this project partially provided by NSF grant DMS-01-04129 and by NASA contract NAS8-39073 (CXC). This chapter is a result of a joint effort of the members of the Astro-Statistics group at Harvard University whose members include A. Connors, D. Esch, P. Freedman, C. Hans, H. Kang, V. L. Kashyap, R. Protassov, D. Rubin, A. Siemiginowska, N. Sourla, and Y. Yu.

3.7 REFERENCES

[1] Carlin, B. P. and Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman & Hall, London.

[2] 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.

[3] Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, 398–409.

[4] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis.* Chapman & Hall, London.

[5] Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in Practice.* Chapman & Hall, London.

[6] Hastings, W. K. (1970). Monte Carlo sampling methods usings Markov chains and their applications. *Biometrika* **57**, 97–109.

[7] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phy.* **21**, 1087–1092.

[8] Protassov, R., van Dyk, D., Connors, A., Kashyap, V., and Siemiginowska, A. (2002). Statistics: Handle with care – detecting multiple model components with the likelihood ratio test. *Astrophysical J.* to appear.

[9] Siemiginowska, A., Elivis, M., Alanna, C., Freeman, P., Kashyap, V., and Feigelson, E. (1997). in *Statistical Challenges in Modern Astronomy II (eds. E. Feigelson and G. Babu)*, 241–253. Springer-Verlag, New York.

[10] 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.

[11] Tierney, L. (1996). in *Markov Chain Monte Carlo in Practice (eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter).* Chapman & Hall, London.

[12] van Dyk, D. A., Connors, A., Kashyap, V., and Siemiginowska, A. (2001). Analysis of energy spectra with low photon counts via Bayesian posterior simulation. *Astrophysical J.* **548**, 224–243.

Commentary by Michael A. Strauss[6]

Astronomers often find themselves tackling complicated likelihood problems. With some basic knowledge of the underlying statistics of a given astronomical problem, and some familiarity with likelihood functions and Bayesian statistics, we often are able to write down a likelihood function in closed form. However, if the problem is complicated enough (read "interesting", as it usually is), we are stymied when it comes time to maximize this likelihood, especially if there is an interesting and complicated parameter space to fit for. This paper describes useful techniques for solving exactly this sort of problem, which are common in astronomy, by a "divide and conquer" approach, doing the problem iteratively. The very nasty problem of deconvolving the effects of "pile-up" in X-ray spectra is a particularly good example of this.

Another problem which may be amenable to this approach is illustrated in Figure 3.3, which shows the spectrum of a quasar from the Sloan Digital Sky Survey (see my contribution to these proceedings). The spectrum shows a blue continuum with strong superposed emission lines. Blueward (to the left) of the Lyα emission line of hydrogen are superposed a large number of absorption lines of Lyα, due to filaments and wisps of hydrogen gas at redshifts between that of the quasar and zero. Astronomers very much want to measure the statistics of the Lyα forest absorption, but are stymied in part because of the lack of complete understanding of the unabsorbed continuum of the quasar itself. That is, the observations represent the convolution of two unknowns: the quasar spectrum, and the Lyα forest absorption spectrum, and it is not clear how optimally to separate the two. It would be interesting to know if the methods described in this paper could allow an optimal solution to this problem.

[6]Princeton University Observatory

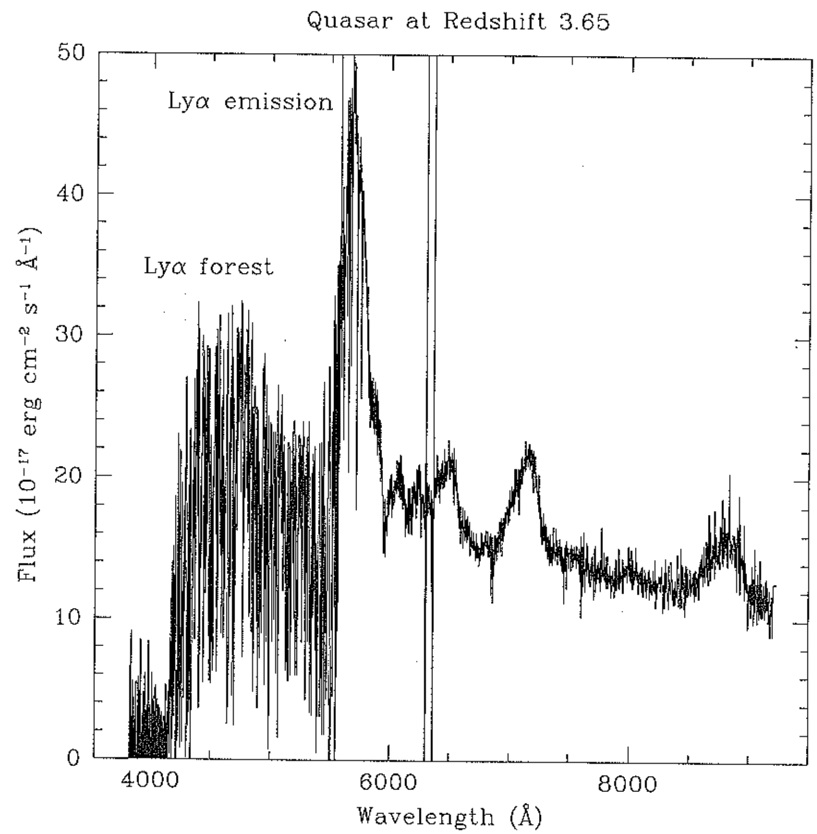

FIGURE 3.3. The spectrum of a high-redshift quasar from the Sloan Digital Sky Survey.