Statistical Science
2004, Vol. 19, No. 2, 275–293
DOI 10.1214/088342304000000314

Highly Structured Models for Spectral Analysis in High-Energy Astrophysics

David A. van Dyk and Hosung Kang

Abstract. The *Chandra X-Ray Observatory*, launched by the space shuttle *Columbia* in July 1999, has taken its place with the *Hubble Space Telescope*, the *Compton Gamma Ray Observatory* and the *Spitzer Infrared Space Telescope* in NASA's fleet of state of the art space-based *Great Observatories*. As the world's premier X-ray observatory, *Chandra* gives astronomers a powerful tool to investigate black holes, exploding stars and colliding galaxies in the hot turbulent regions of the universe. *Chandra* uses four pairs of ultra-smooth high-resolution mirrors and efficient X-ray photon counters to produce images at least 30 times sharper than any previous X-ray telescope. Unlocking the information in these images, however, requires subtle statistical analysis; currently popular statistical methods typically involve Gaussian approximations (e.g., minimum χ^2 fitting), which are not justifiable for the high-resolution low-count data. In this article, we employ modern Bayesian computational techniques (e.g., expectation–maximization-type algorithms, the Gibbs sampler and Metropolis–Hastings) to fit new highly structured models that account for the Poisson nature of photon counts, background contamination, image blurring due to instrumental constraints, photon absorption, photon pileup and source features such as spectral emission lines and absorption features. This application demonstrates the flexibility and power of modern Bayesian methodology and algorithms to handle highly structured models that are convolved with complex data collection mechanisms involving nonignorable missing data.

Key words and phrases: Astrostatistics, Bayesian methods, the *Chandra X-Ray Observatory*, data augmentation, EM algorithm, Markov chain Monte Carlo, missing data, Poisson model, posterior predictive checks, nonignorable missing data, spectral analysis, pileup.

1. SCIENTIFIC BACKGROUND

In his seminal 1979 paper that introduced the likelihood ratio to astrophysicists, Cash (1979) began with the following remark:

> As high energy astronomy matures, experiments are producing data of higher quality in order to solve problems of greater sophistication. With the advent of the *HEAO* satellites, the quality of X-ray astronomy data is being increased again, and it is important that the procedures used to analyze the data be sufficiently sophisticated to make the best possible use of the results.

Twenty-five years later we are in the same situation: Recent advances in astronomical instrumentation allow the collection of high-resolution low-count Poisson data for which standard analysis techniques with their Gaussian approximations are suspect. Indeed, it has become apparent over the past several years that the use of statistical methodology that may

David A. van Dyk is Associate Professor at the University of California, Irvine, California 92697-1250, USA (e-mail: dvd@uci.edu). Hosung Kang is a Ph.D. student, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, USA (e-mail: kang@stat.harvard.edu).

have been quite appropriate 20 years ago is not justifiable for the quality of data that are currently available (Siemiginowska et al., 1997); new statistical methods are needed to take advantage of the precision that technological advances have afforded us. In this article, we describe a highly structured model and corresponding statistical methods that are appropriate for the high-resolution spectral data available with new instrumentation. Before describing the model and its merits, however, we motivate its need.

1.1 High-Energy Image Analysis

X-rays are high-energy electromagnetic waves, that is, photons. Roughly speaking, the production of high-energy electromagnetic waves requires temperatures of millions of degrees and signals the release of deep wells of stored energy such as those in very strong magnetic fields, extreme gravity, explosive nuclear forces and shock waves in hot plasmas. Thus, X-ray telescopes can map nearby stars (like our Sun) that have active magnetic fields, the remnants of exploding stars, areas of star formation, regions near the event horizon of a black hole, very distant but very turbulent galaxies or even the glowing gas that embeds a cosmic cluster of galaxies. The distribution of the energy of the electromagnetic emissions gives insight into the composition, temperatures and relative velocity of an astronomical source. The spatial distribution of the emission reflects physical structures in an extended source, for example, emission jets or the shape of the debris of a stellar explosion. Some sources exhibit temporal variability or periodicity that might result from surface or internal pulsations, eclipses, star spots or magnetic activity cycles. Thus, instrumentation that can precisely measure the energy, sky coordinates and arrival time of X-ray band photons enables astrophysicists to extract subtle clues as to the underlying physics of X-ray sources. Such instrumentation is necessarily space-based, since high-energy photons are absorbed by the Earth's atmosphere. The *Chandra X-Ray Observatory* is an example of the new class of instruments that provides high-precision data. Launched in July of 1999, *Chandra* has already given astronomers a wealth of important new data. The instrumentation aboard *Chandra* includes four pairs of ultra-smooth high-resolution mirrors and efficient X-ray detectors to provide images at least 30 times sharper than any previous X-ray telescope. Data are collected on each X-ray photon that arrives at the detector; the time of arrival, the two-dimensional sky coordinates and the energy are all recorded. Due to instrumental constraints, each of these four variables is discrete. Thus, a data set can be represented by a four-way table of counts with margins corresponding to time, two sky coordinates and energy. In this paper we focus on spectral analysis which models the one-way energy margin; the energy bins are referred to as energy channels. More on the scientific objectives of high-energy astrophysics, the instrumental specifications of *Chandra* and Bayesian methods designed to investigate spectral and spatial structure of the emission appear in van Dyk et al. (2004).

1.2 Statistical Challenges

Throughout this article, we describe a number of statistical challenges that arise when analyzing data from *Chandra* and other high-resolution count-based detectors. It is to address these challenges that we develop new statistical models and methods. To motivate the need for new methodology, we briefly outline some of these challenges.

Chandra's capacity for high-resolution imaging means that it has a much finer discretization of energy than previous instruments. This results in an overall increase in the number of energy channels and leads to lower observed counts in each channel. Thus, Gaussian assumptions that might have been appropriate for data from older instruments are often inappropriate for *Chandra* data. For example, in so-called minimum χ^2 fitting (Lampton, Margon and Bowyer, 1976) one estimates the model parameter θ by computing

$$\hat{\theta} = \arg\min_{\theta} \sum_{l \in \mathcal{L}} \frac{\{n_l - m_l(\theta)\}^2}{\sigma_l^2(\theta)}, \tag{1}$$

where $\mathcal{L}$ is the set of energy channels, n_l is the observed count in energy channel l, $m_l(\theta)$ is the expected count in channel l as a function of the model parameter θ and $\sigma_l^2(\theta)$ is proportional to the sampling variance of n_l. Because of the Poisson nature of the data, $\sigma_l^2(\theta)$ is often taken to be either n_l or $m_l(\theta)$. It is obvious from its functional form that the right-hand side of (1) is an implicit Gaussian assumption. When one observes a relatively large count in each energy channel, this assumption is reasonable and χ^2 fitting may be appropriate. However, the intrinsically low-count data from high-resolution instruments such as those aboard *Chandra* are not approximately Gaussian. Thus, parameter estimates and error bars computed with χ^2 fitting may not be trustworthy. To avoid this problem, one can group the energy channels until

there is a large enough count in each group to justify Gaussian assumptions. Doing so, however, reduces the information in the data and produces a less precise energy spectrum. To take advantage of the information that the new class of instruments provides, a method of analysis is needed that does not rely on large-count Gaussian assumptions.

X-ray count data are generally contaminated with background counts, a number of photons originating somewhere other than the source of interest. To quantify the background contamination, a second data set is collected that is assumed to consist only of background. For example, background count data might be collected around but some distance away from the source. In standard practice, these counts are *directly* subtracted from the source counts. This procedure accounts neither for error in the estimation of the background intensity nor for the fact that the actual background count in the source data varies from the background intensity. Ignoring both of these sources of variability has obvious ramifications in the resulting source minus background count (it may be negative), and uncertain consequences on the parameter estimates and error bars. New methods that properly account for these sources of variability are required.

Pileup occurs in X-ray CCDs (charged coupled devices such as those aboard *Chandra*) when two or more photons arrive at the same area of the detector during the same time frame (i.e., time bin). Such coincident events are counted as a single event with energy equal to the sum of the coincident event energies. The event is lost altogether if the total energy goes above the on-board discriminators. Thus, for bright sources, pileup can seriously distort both the count rate and the energy spectrum. Accounting for pileup in a principled manner requires careful modeling and sophisticated statistical computation.

Finally, the likelihood ratio test, or a Gaussian approximation thereof, is routinely used to test for the presence of additive spectral features. Since the intensity of these features generally is constrained to be positive, the null hypothesis of no feature is on the boundary of the parameter space. Thus, the standard asymptotic reference distribution of the likelihood ratio test is inapplicable and more sophisticated methods are needed for testing such hypotheses.

1.3 Model-Based Solutions

We use model-based Bayesian methods to handle the complexity of *Chandra* data. Multilevel models can be designed with components for both the data collection process (e.g., background contamination and pileup) and the complex spectral structures of the sources themselves. Sophisticated computational methods are required for fitting the resulting highly structured models. We believe a Bayesian perspective is ideally suited to such models in terms of both inference and computation. The models can be parameterized on high-dimensional spaces with numerous nuisance parameters that describe, for example, the data collection process. Bayesian methods offer a straightforward way to handle such parameter spaces. We base inference on marginal posterior distributions of scientifically interesting parameters or groups of parameters.

