Highly-Structured Statistical Models in High-Energy Astrophysics

David A. van Dyk
Department of Statistics, University of California, Irvine, CA 92697-1250

In recent years, an innovative trend has been growing in applied statistics—it is becoming ever more feasible to build application-specific models which are designed to account for the hierarchical and latent structures inherent in any particular data generation mechanism. Such highly-structured models have long been advocated on theoretical grounds, but recently the development of new computational tools (e.g., hardware, software, and algorithms) for statistical analysis has begun to bring such model fitting into routine practice. In this paper, we describe these methods in the context of empirical high-energy astrophysics. A new generation of scientific instruments such as the *Chandra X-ray Observatory* are opening a whole new window to the study of the cosmos. Unlocking the information in the data generated with these complex high-tech instruments, however, requires sophisticated statistical models, methods, and computation. Here we discuss the techniques that the California-Harvard AstroStatistics Collaboration have been using to develop application-specific highly-structured statistical models to address these problems in high-energy astrophysics.

1. THE *CHANDRA X-RAY OBSERVATORY*

The *Chandra X-ray Observatory* took its place along side the *Hubble Space Telescope* and the *Compton Gamma-ray Telescope* as part of NASA's fleet of Great Observatories when it was launched by the Space Shuttle *Columbia* in July 1999. *Chandra* is by far the most precise X-ray telescope ever constructed; it is able to produce images over thirty times sharper than those available from previous X-ray telescopes. Although *Chandra* is a good example of a modern complex scientific instrument, it is but one of a host of such instruments. The complexity of these instruments along with the complexity of the objects that they study and the scientific questions they aim to answer demand sophisticated statistical methods. Off-the-shelf statistical techniques are simply not up to the inferential tasks involved in the scientific exploration of such data. In this paper we use *Chandra* as an example, to show how sophisticated application-specific statistical methods can be designed to meet the scientific challenges posed by modern instrumentation.

Chandra collects data on each photon that arrives at its active detector. The two-dimensional sky coordinates, energy, and time of arrival of each photon are recorded. Because of instrumental constraints, each of these quantities is rounded or binned into a discrete variable. Thus, in principle, the data can be represented by a four-way table of counts. Spectral analysis investigates the one-way marginal table of energy counts; image analysis focuses on the two-way marginal table of coordinates; and timing analysis studies the one-way table of arrival times. More sophisticated analysis might look at joint distributions to study, for example, how the spectrum varies across an extended source. In this paper we confine our attention to spectral analysis and image analysis. As we shall see, even these marginal analysis pose significant challenges.

A typical spectrum is modeled as a mixture of a smooth broad continuum term and a number of narrow emission lines. The continuum is formed by thermal (heat) radiation or by non-thermal processes in relativistic plasmas. The continuum is modeled using a smooth parametric form that includes emission across the entire width of the spectrum. Emission lines, on the other hand are narrow features in the spectrum that can be modeled with Gaussian distribution, Lorentzian distributions, or delta functions. When an electron jumps down from one quantum state of an atom to another, the energy of the electron decreases. This energy is radiated away from the atom in the from of a photon with energy equal to difference of the energies associated with the two quantum states. Unlike the emission that forms the continuum, the energies associated with these differences are discrete and from the emission lines in a spectrum.

Taken together, these features of the spectrum give subtle clues as to aspects such as the temperature and composition of the physical environment of the cosmological source. A stellar corona, for example, is made of numerous ions which can be recognized in a spectrum from their identifying emission lines. If the corona is relatively hot, the emission lines that correspond to more energetic quantum states will be relatively strong. Thus, the relative strength of the emission lines corresponding to a particular ion carries information as to the temperature of the source. Figure 1 shows an ultra high-resolution *Chandra* observation of the spectrum of the star Capella (α Aur). The spectrum is composed of a forest of spectral emission lines. Taken along with prior information obtained from detailed quantum mechanical computations and ground-based laboratory measurements, this data can be used to construct the physical environment of Capella's coronae. These calculations require sophisticated application-specific statistical methods. We do not discuss the details here. Instead we refer the interested reader to van Dyk *et al.* [2004] and detail a much simpler example in Section 3.

