Look at first parag of sec 2.2, for example... (scaffolding, etc.)

The Astrophysical Journal, 808:137 (24pp), 2015 August 1 doi:10.1088/0004-637X/808/2/137

DISENTANGLING OVERLAPPING ASTRONOMICAL SOURCES USING SPATIAL AND SPECTRAL INFORMATION

David E. Jones[1], Vinay L. Kashyap[2], and David A. van Dyk[3]
[1] Statistics Department, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA; djones@fas.harvard.edu
[2] Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA
[3] Imperial College London, Statistics Section, Department of Mathematics, 180 Queen's Gate, London, SW7 2AZ, UK
Received 2014 November 5; accepted 2015 June 2; published 2015 July 28

ABSTRACT

We present a powerful new algorithm that combines both spatial information (event locations and the point-spread function) and spectral information (photon energies) to separate photons from overlapping sources. We use Bayesian statistical methods to simultaneously infer the number of overlapping sources, to probabilistically separate the photons among the sources, and to fit the parameters describing the individual sources. Using the Bayesian joint posterior distribution, we are able to coherently quantify the uncertainties associated with all these parameters. The advantages of combining spatial and spectral information are demonstrated through a simulation study. The utility of the approach is then illustrated by analysis of observations of FK Aqr and FL Aqr with the *XMM-Newton* Observatory and the central region of the Orion Nebula Cluster with the *Chandra X-ray Observatory*.

Key words: methods: statistical – techniques: image processing – X-rays: stars

1. INTRODUCTION

When two or more sources are situated close enough to each other that there is a substantial overlap of their point-spread functions (PSFs), they pose a many-fold problem to astronomical analysis. The first is to recognize that there is an overlap, the second is to determine the number of distinct sources that are involved, the third is to measure their relative intensities, and the fourth is to separate them sufficiently to be able to carry out useful secondary analyses like spectral fitting and variability analysis. These problems are especially complicated for high-energy photon detectors, since the data are firmly in the Poisson regime, background is often a significant component of the data, and the simplifying approximations of a Gaussian process are usually inapplicable. Many researchers have considered the simpler problem of a single source contaminated by background in the low counts regime (e.g., Kraft et al. 1991; Loredo 1992, pp. 275–297; van Dyk et al. 2001; Park et al. 2006; Weisskopf et al. 2007; Laird et al. 2009; Knoetig 2014; Primini & Kashyap 2014), and have generally found that Poisson-likelihood based Bayesian techniques are well suited to address this category of problems.

However, in the case of multiple sources, progress has been slow, and the choices limited. One could construct approximate measures of intensities of the component sources in the Gaussian regime via matrix inversion (Kashyap et al. 1994), or choose to minimize contamination by limiting the sizes of the apertures to cover only the cores of the PSFs (Broos et al. 2010), or carry out full-fledged 2D spatial modeling. All these are approximate or computationally intensive solutions. An important advance was made recently by Primini & Kashyap (2014), who developed a fully Bayesian aperture photometry method that simultaneously models the intensities of the overlapping sources and the intensity of the background. Their method can be applied to any counts image with multiple overlapping sources, with a practical computational limit of up to five sources. Despite this, most of the problems listed above are still extant.

Typically, X-ray data are collected as lists of events, with each event tagged by its location on the detector, its energy,[4] and its arrival time. Binning the positions into images causes a loss of information that could be alleviated by carrying out the analysis on the unbinned event lists. In such a case, it becomes feasible to disentangle individual events and allocate them probabilistically to the several sources that comprise the data set. In the following, we describe an algorithm that directly addresses three of the four problems listed above: it dynamically determines the number of overlapping sources, measures their intensities, and pools individual events into clusters for which follow-up spectral analysis can be carried out. There are related approaches for longer wavelength data originating from an unknown number of sources, for example, Brewer et al. (2013) and Safarzadeh et al. (2014). The former uses Gaussian process models to identify stellar oscillation modes, and the latter uses simulated *Herschel* images based on *Hubble* data to investigate a disentangling method. The principal difference between these methods and our approach is that they conduct analysis at the pixel level, whereas we probabilistically assign individual photons to sources, a key distinction when analyzing low-count X-ray data.