Computational tools such as the expectation–maximization (EM) algorithm (Dempster, Laird and Rubin, 1977), the data augmentation (DA) algorithm (Tanner and Wong, 1987), the Gibbs sampler (e.g., Gelfand and Smith, 1990; Smith and Roberts, 1993) and other Markov chain Monte Carlo (MCMC) methods are ideally suited to highly structured models of this sort; see van Dyk (2003). The modular structure of these algorithms fits hand-in-glove with the hierarchical structure of our models. The Gibbs sampler, for example, samples one set of model parameters from their conditional posterior distribution given all other model parameters. This allows us to sequentially fit one component of the overall model at a time while conditioning on the other components. In this way, a complex model fitting task is divided into a sequence of much easier tasks. Many of these easier tasks involve well-understood procedures. Using an EM algorithm to handle a blurring matrix and background contamination of Poisson data is a good example (Richards, 1972; Lucy, 1974; Shepp and Vardi, 1982; Lange and Carson, 1984; Fessler and Hero, 1994; Meng and van Dyk, 1997). Although this well-known (and often rediscovered) technique is unable to handle the richness of our highly structured model, we utilize it and its stochastic counterpart as a step in our mode finding and posterior sampling algorithms.

The remainder of this article is organized into six sections. We do not attempt to detail how we (or how we expect to) handle all of the modeling, computational and inferential aspects of the analysis of *Chandra* data. Such a task is well beyond the scope of this article. Instead, we outline some of our models and methods to give the reader a flavor of our Bayesian analysis and highly structured models. In some cases, more details can be found in one of several references that are cited in the text; in other cases,

methods are still being developed. We refer interested readers to van Dyk, Connors, Kashyap and Siemiginowska (2001), Protassov et al. (2002), van Dyk and Hans (2002) and van Dyk et al. (2004), as well as several short papers in the conference proceedings *Statistical Challenges in Modern Astronomy III* (Feigelson and Babu, 2003). This article begins in Section 2 with an outline of modelling strategies for both the source spectra and the data collection process. Computation methods that are designed around multiple levels of nonignorable missing data are described in Section 3; an example that illustrates our basic inference techniques appears in Section 4. Section 5 discusses model checking and model diagnostic methods. Statistical methods that account for the important problem of photon pileup appear in Section 6 and concluding remarks appear in Section 7.

2. SPECTRAL MODELS

2.1 The Source Model

The basic goal of high-energy spectral modeling from a statistical perspective is to model the distribution of the energy of high-energy photons (X-ray or γ-ray) from a particular astronomical source. Such a spectral model typically contains several additive components which can be formulated as a finite mixture model. Roughly speaking, the components can be split into two groups: *continuum* terms, which describe the distribution over the entire energy range of interest, and *emission lines*, which are local positive aberrations from the continuum. The finite mixture is then multiplied by an *absorption* factor, which represents stochastic censoring of photons. A proportion of photons is absorbed by matter at the surface of the source or between the source and detector; the probability of absorption varies with energy. In addition to the source model, the data collection mechanism of the detector has several stochastic components that must be accounted for by the data model. Such data distortion is described in Section 2.2. In this section we describe the three components of the spectral model: the continuum, emission lines and absorption features. We both outline the importance of these features in astronomical terms and detail the statistical models that we use to describe them.

Continuum. To understand the astrophysical process we are modelling, consider a source such as a star. The center of the star is composed of very hot gas, which produces copious photons that random walk their way to the surface of the star. This process creates a continuous spectrum, or continuum, of radiated energy and is known as blackbody emission. As another example, consider a high-temperature low-density plasma where photons are not thermalized by repeated collisions with ions in the plasma; the transitions between levels in free electrons, induced by electrostatic interactions with ionized nuclei, result in a so-called thermal Bremsstrahlung continuum. Among other things, the shape of the continuum indicates the temperature of the source. Astrophysicists generally use one of several models or a weighted sum of a number of these models to describe the continuum in some bounded energy range. Table 1 gives the functional form of several common continuum models. These models describe the relative frequencies of the photon energy E. In this sense, the continuum models are akin to probability densities functions, but the continuum models are not normalized to 1. Rather, the integral of the continuum model over the energy range of interest is the expected number of photons due to the particular continuum component. (The *normalization* parameter α is often rescaled so the integral of a particular continuum model represents counts per unit time per unit area of the detector.) The overall continuum model can be written as a finite mixture,

$$f(\theta^C, E) = \sum_{k \in \mathcal{K}^C} f_k(\theta_k^C, E), \tag{2}$$

where $\mathcal{K}^C$ is the index set for the continuum components, f_k is the functional form of continuum term k

TABLE 1
Continuum models

Model	Functional form	Parameter constraints
Power law	$\alpha E^{-\beta}$	$\alpha, \beta > 0$
Broken power law	$\alpha E^{-\beta}(E/E_\star)^{\gamma I(E>E_\star)}$	$\alpha, \beta, E_\star > 0, \gamma > -\beta$
Bremsstrahlung emission	$\alpha e^{-\beta E}$	$\alpha, \beta > 0$
Blackbody emission	$\alpha E^2/(e^{\beta E} - 1)$	$\alpha, \beta > 0$

and $\theta^C = \{\theta_k^C, k \in \mathcal{K}^C\}$ represents the parameters for the set of continuum terms. In our notation, superscripts are generally used to indicate the relevant model component. Thus, C, L, A and B refer to the continuum, (emission) lines, absorption features and background contamination, respectively.

Because of the digital nature of the detector (see Section 2.2), energy is treated as a discrete variable. That is, we consider the photon counts in a number of prespecified energy bins (e.g., as many as 4096 on *Chandra*). To model the counts, we use independent Poisson distributions with intensities determined by the continuum model. In particular, the counts in bin j due to continuum term k are modelled as

$$Y_j^k \sim \text{Poisson}\{\delta_j f_k(\theta_k^C, E_j)\} \quad \text{for } j \in \mathcal{J} \text{ and } k \in \mathcal{K}^C, \tag{3}$$

where δ_j and E_j are the width and mean energy of bin j and $\mathcal{J}$ is the index set of the energy bins. (Here we assume that the binning is fine enough so that the integral of the continuum model over the bin is nearly equal to the model evaluated at the bin mean times the bin width.)

Except for the blackbody emission, we can identify each of the continuum models in Table 1 with generalized linear models with log links. For example, the log of the power law model is a linear function of $\log(E)$ and the log of the broken power law is a linear function of $\log(E)$ and $\log(E/E_\star)I(E > E_\star)$, where I is the indicator function. (Here we assume the location of the break in the power law, $E_\star$, is known. If this location is not known, we model the broken power law given the break point as described here and model the marginal distribution of the location of the break, perhaps with a fully specified proper prior distribution.) The combined counts from all of the continuum terms, $\sum_{k\in\mathcal{K}^C} Y_j^k$, follow a finite mixture of these generalized linear models.

Emission lines. Emission lines are local features added to the continuum and they represent extra emission of photons in a narrow band of energy. Such extra emission is due to photons that are emitted when an electron falls to a lower energy shell of a particular ion; the abundance of the extra emission indicates the abundance of the ion in the source. Thus, analysis of emission lines is informative as to the chemical composition of the surface of the astronomical source. The Doppler shift of the location of a known spectral line (such as a particular hydrogen line) indicates the relative speed of the source. Statistically the emission lines are represented by adding Gaussian, Lorentzian (i.e., a t density with 1 degree of freedom) or delta functions to the continuum. An example of a simple spectral model with a power law continuum and two narrow emission lines appears in the first plot in Figure 1.

We parameterize the intensity in bin $j \in \mathcal{J}$ as a mixture of the continuum term and the emission lines,

$$\lambda_j(\theta^C, \theta^L) = \delta_j f(\theta^C, E_j) + \sum_{k\in\mathcal{K}^L} \theta_{k,\lambda}^L P_j(\theta_{k,\mu}^L, \theta_{k,\sigma}^L) \quad \text{for } j \in \mathcal{J}, \tag{4}$$

where $\mathcal{K}^L$ is the index set for the emission lines, $\theta_{k,\lambda}^L$ is the expected photon count of emission line k and $P_j(\theta_{k,\mu}^L, \theta_{k,\sigma}^L)$ is the probability that a photon from an emission line with center $\theta_{k,\mu}^L$ and spread $\theta_{k,\sigma}^L$ falls in bin j. These probabilities are obtained from the Gaussian, Lorentzian or delta functions that are used to parameterize the emission lines, all of which can be parameterized in terms of their center and spread. (For a delta function, the spread parameter is fixed at zero.) The collection of parameters, $\theta_k^L = (\theta_{k,\lambda}^L, \theta_{k,\mu}^L, \theta_{k,\sigma}^L)$ for $k \in \mathcal{K}^L$, is represented by θ^L.

Absorption features. Absorption lines correspond to narrow intervals in the energy dimension where fewer photons appear than would be expected from (4). These lines are formed because photons have been absorbed by material in the source or between the source and the observer. Because the specific energies at which photons are absorbed correspond to specific line transitions of ions, absorption lines can, for example, give clues as to the composition of a source. The absorption process begins as photons leave the hot center of the source and move toward the colder region near the surface of the source. Because the continuum photons are in a higher energy state than their colder surroundings, they are readily absorbed by material at the surface of the source to balance the energy of the system. A similar process can occur in the interstellar media (ISM) or in the intergalactic media (IGM).

