THE ASTROPHYSICAL JOURNAL, 548:224–243, 2001 February 10

ANALYSIS OF ENERGY SPECTRA WITH LOW PHOTON COUNTS VIA BAYESIAN POSTERIOR SIMULATION

DAVID A. VAN DYK
Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138; vandyk@stat.harvard.edu
ALANNA CONNORS[1]
Department of Astronomy, Whitin Observatory, Wellesley College, Wellesley, MA 02481; connors@frances.astro.wellesley.edu
AND
VINAY L. KASHYAP AND ANETA SIEMIGINOWSKA
Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138;
kashyap@head-cfa.harvard.edu, aneta@head-cfa.harvard.edu
Received 2000 March 9; accepted 2000 August 25

ABSTRACT

Over the past 10 years Bayesian methods have rapidly grown more popular in many scientific disciplines as several computationally intensive statistical algorithms have become feasible with increased computer power. In this paper we begin with a general description of the Bayesian paradigm for statistical inference and the various state-of-the-art model-fitting techniques that we employ (e.g., the Gibbs sampler and the Metropolis-Hastings algorithm). These algorithms are very flexible and can be used to fit models that account for the highly hierarchical structure inherent in the collection of high-quality spectra and thus can keep pace with the accelerating progress of new space telescope designs. The methods we develop, which will soon be available in the Chandra Interactive Analysis of Observations (CIAO) software, explicitly model photon arrivals as a Poisson process and thus have no difficulty with high-resolution low-count X-ray and γ-ray data. We expect these methods to be useful not only for the recently launched *Chandra X-Ray Observatory* and *XMM* but also for new generation telescopes such as Constellation X, *GLAST*, etc. In the context of two examples (quasar S5 0014+813 and hybrid-chromosphere supergiant star α TrA), we illustrate a new highly structured model and how Bayesian posterior sampling can be used to compute estimates, error bars, and credible intervals for the various model parameters. Application of our method to the high-energy tail of the *ASCA* spectrum of α TrA confirms that even at a quiescent state, the coronal plasma on this hybrid-chromosphere star is indeed at high temperatures (>10 MK) that normally characterize flaring plasma on the Sun. We are also able to constrain the coronal metallicity and find that although it is subject to large uncertainties, it is consistent with the photospheric measurements.

Subject headings: methods: data analysis — methods: statistical

1. INTRODUCTION

The ever-increasing power and sophistication of today's high-energy instruments give access to 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 paper we present an innovative implementation of state-of-the-art statistical methods for fitting high-resolution spectra from the *Chandra X-Ray Observatory*. The common "folk wisdom" of how to bin data, subtract background counts, propagate errors, and, for example, estimate the significance of a spectral line profile are unreliable and can lead to unacceptable results (for discussion see Loredo 1993; Nousek 1993; Feigelson & Babu 1997; Siemiginowska et al. 1997; Zimmerman 1997). For example, binning data sacrifices the resolution of the instrument, subtracting background can lead to negative counts with unpredictable results, and statistical black boxes such as the χ^2 and Cash statistics (Lampton, Margon, & Bowyer 1976; Cash 1979), although often useful, may not be equipped to answer standard questions (e.g., Protassov et al. 2001). Some authors have suggested solutions to such problems, which involve ad hoc adaptations of commonly used methods (e.g., Gehrels 1986; Collura et al. 1987; Mighell 1999). Unfortunately, when such solutions are not rooted in a theoretical framework, they have no justification beyond problems that are more or less the same as the simulation studies that justify them, and we are often forced into additional ad hoc adaptations. This approach is difficult to justify in light of modern statistical methods that address reasonable model assumptions directly. Thus, in recent years, astrophysicists have increasingly turned to likelihood-based (e.g., Lucy 1974; Cash 1979; Schmitt 1985; Sciortino & Micela 1992) and Bayesian methods (e.g., Bijaoui 1971; Richardson 1972; Gregory & Loredo 1992; Loredo 1993; Connors 1997; Siemiginowska 1997; Kashyap & Drake 1998; Freeman et al. 1999; see also Appendix C). The primary purpose of this paper is to illustrate how Bayesian methods can provide practical answers to outstanding real problems that standard methods are not able to handle. The methods described here are equipped with readily available parameter estimates, credible intervals, error bars, model-checking techniques, methods for combining information from multiple sources, etc., all within a flexible theoretical framework and without reliance on asymptotic Gaussian approximations.

We illustrate Bayesian data analysis via two detailed examples. The analysis of quasar S5 0014+813 offers a straightforward introduction to our methods, and the

[1] Alanna Connors is currently affiliated with Eureka Scientific, 2452 Delmer Street, Suite 100, Oakland, CA 94602-3017.

TABLE 1

INDEX OF VARIOUS TERMS DISCUSSED AND DEFINED IN THE PAPER

Terms	Defined and/or Discussed
Absorption	§§ 3.1 and 3.4
Bayes's Theorem	Equation (1)
Conjugate prior distribution	Footnote 12
Credible interval	§ 2.1
Data augmentation	§§ 1, 2.3, and B1
Data augmentation algorithm	§§ 2.3 and A1
Effective area	Footnote 4
Equivalent width	Footnote 15
Gamma distribution	Footnote 8
GLM or generalized linear model	Footnote 5, § 3.2
Gibbs sampler	§§ 2.3 and A3
Hierarchical model	Footnote 6
Hyperparameters	§ 2.1
Improper distribution	Footnote 9
MCMC or Markov chain Monte Carlo	§§ 2.3, A1, A2, A3, and A4
Marginal distribution	§ 2.1, Equation (6)
Maximum effective area	Footnote 10
Model	§ 1
Monte Carlo integration	§ 2.2
Multinomial distribution	Footnote 19
Noninformative prior distribution	§ 2.1
Nuisance parameter	§ 2.1, Equation (6)
Observed data model	§ 1
Poisson distribution	Footnote 3
Posterior distribution	Equation (1)
Prior distribution	Equation (1)
PHA or pulse-height amplitude	Footnote 2
Source model	§ 1

extremely low count hybrid-chromosphere supergiant star α TrA observation shows how we tackle a previously intractable analysis. Together, these examples demonstrate the power of Bayesian methods to handle highly structured models designed to reflect the structure in both the source spectrum and the data collection process. Our methods avoid the binning of counts and thus the sacrificing of high-resolution information required by standard data analysis methods. The analysis of S5 0014+813 is consistent with the available standard analysis, which relies on extra binning and the removal of the high-energy low-count tail.

We emphasize that we model not only the source spectrum but also other stochastic components of data collection and the instrument such as background contamination and instrument response. In general, we refer to our stochastic representation of the entire process as the (*statistical*) *model*. For clarity we refer to the spectral or physical model as the *source model* and to the model for the observed (PHA[2]) counts as the *observed data model*. In our detailed example, we develop a model and algorithms for spectral analysis of high-energy (or other) data using a Poisson[3] process for photon arrivals. We allow for (1) stochastic instrument response via a photon redistribution matrix, (2) the absorption of photons, (3) the effective area[4] of the telescope, and (4) background contamination of the source. In particular, we model information on background emissions as the realization of a second Poisson process (see Loredo 1993), thereby eliminating the need to subtract off directly the background counts and the rather embarrassing resulting problem of negative photon counts. The source energy spectrum is modeled as a mixture of several (Gaussian) line profiles and a generalized linear model[5] (GLM) (e.g., McCullagh & Nelder 1989), which accounts for the continuum. GLMs have become the standard statistical method for incorporating information contained in independent variables (as in regression) into many non-Gaussian models and are thus an obvious but innovative choice in this setting.

In addition to several Markov chain Monte Carlo (MCMC) algorithms, we describe and use data augmentation, an important statistical method for Bayesian (and other) analyses. Data augmentation is an elegant computational construct allowing us to take advantage of the fact that if it were possible to collect additional data, statistical analysis would be greatly simplified. This is true regardless of why the so-called missing data are not observed. For example, if we were able to record the counts due to background contamination in addition to total counts in each

[2] Pulse-height amplitude, originally in proportional counters, the number of electrons produced by a photon, hence the amplitude of the current pulse registered by the detector electronics. The term now refers to the measure of the energy deposited on the detector (as opposed to the true energy).

[3] Recall, a random variable X is said to follow a Poisson distribution with parameter or intensity λ if $\Pr(X = x) = e^{-\lambda}\lambda^x/x!$. In this case $E(X) = \lambda$ and we often write $X \overset{d}{\sim}$ Poisson (λ) (read as X is distributed as Poisson with intensity λ). This representation conditions on the intensity parameter, λ, which in turn may vary.

[4] The effective area of the telescope is the fraction of the true geometric area that the telescope presents to sky. This varies with energy.

[5] In a GLM we assume that a transformation (e.g., log) of the model is linear in a set of independent variables. We emphasize that this is not equivalent to transforming the data and proceeding with linear regression. A generalized linear model utilizes the likelihood of the assumed model, which may not be Gaussian (e.g., the assumed model may be Poisson). See § 3.3 for details.

bin, it would, of course, be a trivial task to account for the background. There is a large class of powerful statistical methods designed for "missing data" problems. With the insight that "true" values of quantities recorded with measurement error can be regarded as "missing data," these methods can usefully be applied to almost any astrophysical problem. In particular, we can treat the true image (before instrument response), the absorbed photon counts, and unbinned energies as "missing data" to account for instrument response, absorption, and binning, respectively.

This introduction foreshadows the tone of the paper: in the process of developing new Bayesian methods, we describe and utilize state-of-the-art statistical reasoning, methods, and algorithms, whereby we explore a larger statistical framework for our problem of interest. Although we introduce many new tools and use terminology that may be unfamiliar, we endeavor to write in a manner accessible to astrophysicists and believe the resulting methods justify the required interdisciplinary work. Table 1 indexes terminology used in the paper that may be unfamiliar to some readers. To aid in the translation from standard statistical notation to standard astrophysical usage, many equations have been written according to both standards when notation is introduced. One notational convention is worthy of mention: we use superscripts to identify model components (e.g., background or absorption).

The paper is organized as follows. After a brief overview of the fundamentals of Bayesian analysis in § 2, we lay out our hierarchical statistical model,[6] which summarizes the photon collection process and parameterizes many relevant aspects of the energy spectrum in § 3. Two examples that aim to illustrate a typical data analysis and the advantages of Bayesian methods in this setting are given in § 4. Section 5 contains brief concluding remarks. Finally, in two appendices, we outline such important general MCMC methods as data augmentation, the Gibbs sampler, the Metropolis-Hastings algorithm, and judging convergence using multiple chains. We also describe in detail how we use these algorithms to fit the hierarchical source model of § 3.

2. BAYESIAN ANALYSIS

In this section we outline several important methodological and computational issues involved with Bayesian analysis using a simple model that accounts for background in a simplified Poisson process to motivate and illustrate ideas. Our introduction is brief, and we encourage interested readers to consult one of the several high-quality recent texts on the subject such as Gelman et al. (1995), Carlin & Louis (1996), Gilks, Richardson, & Spiegelhalter (1996), and Sivia (1996). In § 3 we show how the ideas developed here can be used for a detailed spectral analysis.

2.1. *Prior, Sampling, and Posterior Distributions*

Bayesian probability analysis is fundamentally based on one simple result known as Bayes's Theorem, which allows us to update a probability distribution based on new data or other information. In particular, knowledge about a (vector) model parameter, $\boldsymbol{\theta}$, is summarized by a probability distribution, $p(\boldsymbol{\theta})$, such that $\Pr(\boldsymbol{\theta} \in R) = \int_R p(\boldsymbol{\theta}) d\boldsymbol{\theta}$ for any region[7] R. Bayes's Theorem states

$$p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{I}) = \frac{p(\boldsymbol{Y} \mid \boldsymbol{\theta}, \boldsymbol{I}) p(\boldsymbol{\theta} \mid \boldsymbol{I})}{p(\boldsymbol{Y} \mid \boldsymbol{I})}, \tag{1}$$

where $\boldsymbol{Y}$ are the observed data or other new information pertaining to $\boldsymbol{\theta}$ and $\boldsymbol{I}$ represents any initial information known before $\boldsymbol{Y}$ is observed. Here $p(\boldsymbol{\theta} \mid \boldsymbol{I})$ represents our knowledge prior to observing $\boldsymbol{Y}$ and is called the *prior distribution*. The *sampling distribution* or *likelihood*, $p(\boldsymbol{Y} \mid \boldsymbol{\theta}, \boldsymbol{I})$, represents the likelihood of the data given the model parameters, and $p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{I})$ represents our updated knowledge regarding $\boldsymbol{\theta}$ after observing $\boldsymbol{Y}$ and is called the *posterior distribution*. Finally, $p(\boldsymbol{Y} \mid \boldsymbol{I})$ represents the unconditional distribution of $\boldsymbol{Y}$ and acts as the normalizing constant for $p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{I})$. The functional form of Bayes's Theorem describes how our prior knowledge should be updated in light of information contained in the data. The likelihood or sampling distribution is the basis for many standard statistical techniques, while the prior and posterior distributions are specific to Bayesian analysis.

To illustrate Bayes's Theorem, suppose we have observed counts, Y, contaminated with background in a (source) exposure and have observed a second exposure of pure background. Throughout this section, we assume that the source exposure is τ^S minutes and the pure background exposure is τ^B minutes with both exposures using the same area of the detector. (We generally use superscripts to represent photon "sources," e.g., source or background. Occasionally we use superscripts for powers; for clarity we place powers outside parentheses.) To model the source exposure, we assume that Y follows a Poisson distribution with intensity $\lambda^B + \lambda^S$, where λ^B and λ^S represent the expected counts during the source exposure due to background and source, respectively. Thus, the likelihood is

$$p(Y \mid \lambda^B, \lambda^S, \boldsymbol{I}) = \frac{e^{-(\lambda^B + \lambda^S)} (\lambda^B + \lambda^S)^Y}{Y!} \quad \text{for } Y = 0, 1, 2, \ldots . \tag{2}$$

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. As is detailed below, an important advantage of Bayesian methods is their ability to handle nuisance parameters by computing the *marginal* posterior distribution of the parameters of interest. The name "marginal" distribution originates with two-way tables of counts where the table margins sum over one of the variables to give the distribution of the other variable alone (i.e., its marginal distribution). Likewise, the marginal distribution of the parameter of interest is computed by integrating over (i.e., averaging over) the nuisance parameter.

At this point, we specify a prior distribution that allows us to include a priori knowledge (e.g., "allowed parameter ranges") from other experiments or other scientific information. One of the primary advantages of Bayesian analysis is a well-defined mechanism for the inclusion of information

[6] A hierarchical (statistical) model is formulated in terms of unobserved quantities, which are themselves statistically modeled. For example, we may assume that photons first arrive at a detector according to a Poisson process and then are randomly redistributed according to a photon redistribution matrix. A hierarchal model separates these two random processes into two levels of a structured model.