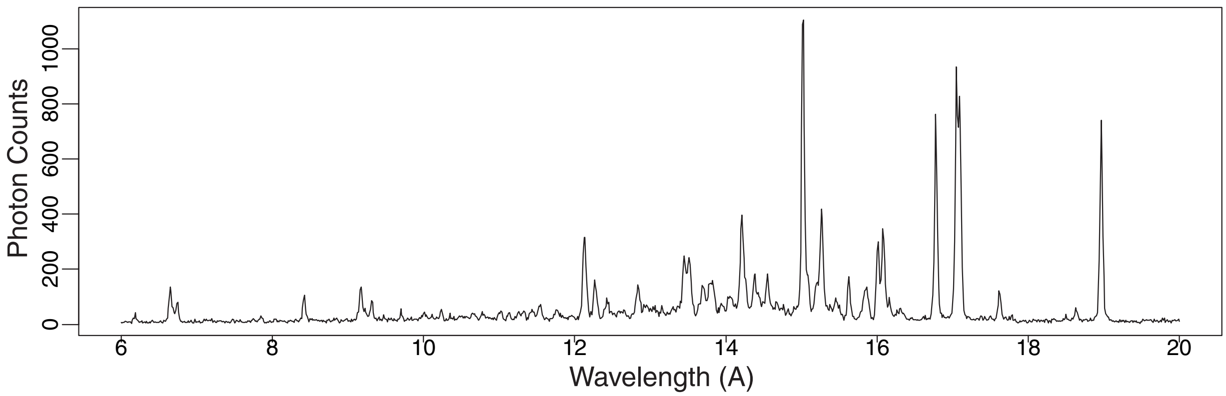

Figure 1: The Spectrum of Capella (α Aur). This high-resolution spectrum was collected using *Chandra*'s high-resolution camera along with its low-energy transmission grating spectrometer. Notice the numerous emission lines that compose the spectrum. A scientific goal is to use this forest of emission lines to reconstruct the composition and the distribution of the temperature of Capella's coronae, where the X-ray emission is produced.

Like spectra X-ray images can be composed of extended smooth features along with local bright features. At one extreme an image might reveal a smooth extended source such as a nebula without bright stars. There may also be a few bright point sources in the smooth extended emission, or the extended emission may be peppered with numerous point sources. Irregular and unpredictable structure is the rule rather than the exception when examining cosmological images. Figure 2 is a *Chandra* image of the central region of the galaxy NGC 6240, the product of the collision of two smaller galaxies. The center of this galaxy is dominated by two massive black holes; one is clearly visible as a white pixel in Figure 2. There appears to be additional structure in the extended source. A loop of hot gas appears toward the upper right of the image, and a larger fainter loop appears off to the right. Because of the high variability of the low-count per pixel data that typifies *Chandra*'s high-resolution images, however, it is difficult to distinguish features in the galaxy from artifacts of the statistical noise. As we shall discuss next, the situation is further complicated by a number of processes inherent in the data collection mechanism that degrade the quality of the data.

Both spectral and spatial characteristics of the data are degraded in a number of ways that must be accounted for in any principled data analysis. For example, the *effective area* of the detector varies with the energy of the photon. *Chandra* focuses X-rays with mirrors. Unfortunately, high-energy photons do not reflect uniformly and simply; some are absorbed and some pass right through the reflector, with a probability that is a function of their energy. A similar process occurs before the photon reaches the detector; lower energy photons are more likely to be absorbed by inter-stellar or inter-galactic media. Thus, the probability that an X-ray reaches the detector depends on the X-ray's energy. In statistical terms, we refer to the photons that are absorbed or undetected because of a relatively small effective area as *missing data*. Because ignoring the missing data mechanism would result in biased spectral analysis, it is called non-ignorable missing data Rubin [1976]. The likelihood that a photon is recorded also depends on where it lands on the detector. Photons landing near the boundary of the CCDs, for example, are less likely to be recorded. This effect is calibrated by the so called *exposure map*.

Because the focusing of the mirrors is not perfect the image of a point source is blurred; the character of the blurring is recorded in the *point spread function*. Another form of data degradation is due to a detector response, which results in a blurring of the photon energies. The recorded energy of a photon that arrives with a particular energy and location on the sky has a probability distribution. Finally, the source photons are generally contaminated by background counts. Common methods for handling data distortion can be quite ad hoc. For example, in spectral analysis a second data set is collected that is assumed to consist only of background counts. This background data is often *directly* subtracts from the source data and the result is analyzed as if it were a source observation free of background contamination. This procedure can lead to negative counts and estimates with questionable statistical properties.