There are various functional forms for absorption features, which can be formulated in terms of the probability that a photon is not absorbed as a function of the photon energy. Multiple absorption features are assumed to be independent and thus these probabilities are multiplied. In particular,

$$\pi(\theta^A, E_j) = \prod_{k\in\mathcal{K}^A} \pi_k(\theta_k^A, E_j), \tag{5}$$

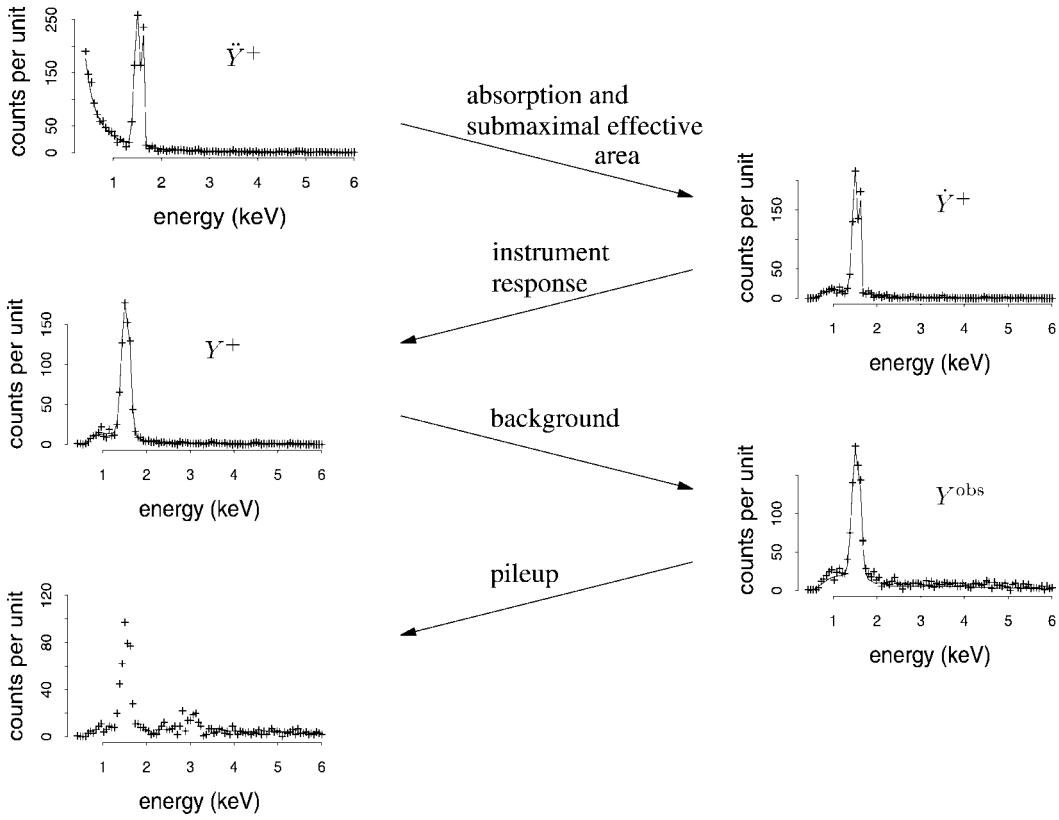

FIG. 1. *The degradation of counts. The various physical processes that significantly degrade the source model and result in the observed channel counts are illustrated. In particular, an artificial data set is used to illustrate* (1) *the absorption of* (*mostly low-energy*) *counts,* (2) *the blurring of spectral features due to instrument response,* (3) *the masking of features due to background contamination and* (4) *the shadows caused by pileup. The solid lines represent the assumed model* (*in the first four plots*) *and the* + *sign represents the simulated data. The first plot illustrates the counts per maximum effective area per total exposure time per bin; the remaining plots illustrate degraded counts per effective area per total exposure time per bin. Note that the effects of pileup are included here for the sake of completeness; we discuss pileup in Section* 6. *The symbols in the upper right of each plot are defined in Table* 2.

where $\mathcal{K}^A$ is the index set for the absorption features, $\theta^A = \{\theta_k^A, k \in \mathcal{K}^A\}$ with θ_k^A the parameters of absorption feature k, and π_k represents the probability of not being absorbed by feature k. The absorption process can be modeled using binomial generalized linear models; many of the standard forms of $\pi_k(\theta_k^A, E_j)$ can be handled with log or complementary log-log links (van Dyk and Hans, 2002; Hans and van Dyk, 2003). For example, an important functional form for absorption lines (Freeman et al., 1999) can be expressed by setting

$$\pi_k(\theta_k^A, E_j) = \exp[-\theta_{k,\lambda}^A \exp\{-(E_j - \theta_{k,\mu}^A)^2/2\theta_{k,\sigma}^A\}], \tag{6}$$

where the components of $\theta_k^A = (\theta_{k,\lambda}^A, \theta_{k,\mu}^A, \theta_{k,\sigma}^A)$ represent the intensity, location and spread of the absorption line, respectively. Alternatively, absorption features can be modeled as "edges," which begin sharply at some fixed energy $E_\star$ and die off slowly with increasing energy. Specifically, we model

$$\pi_k(\theta_k^A, E_j) = \begin{cases} 1, & \text{if } E_j < E_\star, \\ \exp\{-\theta_{k,\lambda}^A (E_j/E_\star)^{-3}\}, & \text{if } E_j \geq E_\star, \end{cases} \tag{7}$$

where $\theta_{k,\lambda}^A$ is the intensity of the edge. The complementary log-log link linearizes both model (6) and model (7).

In addition to localized absorption features, we must account for so-called *continuum absorption*, which can affect a wide range of energies in a high-energy spectrum. This absorption occurs when the absorbed photon frees an electron from an ion and is thus not associated with a specific line transition of the ion. Continuum absorption can sometimes be approximated by

so-called *exponential absorption* with $\pi_k(\theta_k^A, E_j) = \exp(-\theta_k^A/E_j)$. A better model sets $\pi_k(\theta_k^A, E_j) = \exp\{-\theta_k^A X(E_j)\}$, where $X(E_j)$ is a tabulated value; see Morrison and McCammon (1983) for details.

2.2 Data Distortion Model

Unfortunately, due to instrumental constraints, the photon counts are degraded in a number of ways. For example, the effective area of the detector varies with the energy of the photon. Heuristically the instrument works like a prism, which bends an X-ray by an angle that depends on its energy. Some of the photons are bent so far that they miss the detector altogether. Thus, the probability d_j that an X-ray is not refracted off the detector depends on its energy. A second form of data degradation is due to instrument response, which is a characteristic of the detector that results in blurring of the photon energies. A photon that arrives with energy that corresponds to bin j has probability M_{lj} of being recorded in detector channel $l \in \mathcal{L}$, where $\mathcal{L}$ is the index set of the channels. (Here we use the term "bin" to refer to the ranges of energy that correspond to photon counts in an ideal instrument that is not subject to blurring of the photon energies. The term "channel" refers to the energy ranges that correspond to the observed data. The energy ranges for the bins and the channels need not coincide.) Like the effective area vector $d = (d_1, \ldots, d_J)$, the matrix $M = \{M_{lj}\}$, which may not be square, is evaluated by calibration of the detector and is presumed known.

As discussed in Section 1.2, the source spectrum is generally also contaminated by background counts that originate somewhere other than the source of interest. The plots in Figure 1 illustrate the effects of the effective area, instrument response and background. The final plot illustrates the effect of photon pileup, a topic we ignore until Section 6.

Because of these degradations, we model the observed counts as independent Poisson variables with intensity

$$\xi_l(\theta) = \sum_{j \in \mathcal{J}} M_{lj} \lambda_j(\theta) d_j \pi(\theta^A, E_j) + \theta_l^B, \qquad l \in \mathcal{L}, \tag{8}$$

where θ_l^B is the Poisson intensity of the background in channel l and $\theta = (\theta^C, \theta^L, \theta^A, \theta^B)$. In Section 3.1, we describe how the method of data augmentation can be used to construct simple, stable and fast algorithms for fitting this model.

2.3 Specification of Prior Distributions

Whenever possible, we use semiconjugate prior distributions (e.g., on the means, variances and Poisson intensities of Gaussian emission lines). Gaussian prior distributions are used on the coefficients in generalized linear models. These prior distributions are easily incorporated into iteratively reweighted least squares algorithms for computing posterior modes. We use a similar strategy to compute the Student t jumping distribution for a Metropolis–Hastings step when sampling from the posterior distribution; see van Dyk et al. (2001) for details.

Often it is possible to use relatively noninformative prior distributions. The parameters of the continuum, for example, are often well constrained by the data. In some cases, however, informative prior distributions are either necessary or desirable. Prior information on the location and width of weak emission or absorption lines, for example, can greatly improve the quality of the inference. Luckily such information is often scientifically forthcoming, since certain lines are expected or typical in particular classes of astronomical objects. In fact it may be desirable to include information from previous observations or other sources in an analysis of new data, and prior distributions offer an avenue for a unified analysis. In practice, quantifying informative prior information regarding (highly) multivariate parameters can be challenging. We tend to use the seemingly conservative strategy of independently combining univariate prior distributions using semiconjugate forms with appropriate moments or quantiles whenever necessary and noninformative prior distributions whenever possible.

3. NONIGNORABLE MISSING DATA AND STATISTICAL COMPUTATION

3.1 Data Augmentation Strategies

As described in Section 2, data collected with *Chandra* are complex in terms of both the underlying source models and the data collection process; this complexity is reflected not only in statistical models, but also in the computational tools required for model fitting. Ideally, data would be available that fulfill the following criteria:

1. Data are not subject to absorption or the varying effective area of the detector.
2. Data quantify the exact energy of each arriving photon without blurring or binning.

TABLE 2
Data augmentation in the spectral model; for all variables, $j \in \mathcal{J}$, $l \in \mathcal{L}$ and $s \in \mathcal{S}$

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

3. Data record the source of the photon, that is, whether the photon is due to a particular continuum component, a particular emission line or background contamination.

For the purpose of statistical computation, we treat this ideal data set as missing data. The ideal data are in fact nonignorable missing data. For example, since the absorption rate varies with energy, ignoring absorption clearly biases inference for the continuum components and emission lines. With the ideal data, on the other hand, model fitting is greatly simplified. For example, we could gather the photons from each emission line, whether or not they were absorbed or otherwise lost, and use standard statistical methods to learn about the location, width and intensity of the emission line. By adding additional levels of missing data, we can model the data distortion processes. Thus, we might also hope that the photons that were absorbed or lost to the submaximal effective area of the instrument would be recorded along with their energies, their source component and a variable that indicates why they were lost (e.g., to which component of the absorption model). Absorption features could be studied by examining the energy distribution before and after absorption, and modeling the probability of absorption as a function of energy.