[7] The notation $\Pr(\boldsymbol{\theta} \in R)$ represents the probability that $\boldsymbol{\theta}$ is in the region R and is computed as the integral of the probability distribution of $\boldsymbol{\theta}$ over R.

outside the current data set. In the absence of prior information, we use diffuse or so-called noninformative priors, which are ordinarily flat and have minimal influence on the final analysis. The prior distribution itself may be conveniently parameterized using a set of *hyperparameters* that can be varied to represent the researcher's knowledge about the value of the model parameters and the degree of certainty of this knowledge. For example, we use the γ distribution[8] to parameterize prior information for λ^B and λ^S:

$$\lambda^B \mid \boldsymbol{I} \overset{d}{\sim} \gamma(\alpha^B, \beta^B)\,,$$
$$\lambda^S \mid \boldsymbol{I} \overset{d}{\sim} \gamma(\alpha^S, \beta^S)\,, \tag{3}$$

that is,

$$p(\lambda^B \mid \boldsymbol{I}) = \frac{(\beta^B)^{\alpha^B}(\lambda^B)^{\alpha^B-1}e^{-\beta^B\lambda^B}}{\Gamma(\alpha^B)}\,,$$
$$p(\lambda^S \mid \boldsymbol{I}) = \frac{(\beta^S)^{\alpha^S}(\lambda^S)^{\alpha^S-1}e^{-\beta^S\lambda^S}}{\Gamma(\alpha^S)}\,, \tag{4}$$

where the notation $\overset{d}{\sim}$ is read "follows the distribution" and λ^B and λ^S are assumed a priori independent. The γ prior on λ^B is mathematically equivalent to a Poisson likelihood resulting from a count equal to $\alpha^B - 1$ obtained with an exposure of β^B times that of the source exposure. (By mathematically equivalent, we mean that the prior on λ^B is proportional to a Poisson likelihood as a function of λ^B.) This leads to a natural choice of $p(\lambda^B \mid \boldsymbol{I}) = \gamma$ ($\alpha^B = Y^B + 1$, $\beta^B = \tau^B/\tau^S$), where Y^B are the counts from the background exposure. Notice that here (and throughout the paper) we explicitly incorporate information from the background exposure into the analysis via the prior distribution on λ^B. Thus, the counts from the source exposure, Y, are treated as the observed data $\boldsymbol{Y}$ in equation (1). We refer interested readers to Gelman et al. (1995), § 2.7, for further discussion and examples of the γ prior distribution with Poisson data.

The equivalence of the γ prior for λ^S and $\alpha^S - 1$ counts during an exposure of $\tau^S\beta^S$ minutes leads to a natural interpretation of the hyperparameters: for a relatively noninformative prior we choose β^S much less than 1. To illustrate this, we consider two priors: one noninformative and improper,[9] $p(\lambda^S \mid \boldsymbol{I})^{[1]} = \gamma(1, 0) \propto 1$ (*dotted line* in the first plot of Fig. 1); and one informative, where, let us say, we know from other means that three counts are to be expected in the same exposure time, hence $p(\lambda^S \mid \boldsymbol{I})^{[2]} = \gamma(4, 1)$ (*solid line* in the first plot of Fig. 1). This choice of informative prior is only an example: $\gamma(4, 1)$ corresponds to Poisson likelihood resulting from three counts with an exposure time equal to the source exposure (10 minutes). This is a rather informative prior distribution and is chosen to illustrate the effect of very informative prior. The noninformative prior contains information equivalent to zero counts in an exposure of 0 minutes.

Using Bayes's Theorem, with $\boldsymbol{\theta} = (\lambda^B, \lambda^S)$, we can combine the γ priors and the likelihood given in equation (2) to compute the posterior distribution,

$$p(\lambda^B, \lambda^S \mid Y, \boldsymbol{I}) \propto e^{-[\lambda^B(\beta^B+1)+\lambda^S(\beta^S+1)]} \times (\lambda^B + \lambda^S)^Y(\lambda^B)^{\alpha^B-1}(\lambda^S)^{\alpha^S-1}\,, \tag{5}$$

for $\lambda^B \geq 0, \lambda^S \geq 0$.

Nuisance parameters such as λ^B pose a monumental difficulty for classical statistical analysis, which often relies on fixing nuisance parameters at estimated values. Unfortunately, this does not account for uncertainty in their estimates and thus tends to be anticonservative. (Likewise, floating nuisance parameters or "propagating errors" when computing error bars are essentially a Gaussian assumption, which can lead to unpredictable results when such an

[8] The γ (α, β) 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.

[9] An improper distribution is a distribution that is *not* integrable and thus is not technically a distribution. One should use improper prior distributions only with great care since in some cases they lead to improper posterior distributions which are uninterpretable.

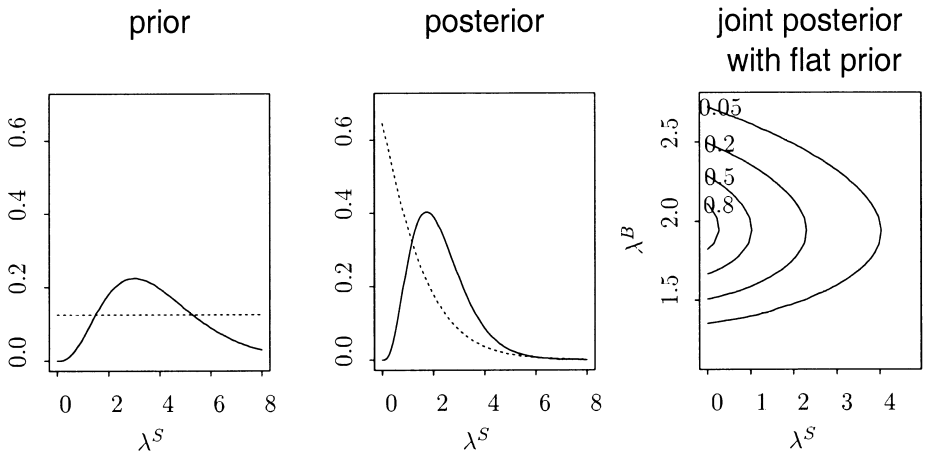

FIG. 1.—Combining information. The figure illustrates the combination of the information contained in the data and the prior distribution into the posterior distribution. The less informative dotted prior has less influence on its (dotted) posterior, which matches the low source count more closely than does the solid posterior. The joint posterior indicates the region of high posterior probability for both parameters under the noninformative prior for λ^S.

assumption is not justified.) The Bayesian solution averages over the posterior distribution (i.e., the uncertainty) of the nuisance parameter by computing the marginal (posterior) distribution of the parameters of interest without Gaussian approximations. For example, the marginal posterior distribution of λ^S can be computed by (numerical) integration,

$$p(\lambda^S \mid Y, \boldsymbol{I}) = \int_0^\infty p(\lambda^B, \lambda^S \mid Y, \boldsymbol{I}) d\lambda^B , \quad (6)$$

and is illustrated for the two priors for λ^S in the second plot of Figure 1, where we assign $Y = 1$ and $Y^B = 48$ with $\tau^s = 10$ minutes and $\tau^B = 2$ hr. In this example, direct subtraction of background would leave a "negative count" of -1; no such difficulty occurs with the Bayesian analysis. (See Loredo 1993 for another derivation of the marginal distribution of λ^S in this setting.)

Since the source count is small relative to the background count, we expect a small λ^S. Although this is evident in both posterior distributions in Figure 1, the highly informative prior distribution centered at $\lambda^S = 3$ pulls the (solid) posterior toward higher values, thus illustrating the effect of an informative prior distribution. Such sensitivity analyses often play an important part in Bayesian (or other) data analyses, since they investigate the sensitivity of the results to the statistical assumptions (e.g., the choice of prior distribution).

The posterior distributions should be interpreted as probability distributions representing the combined information in the prior and data. For example, a region, R, such that $\int_R p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{I}) d\boldsymbol{\theta} = \zeta$ is called a ζ-level credible interval (or credible region if $\boldsymbol{\theta}$ is multidimensional), and we can say $\Pr(\boldsymbol{\theta} \in R \mid \boldsymbol{Y}, \boldsymbol{I}) = \zeta$ (e.g., a 67%, 90%, or 95% credible region). The 90% credible regions for the posterior distributions illustrated in Figure 1 are (0.77, 4.24) using the informative prior and (0.04, 3.84) using the noninformative prior. Such probability statements are measures of our information regarding the value of the parameter $\boldsymbol{\theta}$, given the data and prior information. This is in contrast to the more traditional frequentist definition of probability, which defines a probability to be the long-term frequency of an event generally involving the data given $\boldsymbol{\theta}$. The posterior distribution is a complete summary of our information but is often summarized by its mean, $\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{I})$, and variance, $\mathrm{var}(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{I})$, or its modes and the curvatures at these modes. (The curvatures are most useful when the posterior is [locally] approximately Gaussian, as is asymptotically true under certain regularity conditions; e.g., see Gelman et al. 1995) In the following two sections we describe Monte Carlo methods for computing posterior means, posterior variances, and credible regions. To compute posterior modes (e.g., maximum likelihood estimates), van Dyk (2001) develops several expectation maximization (EM) algorithms for use in astrophysical applications. Posterior modes are often used to compute starting values for the more robust but computationally demanding Monte Carlo methods (see Appendix A, § A2). The EM algorithm gets its name because it iteratively *maximizes* the *expected* log posterior distribution of $\boldsymbol{\theta}$ given the augmented data.

Although a detailed description is beyond the scope of this paper, Bayesian methodology is well equipped for problems involving model selection. Methods based on Bayes's factors, computing the relative posterior probabilities of various competing models, and Bayesian "p-values" are all important and remain areas of active statistical research (e.g., Gregory & Loredo 1992; Protassov et al. 2001).

2.2. *Evaluating the Posterior via Monte Carlo Sampling*

For univariate or small dimensional parameter spaces, we can usually compute the posterior mean, variance, credible regions, and other summaries either analytically or via nonstochastic numerical methods (e.g., Gaussian quadrature or Laplace's method). In higher dimensions, however, these methods can be difficult to implement partially because of the difficulty in finding the region where the integrand is significantly greater than zero. Thus, we often resort to Monte Carlo integration. In particular, if we can obtain a sample from the posterior, $\{\boldsymbol{\theta}_{[t]}, t = 1, \ldots, T\}$, Monte Carlo integration approximates the mean of any function, g, of the parameter with

$$E(g(\boldsymbol{\theta}) \mid \boldsymbol{Y}, \boldsymbol{I}) \approx \frac{1}{T} \sum_{t=1}^{T} g(\boldsymbol{\theta}_{[t]}) , \quad (7)$$

where we assume $E(g(\boldsymbol{\theta}) \mid \boldsymbol{Y}, \boldsymbol{I})$ exists. For example, $g(\boldsymbol{\theta}) = \boldsymbol{\theta}$ and $g(\boldsymbol{\theta}) = [\boldsymbol{\theta} - E(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{I})][\boldsymbol{\theta} - E(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{I})]'$ 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 the value of 1 if the condition in curly brackets holds and 0 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. A general strategy is first to sample from the posterior distribution and then approximate various integrals of interest via Monte Carlo.

2.3. *Obtaining a Sample from the Posterior*

The Monte Carlo approximation methods depend on our ability to obtain a sample from the posterior distribution. Although in some cases the posterior distribution is a well-known distribution and trivial to sample, we must often use sophisticated algorithms to obtain a posterior sample. In Appendix A, we discuss three algorithms that have proven widely applicable in practice, the data augmentation algorithm (Tanner & Wong 1987), the Gibbs sampler (Metropolis et al. 1953), and the Metropolis-Hastings algorithm (Hastings 1970). 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 that are identically (but not independently) distributed according to the posterior distribution. These samples can then be used for Monte Carlo integration as described above; hence, these algorithms are known as MCMC methods (see Tierney 1996 for regularity conditions for using eq. [7] with MCMC draws). 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. These issues are addressed in § A2.

The algorithms used to fit the models described in § 3 rely on the method of data augmentation. The term "data augmentation" originated with computational methods designed to handle missing data, but the method is really quite general and often useful when there is no missing data per se. In particular, for Monte Carlo integration we aim to obtain a sample from the posterior distribution, $p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{I})$.

In some cases, we can augment the model to $p(\boldsymbol{\theta}, \boldsymbol{X} | \boldsymbol{Y}, \boldsymbol{I})$, where $\boldsymbol{X}$ may be missing data or any other unobserved quantity (e.g., counts due to background). With the judicial choice of $\boldsymbol{X}$, it may be much easier to obtain a sample from $p(\boldsymbol{\theta}, \boldsymbol{X} | \boldsymbol{Y}, \boldsymbol{I})$ than directly from $p(\boldsymbol{\theta} | \boldsymbol{Y}, \boldsymbol{I})$. Once we have a sample from $p(\boldsymbol{\theta}, \boldsymbol{X} | \boldsymbol{Y}, \boldsymbol{I})$, we simply discard the sample of $\boldsymbol{X}$ to obtain a sample from $p(\boldsymbol{\theta} | \boldsymbol{Y}, \boldsymbol{I})$. In Appendix B, § B1, we describe how this method can be used for fitting the models described in § 3.

3. FITTING HIGH-RESOLUTION LOW-COUNT SPECTRA

3.1. *Model Overview*

In this section we describe a new class of (statistical) structured models, which simultaneously describes high-resolution source spectra using Gaussian line profiles and a GLM for the continuum and accounts for background contamination of the image, instrument response, and absorption. The model may easily be generalized to account for different line profiles such as the Lorentzian distribution (e.g., Meng & van Dyk 1999). The statistical model is designed to summarize the distribution of photon energies arriving at a detector, which are recorded as counts in a number of energy channels (e.g., as many as 4096 on *Chandra*/ACIS). Newly developed detectors have much higher resolution than their predecessors and thus smaller expected counts per bin. Independent Poisson distributions are therefore more appropriate for the counts than the commonly used Gaussian approximation (e.g., χ^2 fitting). We parameterize the intensity in bin $j \in \mathscr{J} = \{1, \ldots, J\}$ as the sum of a continuum term and K Gaussian lines. That is, the expected true counts per bin for a "perfect" instrument with effective area everywhere equal to the maximum possible effective area[10] are

$$\text{model intensity} = [\text{continuum} + \text{lines}] \text{ absorption},$$

for each energy bin, or more formally,

$$\lambda_j(\boldsymbol{\theta}) = [dE_j \, f(\boldsymbol{\theta}^C, E_j) + \sum_{k=1}^{K} \tilde{\lambda}^k p_j(\mu^k, v^k)] u(\boldsymbol{\theta}^A, E_j) \tag{8}$$

for $j \in \mathscr{J}$, where dE_j is the known width of bin j; $f(\boldsymbol{\theta}^C, E_j)$ is the expected number of counts per keV per maximum effectivearea from the continuum and is a function of the continuum parameter, $\boldsymbol{\theta}^C$; E_j is the known mean energy in bin j; $\tilde{\lambda}^k$ are the expected counts per maximum effective area from line k; $p_j(\mu^k, v^k)$ is the probability that a Gaussian random variable with mean μ^k and variance v^k falls in bin j; and $u(\boldsymbol{\theta}^A, E_j)$ is the probability that a photon in bin j is *not* absorbed. Specific forms for the continuum and absorption terms are discussed below in §§ 3.2 and 3.4, respectively. The superscripts on the model parameters ($\boldsymbol{\theta}$) are mnemonic and represent absorption (A), background (B), continuum (C), and the lines $k = 1, \ldots, K$. The collection of parameters, $\boldsymbol{\theta}^C$, $\boldsymbol{\theta}^k = (\tilde{\lambda}^k, \mu^k, v^k)$ for $k \in \mathscr{K} = \{1, \ldots, K\}$, and $\boldsymbol{\theta}^A$ (along with $\boldsymbol{\theta}^B$ defined below), are represented by $\boldsymbol{\theta}$. An artificial example with power-law continuum, two spectral lines, and no absorption appears in the first plot of Figure 2.

Since data collection is degraded by effective area, instrument response, and background contamination (see Fig. 2), we model the observed counts as independent Poisson vari-

[10] We use the maximum value of the effective area over the spectral energy range of interest in this stage of the analysis. This is only a matter of convenience, and the full effective area variations are included in eq. (9).