The complexity of the cosmological sources, of the instrumentation, and of the scientific questions combine to result in sophisticated data analytic challenges. In the following sections we discuss how we propose to address these challenges using sophisticated application-specific highly-structured models. More details about *Chandra* and the analysis of *Chandra* data can be found in van Dyk *et al.* [2004]. The application of our methods to spectral analysis is the topic of van Dyk *et al.* [2001], Protassov *et al.* [2002], and van Dyk and Kang [2003]; image analysis is dis-

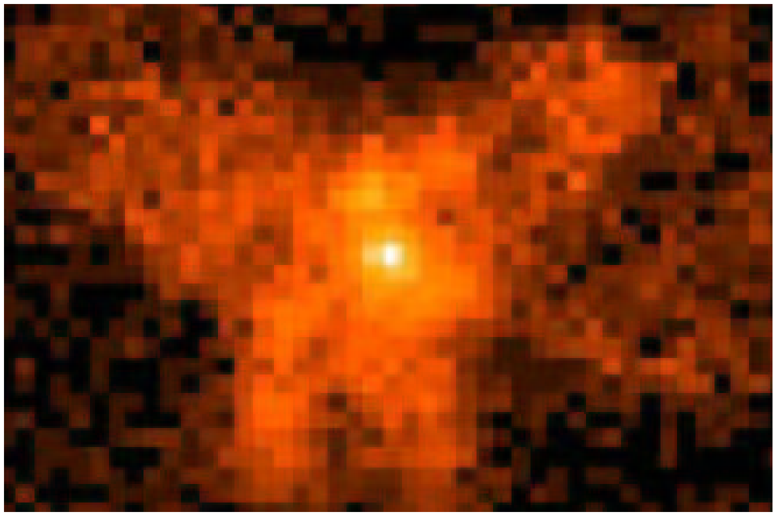

Figure 2: An X-ray Image of NGC 6240. The bright spot at the center of this galaxy is massive black hole. A second black hole appears above and a bit to the left of the brighter black hole. A loop of hot gas appears in the upper right quadrant of this image, and there appears to be a second larger loop off to the right. Whether these are actual features in this galaxy or artifacts of the highly-variable low-count Poisson data is an important astrostatistical question.

cussed in van Dyk and Hans [2002] and Esch *et al.* [2004].

2. APPLICATION-SPECIFIC STATISTICAL MODELS

Any principled analysis of *Chandra* data must account for the complexity of the data generation mechanism. Thus, we propose designing application-specific models that accounts not only for the spectra or images that are of primary scientific interest, but also for the complexities of the instrumentation. These models can be formulated via a hierarchical set of levels that allow us to separate a complex data generation mechanism into a number of well-understood components each of which on its own can be addressed via standard statistical techniques. When these levels are combined they form a highly-structured model that specifically addresses the complexities of the problem at hand. This multi-level approach not only allows us to formulate a highly-structured model using simple tools, but also gives us access to computational techniques that take advantage of the multi-level structure to combine a number of simple steps to fit a highly-structured model. In this article we outline the use of highly-structured multi-level models; a more thorough introduction can be found in Gelman *et al.* [2003].

One way to formulate these models in terms of *missing data*. In a spectral analysis, for example, we might consider the *ideal counts* to be a mixture of continuum and emission line counts that are unaffected by the effective area of the instrument, by photon absorption, by blurring of the photon energies, or by background contamination. These ideal counts are unobserved, and, in this sense can be regarded as missing data. Each ideal count can be modeled as a finite mixture of Poisson random variables, each of which corresponds to one of the emission lines or the continuum term. The key here is that this model can be formulated ignoring all of the mechanisms that degrade the data, thereby separating the complexity of the instrumentation from that of the cosmological sources. Thus, we are able to separate the task of modeling a sophisticated data generation mechanism into a sequence of simple tasks.

Adding a level to this model, we can account for photon absorption by modeling the ideal counts subject to absorption given the ideal counts. There is a parameterized absorption probability in each bin that depends on the energy corresponding to the bin. Because absorption operates independently on each photon, the counts in each bin after absorption are binomial given the corresponding ideal counts. In statistical terms, we use a generalized linear model that is akin to logistic regression for this level of the model. What is important here is that this is again a standard well-understood statistical model.