1.1. Statistical Approach

Here we use finite mixture distributions to model several overlapping sources of photons in a high-energy image. Finite mixture distributions are a useful class of statistical models for data that are drawn from a mixture of several subpopulations; these models are finite in that the (possibly unknown) number of subpopulations is a finite positive integer. (See Titterington et al. 1985 and McLachlan & Peel 2004 for comprehensive discussion of finite mixture distributions, and, for example,

[4] The detector records the pulse height amplitude (PHA), which is roughly proportional to the energy of the incoming photon. These values are often reported as pulse-invariant (PI) gain-corrected PHAs. The distribution of PI for a photon at a given energy is encoded in the detector's Redistribution Matrix Function (RMF). In the following, we use "energy" as a synonym for this recorded PI, and clarify only if there is any ambiguity.

Mukherjee et al. 1998 for a previous application in astronomy.) We take a Bayesian perspective that allows joint inference for the parameters that describe the photon sources (e.g., their number, intensities, and locations), the basic shape of their spectra, and the probability that any particular photon originated from each source, given its recorded location and energy.

Performing inference jointly on the image and spectra improves the precision of the fitted parameters, and also provides more coherent measures of uncertainty than would be available if the spatial and spectral data were analyzed separately. Furthermore, unlike other methods for overlapping sources, our approach quantifies uncertainty about the number of sources. Whether we are ultimately interested in spatial or spectral aspects of sources, identifying the correct number of sources is clearly fundamental. Consequently, a coherent measure of the uncertainty associated with the fitted number of sources is critical to the appropriate interpretation of the fitted parameters of the individual sources.

In some applications inference for the number of sources may seem unnecessary because the sources are clearly identifiable. For instance, the *XMM-Newton* observation of FK Aqr and FL Aqr analyzed in Section 6 has relatively weak background noise, and the sources overlap only moderately. In such cases, the main advantage of the proposed method is that it precisely quantifies the uncertainties associated with the positions, intensities and spectral shapes of the sources. As already mentioned, finite mixture analysis also yields, for each observed photon, the probability that it originated from each inferred source (or the background). In this way we do not deterministically assign photons to sources, but rather properly assess the uncertainty of their origins. This is in contrast to other methods, such as those based on source regions, which deterministically assign photons to nearby sources, and therefore do not properly quantify uncertainties in fitted source parameters.

There is a potential for overfitting in finite mixture models if the number of sources is unknown. This is mitigated when substantial prior information regarding the shape of the PSF or the number of sources is available, or both. In practice, we have detailed information about the PSF, and hence know exactly what the distribution of the recorded photon locations should be for each source. (For point sources this is trivial, but even for extended sources one can easily convolve the source model with the PSF.) Even if the PSF varies across the field, the shape of the photon scatter is completely determined by the location of the source. With this complete knowledge of the PSF, there is only a small risk of overfitting, even with limited prior information regarding the number of sources and their spectral shapes. Indeed, our results do not strongly depend on the choice of prior distribution for the number of sources (see Section 5.1).

Our method is designed for analyzing images composed of an unknown number of point sources that are contaminated with background. However, it can be applied to extended sources, with some modifications to account for spatial variations in intensity and spectra. We also mention that the success of our method depends partly on our ability to use spectra to distinguish point sources from the background, which is possible because a typical X-ray point source spectrum is more peaked than the background. Because of this, we are able to use basic models that capture the rough spectral shape in order to exploit spectral information while conserving computational resources. In the X-ray band, this approach offers substantial improvements over analyses using only spatial data without the cost of precisely modeling the spectra. However, the utility of the method in other wavelength bands will depend somewhat on the nature of the spectra typical of those bands.

Table 1
Symbols Used in This Work

Symbol	Definition
(x_i, y_i)	Location of photon i on the detector
E_i	Energy (PI channel) of photon i
μ_j	True location of source j (2D coordinates)
f_{μ_j}	Point-spread Function centered at μ_j
α_j	Spectral shape parameter for source j (full model)
γ_j	Spectral mean parameter for source j (full model)
α_{jl}	Spectral shape parameter l for source j (extended full model)
γ_{jl}	Spectral mean parameter l for source j (extended full model)
π_{jl}	Weight in [0, 1] of *gamma* component l in source j spectral model (extended full model)
$E_{\min}$, $E_{\max}$	Minimum and maximum detected energy (PI channel)
w_j	Relative intensity of source j ($j = 0$ for background)
K	True number of sources (K_{true} for emphasis)
k	A possible value of K
κ	Prior mean of K
s_i	The true source of photon i (takes the values $0,...,K$, with 0 indicating background)
n_j	True number of photons detected from source j ($j = 0$ for background)
n	Total number of photons detected i.e., $\sum_{j=0}^{K} n_j$
θ_j	Full model parameters for source j i.e., $\{w_j, \mu_j, \alpha_j, \gamma_j\}$
Θ_K	All full model source specific parameters i.e., $\{\theta_0,..., \theta_K\}$ where $\theta_0 = w_0$
L^{full}, L^{sp}, L^{ext}	Likelihood function of the full, spatial-only, and extended full models, respectively
$\psi^{(t)}$	The value of generic parameter ψ in iteration t of the algorithm
x, y, E, s	Vectors of the corresponding photon specific variables (see earlier table entries)
I_A	Indicator function equal to 1 if the event A occurs (e.g., $K = 3$) and 0 otherwise