The method of data augmentation takes advantage of how simple model fitting would be, were such extensive data available. Both the EM algorithm and the DA algorithm are well-known examples that we use along with their generalizations to fit the spectral model described in Section 2. To formalize this, we introduce a hierarchy of augmented data structures that are outlined in Figure 1 and Table 2. As noted above, the first plot in Figure 1 represents a data set that is free of blurring, absorption and background contamination, and has constant effective area. The source model consists of a power law continuum with two strong emission lines. The energies are binned in this plot and we have mixed the photons from the three sources (i.e., the two emission lines and one continuum term). Thus, this data set represents less data augmentation than the ideal data set; the ideal data are represented by level 1 in Table 2, while the first plot in Figure 1 is represented by level 3 of the table. In the notation of Table 2, we use more dots in the accent above a variable to represent a greater degree of augmentation; variables with fewer dots in the accent are (sometimes stochastic) functions of those with more dots. The set $\mathcal{S}$ is the collection of continuum and emission line photon sources; a superscript $+$ indicates a mixture of all the sources in $\mathcal{S}$. Levels 4, 5 and 6 in Table 2 correspond to the second, third and fourth plots in Figure 1.

Reading top to bottom in Table 2, the relationships among the variables are mostly self-explanatory from the description of the model in Section 2. For example, going from level 3 to level 4 accounts for absorption, which works independently on photons and with constant probability within each energy bin; thus,

$$\dot{Y}_j^+ | \ddot{Y}_j^+, \theta \sim \text{Binomial}\big(\ddot{Y}_j^+, d_j \pi(\theta^A, E_j)\big), \quad j \in \mathcal{J}.$$

The effect of the blurring of energy is modeled as multinomial for each energy bin. Summing over the bins, the distribution of energy channel counts (i.e., level 5) is

$$Y^+ | \dot{Y}^+, \theta \sim \sum_{j \in \mathcal{J}} \text{Multinomial}(\dot{Y}_j^+, M_j),$$

where $Y^+ = \{Y_l^+, l \in \mathcal{L}\}$, $\dot{Y}^+ = \{\dot{Y}_j^+, j \in \mathcal{J}\}$ and M_j is the jth column of M, $j \in \mathcal{J}$. Finally, the observed data are the background contaminated version of level 5. That is, the distribution of level 6 given level 5 is

$$Y_l^{\text{obs}} | Y_l^+, \theta \sim Y_l^+ + \text{Poisson}(\theta_l^B), \quad l \in \mathcal{L}.$$

As described above, given the several augmented data sets described in Table 2, statistical inference for the unknown model parameters is straightforward. Given the model parameters and the observed data, the augmented data sets in Table 2 also follow simple standard distributions. For example, stochastically separating background from source counts corresponds to a binomial distribution,

$$Y_l^+ | Y_l^{\text{obs}}, \theta \sim \text{Binomial}\left(Y_l^{\text{obs}}, \frac{\xi_l(\theta) - \theta_l^B}{\xi_l(\theta)}\right), \quad l \in \mathcal{L},$$

as is typical when stochastically dividing observations among components of a finite mixture model. Accounting for absorption and the varying effective area of the detector corresponds to an added Poisson variable,

$$\ddot{Y}_j^+ | \dot{Y}_j^+, \theta \sim \dot{Y}_j^+ + \text{Poisson}\big[\lambda_j(\theta^C, \theta^L)\{1 - d_j \pi(\theta^A, E_j)\}\big], \quad j \in \mathcal{J}.$$

In this way, we can sample each level of the augmented data described in Table 2 given the data in the rows below it in Table 2 and the model parameters. Because, given the parameters, the missing data follow a series of standard distributions, and given the missing data, model fitting is straightforward, we can construct iterative sampling and mode finding algorithms such as the DA and EM algorithms to fit the spectral model. Incorporating proper prior information, which is typically important, for example, for the emission line parameters, is described in detail by van Dyk (2000a) and van Dyk et al. (2001) for EM and DA, respectively.

3.2 Efficient Computation

By far the most time-consuming aspect of an iteration of the EM or DA algorithms for fitting the spectral model is removal of background counts and deblurring of source counts (i.e., sampling or computing the expectation of Y_l^+ for each $l \in \mathcal{L}$ and $\dot{Y}_j^+$ for each $j \in \mathcal{J}$). These computations involve looking up many values in the typically large matrix M; this process generally consumes a significant proportion of the computing time, even when using sophisticated sparse matrix techniques. Thus, if the data were available without background contamination or blurring, namely $\dot{Y}^+ = \{\dot{Y}_j^+, j \in \mathcal{J}\}$, model fitting would be computationally much less demanding. The nested EM algorithm (van Dyk, 2000b) takes advantage of this by running several iterations of an EM algorithm that treats $\dot{Y}^+$ as observed data and then updates $\dot{Y}^+$ in a standard E step. This strategy effectively nests an EM algorithm within an EM algorithm. In particular, after running one complete EM iteration, we might run several iterations fixing $\dot{Y}^+$ and update only the quantities in the first three rows of Table 2 in the (inner) E step and θ in the M step. If this inner EM iteration converges slowly, as might be the case when there are many line profiles, a relatively large number of partial updates (e.g., 10) may substantially improve the overall speed of the algorithm as compared to the standard EM algorithm. An analogous strategy involves sampling Y_l^+ for $l \in \mathcal{L}$ and $\dot{Y}^+$ only every so many iterations of a DA sampler and may result in a larger effective sample size per unit time.

The computational advantage of the nested EM algorithm is illustrated using a spectrum of the high redshift quasar PG1637 + 706 collected with *Chandra*. The data were modeled using a power law continuum, Morrison and McCammon's (1983) absorption model, a power law model for the background, and a single Gaussian line with fitted center and intensity, for a total of seven free parameters. The solid line in Figure 2 shows the CPU time required for convergence as a function of the number of inner EM iterations; an algorithm with one inner iteration corresponds to the standard EM algorithm. Convergence was determined when the log-likelihood increased by less than 10^{-8} in one complete iteration. With about 10 inner iterations, the nested EM algorithm converges in about a quarter of the time required by the standard EM algorithm. The dotted line in Figure 2 will be described shortly; we return to this example in Section 4.

To further improve the algorithms, we reduce the augmented information for θ using conditional augmentation (Meng and van Dyk, 1997; van Dyk and Meng, 2001) by reducing the counts attributed to the absorbed photons in each emission line, $\ddot{Y}_j^k - \dot{Y}_j^k$, $k \in \mathcal{K}^L$, where $\dot{Y}_j^k$ is the proportion of $\dot{Y}_j^+$ attributed to emission line k; see Table 2. Recall that absorption does not occur uniformly across the energy of an emission line and that the energies of the observed photons are biased toward areas of low absorption, complicating parameter estimation. It is important to note, however, that we need not account for (i.e., augment) all

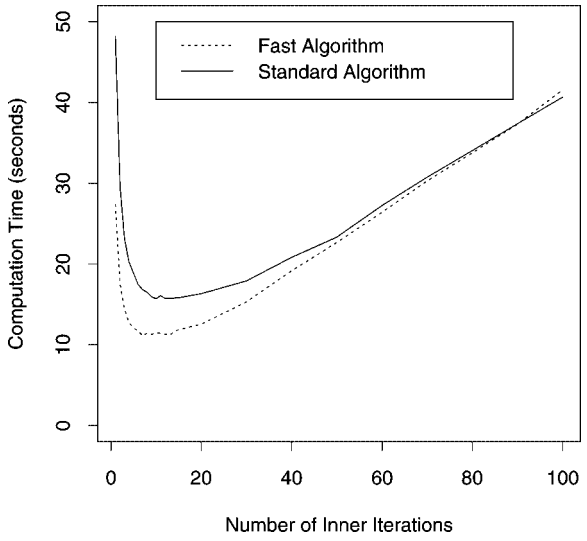

FIG. 2. *Various EM-type algorithms for fitting the spectral model. The figure illustrates the effect of the number of inner iterations in a nested EM algorithm on the required CPU time for convergence with both the standard and fast* (*i.e., conditional augmentation*) *algorithms. The optimal algorithm is the fast algorithm with about* 10 *inner iterations and it requires only about a quarter of the CPU time of the standard EM algorithm, which has one inner iteration.*

the absorbed photons, but rather we need absorption only to be uniform across the support energies of the emission line. In particular, suppose $a_{\min}$ is the lowest absorption rate, $a_{\min} = \min_j\{1 - d_j g(\theta^A, E_j)\}$, where j varies over the support of emission line k. When we compute $\ddot{Y}_j^k$, we act as if the absorption rate were $1 - d_j g(\theta^A, E_j) - a_{\min}$. Thus, we add fewer counts to each bin. In particular, if line k is a delta function, we need not account for absorption at all when updating θ_k^L. This is the strategy used in the fast EM and fast nested EM algorithms illustrated in Figure 2. The fast EM algorithm offers additional computational savings over nesting; see van Dyk and Meng (2000) for another example involving the spectral model.

4. EXAMPLE

In this section we use our spectral model to study the quasar mentioned in Section 3.2; see also Sourlas et al. (2003) and van Dyk et al. (2001) for other examples of the application of this model.

Quasars are the most distant distinct detectable objects in the universe. They are believed to be supermassive black holes, whose masses exceed that of the Sun by a million times. They are powered by the gravitational potential energy of gas and stars falling into the central black hole, which results in emission across the electromagnetic spectrum. Because they are so distant, they give us a glimpse into the very distant past; the light that is now reaching the Earth left the quasar when the universe was as little as 10% of its current age, measured from the Big Bang. The study of quasars therefore has important consequences for cosmological theory.