Similarly we can add levels to the model to account for the effective area of the instrument, the blurring of the phonon energies, and background contamination. Each of these levels incorporates what is known about the instrumentation (e.g., from instrument calibration) into a standard statistical model.

The discussion in this section is a broad overview of how multi-level models can be built for spectral analysis of *Chandra* data; details of this approach in this specific application can be found in van Dyk *et al.* [2001] and van Dyk and Kang [2003]. The key here, however, is the principle that a statistical model can be designed to incorporate specific features of any scientific data generation mechanism. The goal should be to formally model as much as this mechanism as possible. The mathematics of probability modeling ensures that properly modeled variability in the data generation mechanism will be reflected in the resulting uncertainty in the fitted model and the error bars on the model parameters. Preprocessing the data and ad hoc data manipulation do not properly account for variability and can have unpredicted consequences on the statistical properties (e.g., bias and coverage) of the resulting estimates. In practice, there is always some data preprocessing that must occur, but it is important to consider the effects on model uncertainty and generally to avoid preprocessing when possible.

A Simple Example. For the purpose of illustration, we consider a simple example; we emphasize that this example is not meant to illustrate the power of highly-structured models, but rather to show how they work. More sophisticated examples appear in the papers cited in this article.

Suppose we observe a single count, Y that is background contaminated, i.e., Y is a mixture of source and background counts. We also observe a second count, Z, that is a pure background count. We allow the exposure times for the two counts to be different and label them τ_S and τ_Z, respectively. The goal is to estimate the source count rate. We can easily formulate this problem in terms of missing data by supposing $Y = Y_S + Y_B$, where Y_S and Y_B are the source and background counts in the initial exposure. Clearly, Y_S and Y_B are unobserved quantities; we refer to these quantities as missing data. If the missing data were observed, it would be easy to estimate the source count rate, Y_S/τ_S. Likewise if the source and background count rates were known, we could easily split Y into Y_S and Y_B based on the relative intensities of the two rates. Thus, identifying Y_S and Y_B as missing data simplifies the relationships among the data and the quantities of scientific interest. Although this discussion is heuristic, it can be formalized to formulate algorithms for maximum likelihood fitting and Bayesian methods. Details in this particular example can be found in van Dyk [2003], a general discussion of these topics are the subject of Section 3 and 4.

3. STATISTICAL INFERENCE

Fitting a statistical model involves not only statistical computation, the subject of Section 4, but also the selection of a criterion for the fit. Common methods include χ^2 fitting, maximum likelihood, and Bayesian methods.

The method of χ^2 fitting ignores the variance structure inherent in the data by essentially making large sample Gaussian assumptions on the errors. As such, this method is especially inappropriate for high-resolution low-count data which exhibit Poisson errors.

Methods based on the likelihood are more appropriate in that they can explicitly account for error structures in the data. Bayesian methods take this one step further by allowing statistical inference that combines other scientific information with the data . The prior distribution is used to quantify information outside the data, the likelihood function quantifies information in the data, and these are combined via Bayes Theorem to form the posterior distribution. From a Bayesian perspective, the posterior distribution is a compete summary of the available information.

In practice, the prior distribution can be used to quantify information available from other data sources, from instrumental calibration, or from analytical physical calculations. Prior distribution are often used to quantify what is know about the values of parameters that are not of primary interest, and in some cases are used to quantify what is known about the likely values of parameters of direct scientific interest. Alternatively, prior distributions can be used to introduce structure on groups parameters. For example, it might be known that the values of a group of parameters are related to each other. This is sometimes the case with the wavelength of a group of emission lines associated with a particular ion. In image analysis, we can use prior distributions to encourage a smooth reconstruction of extended emission. Thus, we emphasizes that despite their reputation for being subjective and unscientific, prior distribution can be used in an objective manner to quantify model assumptions or concrete scientific information.

In Figure 3, we illustrate prior and posterior distributions under the simple background contamination example introduced in Section 2. The figure corresponds to a simulated data set with $Y = 1$, $Z = 48$, $\tau_S = 1$, and $\tau_B = 24$. The first plot illustrates two possible prior distributions on the source rate; one is flat and the other prefers values near three. The corresponding posterior distributions appear in the second plot. Since both the source and background rates are unknown parameters, the posterior distributions for the source rate are marginal distributions that result from integrating the joint posterior distribution over the background rate. An attractive feature of Bayesian methods is a simple principled prescription for handling nuisance parameters: They can be integrated out, leaving the marginal posterior distribution of the parameters of interest.