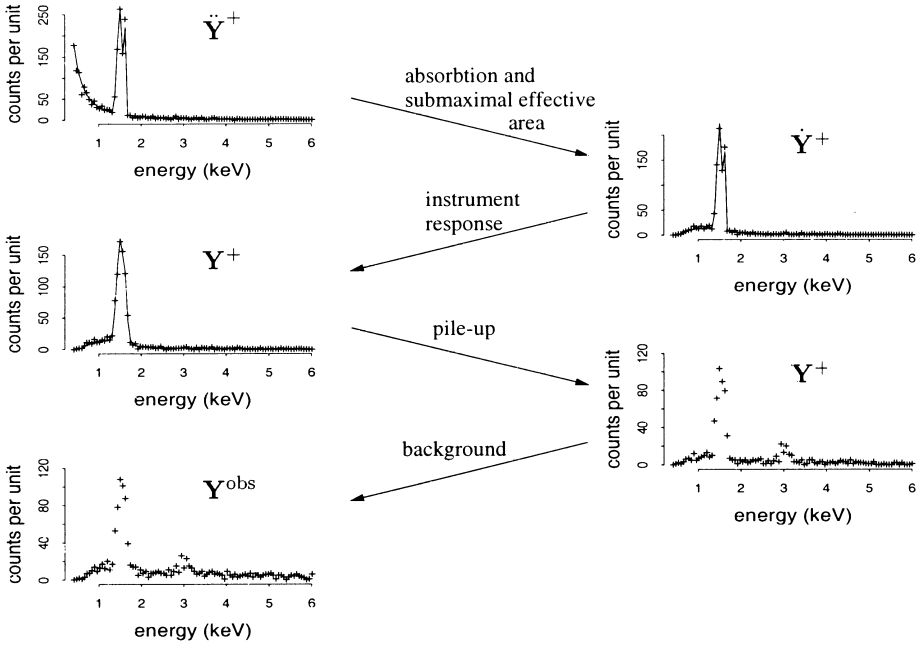

FIG. 2.—Degradation of counts. The figure illustrates the various physical processes that significantly degrade the source model and result in the observed PHA counts. In particular, an artificial data set is used to illustrate (1) the absorption of (mostly low-energy) counts, (2) the blurring of spectral features due to instrument response, (3) the shadows caused by pile-up, and (4) the masking of features due to background. The solid lines represent the assumed model (in the first three plots) and the plus sign 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 pile-up are included here for the sake of completeness; we do not deal with this aspect of the analysis in this paper. The symbols in the upper right of each plot are defined in § B.1.

ables with intensity

observed
intensity

$$= \begin{matrix}\text{instrument}\\ \text{response}\end{matrix}\left(\begin{matrix}\text{model}\\ \text{intensity}\end{matrix}\times\begin{matrix}\text{effective}\\ \text{area}\end{matrix}\right)+\text{background}\,,$$

for each energy channel, or more formally,

$$\xi_l(\boldsymbol{\theta}) = \sum_{j\in\mathscr{J}} M_{lj}\lambda_j(\boldsymbol{\theta})d_j + \lambda_l^B(\boldsymbol{\theta}^B) \quad (9)$$

for $l \in \mathscr{L}$. $\mathscr{L} = \{1, \ldots, L\}$, where the $L \times J$ matrix $\boldsymbol{M} = \{M_{lj}\}$ represents instrument response: a photon arriving in bin j has probability M_{lj} of being detected in observed bin l; $\boldsymbol{d} = (d_1, \ldots, d_J)$ is the effective area of bin j, normalized so that $\max_{j\in\mathscr{J}} d_j = 1$; and $\lambda_l^B(\boldsymbol{\theta}^B)$ is the expected counts due to the background that may be known from calibration in space or parameterized in terms of $\boldsymbol{\theta}^B$. As with $\mathscr{J}$, $\mathscr{L}$ may be any subset of detector bins. In general, the counts are also degraded by pile-up (e.g., Knoll 1989; see also Fig. 2). Here we ignore pile-up, which is justifiable for low-intensity or spatially diffuse sources (see the discussion in § 5).

In the next several sections we describe the stochastic models for each of the sources of photons in turn. This includes both likelihoods that describe the sampling distribution of the data (parameterized by $\boldsymbol{\theta}$) and prior distributions that allow us to incorporate scientific information about the likely parameter values. As described below, the prior distributions are parameterized using the hyperparameter ϕ.

3.2. *The Continuum*

The photon counts due to the continuum are modeled via a GLM (McCullagh & Nelder 1989), specifically a loglinear model. That is, the log expected counts per keV per maximum effective area are assumed to be a linear function of a set of independent variables, X_j^C, which in turn are typically functions of E_j, hence the notation $f(\boldsymbol{\theta}^C, E_j)$. In particular, we model the counts in bin j due to the continuum, denoted Y_j^C, as

$$Y_j^C \overset{d}{\sim} \text{Poisson}\,(dE_j\, f(\boldsymbol{\theta}^C, E_j))\,, \quad (10)$$

i.e.,[11]

$$p(Y_j^C \mid \boldsymbol{\theta}^C) = \frac{e^{-\lambda_j^C}(\lambda_j^C)^{Y_j^C}}{(Y_j^C)!} \quad \text{with } \lambda_j^C = f(\boldsymbol{\theta}^C, E_j)dE_j\,, \quad (11)$$

independently for $j \in \mathscr{J}$. Here $\log f(\boldsymbol{\theta}^C, E_j) = \boldsymbol{X}_j^C\boldsymbol{\theta}^C$, with $\boldsymbol{\theta}^C$ a $(P^C \times 1)$ vector parameter, $\boldsymbol{X}_j^C$ a $(1 \times P^C)$ vector of independent variables, and P^C the number of parameters in the continuum model. Note that we are explicitly using a Poisson process for the photon counts as opposed to an often poor Gaussian approximation.

The flexible framework of the GLM allows us to adjust the expected counts in bin j for any set of independent variables. For example, several standard continuum models are easily available. In particular, a power-law model is obtained by setting $\boldsymbol{X}_j = [1, \log(E_j)]$ for $j \in \mathscr{J}$ so that

$$f(\boldsymbol{\theta}^C, E_j) = e^{\theta_1^C}E_j^{\theta_2^C} = \alpha E_j^{-\beta} \quad \text{for } j \in \mathscr{J}\,, \quad (12)$$

where the familiar form of the power-law model in the last expression is obtained by identifying (α, β) with $(e^{\theta_1 C}, -\theta_2^C)$. It is easy to generalize this to handle more complicated models. A break in the power law (i.e., a change point) can be added at E_* by setting $\boldsymbol{X}_j = [1, \log(E_j), \log(E_j/E_*)I\{E_j > E_*\}]$, so that

$$f(\boldsymbol{\theta}^C, E_j) = \begin{cases} e^{\theta_1^C}E_j^{\theta_2^C} & \text{for } E_j \le E_* \\ e^{\theta_1^C}E_j^{\theta_2^C+\theta_3^C}E_*^{-\theta_3^C} & \text{for } E_j > E_* \end{cases} \quad \text{for } j \in \mathscr{J}\,. \quad (13)$$

The factor $E_*^{-\theta_3^C}$ ensures that $f(\boldsymbol{\theta}^C, E_j)$ is continuous at $E_j = E_*$. As a final example, we obtain an exponential continuum representing bremsstrahlung emission by setting $\boldsymbol{X}_j^C = (1, -E_j)$ so

$$f(\boldsymbol{\theta}^C, E_j) = e^{\theta_1^C}e^{-E_j\theta_2^C} = \frac{\alpha}{\sqrt{T}}\,e^{-E_j/kT} \quad \text{for } j \in \mathscr{J}\,, \quad (14)$$

where $(\alpha, T) = [e^{\theta_1^C}(k\theta_2^C)^{-1/2}, 1/k\theta_2^C]$.

It is convenient to assume that the prior distribution on $\boldsymbol{\theta}^C$ is multivariate Gaussian with a diagonal variance matrix. That is, $\theta_p^C \overset{d}{\sim} N(\mu_p^C, \upsilon_p^C)$ with $\phi^C = \{(\mu_p^C, \upsilon_p^C)$ for $p = 1, \ldots, P^C\}$. The hyperparameter, ϕ^C, is set by the user where μ_p^C is a "best guess" of θ_p^C and υ_p^C is a measure (in squared standard deviations) of the error of this "best guess." Large values of υ_p^C reflect little prior information for θ_p^C.

3.3. *Emission Lines*

Lines reflect deviation in the smooth spectrum due to the continuum because of photon emissions from various ions present in the source. In particular, we model the energies of photons due to line $k \in \mathscr{K}$, denoted Y_i^k, as

$$Y_i^k \overset{d}{\sim} N(\mu^k, \upsilon^k)\,, \quad (15)$$

i.e.,

$$p(Y_i^k \mid \mu^k, \upsilon^k) = \frac{1}{\sqrt{2\pi\upsilon^k}}\,e^{-(Y_i^k-\mu^k)^2/2\upsilon^k}\,, \quad (16)$$

independently for $i = 1, \ldots, N^k$. Equation (15) represents a line with intensity normalized to 1. The total line counts for a perfect instrument (i.e., with effective area everywhere equal to its maximum possible value) are denoted $(N^1, \ldots, N^K)$ and assumed to be independent Poisson random variables,

$$N^k \overset{d}{\sim} \text{Poisson}\,(\tilde{\lambda}^k) \quad \text{independently for } k \in \mathscr{K}\,. \quad (17)$$

Proper prior information for the lines and the continuum is important for a reasonable fit when the spectral model includes emission lines. In particular, prior information is especially important for relatively weak lines, since it is difficult to distinguish a weak line from a chance fluctuation in the continuum. Luckily, such prior information is often scientifically forthcoming in the form of knowledge (e.g., laboratory measurements and physics theory) of probable sizes and locations of the various lines. We begin with the line location and width (actually the variance), (μ^k, υ^k), for which priors[12] are assigned independently for each line

[11] Here and in the remainder of the paper we suppress the conditioning on the initial information, ***I***. That is, it should be understood that all distributions implicitly condition on ***I***.

[12] We choose this prior distribution partially because it is the so-called *conjugate* prior distribution, i.e., the resulting posterior distribution is from the same family as the prior distribution (e.g., Gaussian with updated parameters). This property significantly simplifies model fitting with no cost in terms of the accuracy of parameter estimation.

$k \in \mathscr{K}$,

$$v^k \overset{d}{\sim} \frac{v_0^k v_0^k}{\chi^2_{v_0}}, \quad \text{and} \quad p(\mu^k \mid v^k) = \mathrm{N}\left(\mu_0^k, \frac{v^k}{\kappa_0^k}\right), \tag{18}$$

where $\chi^2_{v_0}$ is a variable that follows the χ^2 distribution with v_0 degrees of freedom. We interpret the hyperparameter, $\phi^k = (\mu_0^k, v_0^k, \kappa_0^k, v_0^k)$, using the mean and variance of the distributions in equation (18). For example,

$$E(v^k \mid \phi^k) = \frac{v_0^k v_0^k}{v_0^k - 2} \quad \text{for } v_0^k > 2 \tag{19}$$

and

$$\text{var}\,(v^k \mid \phi^k) = 2(v_0^k v_0^k)^2/(v_0^k - 2)^2(v_0^k - 4) \quad \text{for } v_0^k > 4\,. \tag{20}$$

(Recall that the units here are keV for means and keV^2 for variances.) Thus, the mean and variance of the prior for v^k may be tuned using v_0^k and v_0^k; a small value of v_0^k results in a wide, relatively noninformative prior. Since the data are discrete, a priori we cannot allow the standard deviation of the line to become too small (say below the PHA bin width of the bin that contains the center of μ^k) since there is not information in the data about the width of a line that is narrower than one PHA channel. This is accomplished by truncating the prior distribution of v^k. For the prior on μ^k, the mean and variance are given by μ_0^k and κ_0^k; μ_0^k is the most probable location of the kth line, and κ_0^k calibrates the uncertainty in the location of the kth line relative to the width of the line.

An alternative interpretation of the priors is in terms of additional hypothetical photons. Heuristically, the effect of the prior on μ^k if v^k were known would be the same as κ_0^k photons all known to be from line k and equal to μ_0^k. Likewise, the effect of the prior on v^k is the same as adding v_0^k photons with average squared deviation from the center of the line equal to v_0^k.

We now turn to the prior distribution on $\tilde{\lambda}^k$ and set (independently) $\tilde{\lambda}^k \overset{d}{\sim} \gamma(\phi_1^k, \phi_2^{\tilde{\lambda}})$,[13] which has mean $\phi_1^k/\phi_2^{\tilde{\lambda}}$ and variance $\phi_1^k/(\phi_2^{\tilde{\lambda}})^2$. Roughly speaking, the γ prior contains the same information as $\phi_2^{\tilde{\lambda}}$ Poisson observations (with exposure equal to the source exposure) with a total of $\phi_1^k - 1$ counts. Since the data consist of a single observation (for each bin), $\phi_2^{\tilde{\lambda}}$ can be interpreted as the weight put on the prior relative to the data; $\phi_2^{\tilde{\lambda}} = 1$ induces a prior as influential as the data in the absence of absorption, blurring, background, and lines. Thus, values of $\phi_2^{\tilde{\lambda}} \ll 1$ are typically recommended for noninformative priors. The hyperparameters ϕ_1^k can be interpreted as the prior relative sizes of the lines. That is, $\phi_1^k/\sum_{\tilde{k}} \phi_1^{\tilde{k}}$ is the prior proportion of line photons from line k. We define the hyperparameter for line k as $\phi^k = [\mu_0^k, \kappa_0^k, v^k, (\sigma_0^2)^k, \phi_1^k, \phi_2^{\tilde{\lambda}}]$; the last element is not indexed by k since it is constant for $k \in \mathscr{K}$.

3.4. *Absorption and Correction for Effective Area*

From the viewpoint of our statistical algorithm, both the telescope effective area and astrophysical absorption (e.g., absorption due to the ISM) are handled in the same way. These two processes act independently on individual photons and randomly prevent an (energy-dependent) proportion of photons from being observed. The only essential statistical difference is that only the absorption process has unknown parameters. In particular, we suppose that the probability that a photon is *not* absorbed (statistically speaking "censored") by either of these two processes is

$$d_j u(\boldsymbol{\theta}^A, E_j)\,, \quad \text{where } \log u(\boldsymbol{\theta}^A, E_j) = \mathbf{X}_j^A \boldsymbol{\theta}^A \quad \text{for } j \in \mathscr{J}\,, \tag{21}$$

where $\boldsymbol{\theta}^A$ is a $(P^A \times 1)$ parameter, $\boldsymbol{X}_j^A$ is a $(1 \times P^A)$ vector of independent variables, and P^A is the number of parameters in the absorption model, $u(\boldsymbol{\theta}^A, E_j)$. As an example, simple exponential absorption can be written in this linear form with θ^A a scalar and $\boldsymbol{X}_j^A = -1/E_j$, i.e., $u(\theta^A, E_j) = e^{-\theta^A/E_j}$. For more complicated absorption models, $\boldsymbol{X}_j^A$ typically consists of a tabulated absorption function.

The prior for $\boldsymbol{\theta}^A$ is multivariate Gaussian, $\theta_p^A \overset{d}{\sim} N(\mu_p^A, v_p^A)$, independently for $p = 1, \ldots, P^A$. The prior is interpreted similarly to that for the continuum parameter $\boldsymbol{\theta}^C$. We, however, truncate this prior to ensure $\exp\{\boldsymbol{X}_j^A \boldsymbol{\theta}^A\} < 1$ for each j, to ensure that the proportion of photons not absorbed is less than 1. With appropriately chosen $\boldsymbol{X}_j^A$, this can be accomplished by assuming that each component of $\boldsymbol{\theta}^A$ is negative.