In this example we focus on an emission line in this energy spectrum of the high redshift quasar PG1637+706. By measuring the location of the emission line in the spectrum and accounting for the expansion of the universe, we can estimate the distance of the quasar from the Earth. The wavelengths of electromagnetic waves originating from objects moving away from us appear to be elongated and hence lowered in energy when they reach us. By measuring the change in energy, we can recover this recession velocity. In a uniformly expanding universe, the recession velocity is a direct measure of distance.

We fit a spectral model consisting of a power law continuum, $f(\theta^C, E_j) = \alpha^C E_j^{-\beta^C}$, with the absorption model of Morrison and McCammon (1983) to account for absorption due to the ISM and IGM, and a power law continuum for background counts, $f(\theta^B, E_j) = \alpha^B E_j^{-\beta^B}$. We consider three models for the emission line.

MODEL 0. There is no emission line.

MODEL 1. There in an emission line with fixed location in the spectrum, but unknown intensity.

MODEL 2. There is an emission line with unknown location and intensity.

We use a Gaussian line profile for the emission line with standard deviation fixed at 0.125 keV throughout.

Initially, there was only a suspicion that there might be an emission line in the spectrum and we had no prior information as to the likely location for the line. To find candidate locations, we fit the model via maximum likelihood using the EM algorithm with 51 different starting values evenly spaced between 1.0 and 6.0 keV. We begin with the EM algorithm, because fitting the line location via the Gibbs sampler can be dangerous. The posterior distribution has several modes, corresponding to potential line locations, and the Gibbs sampler is generally unable to jump between these modes. Moreover, if the sampler is started far from any of the modes and a flat prior distribution is used for the line location, there may be no counts attributed to the line when the missing data are drawn,

TABLE 3
Multiple modes for the emission line location fitted using the EM algorithm

Mode (keV)	Domain of convergence (keV)	Log-likelihood
1.059	1.0–1.3	2589.31
1.776	1.4–2.0	2590.37
2.369	2.1–2.3	2590.19
2.807*	2.4–3.7	2594.94
4.216	3.8–4.7	2589.57
5.031	4.8–5.2	2589.31
5.715	5.3–5.9	2589.74

NOTE: The log-likelihood values imply that the largest mode corresponds to an emission line at 2.807 keV. (The asterisk indicates the principal mode.) The maximum log-likelihood for the model when no emission line is included is 2589.31.

leading to numerical difficulties when the line location is updated. The results of the 51 runs of the EM algorithm appear in Table 3. Judging from Table 3, the line is most likely to be located at about 2.81 keV. After consulting with experts, we found that the EM result agrees completely with expert knowledge: the line is most likely to be between 2.74 and 2.87 keV.

To further investigate the posterior distribution near the principal mode, we ran three Gibbs sampler chains, each starting near the mode, for 4000 iterations with 3 inner iterations per iteration. Gelman and Rubin's (1992) $\hat{R}$ statistic indicated adequate convergence of the chains for this mode. (As expected the chains never jumped between modes.) The marginal posterior distributions for the seven parameters all appear to be symmetric and unimodal; marginal summary statistics are given in Table 4. The 95% posterior interval for the line location is (2.66 keV, 2.94 keV), which in principle can be used to compute an interval for the relative velocity and the distance of the quasar. In the following section, we discuss the fit of the model and the strength of the evidence for including the emission line in the model.

5. MODEL CHECKING AND SELECTION

The family of models described in Section 2 represents a highly structured abstraction of the observed data with multiple levels of latent variables. Although the models are motivated by physical principles and instrumental specifications, checking the fit and verifying the latent structure of such complex models is a challenging task. In this section we discuss two strategies for checking the self-consistency of the model, that is, the ability of the fitted model to predict the data to which the model was fitted. First we discuss graphical methods based on residuals and then we discuss more formal tests based on the posterior predictive distribution.

5.1 Model Checking and Diagnostics

Graphical model diagnostics can be used to investigate whether the fitted models are consistent with the observed data. The first row of Figure 3 illustrates the fitted model for the source observations associated with quasar PG1637+706 as described in Section 4. The two columns in Figure 3 correspond to the model without an emission line and with an emission line fixed at 2.81 keV, respectively. The fitted models are obtained using $\hat{\theta}$, the posterior mean of the model parameter. In general, we transform each component of θ separately to symmeterize its marginal posterior distribution and to compute the posterior mean via Monte Carlo on this scale. The expected counts per channel, $\xi_l(\hat{\theta})$ [see (8)], along with the approximate error bars $\pm 2\sqrt{\xi_l(\hat{\theta})}$, are plotted against energy channel and compared with the observed counts. Some of the structure in $\xi_l(\hat{\theta})$ as a function of energy is due to the effects of the response matrix and the effective area of

TABLE 4
Summary statistics for the marginal posterior distributions for the analysis in Section 4

Parameter	2.5%	25%	Median	75%	97.5%	Mean	Std. dev.
α^C	3.499e −04	3.751e −04	3.890e −04	4.034e −04	4.317e −04	3.895e −04	2.084e −05
β^C	1.15683	1.28163	1.34854	1.41392	1.53951	1.34819	0.09822
θ^A	−1.13618	−0.8639	−0.72117	−0.57765	−0.30594	−0.7213	0.21244
α^B	−0.72395	−0.4071	−0.25793	−0.11736	0.14292	−0.26616	0.22158
β^B	−1.32096	−1.06123	−0.92721	−0.7889	−0.52515	−0.92561	0.20302
$\theta^L_{1,\lambda}$	33.9036	77.659	104.127	133.295	205.525	107.83115	43.46528
$\theta^L_{1,\mu}$	2.65657	2.75375	2.7948	2.83581	2.9422	2.79551	0.07121

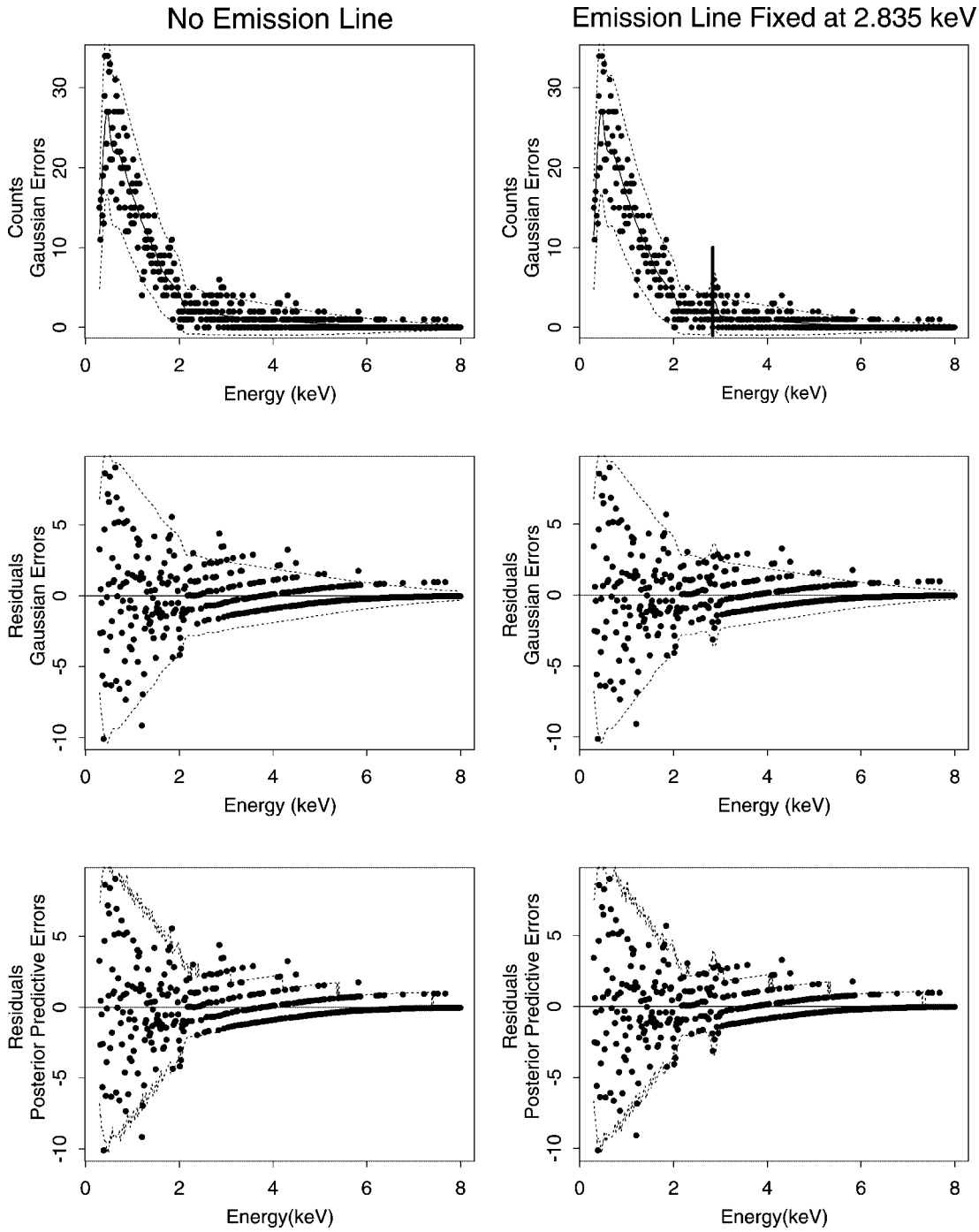

FIG. 3. *Model diagnostic plots. The first two rows of the figure show the data and residuals with predictive errors based on a Gaussian approximation. The third row shows the residuals with errors based on the posterior predictive distribution. The two columns correspond to Model* 0 *and Model* 1*, respectively. The advantage of the posterior predictive errors is evident for the low counts in the high-energy tail of the spectra. The excess counts near* 2.81 *keV are apparent in the first column and are corrected for in the second column, indicating evidence for including the emission line in the model.* (*The location of the emission line is represented by a vertical line in the upper right-hand plot.*)

the instrument. The second row of Figure 3 compares the residuals [i.e., $Y_l^{\text{obs}} - \xi_l(\hat{\theta})$] with an approximation of their error, $\pm 2\sqrt{\xi_l(\hat{\theta})}$. Although these error bars are easy to compute, they are based on a Gaussian approximation and do not account for the posterior variability of θ. A better strategy is to compute intervals based on the posterior predictive distribution. For example, using Monte Carlo, we can easily compute the high-

est 95% posterior predictive interval independently for each channel count. These intervals (based on 20,000 Monte Carlo draws) are compared with the observed counts in the third row of Figure 3. Comparing the second and third rows, we observe that the posterior predictive intervals cover the low intensity counts much better than the Gaussian approximations. Figure 3 also illustrates the evidence for the inclusion of the emission line. There are three points above the upper limit of the predictive intervals in channels near 2.81 keV in the first column of the figure. When we add the emission line to the model, the error bars widen to include these three points. In the next section we quantify the strength of this evidence.