The posterior distributions in the second plot represent a compromise between the data and the prior distributions. The data from the background exposure alone, $Z = 48$ with $\tau_B = 24$ suggests a background rate of two. Given that Y is only one, direct background subtraction would result in a nega-

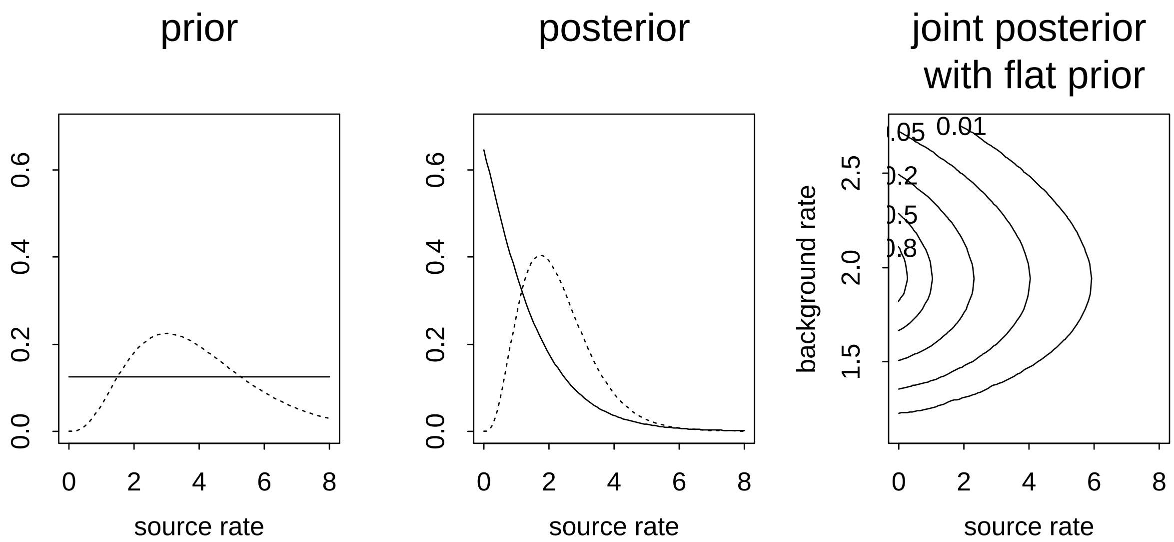

Figure 3: Sensitivity Analysis. The first plot shows two possible prior distributions on the source rate parameter. The second plot represents the resulting marginal posterior distributions on the same parameter. Notice that in this case the posterior distribution is sensitive to the choice of the prior distribution. The final plot shows the joint posterior distribution of the source and background rates under the flat prior distribution. The marginal posterior plotted with a solid line in the second plot is the integral of that in the third plot over the background rate.

tive source rate of -1.[1] Thus, the data favors small values of the source rate; the maximum likelihood estimate is zero. This is reflected in the solid posterior distribution, which has its mode at zero. The dotted posterior distribution, on the other hand, is a compromise between the dotted prior distribution which favors slightly larger values of the source rate and the data. The final plot illustrates the contours of the joint posterior distribution under the flat prior on the source rate. Integrating this joint posterior distribution over the background rate yields the solid posterior distribution in the second plot.

From a Bayesian perspective, the marginal posterior distribution of the source rate is a complete summary of the information available for this parameter. The posterior mode or mean are often used as the fitted values of the parameters, while some measure of the posterior variability is used to generate confidence intervals or error bars. Any such summary of the posterior distribution, however, is an imperfect representation and is less informative than the posterior distribution itself. Summaries of this sort are especially problematic when the posterior distribution is multi modal or highly skewed. This is illustrated by the marginal posterior distribution plotted by the solid line in the second plot of Figure 3. Although the mode of this distribution is zero, this value does not appear to be an adequate summary of the distribution. Thus, one of the primary advantages of the Monte Carlo methods described in Section 4 is that they summarize the entire posterior distribution.

[1]The fact that this ad hoc technique results in a negative count rate is an indication that such ad hoc methods can have unexpected and uninterpretable results. Thus, these methods should be avoided and model based methods such as maximum likelihood or Bayesian methods should be preferred.