3.5. *Background*

We assume the availability of a separate observation containing background counts that can be used to model the background spectrum and correct the source spectrum. Rather than simply subtracting off (a scalar multiple of) the background counts, however, we account for the variation due to the Poisson character of the counts. In particular, we suppose that the background count in PHA channel l is

$$Y_l^B \overset{d}{\sim} \text{Poisson}\,(\theta_l^B) \quad \text{independently for } l \in \mathscr{L}\,, \tag{22}$$

where the unobserved quantity, Y_l^B, is the counts in PHA channel l that are due to background. We parameterize the prior for θ_l^B as $\gamma(\phi_{l,1}^B, \phi_{l,2}^B)$, which we expect to be informative based on a pure background exposure. In particular, a reasonable prior based on $Y_l^{\text{obs},B}$ background counts would be $\gamma(Y_l^{\text{obs},B} + 1, \tau)$, where τ is the background exposure time and area relative to the source exposure time and area.

In an extreme case, when the background is very well determined, e.g., via a very long exposure, we may fix $\theta_l^B = Y_l^{\text{obs},B}/\tau$ and discard the prior distribution; here we are effectively setting the prior variance to 0. Note this is not equivalent to subtracting off the background because we still allow for Poisson variability in the background counts that contaminate the source counts. An alternative strategy is to fit a parameterized model to the background. For example, we might assume $Y_l^B \overset{d}{\sim}$ Poisson $(\lambda_l^B(\boldsymbol{\theta}^B, \tilde{E}_l))$, where $\log \lambda_l^B(\boldsymbol{\theta}^B, \tilde{E}_l) = \boldsymbol{X}_l^B \boldsymbol{\theta}^B$ for $l \in \mathscr{L}$ with $\boldsymbol{X}_l^B$ a row vector of independent variables depending on the energy of PHA channel l, $\tilde{E}_l$. This allows the background counts to be modeled as a power law, broken power law, or any other loglinear model.

4. APPLICATIONS

In this section we illustrate our methods and algorithms using two data sets. We first analyze *ASCA*/SIS data of high-redshift ($z = 3.384$) quasar S5 0014+813 (Elvis et al. 1994) to illustrate the various summaries available in a relatively straightforward MCMC analysis. The second analysis involves an extremely low count stellar coronal source (α TrA) and illustrates the power of Bayesian

[13] We choose a γ prior partially because it is conjugate to the Poisson distribution.

TABLE 2

FITTED VALUES AND CREDIBLE REGIONS FOR THE QUASAR DATA

Parameter	Estimate	95% Interval
Power law	2.23	(1.90, 2.58)
Normalization.....	2.47E−3	(1.37E−3, 4.55E−3)
Absorption	2.05	(1.43, 2.71)
Equivalent width (keV).....	0.0282	(0.0023, 0.0788)

methods to combine information from various sources and quantify the weak information available in this data.

4.1. *Quasar S5 0014+813*

A typical quasar X-ray spectrum can be described by an absorbed power law. A fluorescent iron line (Fe Kα) emitted at energy between $\sim$6.4 and 6.8 keV if detected can be a signature of a reflection component and its ionization state (George et al. 2000). Quasar S5 0014+813 (Kühr et al. 1981) at redshift $z = 3.384$ is among the highest X-ray flux quasars known with $z \sim 3$. S5 0014+813 was observed with *ASCA* on 1993 October 29 with an exposure time of 22.8 ks in the SIS0 detector (Elvis et al. 1994). Here we apply our model to this data to illustrate the method and look for signatures of the iron emission line.

The spectral data were extracted with the standard screening criteria (Elvis et al. 1994), and standard response matrices were used.[14] We use all of the original 512 PHA instrument channels except the unreliable channels below $\sim$0.5 keV and above $\sim$10 keV. In addition, we do not group any channels. (Channels are usually grouped in order to justify the use of the default χ^2 techniques with their Gaussian assumptions.) As is allowed with a Poisson model, we instead use only the original PHA bins. The source model included the exponential shape of Galactic absorption (see § 3.4) and a power-law continuum (i.e., eq. [12]) with a narrow emission line at 1.45 keV (observed frame; $\sim$6.7 keV rest frame). We accounted for background using a background Poisson process with intensity equal to the (rescaled) background counts in each PHA channel. Flat priors were used on all model parameters.

To estimate the four model parameters (i.e., the power-law, normalization, and exponential absorption parameters and the equivalent width[15] of the line), a sample from their posterior distribution was obtained by running three MCMC chains using dispersed starting values. The chains showed excellent mixing (as measured with $\hat{R}^{1/2}$; see § A2) after 2000 draws. In the Monte Carlo evaluation, the second half of each of the chains was used along with an additional run of 2000 draws from each chain, for a total of 9000 draws.

Summaries of the model fit appear in Table 2 and Figures 3 and 4. The parameter estimates are posterior means computed using a transformation that makes the marginal posterior distributions more symmetric and hence the posterior mean a more informative summary [i.e., ln(normalization) and sqrt(equivalent width)]. In particular, if we represent the draws of the normalization parameter as $\{\theta_{[t]}, t = 1, \ldots, 9000\}$, the point estimate of this parameter was computed as

$$\exp\left\{\frac{1}{9000}\sum_{t=1}^{9000}\ln(\theta_{[t]})\right\} = \sqrt[9000]{\prod_{t=1}^{9000}\theta_{[t]}}, \qquad (23)$$

the geometric mean. The credible intervals are computed using the 0.025 and 0.975 quantiles of the draws and are invariant to (monotonic) transformations. Pairwise credible regions appear in Figure 3. The scatter plots illustrate the regions of highest posterior probability by plotting the Monte Carlo draws: Pr ($\boldsymbol{\theta} \in R$) is approximately equal to the proportion of points in that region. The gray-scale images give Monte Carlo estimates of the (*darker*) 50% and (*lighter*) 90% marginal posterior regions. The grainy character of the images is due to the Monte Carlo approximation. Even with this relatively large data set and with the use of transformations, the non-Gaussian character of the posterior is evident. We expect that higher dimensional marginal posterior distributions are even less Gaussian in character. Figure 4 compares the fitted source model corrected for effective area and absorption with the PHA counts and illustrates the residual for each PHA channel and the stability of the estimated continuum.

4.2. *Hybrid-Chromosphere Supergiant Star α TrA*

Unlike the simple power-law spectrum of the quasar in the previous section, stellar coronal spectra are complicated by a bremsstrahlung continuum and the presence of numerous emission lines. Such complex spectra are much more difficult to model, and in addition, the intensity of the bremsstrahlung continuum drops exponentially at high energies, resulting in very few counts. Analyzing such spectra is however crucial to the understanding of coronal structure, mechanisms of coronal heating, etc. A case in point is the corona of the hybrid supergiant star α TrA (HD 150798, K4II, $B-V = 1.44$, $V = 1^{\rm m}.92$), which shows evidence of both strong magnetic activity as indicated by X-ray emission (Brown et al. 1991; Kashyap et al. 1994) and stellar outflow seen in absorption profiles (Hartmann, Dupree, & Raymond 1981). X-ray observations with the *ROSAT*/PSPC (Kashyap et al. 1994) indicate that its corona is dominated by transient, unstable plasma that is confined by magnetic loops that are closed on short length scales (Rosner et al. 1994). Constraining the maximum temperatures present in the corona is therefore of primary importance. Here we use data obtained with *ASCA*, at higher energies than *ROSAT*, to model the spectrum. The low number of counts detected at high energies makes this spectrum difficult to analyze by traditional means, and we must bring to bear the full power of a hierarchical Bayesian analysis in order to constrain the maximum temperatures present in the corona.

Supergiant star α TrA was observed with *ASCA* in 1995 March for $\approx$34 ks. During this observation, the source exhibited no flares. The count rate was steady and corresponded to the quiescent state identified with *ROSAT*.

We model the high-energy region of the *ASCA* spectrum (2.5–7.5 keV) as a combination of a bremsstrahlung continuum

$$\frac{\text{Norm}}{\sqrt{T}}\, e^{-E/k_{\rm B}T}, \qquad (24)$$

where T is the electron temperature, E is the energy in keV, $k_{\rm B}$ is the Boltzmann constant, and Norm is a normalization;

[14] ftp://legacy.gsfc.nasa.gov/caldb/data/asca/sis/cpf/94nov9.

[15] The equivalent width is defined as $\tilde{\lambda}^k/f(\theta^C, \mu^k)$.

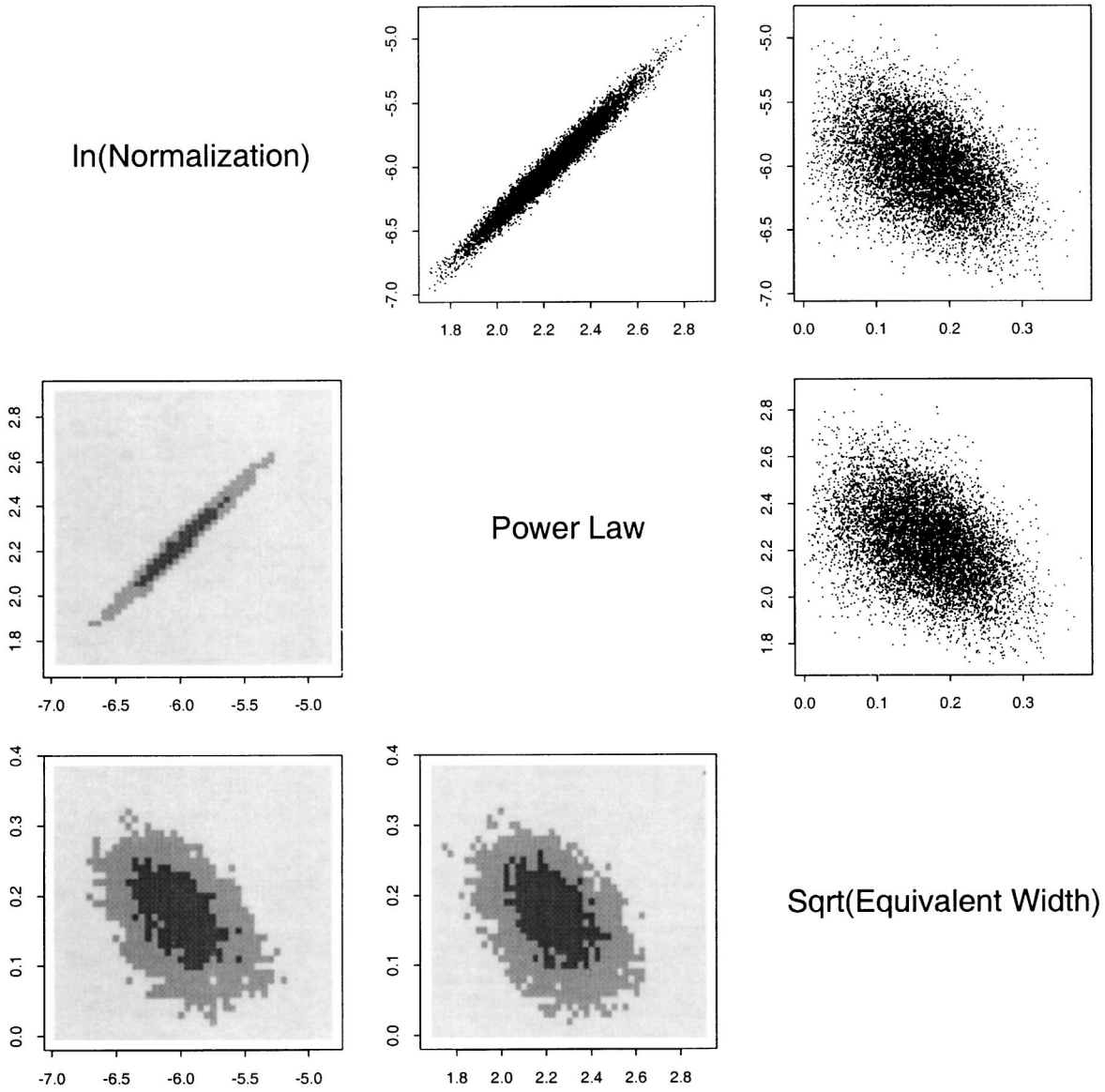

FIG. 3.—Posterior distributions of pairs of parameters obtained via MCMC. The plots show pairwise marginal posterior distributions for the model parameters in the analysis of quasar S5 0014+813. The plots in the upper right are scatter plots of the Monte Carlo draws and indicate areas of highest posterior probability. The plots in the lower left are gray-scale images of the Monte Carlo approximations to 50% (*darker*) and 90% (*lighter*) credible regions. The text along the diagonal labels the axes for each of the plots.

and a number (~ 10) of narrow emission lines located at the positions of known strong lines whose widths and locations[16] are fixed, but intensities are allowed to vary. The units of Norm are counts $\mathrm{keV}^{-1}\ \mathrm{cm}^{-2}\ \mathrm{s}^{-1}$. Because of the low counts, we fit a power law to the background.

We apply the above model to SIS0 data (~ 28 ks) and the combined data from the 2 GIS detectors (~ 33 ks in each). Note that the very low counts present in these data (~ 150 counts in SIS0, ~ 300 in GIS) preclude any "traditional" analysis: it is only by using the full Bayesian machinery that we can derive useful results from such data.

We carry out the analysis in two steps:

1. Choose highly noninformative priors on the parameters to analyze SIS0 data: (1) on normalization, $p(\mathrm{Norm}/T^{1/2})$ is such that it lies in between 10^{-9} and 10^{-1} with a 90% probability; (2) on temperature, $p(1/kT)$ is such that it is always positive and also is nearly flat in the temperature range of interest; and (3) on line intensity, $p(\tilde{\lambda}^k)$ is such that a priori all the lines have the same intensity and the maximum total counts due to lines are 100. (The total of source + background counts for the SIS0 observation is only 154; atomic emission-line models indicate that for the temperature and energy range of interest, 100 corresponds to the maximum possible contribution to the spectrum from lines.) We choose Gaussian forms for the first prior distribution and a gamma prior for the last; these priors are illustrated as solid lines in Figure 5. Thus,

$$p\left(\ln\left(\frac{\mathrm{Norm}}{\sqrt{T}}\right)\right) = N(\mu = -9.21,\ \sigma = 5.58)\,, \qquad (25)$$

$$p\left(\frac{1}{kT}\right) = N(\mu = 0.95,\ \sigma = 2.5)\,, \qquad (26)$$

[16] Line location can be known only to the resolution of the instrument, and hence each of the model lines represents the sum of a large number of lines within the resolution element; we find that we only exclude less than 5% of the line flux in the energy range considered by this approximation.