5.2 Model Selection

Deciding whether to include a specific model component in a spectral model is often of direct scientific interest. Here we discuss the use of tests based on the posterior predictive distribution that aim to check the adequacy of a seemingly parsimonious model with an eye on the possibility of adding model components to better describe the data. Formal tests in this context are an especially challenging statistical task due to the form of the models in question. For example, a standard question of scientific interest is whether the data support the presence of a particular emission line. Suppose the hypothesized line has a known center and spread. Then formally we test the null hypothesis that $\theta^L_{k,\lambda} = 0$ against the alternative $\theta^L_{k,\lambda} > 0$ for some k. Given the form of the model in (4), this test is equivalent to the notoriously difficult task of determining the number of components in a finite mixture model or, more generally, of testing whether a parameter is on the boundary of its parameter space. It has been well known for decades that the standard asymptotic null distribution of the likelihood ratio test is inappropriate for this task (see the discussion in Titterington, Smith and Makov, 1985). Unfortunately, this fact seems little known among astrophysicists, who routinely use the likelihood ratio test or a related F test to check for emission lines or other model components with similar statistical difficulties.

The misapplication of the likelihood ratio and F tests is endemic in the astrophysical literature. A recent survey found over 125 papers published in *The Astrophysical Journal, Astrophysical Letters* or *The Astrophysical Supplement* that used the likelihood ratio test or the related F test in a questionable manner. The survey reviewed 183 papers published between 1995 and mid-2001 that were found using an electronic search for the keywords F statistic, F test or LRT. The results of the survey are summarized in Table 5.

TABLE 5
Results of a survey of papers in The Astrophysical Journal, its Letters and Supplement, published between 1995 and mid-2001 and returned by a search for F statistic, F test or LRT at The Astrophysical Journal website

Type of test	Number of papers
Null space on boundary	106
Comparing nonnested models	17
Other questionable cases	4
Seemingly appropriate use of test	56

We suggest using posterior predictive p values (Rubin, 1984; Meng, 1994; Gelman, Meng and Stern, 1996) to test null hypotheses which lie on the boundary of the parameter space. An appeal of the posterior predictive p value in astrophysics is its strong analogy with the frequentist p value, which is relatively well understood by astrophysicists. In particular, an *arbitrary* statistic $T(y)$ (e.g., the likelihood ratio statistic) is calibrated via its posterior predictive distribution

$$\begin{aligned} p\{T(y_{\text{rep}})|y\} &= \int p\{T(y_{\text{rep}}), \theta|y\}\, d\theta \\ &= \int p\{T(y_{\text{rep}})|\theta\} p(\theta|y)\, d\theta, \end{aligned} \tag{9}$$

where y_{rep} is a replicated data set and the second equality follows because y and y_{rep} are independent given θ. Thus, we average the *sampling distribution* $p\{T(y)|\theta\}$ over the uncertainty quantified by the posterior distribution of θ. The difficulty with the standard likelihood ratio test is that its sampling distribution is not easily calibrated. If not for the dependence of the sampling distribution on the unknown parameter θ, calibration could be accomplished via Monte Carlo methods. Posterior predictive p values overcome this final difficulty by accounting for the uncertainty in θ via its posterior distribution.

The result can sometimes be conservative, especially when the test statistic is poorly suited for detecting the model feature in question (Meng, 1994; Bayarri and Berger, 1999). Thus, Bayarri and Berger (1999) suggested conditioning on sufficient statistics for nuisance parameters to elicit more power. In practice, this suggestion can be mathematically and computationally demanding. The principle advantage of posterior predictive p values is that, although analytical results are typically not available, calibration is easily accomplished via Monte Carlo methods. Specifically,

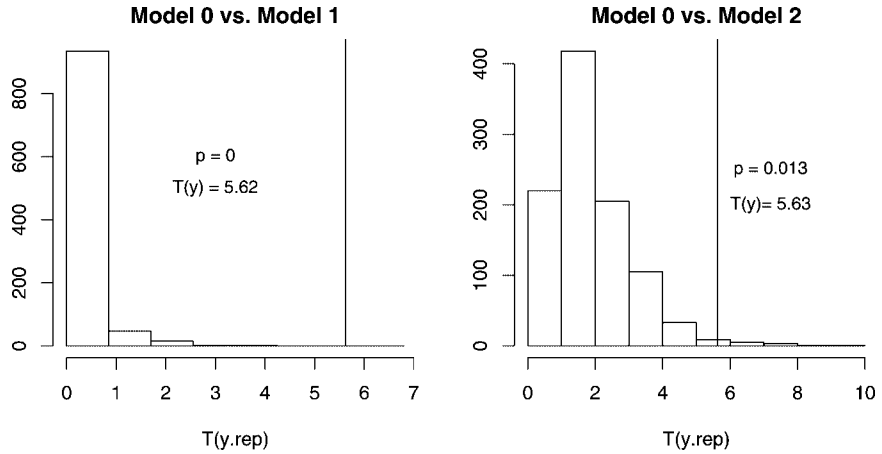

FIG. 4. *The posterior predictive check. The two histograms compare the observed likelihood ratio test statistics* (*vertical lines*) *with* 1000 *simulations from the posterior predictive distribution. The left plot is the comparison between Model* 0 *and Model* 1*, and the right plot is the comparison between Model* 0 *and Model* 2*. Both model checks indicate strong evidence for including the emission line.*

we need only sample θ from its posterior distribution, sample a replicated data set y_{rep} from the sampling distribution given the sampled value of θ and compute the (likelihood ratio) test statistic using the replicated data. The frequency under repeated sampling that this procedure results in a more extreme test statistic than is actually observed is the posterior predictive p value. If this is a very small number, we conclude that the data would be unlikely to have been generated under the posterior predictive distribution, and in terms of the characteristics measured by the test statistic, the model is not adequate for the data. [There are of course numerous other techniques for model checking, e.g., Bayes factors and the Bayesian information criterion; see Protassov et al. (2002) for a more detailed discussion of our preference for posterior predictive p values in this setting.]

To illustrate the use of posterior predictive p values, we return to the example of Section 4 to quantify the evidence for the emission line. We use the likelihood ratio test as the test statistic,

$$T(y_{\text{rep}}) = \log\left\{\frac{\sup_{\theta\in\Theta_i} L(\theta|y_{\text{rep}})}{\sup_{\theta\in\Theta_0} L(\theta|y_{\text{rep}})}\right\}, \quad i = 1, 2,$$

where Θ_0, Θ_1 and Θ_2 represent the parameter spaces for Model 0, Model 1 and Model 2, respectively; see Section 4. We use the EM algorithm to compute $T(y_{\text{rep}})$. In particular, after generation of 1000 data sets from the posterior predictive distribution under Model 0, we fit each of the three models to each of the 1000 data sets via maximum likelihood. When we fit Model 2, we used six evenly spaced starting values for the line location over the range (1.0 keV, 4.0 keV); the maximum of the resulting six local maximum likelihood values is taken to be the global maximum likelihood. Although this procedure is not guaranteed to return the global maximum, it is a legitimate statistical procedure that results in a test statistic, whose posterior predictive distribution we investigate. Figure 4 shows the posterior predictive distribution of $T(y_{\text{rep}})$ and posterior predictive p value with both Model 1 and Model 2 as the numerator model. Together the two posterior predictive p values indicate that there is strong evidence for the presence of the emission line in the spectrum. Given the prior belief that the line is near 2.81 keV, it is legitimate to use the first posterior predictive p value, which is essentially zero. Without such prior information, one should use the second value, which is about 0.01. It is evident that the prior information increases the power of the comparison.

6. PILEUP

6.1 The Nature of Pileup

We turn now to photon pileup, a form of data degradation that is much more challenging than the forms discussed in Section 2.2. Pileup occurs in X-ray CCD's when two or more photons arrive at the same location on the detector during the same time frame. Such coincident events are counted as a single higher energy event or lost altogether if the total energy goes above the on-board discriminators. Thus, for bright sources pileup can seriously distort both the count rate and the energy spectrum.

Chandra collects data using an on-board *event* detection algorithm, which at the end of each time frame scans spatial pixels for local maxima in the energy charge. Because the charge from a single photon is often recorded across several adjacent pixels, the charge in neighboring pixels is investigated, for example, in a 3×3 or 5×5 *event detection island*. The events are classified into an *event grade*, which depends on the spread of the charge in the event detection island. Because the event grade distribution depends on the number of photons that are piled (so-called grade migration due to pileup), the grade carries important information for accounting for pileup. A typical strategy is to discard events that are likely to be piled based on their grade. Since the likelihood of being piled may vary with energy, such strategies have the potential to bias results. The exact nature of the classification of the events into grades is quite complex and beyond the scope of this paper; more details can be found in Ballet (1999).