Note. Notation used only in a single section is defined where it appears and is not included in this table.

The remainder of the paper is organized into seven sections. Section 2 develops the statistical model for isolated sources in the context of high-energy data sets, and describes how these models are combined in the case of multiple sources. Section 3 uses a motivating example to illustrate the method and the benefits of incorporating spectral models for the sources. The beginning of Section 4 gives a brief review of Bayesian inference. The remainder of Section 4 describes the details of the proposed Bayesian analysis and computational approach. Section 5 presents two simulation studies. The first illustrates that inference for the number of sources is insensitive to the choice of prior distribution, and the second more thoroughly studies the advantages of using the spectral data. Sections 6 and 7 present the results of our analysis of observations from the *XMM-Newton* and *Chandra X-Ray Observatory*. The

XMM observation is of the apparent visual binary FK Aqr and FL Aqr, and the *Chandra* observation is of approximately 14 sources from near the center of the Orion nebula. We summarize in Section 8 and computational details are in the appendices. Our *Bayesian Separation of Close Sources (BASCS)* software is available on GitHub at https://github.com/astrostat/BASCS.

2. DATA AND STATISTICAL MODELS

2.1. Structure of the Data

High-energy detectors record directional coordinates (x_i, y_i) and energy E_i for each detected photon, where $i = 1,...,n$ indexes the photons. As mentioned, in practice, the PI channel is used to quantify energy. We denote the full set of spatial and spectral information for n detected photons by (x, y, E). These observed quantities are subject to the effects of the PSF and the spectral Redistribution Matrix Function (RMF). We explicitly account for PSF effects in our model, but model the *observed* spectra, the convolution of the source spectra and the RMF. This strategy does not allow us to fit source spectral models, but does allow us to leverage spectral data to separate the sources. Even though all the attributes are recorded digitally and are binned quantities, we treat them as continuous variables for simplicity, since this binning is at scales that heavily oversample the PSF.

Each photon is assumed to originate from one of several point sources or the background, but its exact origin is unknown. Furthermore, the number of point sources contributing photons to the data, their locations, intensities, and spectral distributions are all unknown. We assume background is distributed uniformly across the image, its strength and spectral distribution are often not known.

2.2. Prototype Model for a Single Source

To introduce notation and our model in the simplest case, we first suppose that the data consist of photons from a single source, with no background contamination. Statistical models specify a distribution for the observed data conditional on a number of typically unknown parameters; we discuss parameter fitting Section 4.3. In the current case, given the unknown position of the source, the detected photons are assumed to be dispersed according to a PSF. That is,

$$(x_i, y_i) \Big| \mu \sim \text{PSF centered at } \mu \tag{1}$$

for $i = 1,...,n$, where $\mu = (\mu_x, \mu_y)$ is the unknown position of the source.[5] We use the same 2D King profile[6] in all the simulations and data analyses presented, see Read et al. (2010)[7] and King (1962). The King profile density, shown in Figure 15 in Appendix C, has heavy tails and is essentially a bivariate Cauchy distribution. Specific parameter values are detailed in Appendix C. More generally, although our method assumes that the PSF is known given μ, it may vary with μ. Furthermore, the PSF may be any function which can be quickly evaluated analytically or numerically. Even in cases where computationally expensive evaluations are required our method is feasible if the PSF is first tabulated.