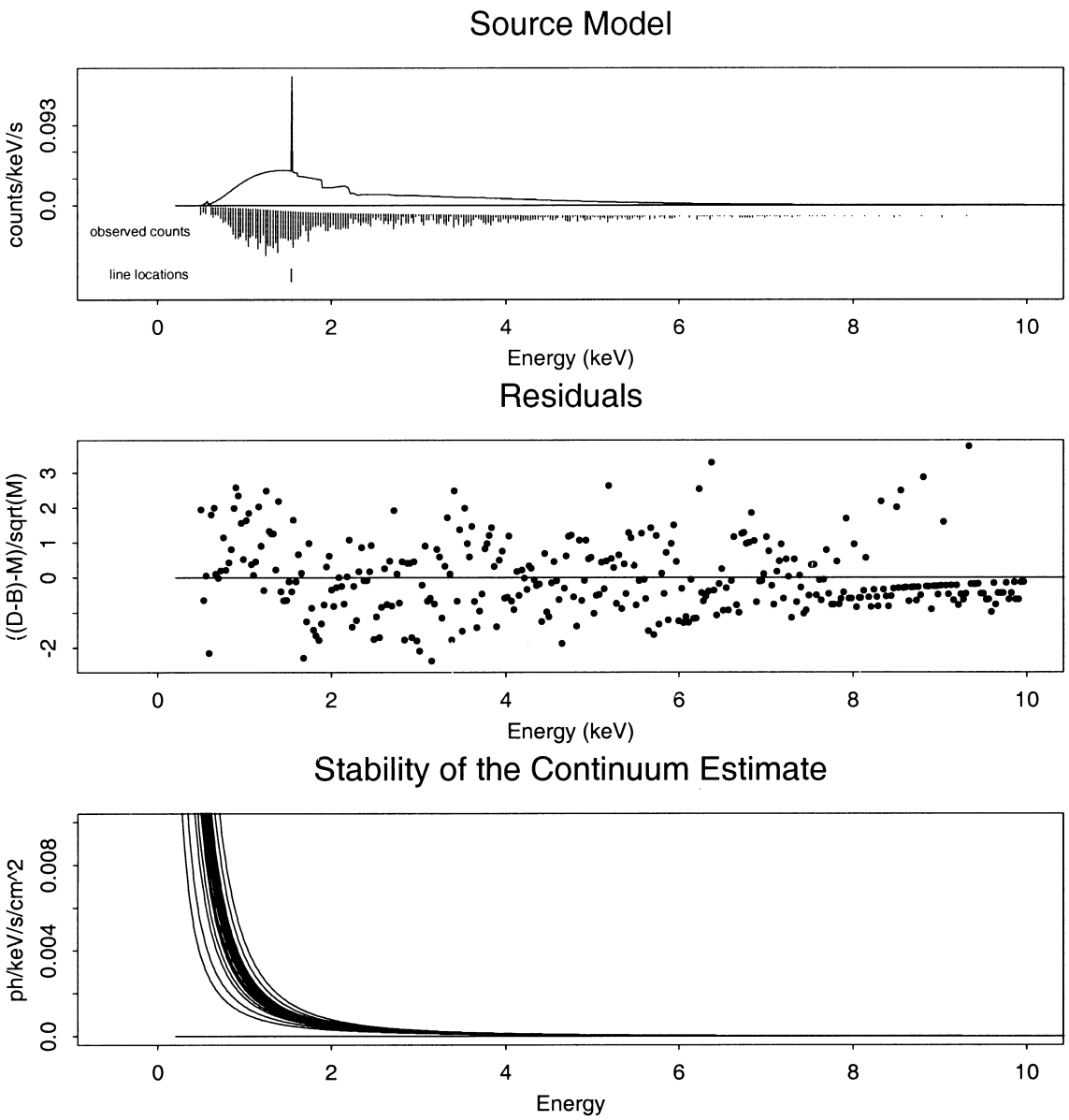

FIG. 4.—Quasar S5 0014+813 model fit. This plot gives an overview of the fitted model. The first panel compares the fitted source model (corrected for effective area and absorption, but not the instrument's photon response matrix) with the observed PHA counts. The second panel gives the residual for each PHA channel, which were computed by subtracting off background and standardizing by the model standard deviation. The final plot illustrates the stability of the continuum estimate.

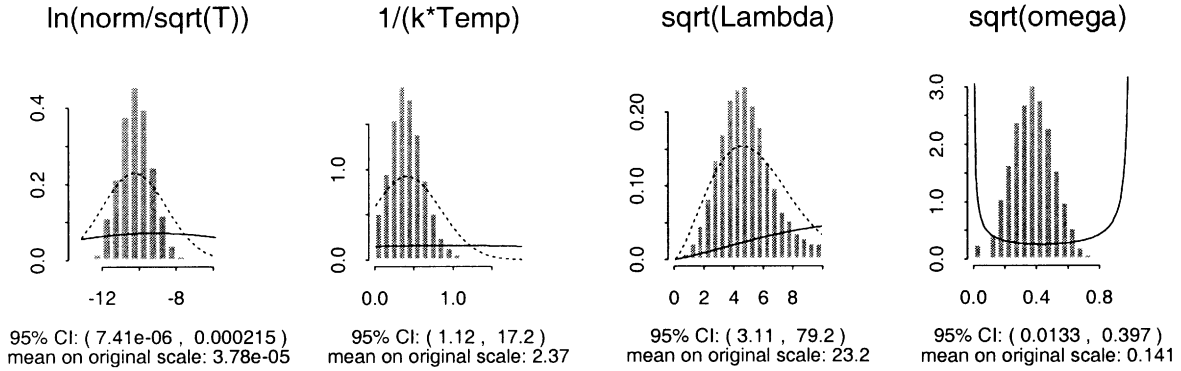

FIG. 5.—Using *ASCA*/SIS to compute the prior. Here Lambda is the expected model counts from lines and omega is the ratio of the total counts in the lines to the total counts in the spectrum (correcting for the effects of absorption and instrument response). Transformations of the parameters that produce a distribution near to the Gaussian are displayed. The listed means and credible intervals, however, refer to the original parameters. The solid lines in these plots represent the relatively diffuse priors used to compute the posterior distributions represented by the histograms based on the SIS0 observation. These posterior distributions were in turn used to choose the priors for the GIS data after some dispersion was added, as represented by the dotted curves. We do not specify the prior for the proportion of source photons from the lines, Ω, but rather this prior is implied by the other priors. The solid line is an approximation based on sampling from the prior and distributing the SIS0 counts to the continuum and lines after correcting for background.

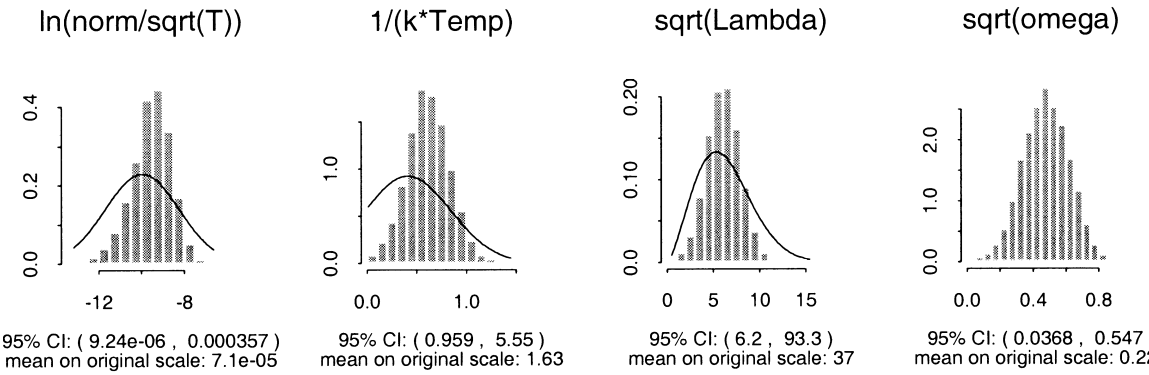

FIG. 6.—Some marginal posterior distributions. Using the priors computed with SIS0 data (*solid lines*) we fitted the source model to the GIS data. The resulting marginal posterior distributions are illustrated here using normalizing transformations. The estimates and credible intervals are on the original scales.

and

$$p(\tilde{\lambda}^k) = \gamma(\phi_1^k = 0.11,\ \phi_2^{\tilde{\lambda}} = 0.0033) \quad \text{for } k = 1,\ \ldots,\ 10\ . \tag{27}$$

2. Use the posterior distribution resulting from the above step to define more informative priors to analyze GIS data. These priors also correct for the difference in exposure time and average effective area between the SIS0 and the GIS

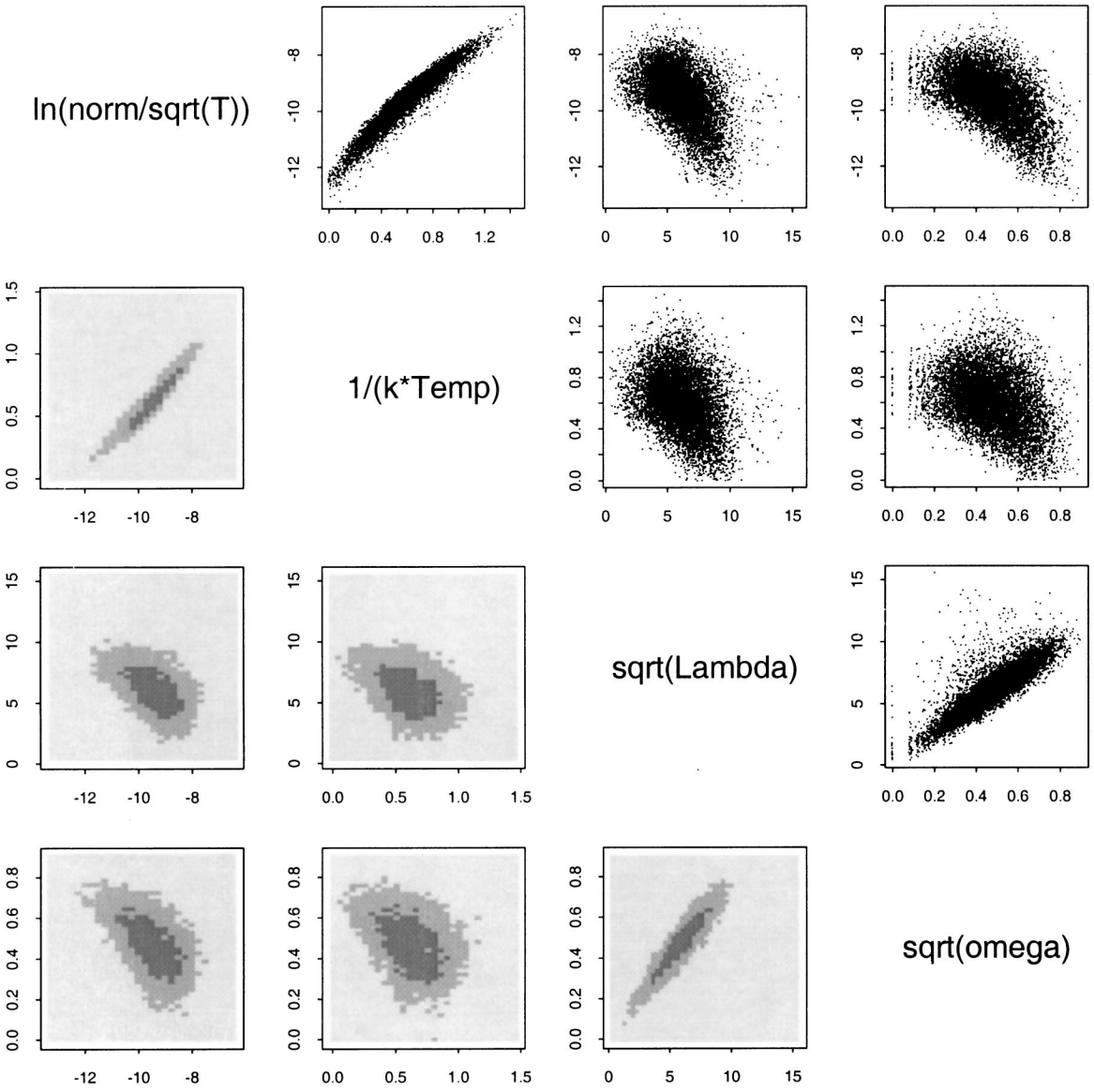

FIG. 7.—Some bivariate marginal posterior distributions. These plots are as described in Fig. 3 and illustrate pairwise credible regions for the various model parameters. Again the text along the diagonal labels the axes for each of the plots.

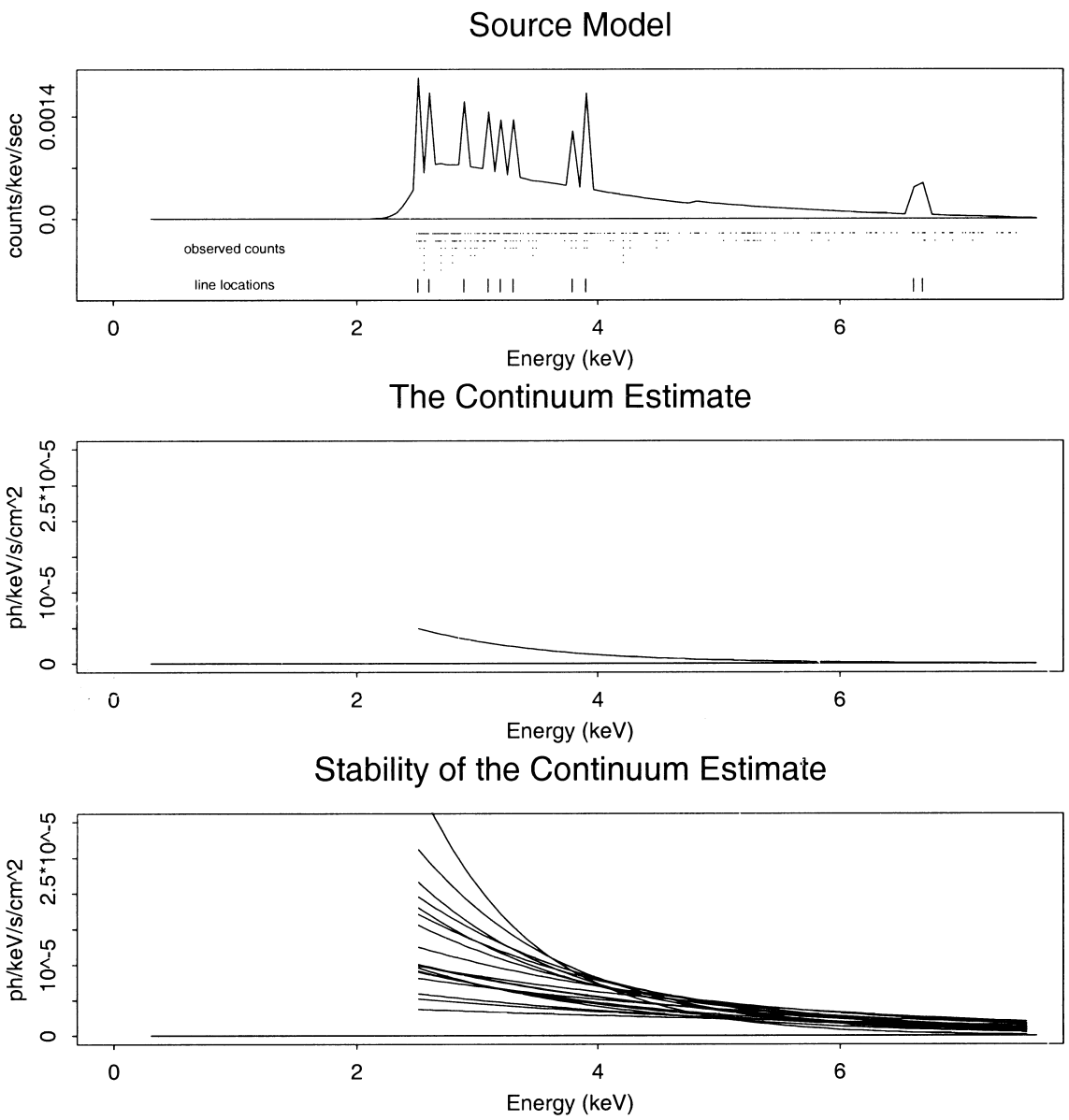

FIG. 8.—Model fit. These plots are as described in Fig. 4, except the plot of residuals is replaced by a plot of the estimated continuum. Note the instability of the continuum due to the low counts.

data. The posterior variances from the initial analysis were increased somewhat when computing the priors for the second analysis. These priors are illustrated as solid lines in Figure 6. Thus,

$$p\left(\ln\left(\frac{\text{Norm}}{\sqrt{T}}\right)\right) = N(\mu = -9.97, \sigma = 1.74)\,, \qquad (28)$$

$$p\left(\frac{1}{kT}\right) = N(\mu = 0.41, \sigma = 0.43)\,, \qquad (29)$$

and

$$p(\tilde{\lambda}^k) = \gamma(\phi_1^k = 0.12, \phi_2^{\tilde{\lambda}} = 0.025) \quad \text{for } k = 1, \ldots, K\,. \qquad (30)$$

We ran three Markov Chains in each analysis to obtain draws from the posterior distribution of $[\ln(\text{Norm}/T^{1/2}), 1/kT, \tilde{\lambda}^1, \ldots, \tilde{\lambda}^{10}]$. In both analyses there was excellent mixing after 6000 draws, and we used the second half of each chain for a total of 9000 Monte Carlo draws.