Accounting for pileup is perhaps the most important outstanding data-analytic challenge involving *Chandra* data. Within a model-based framework, however, there is no conceptual problem in dealing with complicated data generation processes such as those involved in pileup. Specifically, we treat the number of photons that correspond to each event as well as their energies as missing data. The sum of the individual missing energies is the observed event energy. In a Bayesian framework using MCMC, we need to stochastically separate a subset of the observed counts into multiple counts of lower energy using the current iteration of the spectral–spatial model being fitted. [Other statistical methods have been explored to address pileup; see Davis (2001) for an approach based on minimum χ^2 fitting.]

Because of the nature of pileup it is impossible to appropriately model a piled spectrum without accounting for the spatial characteristics of the image. That is, we cannot concentrate solely on the spectral margin of the data. Consider two sources with the same spectrum, but with different spatial characteristics; in particular, suppose one is a point source, which is highly concentrated on the detector, and the other is an extended source, which is highly dispersed spatially. Since more photons land near each other on the detector, the first source is apt to be much more piled than the second. Any reasonable analysis must take this into account. Nonetheless, many sources of interest are essentially point sources and such sources are subject to the most severe pileup. Thus, to avoid the complication of jointly modeling spatial and spectral data, we tackle the important problem of point sources. This is the first of three simplifying assumptions that we use in our solution.

ASSUMPTION 1. We assume that the source is a point source, blurred uniformly across a region of the detector. [Due to instrument response, there is always some (nonuniform) spatial blurring, which is quantified via the so-called *point spread function*.]

ASSUMPTION 2. We suppose that each event is composed of either one or two actual photon arrivals.

ASSUMPTION 3. We assume the event detection islands are at fixed locations on the detector.

In Assumption 2, we specify that each event corresponds to exactly one or exactly two photons arriving in the event detection island during a single time frame. This assumption is clearly a simplification; there is a (generally smaller) probability that three or more photons pile. Even if there are only two photons involved in the pileup process, their energies may not add up strictly to the observed event energy because of the nature of the on-board event detection algorithm. In Assumption 3, we neglect the fact that the locations of the event detection islands are determined by local maxima in the charge on the detector at each time frame. Although none of the simplifying assumptions is realistic, they serve as an approximation and a starting point for more sophisticated methods. Moreover, as is demonstrated in Section 6.3, real data analyses can be robust to these assumptions.

6.2 Statistical Modelling of Pileup

To account for pileup using an MCMC sampler in the Bayesian framework, we need only stochastically separate a subset of the events into a number of lower energy photons. (Under Assumption 2, we divide events into at most two photons.) This illustrates the power of fitting highly structured hierarchical models via MCMC: a complex model can be fit via a sequence of relatively simple steps. In particular, we have an MCMC sampler composed of two steps:

STEP 1. Given the spectrum, we "unpile" the counts.

STEP 2. Given the unpiled counts, we update the spectrum.

Only Step 1 is new in this procedure; Step 2 contains many substeps and is described in Section 3 in the absence of pileup. Thus, here we focus only on deriving Step 1.

As in Section 2.2, we suppose that in the absence of pileup, the observed counts are Poisson with intensity $\xi_l(\theta)$ for some set of energy channels $l \in \mathcal{L}$. The first panel in Figure 5 illustrates a power law continuum with simple exponential absorption and three spectral lines. For simplicity, in Figure 5 we assume there is no background contamination or instrument response and that the effective area is constant. Accounting for these effects causes no difficulty because of the hierarchical structure of the model, but obscures the ideas involved with pileup.

Under this Poisson model, we can compute the probabilities of the possible one-photon and two-photon events in any particular event detection island. Suppose we observe an event with energy $E_\star$, that is, an event counted in energy channel $l(E_\star)$, the channel that corresponds to $E_\star$. Let $Y = \{Y_l, l \in \mathcal{L}\}$ be the photon counts in each energy channel during the relevant time frame and in the relevant event detection island; the sum of the energies of these counts is the observed energy of the event, $E_\star$. We wish to sample the component counts Y from $p\{Y|E_\star, \xi(\theta)\}$, where $\xi(\theta)$ represents the set of Poisson intensities in a single time frame and in a single event detection island. (Under Assumptions 1 and 3 the intensities per time frame per event detection island are a scalar multiple of the overall intensities.)

By Bayes theorem,

$$p\{Y|E_\star, \xi(\theta)\} \propto p\{E_\star|Y, \xi(\theta)\}p\{Y|\xi(\theta)\}. \tag{10}$$

The first term on the right-hand side of (10) ensures that the energies of the photons that make up the event sum to $E_\star$. Thus, if the event corresponds to a single photon, we must have

$$Y_{l(E_\star)} = 1 \quad \text{and} \quad Y_l = 0 \quad \text{for } l \neq l(E_\star). \tag{11}$$

The situation is somewhat more complicated when an event arises from multiple photons. Because the energies are binned into energy channels, determining what values of Y are possible can be somewhat complicated, especially if the channels are of unequal size, that is, correspond to energy ranges of differing width. In actual data analysis, we use channels that are equal in size and when adding channel energies, we assume that photon energies are equal to the midpoint of the channel energy range. Because of *Chandra's* high resolution, we expect this approximation to be of little consequence. For a two-photon event, we have

$$Y_l = \begin{cases} 1, & \text{for } l = l(E_1) \text{ and } l = l(E_2), \\ 0, & \text{otherwise,} \end{cases} \tag{12}$$

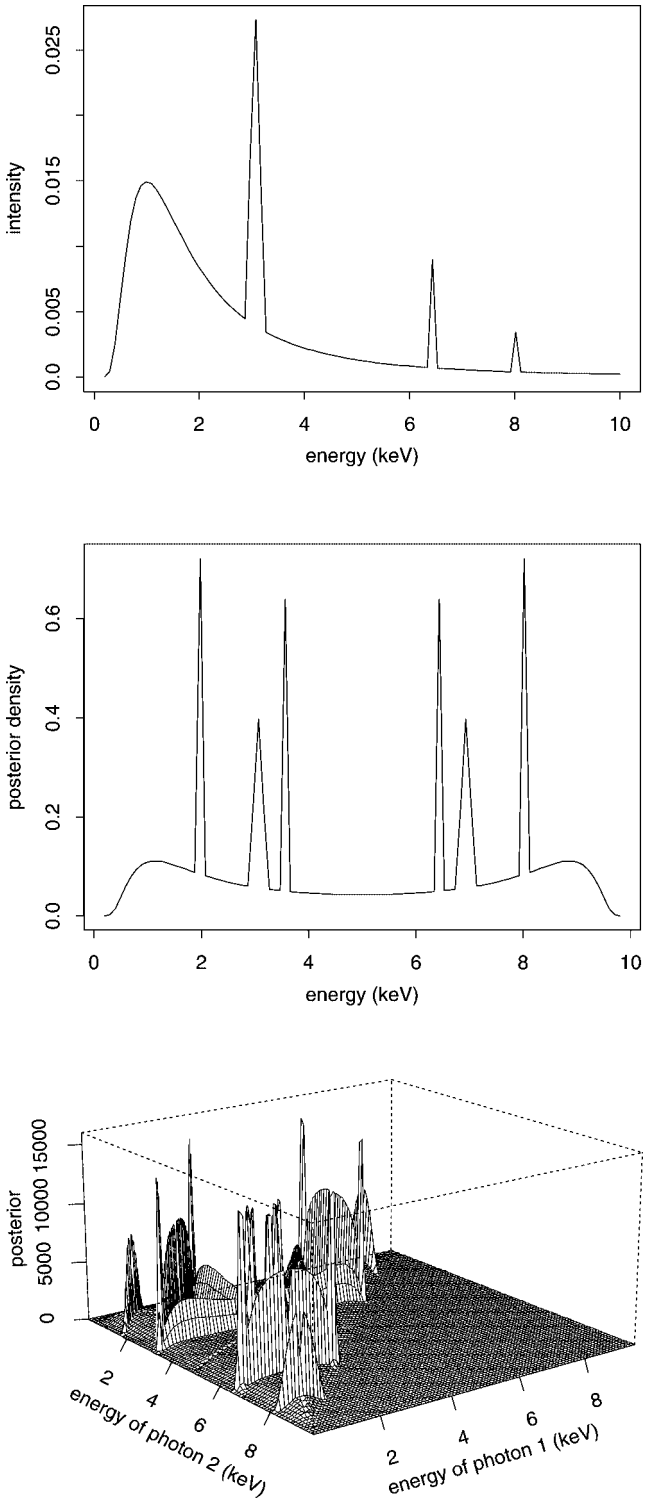

FIG. 5. *Accounting for pileup. The first panel illustrates a power law with exponential absorption and three emission lines. The probability density of the energy of one of two photons with energies that sum to* 10 *keV with this source spectrum is illustrated in the second panel. The final panel shows the joint probability density of the energy of two of three photons with energies that sum to* 10 *keV.*

where E_1 and E_2 are two energies such that $E_1 + E_2 = E_\star$. Alternatively, we may have $Y_{l(E_\star/2)} = 2$ and $Y_l = 0$ for all other channels.

Once the set of possible photon energies is determined, their relative probabilities are computed using the second factor on the right-hand side of (10),

$$\prod_{l \in \mathcal{L}} \frac{\xi_l(\theta)^{Y_l} \exp(-\xi_l(\theta))}{Y_l!}, \tag{13}$$

with Y given, for example, in (11) or (12). The second panel of Figure 5 shows the probability density of the energy of one of two component photons in an event of energy 10 keV based on the spectrum in the first panel of Figure 5. In particular, we plot $\Pr\{E_1, E_2 = 10 - E_1 | E_\star = 10, \xi(\theta)\}$ versus E_1; the probability is also conditional on there being exactly two component photons.