4. STATISTICAL COMPUTATION

In this section we discuss two computational methods for posterior exploration: mode finders and Monte Carlo methods.

Although modes can be misleading summaries of likelihood functions or posterior distributions, mode finders can be useful for initial exploration. For example, a *Chandra* image can easily have tens of thousands of pixel intensities. When working in very high dimensional parameter spaces, algorithms that quickly find areas of high posterior probability are a valuable tool.

There are many well-known strategies for finding modes of high-dimensional posterior distributions; Newton's method, Fisher's scoring, and conjugate gradient are well-known examples. Here we discuss another method that is especially useful with highly-structured models. The EM algorithm (Dempster *et al.* [1977]) is a two-step iterative routine for computing posterior modes (maximum a posterior, MAP, estimates) in problems that are formulated in terms of missing data. Details can be found in van Dyk [2003] or McLaughlan and Krishnan [1997]; here we

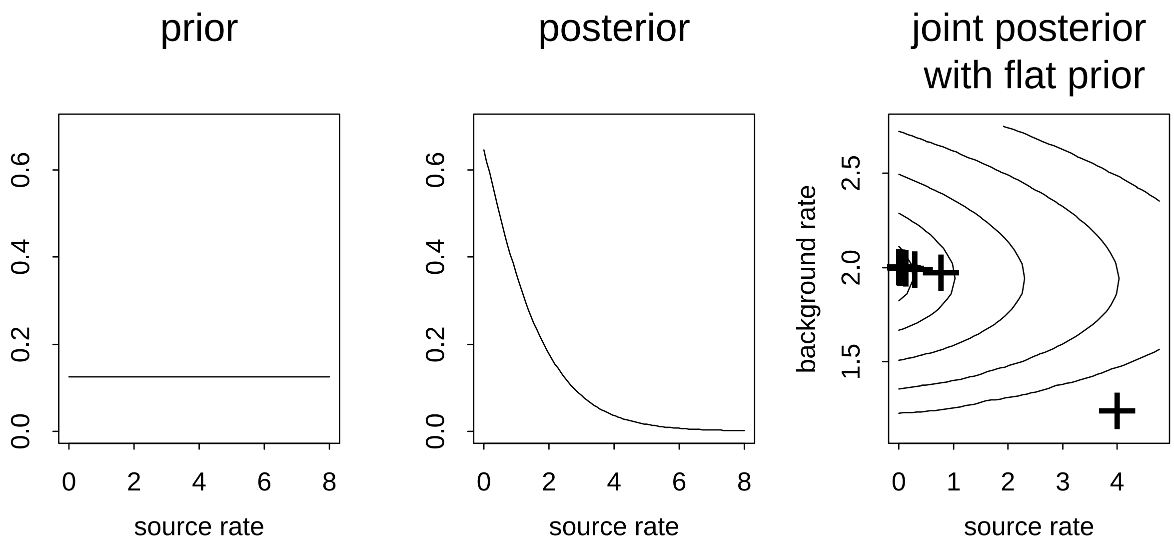

Figure 4: Using the EM Algorithm to Find the Posterior Mode. The first two plots are the same as in Figure 3 except that we consider only the flat prior distribution on the source rate. The final plot illustrates the contours of the corresponding joint posterior distribution along with the steps of an EM algorithm designed to compute the posterior mode. The skewed nature of the posterior distribution illustrates the potentially misleading nature of the modal estimates.

simply discuss how the EM algorithm works in the simple background contamination example described in Section 2. The first step of the EM algorithm, the *Expectation-Step*, replaces the missing values of Y_S and Y_B by their conditional expectation given Y, X, and the source and background rates. Simple probabilistic calculations show that this is accomplished by dividing the counts, Y into Y_S and Y_B using the relative size of the two corresponding rates. In the second step of this EM algorithm, the *Maximization-Step*, the rates are updated treating (Y, Y_S, Y_B, X) as data, i.e., we set the source and background rates equal to Y_S/τ_S and $(Y_B + X)/(\tau_S + \tau_B)$, respectively. The iterates of this EM algorithm using the simulated data set discussed in Section 3 are illustrated in Figure 4.