The results of the analysis are shown in Figures 5–8. (In the figures the parameter Ω refers to the proportion of source photons from the lines and $\lambda = \sum_{k=1}^{10} \tilde{\lambda}^k$.) We find that the plasma temperature is $\sim 19^{<64}_{>11} \times 10^6$ K (Fig. 4). Such a large value (cf. $\sim 2 \times 10^6$ K in the quiet solar corona) clearly lends credence to the idea that the corona on α TrA is dominated even in quiescence by flarelike events.

As a by-product of our analysis, we also obtain the flux in the modeled lines relative to the continuum. In principle, this allows us to constrain the metallicity *for the first time* in the corona of α TrA by comparing the observed ratio of the line and continuum fluxes[17] with that derived from thermal emission models computed over the same temperature range (Drake & Kashyap 1998; Kashyap et al. 1998). The photospheric metallicity (Taylor 1999) is [Fe/H] = 0.3, and

[17] Incompleteness in atomic line databases (see Brickhouse 1998) contributes to an error of less than 5% on the line-to-continuum ratios calculated here. They are negligible compared to the measurement error.

we derive for the coronal metallicity $[\mathrm{Fe/H}] = 0.4^{+1.1}_{-0.6}$, where the quoted range represents posterior deviations at 1 σ. While the uncertainty on our measurement is quite large (it is essentially unbounded at high metallicity), it is encouraging that the corona does not appear to be metal abundance deficient (see Drake 1996).

5. DISCUSSION

The power of the Bayesian methods illustrated here lies in their ability to combine information and to model directly the highly structured hierarchical features of the data, both in a principled manner. These features are illustrated in the α TrA example. First, by combining information from several detectors, we are able to extract information from the data regarding the plasma temperature. More generally, Bayesian methods allow for the incorporation of various forms of quantifiable prior information through the prior distribution. Of course, results are then conditional on the prior information: if these priors are not trusted, the conclusions cannot be trusted either. On the other hand, if the prior information is accepted as reasonable, the posterior distribution should be accepted as a conglomeration of prior scientific information and the data. Second, the extremely low counts in the α TrA data, along with many free parameters (10 emission-line intensities and two continuum parameters), illustrate a situation in which methods based on the Gaussian distribution and the central limit theorem are simply without justification. Methods that account for the Poisson (e.g., highly variable) character of the data have a sound mathematical basis and, in contrast to standard methods such as χ^2 fitting, are equipped to handle such data.

The hierarchy in the model described in § 3 can be extended to account for various more complicated features in the data, e.g., absorption lines, pile-up, and joint spatial, spectral, and temporal structure. Dealing with 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 Bayesian framework. After accounting for other features in the data such as instrument response, background, and absorption, we simply need to separate the observed counts into multiple counts of lower or equal energy based on the (current draw of the) spectral and spatial model. 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. Thus, we believe there is great promise in Monte Carlo techniques, which, if carefully designed, can automatically exclude numerous possibilities with minute probability. Although there remains much work to be done, Bayesian methods in conjunction with MCMC algorithms offer a practical and innovative solution to many outstanding data-analytic challenges in astrophysics.

The authors gratefully acknowledge funding for this project partially provided by NSF grant DMS-97-05157, by the *Chandra X-Ray Observatory*, and by NASA grant NAS8-39073. They also thank Rostislav Protassov and C. J. Shen for their work on the programming of the various algorithms, Richard Edgar and Paul Gorenstein for helpful comments on an earlier draft, and the referee, Jeff Scargle, whose careful reading and detailed suggestions have greatly improved the paper.

APPENDIX A

MARKOV CHAIN MONTE CARLO METHODS

A1. THE DATA AUGMENTATION ALGORITHM

The data augmentation algorithm is designed to obtain a sample from the posterior distribution for use in Monte Carlo integration. The strategy of the algorithm is to embed the posterior distribution, $p(\boldsymbol{\theta} \mid \boldsymbol{Y})$, into a distribution in a large space, $p(\boldsymbol{\theta}, \boldsymbol{Y}^{\mathrm{mis}} \mid \boldsymbol{Y})$. If we can obtain a sample from this second distribution, we need only discard the sampled values of $\boldsymbol{Y}^{\mathrm{mis}}$ to obtain the desired sample from the posterior. The quantity $\boldsymbol{Y}^{\mathrm{mis}}$ can be any unobserved quantity; it is referred to as "missing data" for historical reasons. For clarity we denote the observed data $\boldsymbol{Y}^{\mathrm{obs}}$ and the augmented data $\boldsymbol{Y}^{\mathrm{aug}} = (\boldsymbol{Y}^{\mathrm{obs}}, \boldsymbol{Y}^{\mathrm{mis}})$. In order to obtain a sample from $p(\boldsymbol{\theta}, \boldsymbol{Y}^{\mathrm{mis}} \mid \boldsymbol{Y}^{\mathrm{obs}})$, the data augmentation algorithm uses an iterative sampling scheme that samples (1) $\boldsymbol{Y}^{\mathrm{mis}}$ conditional on the model parameters and $\boldsymbol{Y}^{\mathrm{obs}}$ and (2) the model parameters given $\boldsymbol{Y}^{\mathrm{aug}}$. Clearly, the 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:

1. Draw $\boldsymbol{Y}^{\mathrm{aug}}_{[t]}$ from $p(\boldsymbol{Y}^{\mathrm{aug}} \mid \boldsymbol{Y}^{\mathrm{obs}}, \boldsymbol{\theta}_{[t-1]})$.
2. Draw $\boldsymbol{\theta}_{[t]}$ from $p(\boldsymbol{\theta} \mid \boldsymbol{Y}^{\mathrm{aug}}_{[t]})$.

Under certain regularity conditions (for details see Meyn & Tweedie 1993; Roberts 1996; Tierney 1994, 1996) the stationary distribution of the resulting Markov chain is the desired posterior distribution, i.e., for large t, $\boldsymbol{\theta}_{[t]}$ approximately follows $p(\boldsymbol{\theta} \mid \boldsymbol{Y}^{\mathrm{obs}})$.

To illustrate the utility of the data augmentation algorithm, we return to the simple background contamination model introduced in § 2.1. The choice of $\boldsymbol{Y}^{\mathrm{aug}}$ is clear in the example; we set $\boldsymbol{Y}^{\mathrm{aug}} = \{Y, Y^S, Y^B\}$, where Y is the total counts, Y^S is the unobserved source counts from the source exposure, and Y^B is the counts from the pure background observation. (i.e., we can consider Y^S to be the missing data). With this choice of $\boldsymbol{Y}^{\mathrm{aug}}$, both $p(\boldsymbol{Y}^{\mathrm{aug}} \mid \boldsymbol{Y}^{\mathrm{obs}}, \boldsymbol{\theta})$ and $p(\boldsymbol{\theta} \mid \boldsymbol{Y}^{\mathrm{aug}})$ are easy to sample, and thus the data augmentation algorithm is easy to use; here $\boldsymbol{Y}^{\mathrm{obs}} = \{Y, Y^B\}$ and $\boldsymbol{\theta} = (\lambda^B, \lambda^S)$. Given some $\boldsymbol{\theta}_{[0]} = (\lambda^B_{[0]}, \lambda^S_{[0]})$, the two steps of the data augmentation algorithm at iteration t become as follows:

1. Draw $Y^S_{[t]}$ from

$$Y^S \mid \boldsymbol{\theta}_{[t-1]},\ \boldsymbol{Y}^{\text{obs}} \overset{d}{\sim} \text{binomial}\left(Y, \frac{\lambda^S}{\lambda^B + \lambda^S}\right), \tag{A1}$$

i.e.,

$$p(Y^S \mid \boldsymbol{\theta}_{[t-1]},\ \boldsymbol{Y}^{\text{obs}}) = \binom{Y}{Y^S} (Y^S)^{\lambda S/(\lambda B + \lambda S)} (Y - Y^S)^{\lambda B/(\lambda B + \lambda S)}. \tag{A2}$$

2. Draw

$$\lambda^B_{[t]} \mid \boldsymbol{Y}^{\text{aug}}_{[t]} \overset{d}{\sim} \gamma(\alpha^B + \mathrm{Y} - \mathrm{Y}^S,\ \beta^B + 1), \tag{A3}$$

i.e.,

$$p(\lambda^B_{[t]} \mid \boldsymbol{Y}^{\text{aug}}_{[t]}) = \frac{[\lambda^B_{[t]}(\beta^B + 1)]^{(\alpha^B + Y - Y^S)} e^{-\lambda^B_{[t]}(\beta^B + 1)}}{\lambda^B_{[t]}\,\Gamma(\alpha^B + Y - Y^S)}, \tag{A4}$$

where α^B and β^B are typically chosen using the pure background observation as described in § 2.1, and

$$\lambda^S_{[t]} \mid \boldsymbol{Y}^{\text{aug}}_{[t]} \overset{d}{\sim} \gamma(\alpha^S + Y^S,\ \beta^S + 1). \tag{A5}$$

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 that the iteration converges to the desired draws from the posterior distribution.

By selecting a starting value and iteratively sampling according to equations (A1), (A3), and (A5), we obtain a Markov chain that delivers a dependent sample from the posterior distribution upon convergence. In the next section we use the data augmentation algorithm to illustrate the important practical issues of selecting starting values, detecting convergence, and accounting for the dependency in the sample.

A2. STARTING VALUES, CONVERGENCE, AND MULTIPLE CHAINS

An important and difficult aspect of MCMC methods in practice is ascertaining convergence to stationarity. Since the stationary distribution of the Markov chain is the posterior distribution of interest, we can consider $\{\boldsymbol{\theta}_{[t]},\ t > T_0\}$ to be a (dependent) posterior sample, which can be used for Monte Carlo integration. Thus, determining $\boldsymbol{\theta}_{[0]}$ and T_0 is critical for valid inference. There is a large and growing literature on these related subjects, and we refer interested readers to recent texts on the subject by Gelman et al. (1995), Carlin & Louis (1996), and Gilks et al. (1996), as well as the review article on convergence by Cowles & Carlin (1996). Here we briefly outline the approach that we find most fruitful.

As proposed by Gelman & Rubin (1992), we suggest running multiple Markov chains with a variety of starting values spread throughout the parameter space. This is a useful procedure since a single chain can appear to have converged when actually it has only settled temporarily in one region of the parameter space. This is illustrated with the Markov chain in Figure 9. The three chains show the draws of a variance parameter for a random effects model (for details see van Dyk & Meng 2001). Note that although chain 3 appears relatively stable, it is far from convergence during the first 10,000 draws. This is evident when it is compared with the other chains, but less so when we look only at the beginning of chain 3. It is recommended that the starting values for the several chains be spread broadly in the parameter space (relative to the region of high posterior probability). This can often be accomplished by roughly mapping the posterior, for example, using estimates and errors based on the χ^2 estimates, posterior modes, or maximum likelihood estimates. (See van Dyk 2001 for details on the computation of posterior modes and maximum likelihood estimates for our spectral model.) Once such "overdispersed" starting values are obtained, we can run the several chains until all converge to the same region of the parameter space. (There may be more than one mode in the posterior, in which case the chains may converge to different modes, i.e., different regions

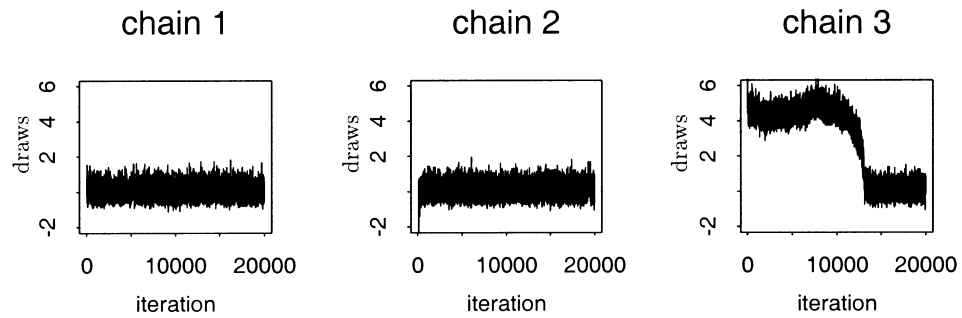

FIG. 9.—Several chains from a random effects model. Notice that chain 3 appears to have converged during the first 10,000 iterations. Comparison with chain 1 and chain 2, however, makes it clear that chain 3 did not converge until after iteration 10,000.

of the parameter space.) The $\hat{R}^{1/2}$ statistic of Gelman & Rubin (1992) measures the relative size of the total variance in the draws of a univariate function of the parameter and the average within chain variance of the same function, i.e.,

$$\sqrt{\hat{R}} = \sqrt{\frac{[(T-1)/T]W + (1/T)B}{W}}, \tag{A6}$$

where B is the between chain variance, W is the mean within chain variance, and T is the number of draws. If the variance within each chain is as great as the total variance in all the draws, i.e., $\hat{R}^{1/2}$ is near 1, then we can be confident that all the chains have converged to the same region of the parameter space. Typically we compute $\hat{R}^{1/2}$ using the last half (or two-thirds) of each of the chains. Once an acceptable level of $\hat{R}^{1/2}$ is obtained (say below 1.2), we omit the first half (or third) of the chain in all further analysis. If we have several starting values that cover a large enough region of the parameter space, we can be confident that the chains sample all areas with high posterior probability and thus the Monte Carlo approximations are unbiased estimators of the quantities they estimate.

The variance of the Monte Carlo approximations is a function of the posterior variance of the quantity being approximated, the posterior sample size (i.e., $T - T_0$), and the autocorrelation function of the Markov chain. Typically Monte Carlo errors are small relative to the posterior variance with several thousand posterior draws and thus are of little consequence. Monte Carlo error can be quantified by repeating the analysis for the first half and second half of the Markov chain and noting if the results are substantively different. See Roberts (1996) and references therein for details and extensions.

A3. THE GIBBS SAMPLER

In this and the next section we describe two additional MCMC methods, which are designed to deliver a sample from the posterior distribution and are often useful when the data augmentation algorithm is not practical. The Gibbs sampler can be viewed as an extension of the data augmentation algorithm in which we wish to sample from $p(\boldsymbol{\theta} \mid \boldsymbol{Y}^{\rm obs})$, and the vector, $\boldsymbol{\theta}$, can be viewed as a combination of model parameters and "missing data." (In many instances, there is no "missing data.") We partition $\boldsymbol{\theta}$ into $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_P)$, where $\boldsymbol{\theta}_p$ may be a scalar or vector quantity for each p. The Gibbs sampler again starts with some starting value $\boldsymbol{\theta}_{[0]}$ and at iteration t samples according to the following conditional distributions:

$$\begin{aligned}
&1.\ \text{Draw } (\boldsymbol{\theta}_1)_{[t]} \overset{d}{\sim} p(\boldsymbol{\theta}_1 \mid (\boldsymbol{\theta}_{-1})_{[t]}, \boldsymbol{Y}^{\rm obs}),\\
&2.\ \text{Draw } (\boldsymbol{\theta}_2)_{[t]} \overset{d}{\sim} p(\boldsymbol{\theta}_2 \mid (\boldsymbol{\theta}_{-2})_{[t]}, \boldsymbol{Y}^{\rm obs}),\\
&\qquad\vdots\\
&P.\ \text{Draw } (\boldsymbol{\theta}_P)_{[t]} \overset{d}{\sim} p(\boldsymbol{\theta}_P \mid (\boldsymbol{\theta}_{-P})_{[t]}, \boldsymbol{Y}^{\rm obs}),
\end{aligned}$$

where $(\boldsymbol{\theta}_{-p})_{[t]} = [(\boldsymbol{\theta}_1)_{[t]}, \ldots, (\boldsymbol{\theta}_{p-1})_{[t]}, (\boldsymbol{\theta}_{p+1})_{[t-1]}, \ldots, (\boldsymbol{\theta}_P)_{[t-1]}]$. That is, we draw each component of $\boldsymbol{\theta}$ in turn conditional on the current values of the rest of $\boldsymbol{\theta}$ and the data.