Although in our solution we assume there are at most two piled photons in each event, there is no conceptual problem with handling three or more photons. The difficulty lies in computation. For example, if we want to split an event into three photons, we have to evaluate all possible combinations of three lower energies that sum to the observed event energy. As an example, we again consider the spectrum in Figure 5. Suppose that a 10-keV event actually consists of three photons; the third panel in Figure 5 illustrates the distribution of the energies of two of the photons. (The third energy is determined because the three energies must sum to 10 keV.) Efficient sampling from this highly structured distribution would require sophisticated Monte Carlo methods.

6.3 Unpiling 3c273 ACIS-S Spectrum

In this section, we illustrate and validate our pileup procedures using a pair of *Chandra* observations of 3c273, a strong X-ray point source. The first observation (denoted ACIS-S) is a standard *Chandra* observation and is highly piled because of the intensity of the source. The second observation (denoted ACIS-S/HETG) was obtained using a grating on *Chandra*; the grating spreads the source across the detector according to the energy of the photons. Because more of the detector surface is used, the grating spectrum should exhibit significantly less pileup. (The higher spectral resolution of grating data comes at the cost of the spatial resolution.) Because the grating data are essentially unpiled, they offer an ideal test of our procedures for handling pileup. If our procedures work well, they should return essentially the same fit for both data sets (Kang et al., 2003).

TABLE 6
Summaries of fitted models for the analysis in Section 6.3

Data	Pileup in model?	Fitted model parameters[a]		
		Γ[b] < 2 keV	Γ > 2 keV	% piled
ACIS-S/HETG	No	1.70 ± 0.06	1.05 ± 0.05	n/a
ACIS-S/HETG	Yes	1.70 ± 0.05	1.07 ± 0.05	00.6%
ACIS-S	No	1.53 ± 0.03	1.12 ± 0.04	n/a
ACIS-S	Yes	1.69 ± 0.03	1.29 ± 0.05	14.3%

[a]Error bars are one posterior standard deviation.
[b]The power law parameter is represented by Γ.

We fit a spectral model that consists of a broken power law continuum with a break at 2 keV. The energy channels from 0.5 to 8.0 keV were used to fit the model to both data sets in each of two ways: ignoring and accounting for pileup. The fitted models are summarized in Table 6. Since we expect little pileup in the ACIS-S/HETG data, it is not a surprise that accounting for pileup did not change the fit significantly. For the ACIS-S data, however, the fit does change significantly. When we account for pileup in the ACIS-S data, the fit matches the fits for the ACIS-S/HETG data, but only for the low energies. The problem for higher energies is that when higher energy photons are piled, they are often recorded as events of energy greater than 8 keV. When we analyze only the data up to 8 keV, we miss these events and underestimate the expected counts in the energy channels from 2 to 8 keV; thus, the power law parameter is overestimated above 2 keV. In principle, this bias can be accounted for by adding another level of missing data in the formulation of the model: If photons are piled to energies greater than 8 keV, they are not recorded. Extending the model in this direction is an area of current research.

7. FUTURE WORK

In this article we have outlined an important component of the collaborative work of the California-Harvard Astro-Statistics Collaboration (CHASC), URL: www.ics.uci.edu/˜dvd/astrostat.html. There are many related projects. For example, we are developing multiscale methods for image analysis of spatial data (van Dyk and Hans, 2002; Esch, 2003; van Dyk et al., 2004; Esch, Connors, Karovska and van Dyk, 2004). Ultimately, we hope to construct models for joint spatial–spectral analysis that can both describe how the spectrum changes over an extended source and correctly account for pileup. Some sources have

a multitude of important emission lines, and carefully designed hierarchical structure can be used both to describe the population of emission lines and to reduce dramatically the dimension of the parameter space on the bottom level of the model (van Dyk et al., 2004). A similar strategy is being developed to describe the distribution of the intensity or flux of X-ray sources across the sky. In addition to the data distortion mechanisms described in Section 2.2, this model must account for the propensity of detection as a function of source intensity.

As is evidenced by the other articles in this issue and in the recent conference, *Statistical Challenges in Modern Astronomy III* and its proceedings (Feigelson and Babu, 2003), there is a multitude of fruitful areas for astronomers and statisticians to work together. It is our hope that we have whetted the appetites of some readers who will join us in this exciting area of scientific collaboration.

ACKNOWLEDGMENTS

The authors gratefully acknowledge funding for this project provided in part by NSF Grants DMS-01-04129 and DMS-04-38240 and by NASA Contract NAS8-39073 (CXC). This work is a product of the collaborative effort of the California-Harvard Astro-Statistics Collaboration (CHASC), URL: www.ics.uci.edu/˜dvd/astrostat.html whose members include A. Connors, D. Esch, P. Freeman, C. Hans, V. L. Kashyap, R. Protassov, A. Siemiginowska, N. Sourlas and Y. Yu.

REFERENCES

BALLET, J. (1999). Pile-up on X-ray CCD instruments. *Astronomy and Astrophysics Suppl. Ser.* **135** 371–381.

BAYARRI, M. J. and BERGER, J. O. (1999). Quantifying surprise in the data, and model verification. In *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 53–82. Oxford Univ. Press, London.

CASH, W. (1979). Parameter estimation in astronomy through application of the likelihood ratio. *Astrophysical J.* **228** 939–947.

DAVIS, J. E. (2001). Event pileup in charge-coupled devices. *Astrophysical J.* **562** 575–582.

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** 1–38.

ESCH, D. N. (2003). Extensions and applications of three statistical models. Ph.D. dissertation, Dept. Statistics, Harvard Univ.

ESCH, D. N., CONNORS, A., KAROVSKA, M. and VAN DYK, D. A. (2004). An image restoration technique with error estimates. *Astrophysical J.* **610** 1213–1227.

FEIGELSON, E. D. and BABU, G. J., eds. (2003). *Statistical Challenges in Modern Astronomy III*. Springer, New York.

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

FREEMAN, P., GRAZIANI, C., LAMB, D., LOREDO, T., FENIMORE, E., MURAKAMI, T. and YOSHIDA, A. (1999). Statistical analysis of spectral line candidates in gamma-ray burst GRB 870303. *Astrophysical J.* **524** 753–771.

GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409.

GELMAN, A., MENG, X.-L. and STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statist. Sinica* **6** 733–807.

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.* **7** 457–472.

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.

KANG, H., VAN DYK, D. A., YU, Y., SIEMIGINOWSKA, A., CONNORS, A. and KASHYAP, V. (2003). New MCMC methods to address pile-up in the Chandra X-ray observatory. In *Statistical Challenges in Modern Astronomy III* (E. Feigelson and G. Babu, eds.) 449–450. Springer, New York.

LAMPTON, M., MARGON, B. and BOWYER, S. (1976). Parameter estimation in X-ray astronomy. *Astrophysical J.* **208** 177–190.

LANGE, K. and CARSON, R. (1984). EM reconstruction algorithms for emission and transmission tomography. *J. Computer Assisted Tomography* **8** 306–316.

LUCY, L. B. (1974). An iterative technique for the rectification of observed distributions. *Astronomical J.* **79** 745–754.

MENG, X.-L. (1994). Posterior predictive p-values. *Ann. Statist.* **22** 1142–1160.

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.

MORRISON, R. and MCCAMMON, D. (1983). Interstellar photoelectric absorption cross sections, 0.03–10 keV. *Astrophysical J.* **270** 119–122.

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. *Astrophysical J.* **571** 545–559.

RICHARDS, W. H. (1972). Bayesian-based iterative method of image restoration. *J. Opt. Soc. Amer.* **62** 55–59.

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172.

SHEPP, L. A. and VARDI, Y. (1982). Maximum likelihood reconstruction for emission tomography. *IEEE Trans. on Medical Imaging* **1** 113–122.

SIEMIGINOWSKA, A., ELVIS, M., ALANNA, C., FREEMAN, P., KASHYAP, V. and FEIGELSON, E. (1997). AXAF data analysis challanges. In *Statistical Challenges in Modern Astronomy II* (G. Babu and E. Feigelson, eds.) 241–258. Springer, New York.

SMITH, A. F. M. and ROBERTS, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B* **55** 3–23.

SOURLAS, N., VAN DYK, D. A., KASHYAP, V., DRAKE, J. and PEASE, D. (2003). Bayesian spectral analysis of "MAD" stars. In *Statistical Challenges in Modern Astronomy III* (E. Feigelson and G. Babu, eds.) 489–490. Springer, New York.

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.

TITTERINGTON, D. M., SMITH, A. F. M. and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

VAN DYK, D. A. (2000a). Fast new EM-type algorithms with applications in astrophysics. Technical report.

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

VAN DYK, D. A. (2003). Hierarchical models, data augmentation, and Markov chain Monte Carlo (with discussion). In *Statistical Challenges in Modern Astronomy III* (E. Feigelson and G. Babu, eds.) 41–56. Springer, New York.

VAN DYK, D. A. and HANS, C. M. (2002). Accounting for absorption lines in images obtained with the Chandra X-ray Observatory. In *Spatial Cluster Modelling* (A. Lawson and D. Denison, eds.) 175–198. CRC Press, London.

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.

VAN DYK, D. A., CONNORS, A., ESCH, D. N., FREEMAN, P., KANG, H., KAROVSKA, M., KASHYAP, V., SIEMIGINOWSKA, A. and ZEZAS, A. (2004). Deconvolution in high energy astrophysics: Science, instrumentation, and methods (with discussion). *Bayesian Analysis*. To appear.

VAN DYK, D. A. and MENG, X.-L. (2000). Algorithms based on data augmentation. In *Computing Science and Statistics*: *Proc. 31st Symposium on the Interface* (M. Pourahmadi and K. Berk, eds.) 230–239. Interface Foundation of North America, Fairfax Station, VA.

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