There are a variety of extensions to the EM algorithm that significantly broaden its application in the context of models formulated in terms of missing data (Meng and van Dyk [1997], van Dyk and Meng [2000], McLaughlan and Krishnan [1997]). As our example illustrates, EM-type algorithms are often easy to formulate even in highly-structured models. Another advantage of the EM algorithm is that it exhibits much more predictable and stable convergence than many other mode finders. In particular, it is guaranteed to increase the value of the posterior distribution at each iteration, i.e., it converges monotonically, see Figure 4. The primary disadvantage of the EM algorithm is that it sometimes can be slow to converge. Several of the extensions of EM, however, can be used to significantly improve its rate of convergence (van Dyk and Meng [2000], McLaughlan and Krishnan [1997]).

Although mode finders are useful for initial exploration of a posterior distribution, more sophisticated methods are required for thorough exploration. Figure 5 shows the same prior and posterior distributions as Figure 4, but includes a Monte Carlo sample from the posterior distribution. The Monte Carlo sample can be used to summarize the full posterior distribution and to easily represent marginal distributions of interest. Thus, the histogram of the Monte Carlo sample in the second plot of Figure 5 contains the same information as the plotted marginal posterior distribution. Likewise, the scatterplot of the sample in the third plot conveys the same information as the contours of the joint posterior distribution. The advantage of the Monte Carlo sample is clear when one considers high-dimensional parameter spaces. With a Monte Carlo sample from the joint posterior distribution one can easily plot histograms of the relevant marginal distributions even if the dimension of the joint distribution is in the tens, hundreds, thousands, or larger. Thus, with a Monte Carlo sample we can numerically integrate a distribution that would be impossible to integrate with any other numerical method.

There is a large statistical literature on methods to obtain a Monte Carlo sample from posterior distributions. One method that has proved to be very useful is to construct a Markov chain with stationary distribution equal to the target posterior distribution. Upon convergence, the Markov chain will deliver a (correlated) Monte Carlo sample from the posterior

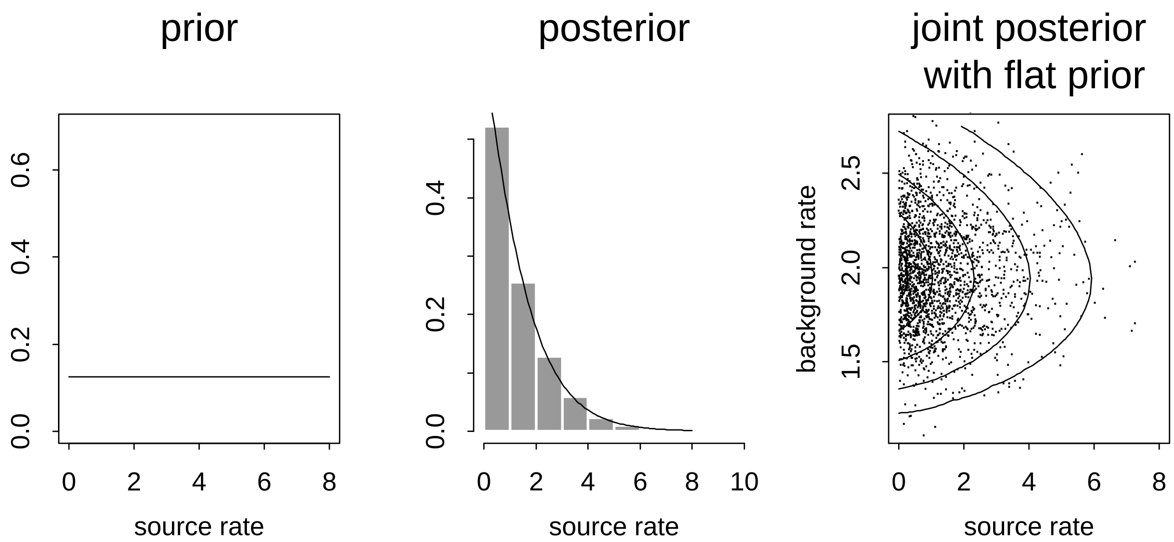

Figure 5: Monte Carlo Samples from the Posterior Distribution. The prior and posterior distributions in the three figures are the same as those in Figure 4. The second two plots illustrate Monte Carlo samples from the posterior distributions. The histogram conveys the same information as plotted marginal posterior distribution. Likewise, the scatter plot in the third plot contains information equivalent to that in the contour plot.