The advantage of the Gibbs sampler over the data augmentation algorithm is that in many settings additional conditioning results in simpler draws. The disadvantage is that the resulting Markov chains tend to have higher autocorrelation and are slower to converge to stationarity as P, the number of steps per iteration, increases.

A4. THE METROPOLIS-HASTINGS ALGORITHM

As a final extension, we consider the case in which one (or more) of the steps in the Gibbs sample involves a conditional distribution that is not easy to sample. The Metropolis-Hastings algorithm (Metropolis & Ulam 1949; Metropolis et al. 1953; Hastings 1970) replaces the conditional distribution by some convenient "jumping rule" that approximates the conditional distribution. A proposal draw is sampled according to the jumping rule and is either accepted or rejected (in which case, the Markov chain is fixed at the previous draw) according to a rule that maintains the desired stationary distribution (see, e.g., Gelman et al. 1995 for details).

APPENDIX B

DETAILS OF THE MCMC ALGORITHM

B1. DATA AUGMENTATION

The algorithms used to fit the model described in § 3.1 rely on the method of data augmentation. In this section we detail the layers of the data augmentation scheme we use. We aim to construct an idealized data set for which model fitting is a relatively easy task. That is, given the augmented data, we can easily sample the model parameters. Likewise, given the model parameters, we can easily sample the augmented data, and thus we can construct a data augmentation algorithm as described in § A1. Suppose, for example, that a data set uncontaminated by background or instrument response were available. Clearly, model fitting would be easier. We define an even larger data set that contains the unbinned, true, and blurred energies of all photons that would have arrived at the detector if there had been no absorption and if we were using a perfect instrument with the effective area equal to its maximum value over all energies used. This data set also includes a variable indicating

absorption and loss to reduced effective area[18] and a variable indicating the source of each photon, i.e., background (B), continuum (C), and each of the K line profiles; the set of sources is denoted as $\mathscr{S} = \{B, C, 1, 2, \ldots, K\}$. This idealized data set is summarized in Table 3.

The data augmentation scheme is illustrated in Figure 10, in which squares and circles represent observed and unobserved ("augmented") quantities, respectively. Given the model parameter $\boldsymbol{\theta}$, we obtain a sample set of photon energies, $\ddot{\boldsymbol{Y}}^s = (\ddot{Y}^s_1, \ldots, \ddot{Y}^s_{N^s})'$ for $s \in \mathscr{S}$ (see the third column of Fig. 10), representing the undegraded "augmented" data; N^s is the total count for source s. (As a mnemonic device, more dots in the accent above Y signifies further removal of a quantity from actual observable quantities.) Here $\ddot{\boldsymbol{Y}}^k$ contains the exact energy of all photons attributed to line k before absorption, with maximum effective area and no background contamination. (The background photon energies, $\ddot{\boldsymbol{Y}}^B$, do not appear in Fig. 10 because we model the detected counts [e.g., in PHA channels] rather than true counts; see § 3.5.) The first two columns of Figure 10 represent the hyperparameters and model parameters detailed in §§ 3.2–3.5.

[18] Absorption and effective area are handled together, so we need only one indicator variable; see § 3.4.

TABLE 3

VARIABLES ASSOCIATED WITH EACH PHOTON, FOR $i = 1, \ldots, N$

Variable	Notation	Range
Photon energy	$\ddot{Y}_i$	Positive, measured in keV
Indicator for background	Z_i^B	1 for background photons 0 for other photons
Indicator for continuum	Z_i^C	1 for continuum photons 0 for other photons
Indicator for line k, for $k = 1, \ldots, K$	Z_i^k	1 for photons from line k 0 for other photons
Indicator for absorption	Z_i^A	1 for absorbed photons 0 for other photons

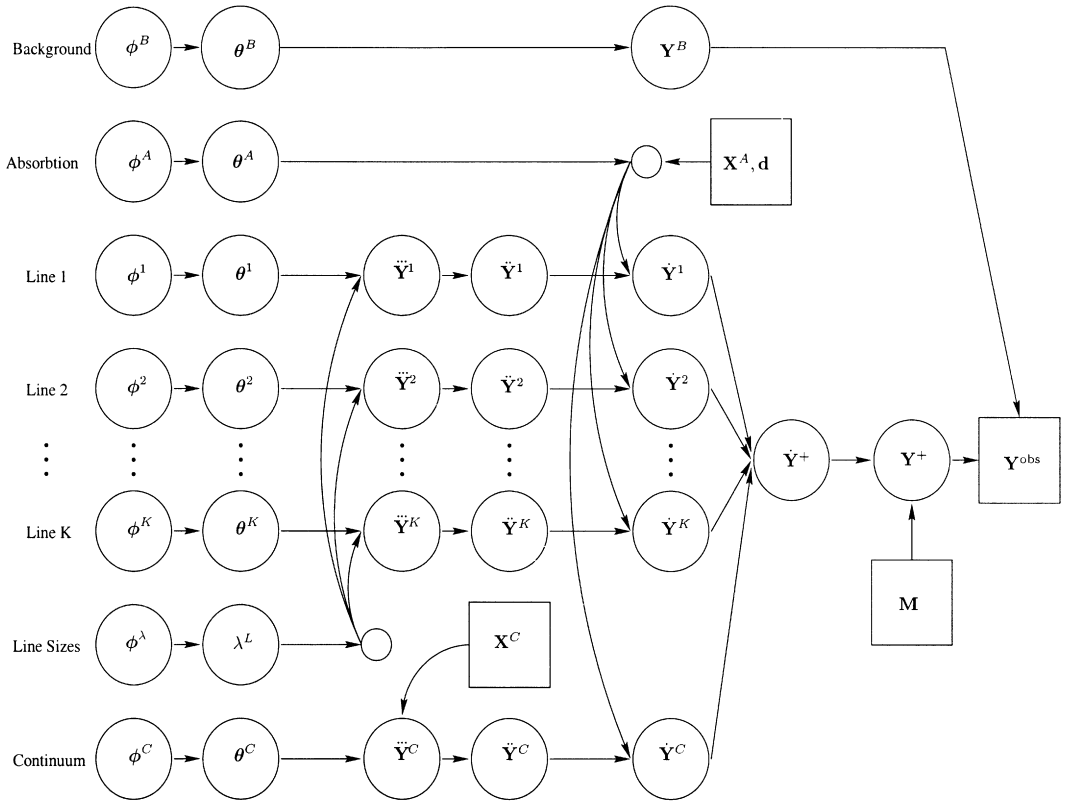

FIG. 10.—Graphical representation of the data augmentation scheme. Here ϕ represents hyperparameters, $\boldsymbol{\theta}$ model parameters, $\ddot{\boldsymbol{Y}}$ true photon energies, $\dot{\boldsymbol{Y}}$ binned energies, $\dot{\boldsymbol{Y}}$ binned true photon energies after absorption accounting for effective area, $\boldsymbol{Y}$ source counts in PHA channels, $\boldsymbol{Y}^{\rm obs}$ the observed counts, $\boldsymbol{M}$ the instrument response matrix, $\boldsymbol{d}$ the effective area vector, and $\boldsymbol{X}^A$ and $\boldsymbol{X}^C$ independent variables describing absorption and continuum, respectively; circles represent unobserved quantities, and squares observed quantities; details of the subscripts and superscripts are given in the text. The figure illustrates the interplay of the various model parameters, hyperparameters, observed quantities, and data augmentation. As an example, the first arrow in the row labeled "background" corresponds to the relationship between the background hyperparameters, ϕ^B, and the background intensities, θ^B, e.g., $\theta_i^B \overset{\mathrm{d}}{\sim} \gamma(\phi^{\rm B}_{i,1}, \phi^{\rm B}_{i,2})$. The second arrow corresponds to the Poisson nature of the background counts (see eq. [22]). The final background arrow illustrates the background contamination of the observed PHA counts.

TABLE 4

SUMMARY STATISTICS FOR THE SPECTRAL MODEL

Variable	Notation	Range
Unbinned energies	$\ddot{\boldsymbol{Y}}^s$	Positive, measured in keV, $s \in \mathscr{S}$
Binned energies	$\ddot{\boldsymbol{Y}}^s_j$	Counts for $j \in \mathscr{J}\mathscr{S} \in \mathscr{S}$
Binned energies after absorption	$\dot{\boldsymbol{Y}}^s_j$	Counts for $j \in \mathscr{J}\mathscr{S} \in \mathscr{S}$
Blurred PHA counts without background	$\boldsymbol{Y}^+_l$	Counts for $l \in \mathscr{L}$
Observed data	$\boldsymbol{Y}^{\rm obs}_l$	Counts for $l \in \mathscr{L}$

The array of energies represented by $\ddot{\boldsymbol{Y}}^s$ are binned into instrument-specific energy bins to obtain a sample spectrum, $\ddot{\boldsymbol{Y}}^s = (\ddot{Y}^s_1, \ldots, \ddot{Y}^s_J)'$ (see the fourth column in Fig. 10). In particular,

$$\ddot{Y}^s_j = \sum_{i=1}^{N^s} I\{\ddot{\boldsymbol{Y}}^s_i \in B_j\} \qquad \text{for } j \in \mathscr{J} \text{ and } s \in \mathscr{S}\ , \tag{B1}$$

where B_j is the jth energy bin. The first plot in Figure 2 illustrates the undegraded counts from the continuum and lines, $\ddot{Y}^+_j = \sum_{s\in\mathscr{S}\backslash B} \ddot{Y}^s_j$ for the artificial data set, where the notation $\mathscr{S}\backslash B$ indicates set subtraction, i.e., the set $\mathscr{S}$ with B removed. The solid line represents $E(\ddot{Y}^+_j \mid \boldsymbol{\theta})$, which equals the term in square brackets in equation (8). Because of absorption and effective area, a portion of these photons are not detected. The sample counts *after* absorption (and accounting for effective area) are depicted in the fifth column of Figure 10, by $\dot{\boldsymbol{Y}}^s = (\dot{Y}^s_1, \ldots, \dot{Y}^s_J)'$ with

$$\dot{Y}^s_j = \sum_{i=1}^{N^s} I\{\ddot{\boldsymbol{Y}}_i \in B_j\}(1 - Z^A_i) \qquad \text{for } j \in \mathscr{J} \text{ and } s \in \mathscr{S}\ . \tag{B2}$$

As described in Table 3, Z^A_i is 1 if photon i is absorbed and 0 otherwise. The second plot in Figure 2 represents $\dot{\boldsymbol{Y}}^s$ with $E(\dot{Y}^+_j \mid \boldsymbol{\theta}) = \lambda_j(\boldsymbol{\theta})d_j$ plotted as the solid line (see eq. [8]). The next two circles in Figure 10 represent the adding of sources and the blurring (i.e., instrument response) process. In particular, $\dot{\boldsymbol{Y}}^+ = \sum_{s\in\mathscr{S}\backslash B} \dot{\boldsymbol{Y}}^s$. The blurred data, $\boldsymbol{Y}^+ = (Y^+_1, \ldots, Y^+_L)'$, are a stochastic function of $\dot{\boldsymbol{Y}}^+$ (i.e., a multinomial distribution[19]),

$$\boldsymbol{Y}^+ \overset{d}{\sim} \sum_{j\in\mathscr{J}} \text{multinomial}\ (\dot{Y}^+_j, \boldsymbol{M}_j)\ , \tag{B3}$$

where $\dot{\boldsymbol{Y}}^+ = (\dot{Y}^+_1, \ldots, \dot{Y}^+_J)'$ and $\boldsymbol{M}_j$ is the jth column of $\boldsymbol{M}$; $\boldsymbol{Y}^+$ appears in the third plot of Figure 2 with $E(Y^+_l \mid \boldsymbol{\theta}) = \sum_{j\in\mathscr{J}} M_{lj}\lambda_j(\boldsymbol{\theta})d_j$. The counts due to background contamination are denoted as $\dot{\boldsymbol{Y}}^B = (\dot{Y}^B_1, \ldots, \dot{Y}^B_L)'$, and the observed data are denoted as $\boldsymbol{Y}^{\rm obs} = (Y^{\rm obs}_1, \ldots, Y^{\rm obs}_L)'$, with $Y^{\rm obs}_l = Y^B_l + Y^+_l$ for $l \in \mathscr{L}$; $\boldsymbol{Y}^{\rm obs}$ is illustrated in the final plot of Figure 2 and has expectation $\xi(\boldsymbol{\theta})$ (see eq. [9]).

B2. THE ALGORITHMS

In this section we present the details of the MCMC algorithm that we use to sample from the posterior distribution for our spectral model. We use an algorithm that alternately draws the "missing data" given the model parameters and the parameters given the "missing data." Both draws are conditional on the observed photon counts and the prior hyperparameters, $\phi = \{\phi^A, \phi^s, s \in \mathscr{S}\}$. In particular, we define two groups: (1) the augmented data, $\boldsymbol{Y}^{\rm aug} = \{\boldsymbol{Y}^{\rm obs}, \boldsymbol{Y}^B, \boldsymbol{Y}^+, \dot{\boldsymbol{Y}}, \ddot{\boldsymbol{Y}}, \ddot{\boldsymbol{Y}}\}$, where $\dot{\boldsymbol{Y}} = \{\dot{\boldsymbol{Y}}^s, s \in \mathscr{S}\backslash B\}$ are the binned true energies, after absorption and accounting for effective area, $\ddot{\boldsymbol{Y}} = \{\ddot{\boldsymbol{Y}}^k, k \in \mathscr{K}\}$ are the binned true energies, and $\ddot{\boldsymbol{Y}} = \{\ddot{\boldsymbol{Y}}^k, k \in \mathscr{K}\}$ are the (unbinned) true energies; and (2) $\boldsymbol{\theta} = \{\boldsymbol{\theta}^A, \boldsymbol{\theta}^s, s \in \mathscr{S}\}$ consists of the various model parameters. Using Bayes's Theorem, we are able to derive the necessary conditional distributions, which are described below.

First, we draw $\boldsymbol{Y}^{\rm aug}$ from $p(\boldsymbol{Y}^{\rm aug} \mid \boldsymbol{Y}^{\rm obs}, \boldsymbol{\theta})$; the draw is broken into the following five steps:

1. Independently separate the background counts,

$$Y^B_l \mid \boldsymbol{Y}^{\rm obs}, \boldsymbol{\theta} \overset{d}{\sim} \text{binomial}\left(Y^{\rm obs}_l, \frac{\theta^B_l}{\xi_l(\boldsymbol{\theta})}\right) \quad \text{for } l \in \mathscr{L}\ . \tag{B4}$$

2. Restore the blurred photons,

$$\dot{\boldsymbol{Y}}^+ \mid \boldsymbol{Y}^B, \boldsymbol{Y}^{\rm obs}, \boldsymbol{\theta} \overset{d}{\sim} \sum_{l\in\mathscr{L}} \text{multinomial}\left(Y^+_l, \frac{d_1\lambda_1(\boldsymbol{\theta})M_{1l}, \ldots, d_J\lambda_J(\boldsymbol{\theta})M_{Jl}}{\sum_{j\in\mathscr{J}} d_j\lambda_j(\boldsymbol{\theta})M_{jl}}\right), \tag{B5}$$

where $Y^+_l = Y^{\rm obs}_l - Y^B_l$.