distribution. This technique is known as Markov chain Monte Carlo or MCMC. There are a number of technical issues that arise when using MCMC. It is more difficult to determine when a Markov chain has reached its stationary distribution than when a mode finder has reached a mode. Multi-modal posterior distributions pose extra challenges because Markov chains can easily be caught in one of the modes. This again highlights the advantage of identifying the modes using a mode finder *before* running a MCMC sampler. Severe correlation among the draws can also complicate Monte Carlo integration and evaluation of the posterior distribution. Thus, numerous strategies have been developed to improve the convergence and to reduce the autocorrelation of MCMC samplers. We do not attempt to address these issues here. Instead we point the interested reader to a number of references on the subject (Gelman *et al.* [2003], Gilks *et al.* [1996], van Dyk [2003], van Dyk and Meng [2001]) and describe how the Gibbs sampler can be used to construct a Markov chain with stationary distribution equal to the joint posterior distribution illustrated in Figure 5.

The Gibbs sampler constructs a Markov chain by partitioning the vector of unknown quantities (e.g., model parameters and missing data) into a number of subvectors. Each of these subvectors is updated by sampling from its conditional distribution given the most recent draw of the other subvectors and the observed data. In our simple example, we wish to sample from the posterior distribution of Y_S, Y_B, and the source and background rates given Y and X. We start by sampling Y_S and Y_B given the two rates and the observed data. That is, we stochastically separate Y into source and background counts. It can be shown that this distribution is a simple binomial distribution with probability determined by the relative sizes of the source and background rates and the number of trials equal to Y. In the second step, we sample the rates given Y_S, Y_B, Y, and X. Under this conditional distribution, the rates are independent and both follow gamma distributions, the posterior distribution of a Poisson rate parameter under the standard Bayesian prior distribution. Thus, we divide the unknown quantities into two groups: the missing data and the rate parameters. By iteratively sampling each from their corresponding standard conditional distributions, we construct a Markov chain with stationary distribution equal to the target posterior distribution. The result under our simulated data set is plotted in the histogram and scatter plot in Figure 5.

5. SUMMARY

In this article we have outlined a framework for statistical inference that designs application-specific highly-structured statistical models, uses a Bayesian paradigm for statistical inference, and utilizes sophisticated computational methods such as EM-type algorithms and MCMC. Although space does not permit us to illustrate the power of these methods in real problems, we hope interested readers will refer to the several papers cited in this article that use these methods to solve outstanding challenges in empirical high-energy astrophysics.

Acknowledgments

The author gratefully acknowledges funding for this project partially provided by NSF grant DM-01-04129 and by NASA Contract NAS8-39073 (CXC). This work is the product of the collaborative efforts of the California-Harvard AstroStatistics Collaboration (CHASC, URL: `http://www.ics.uci.edu/~dvd/astrostat.html`) whose members include A. Connors, D. Esch, P. Freeman, H. Kang, V. L. Kashyap, X. L. Meng, T. Park, A. Siemiginowska, E. Sourlas, Y. Yu, and A. Zezas.

References

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological* **39**, 1–37.

Esch, D. N., Connors, A., Karovska, M., and van Dyk, D. A. (2004). An image reconstruction technique with error estimates. *Submitted to The Astrophysical Journal* .

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. Chapman & Hall, London, 2nd edn.

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

McLaughlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. John Wiley & Sons, New York.

Meng, X.-L. and van Dyk, D. A. (1997). The EM algorithm – an old folk song sung to a fast new tune (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological* **59**, 511–567.

Protassov, R., van Dyk, D. A., Connors, A., Kashyap, V., and Siemiginowska, A. (2002). Statistics: Handle with care – detecting multiple model components with the likelihood ratio test. *The Astrophysical Journal* **571**, 545–559.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

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

van Dyk, D. A., Connors, A., Esch, D. N., Freeman, P., Kang, H., Karovska, M., Kashyap, V., Siemiginowska, A., and Zezas, A. (2004). Deconvolution in high energy astrophysics: Science, instrumentation, and methods. *Bayesian Analysis* submitted.

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

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

van Dyk, D. A. and Kang, H. (2003). Highly structured models for spectral analysis in high energy astrophysics. *Statistical Science, to appear* .

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

van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation. *The Journal of Computational and Graphical Statistics* **10**, 1–111.