3. Independently separate the counts into line and continuum counts,

$$(\dot{\boldsymbol{Y}}^C, \dot{\boldsymbol{Y}}^1, \ldots, \dot{\boldsymbol{Y}}^K) \mid \dot{\boldsymbol{Y}}^+, \boldsymbol{Y}^B, \boldsymbol{Y}^{\rm obs}, \boldsymbol{\theta} \overset{d}{\sim} \text{multinomial}\left(\dot{Y}^+_j, \frac{[d_j f(\boldsymbol{\theta}^C, E_j), \tilde{\lambda}^1 p^1_j \ldots, \tilde{\lambda}^K p^K_j]}{d_j f(\boldsymbol{\theta}^C, E_j) + \sum_{k=1}^K \tilde{\lambda}^k p^k_j}\right) \quad \text{for } j \in \mathscr{J}\ , \tag{B6}$$

where $p^k_j = p_j(\mu^k, \upsilon^k)$; see equation (8).

[19] The multinomial $(n, \boldsymbol{p})$ distribution is a distribution for nonnegative integer valued random vectors and generalizes the binomial distribution. In particular, a vector randomly selected from this distribution sums to n and its expected value is $n\boldsymbol{p}$, where $\boldsymbol{p}$ is a probability vector that sums to 1.

4. Independently restore the absorbed counts in the lines,

$$\ddot{Y}_j^k \mid \dot{\boldsymbol{Y}},\ \boldsymbol{Y}^B,\ \boldsymbol{Y}^{\rm obs},\ \boldsymbol{\theta} \overset{d}{\sim} \dot{Y}_j^k + \text{Poisson}\ (\tilde{\lambda}^k p_j^k [1 - d_j\, u(\boldsymbol{\theta}^A,\ E_j)])\ , \tag{B7}$$

for $j \in \mathscr{J}$ and $k \in \mathscr{K}$.

5. Independently deround the photon energies from the lines,

$$\ddot{\ddot{Y}}_i^k \mid \ddot{\boldsymbol{Y}},\ \dot{\boldsymbol{Y}},\ \boldsymbol{Y}^B,\ \boldsymbol{Y}^{\rm obs},\ \boldsymbol{\theta} \overset{d}{\sim} N(\mu^k,\ \upsilon^k),\ \text{truncated to}\ B_j\ , \tag{B8}$$

for $i = 1,\ \ldots,\ \ddot{Y}_+^k$ and $k \in \mathscr{K}$ with $\ddot{\boldsymbol{Y}}_+^k = \sum_{j \in \mathscr{J}} \ddot{\boldsymbol{Y}}_j^k$. We note that this draw is omitted for line k if υ^k is fixed at zero, i.e., if line k is a delta function. (In this case μ^k is not fitted by the algorithm).

Second, we draw $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta} \mid \boldsymbol{Y}^{\rm aug},\ \phi)$, taking advantage of conditional independence among several vector components of $\boldsymbol{\theta}$:

1. Independently draw the background model parameters,

$$\theta_l^B \mid \boldsymbol{Y}^{\rm aug},\ \phi \overset{d}{\sim} \gamma(\phi_{l,1}^B + \boldsymbol{Y}_l^B,\ \phi_{l,2}^B + 1)\ . \tag{B9}$$

2. Draw the variance and mean independently for each line profile,

$$\upsilon^k \mid \boldsymbol{Y}^{\rm aug},\ \phi \overset{d}{\sim} \frac{1}{\chi^2_{\nu_0^k + \ddot{Y}_+^k}} \left[\nu_0^k\, \upsilon_\phi^k + \sum_{i=1}^{\ddot{Y}_+^k} \left(\ddot{\ddot{Y}}_i^k - \frac{\ddot{\ddot{Y}}_+^k}{\ddot{Y}_+^k} \right)^2 + \frac{\kappa_0^k\, \ddot{Y}_+^k}{\kappa_0^k + \ddot{Y}_+^k} \left(\mu_0^k - \frac{\ddot{\ddot{Y}}_+^k}{\ddot{Y}_+^k} \right)^2 \right], \tag{B10}$$

$$\mu^k \mid \upsilon^k,\ \boldsymbol{Y}^{\rm aug},\ \phi \overset{d}{\sim} N\left(\frac{\kappa_0^k\, \mu_0^k + \ddot{\ddot{Y}}_+^k}{\kappa_0^k + \ddot{Y}_+^k},\ \frac{\upsilon^k}{\kappa_0^k + \ddot{Y}_+^k} \right), \tag{B11}$$

where $\ddot{\ddot{Y}}_+^k = \sum_{i=1}^{\ddot{Y}_+^k} \ddot{\ddot{Y}}_i^k$. Again, this step is omitted for line k if it is assumed to be a delta function.

3. Draw the line intensities $\tilde{\lambda}^k$, independently for each k,

$$\tilde{\lambda}^k \mid \boldsymbol{Y}^{\rm aug},\ \phi \overset{d}{\sim} \gamma(\ddot{Y}_+^k + \phi_1^k,\ 1 + \phi_2^{\tilde{\lambda}}) \qquad \text{independently for}\ k \in \mathscr{K}\ . \tag{B12}$$

4. Draw the parameters for the GLM for the continuum and absorption models. For this final step, we condition only on $\boldsymbol{Y}^{\rm obs}$, $\boldsymbol{Y}^B$, $\boldsymbol{Y}^+$, and $\dot{\boldsymbol{Y}}^+$, rather than $\boldsymbol{Y}^{\rm aug}$. We expect this substitution to improve the rate of convergence of the sampler. Because the conditional distribution is not from a standard family, we use a Metropolis-Hastings step. In particular, we note that

$$\dot{Y}_j^+ \mid \boldsymbol{\theta},\ \tilde{\boldsymbol{\lambda}} \overset{d}{\sim} \text{Poisson}\left(d_j \exp\ (\boldsymbol{X}_j^A\, \boldsymbol{\theta}^A) \sum_{k=1}^{K} \tilde{\lambda}^k p_j^k \right) \qquad \text{for}\ k \in \mathscr{K}\ , \tag{B13}$$

$$\dot{Y}_j^C \mid \boldsymbol{\theta},\ \tilde{\boldsymbol{\lambda}} \overset{d}{\sim} \text{Poisson}\ (dE_j d_j \exp\ (\boldsymbol{X}_j^C\, \boldsymbol{\theta}^C + \boldsymbol{X}_j^A\, \boldsymbol{\theta}^A)) \qquad \text{for}\ j \in \mathscr{J}\ , \tag{B14}$$

where $\tilde{\boldsymbol{\lambda}} = (\tilde{\lambda}^1,\ \ldots,\ \tilde{\lambda}^K)$. Since given $(\boldsymbol{\theta}^1,\ \ldots,\ \boldsymbol{\theta}^K)$ the log of each Poisson parameter differs from a linear combination of $\boldsymbol{\theta}^C$ and $\boldsymbol{\theta}^A$ by a known constant, the conditional posterior mode can easily be computed using a minor modification of an iteratively reweighted least-squares algorithm (e.g., Thisted 1988). This algorithm can also account for the prior information described in § 3.2 and reports the curvature of the log posterior at the mode. A multivariate t-distribution with 4 degrees of freedom with the appropriate mode and (perhaps inflated) curvature can be used as a jumping distribution to generate a proposal for the next sample from the conditional distribution. The relative mass of the jumping distribution and actual conditional distribution of $\boldsymbol{\theta}^C$ and $\boldsymbol{\theta}^A$ at the previous draw and proposed draw are combined to determine if the proposal should be accepted or rejected (in which case the previous draw is reused). Several (five to 10) proposals are drawn at each iteration. We note that the same procedure can be used to fit a GLM to the background counts (as was done in § 4.2).

Although the MCMC methods detailed above may seem inhibiting as a whole, each of the required steps is quite simple. The power of the MCMC methods described here (e.g., the Gibbs sampler) lies in their ability to break complicated model-fitting tasks into a succession of relatively simple tasks. Our general strategy is to use Bayes's Theorem to derive a posterior distribution that hierarchically accounts for the complexity both in the posited model and in data collection. We then use modern statistical algorithms that devolve model fitting into a sequence of relatively simple steps. We believe that this is a powerful strategy for dealing with the ever-increasing power and sophistication of today's astronomical instruments.

APPENDIX C

INTERNET RESOURCES

There are several internet sites where one can find papers describing Bayesian methods and related software. The MCMC preprint service[20] and STATLIB[21] Web sites are both large general statistical sites that offer various software, preprints, and links that may be of interest to astrophysicists. Three Web sites (that we know of) aim specifically at the interface of astrophysics and statistics.[22]

[20] http://www.mcs.surrey.ac.uk/Personal/S.Brooks/MCMC.
[21] http://lib.stat.cmu.edu.
[22] http://www.fas.harvard.edu/~vandyk/astrostat.html, http://astrosun.tn.cornell.edu/staff/loredo/bayes, http://www.astro.psu.edu/statcodes.

REFERENCES

Bijaoui, A. B. 1971, A&A, 13, 226
Brickhouse, N. 1998, in ASP Conf. Ser. 154, 10th Cambridge Workshop, Cool Stars, Stellar Systems, and the Sun, ed. R. Donahue & J. Bookbinder (San Francisco: ASP), 487
Brown, A., Drake, S., Van Steenberg, M., & Linsky, J. 1991, ApJ, 373, 614
Carlin, B. P., & Louis, T. A. 1996, Bayes and Empirical Bayes Methods for Data Analysis (Chapman & Hall: London)
Cash, W. 1979, ApJ, 228, 939
Collura, A., Maggio, A., Sciortino, S., Serio, S., Vaiana, G. S., & Rosner, R. 1987, ApJ, 315, 340
Connors, A. 1997, in Data Analysis in Astronomy, ed. V. Gesu, M. J. B. Duff, A. Heck, M. C. Maccarone, L. Scarsi, & H. U. Zimmermann (Singapore: World Scientific), 251
Cowles, M. K., & Carlin, B. P. 1996, J. Am. Stat. Assoc., 91, 883
Drake, J. 1996, in ASP Conf. Ser. 109, 9th Cambridge Workshop, Cool Stars, Stellar Systems, and the Sun, ed. R. Pallavicini & A. Dupree (San Francisco: ASP), 203
Drake, J., & Kashyap, V. 1998, in ASP Conf. Ser. 154, 10th Cambridge Workshop, Cool Stars, Stellar Systems, and the Sun, ed. R. Donahue & J. Bookbinder (San Francisco: ASP), 1014
Elvis, M., Matsuoka, M., Siemiginowska, A., Fiore, F., Mihara, T., & Brinkmann, W. 1994, ApJ, 436, L55
Feigelson, E. D., & Babu, G. J. 1997, in Data Analysis in Astronomy, ed. V. Gesu, M. J. B. Duff, A. Heck, M. C. Maccarone, L. Scarsi, & H. U. Zimmermann (Singapore: World Scientific), 281
Freeman, P. E., Graziani, C., Lamb, D. Q., Loredo, T. J., Fenimore, E. E., Murakami, T., & Yoshida, A. 1999, ApJ, 524, 753
Gehrels, N. 1986, ApJ, 303, 336
Gelfand, A. E., & Smith, A. F. 1990, J. Am. Stat. Assoc., 85, 398
Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. 1995, Bayesian Data Analysis (London: Chapman & Hall)
Gelman, A., & Rubin, D. B. 1992, Stat. Sci., 7, 457
George, I. M., Turner, T. J., Yaqoob, T., Netzer, H., Laor, A., Mushotzky, R. F., Nandra, K., & Takahashi, T. 2000, ApJ, 531, 52
Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. 1996, Markov Chain Monte Carlo in Practice (London: Chapman & Hall)
Gregory, P. C., & Loredo, T. J. 1992, ApJ, 398, 146
Hartmann, L., Dupree, A., & Raymond, J. 1981, ApJ, 246, 193
Hastings, W. K. 1970, Biometrika, 57, 97
Kashyap, V., & Drake, J. J. 1998, ApJ, 503, 450
Kashyap, V., Drake, J. J., Pease, D. O., & Schmitt, J. H. M. M. 1998, Am. Astron. Soc. Meeting, 192, 8201
Kashyap, V., Rosner, R., Harnden, Jr., F. R., Maggio, A., Micela, G., & Sciortino, S. 1994, ApJ, 431, 402
Knoll, G. F. 1989, Radiation Detection and Measurement (New York: Wiley)
Kühr, H., Witzel, A., Pauliny-Toth, I., & Nauber, U. 1981, A&AS, 45, 367
Lampton, M., Margon, B., & Bowyer, S. 1976, ApJ, 208, L177
Loredo, T. J. 1993, in Statistical Challenges in Modern Astronomy, ed. G. J. Babu & E. D. Feigelson (New York: Springer), 275
Lucy, L. B. 1974, AJ, 79, 745
McCullagh, P., & Nelder, J. A. 1989, Generalized Linear Models (2d ed.; London: Chapman & Hall)
Meng, X.-L., & van Dyk, D. A. 1999, Biometrika, 86, 301
Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., & Teller, E. 1953, J. Chem. Phys., 21, 1087
Metropolis, N., & Ulam, S. 1949, J. Am. Stat. Assoc., 44, 335
Meyn, S. P., & Tweedie, R. L. 1993, Markov Chains and Stochastic Stability (New York: Springer)
Mighell, K. J. 1999, ApJ, 518, 380
Nousek, J. A. 1993, in Statistical Challenges in Modern Astronomy, ed. G. J. Babu & E. D. Feigelson (New York: Springer), 307
Protassov, R., van Dyk, D. A., Connors, A., Kashyap, V. L., & Siemiginowska, A. 2001, Tech. Rep.
Richardson, W. H. 1972, J. Opt. Soc. Am., 62, 55
Roberts, G. O. 1996, in Markov Chain Monte Carlo in Practice, ed. W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (London: Chapman & Hall), 45
Rosner, R., Musielak, Z., Cattaneo, F., Moore, R., & Suess, S. 1994, ApJ, 442, L25
Schmitt, J. H. M. M. 1985, ApJ, 293, 178
Sciortino, S., & Micela, G. 1992, ApJ, 388, 595
Siemiginowska, A. 1997, in Data Analysis in Astronomy, ed. V. Gesu, M. J. B. Duff, A. Heck, M. C. Maccarone, L. Scarsi, & H. U. Zimmermann (Singapore: World Scientific), 221
Siemiginowska, A., Elvis, M., Connors, A., Freeman, P., Kashyap, V., & Feigelson, E. 1997, in Statistical Challenges in Modern Astronomy II, ed. G. J. Babu & E. D. Feigelson (New York: Springer), 241
Sivia, D. S. 1996, Data Analysis: A Bayesian Tutorial (Oxford: Oxford Univ. Press)
Tanner, M. A., & Wong, W. H. 1987, J. Am. Stat. Assoc., 82, 528
Taylor, B. 1999, A&AS, 134, 523
Thisted, R. A. 1988, Elements of Statistical Computing (London: Chapman & Hall)
Tierney, L. 1994, Ann. Stat., 22, 1701
———. 1996, in Markov Chain Monte Carlo in Practice, ed. W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Chapman & Hall: London), 59
van Dyk, D. A. 2001, Tech. Rep.
van Dyk, D. A., & Meng, X.-L. 2001, J. Comput. Graph. Stat., in press
Zimmerman, H. U. 1997, in Data Analysis in Astronomy, ed. V. Gesu, M. J. B. Duff, A. Heck, M. C. Maccarone, L. Scarsi, & H. U. Zimmermann (Singapore: World Scientific), 53