An important feature of our overall approach is that it also utilizes the spectral data to better assess the likely origin of each photon (when background or more than one source is present). With this end in mind, we propose a simple and computationally practical model for the basic shape of the source spectrum. In particular, we model photon energies using a *gamma* distribution,[8]

$$E_i | \alpha, \gamma \sim \ gamma\ (\alpha, \alpha/\gamma) \tag{2}$$

for $i = 1,...,n$. Here, α and γ are the unknown shape and mean parameters used to describe the basic spectral distribution.[9] The *gamma* distribution allows flexible modeling of positive quantities with right skewed distributions.[10] We emphasize that we aim to summarize the essential shape of the spectral distribution, rather than to model the details of emission lines and other spectral features. This is practical because for high-energy missions, the effective areas are typically small at low and high energies, with a broad peak in the middle; the resulting counts spectrum is reasonably modeled by a single- or double-component *gamma* distribution (particularly since we ignore the RMF). Our goal is to identify sources and divide photons among them, not to carry out detailed spectral analysis. However, our algorithm allows for complex spectral models to be built in if necessary. In addition, and computationally more feasible, once the *gamma* model has fulfilled its role in separating sources, a more sophisticated spectral model may then be used to draw scientific conclusions about the spectral distributions of the disentangled sources. This final stage will be discussed in Section 7.2.

2.3. Prototype Model for Multiple Sources

In practice there are multiple sources and background contamination, hence we introduce a *finite mixture model*. Let K be a parameter denoting the number of sources and $\mu = (\mu_1,..., \mu_K)$ be a $2 \times K$ matrix giving the source positions i.e., $\mu_j = (\mu_{jx}, \mu_{jy})$, for $j = 1,..., K$. If we knew the origin of every photon then, we could model the spatial and spectral data associated with each point source as we did in Section 2.2. We thus introduce a new variable s_i which indicates the source number associated with photon i. Each s_i takes on a value between 1 and K, and we let s denote the vector $(s_1,..., s_n)$. Note that s_i is never actually observed and thus is a *latent variable*. A latent variable is essentially an unknown parameter which is useful for modeling, but may not be of direct interest in itself. Here, we have introduced s_i to simplify the model and to facilitate the algorithms used for inference, which are described in Section 4.3.

[5] The notation $x|z \sim F$ means that, the variable x has the distribution denoted by F if z is fixed and known, and we say that x, given z, follows the distribution F. Throughout, when we use this notation we mean that repeated realizations of x are independent given z.

[6] The `beta2d` model in CIAO/*Sherpa*.

[7] http://xmm2.esac.esa.int/external/xmm sw cal/calib/rel notes/

[8] A standard parameterization of the *gamma* (α, β) distribution yields the density $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} = \frac{\alpha^\alpha}{\gamma^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{\alpha}{\gamma} x}$, $x > 0$. Here α and $\beta = \alpha/\gamma$ are the shape and rate parameters, respectively.

[9] We parameterize the *gamma* distribution using the shape and mean, instead of the shape and rate, for interpretability and because computationally it is best to avoid rates, which in our applications tend to be close to the parameter space boundary at zero.

[10] Indeed, the Exponential and Chi-squared distributions are special cases, and a *gamma* can also closely resemble a (truncated) Gaussian distribution.

As a parameter, s_i is also conditioned on in our spatial model, which now becomes

$$(x_i, y_i)\Big|\big(\mu, s_i = j\big) \sim \text{PSF centered at } \mu_j \tag{3}$$

for $i = 1,...,n$. As an unknown parameter, s_i, plays a role similar to μ; it is "given" in Equation (3). The spectral model can also be straightforwardly generalized to the multiple source case. We have

$$E_i\Big|\big(\alpha_j, \gamma_j, s_i = j\big) \sim gamma\,(\alpha_j, \alpha_j/\gamma_j) \tag{4}$$

for $i = 1,...,n$, where the parameters α_j and γ_j usually differ among the sources.

In addition to point sources, we must model the background. To this end we extend the set of possible values of s_i to include 0. Throughout, symbols indexed by 0 refer to the background. We assume that photons originating from the background are uniformly distributed across the image,

$$(x_i, y_i)\Big|\big(\mu, s_i = 0\big) \sim \text{Uniform} \tag{5}$$

for i such that $s_i = 0$. Instrument effects may cause the background to be non-uniform, and a refinement would be to model such effects.

The background spectrum is also assumed to be flat over the energy range of the source spectra. That is, it is assumed to have a uniform distribution on $(E_{\min}, E_{\max})$, where $E_{\min}$ and $E_{\max}$ are the minimum and maximum photon energy observed. This is a good approximation because the background spectrum is expected to be less peaked than that of a point source.

So far we have not considered the intensities of the different sources and the background. Naturally there should be a parameter for each source, and one for the background, to specify the intensities. Let n_j denote the number of photons originating from source j, for $j = 0,...,K$ (with zero denoting the background), mathematically,[11] $n_j = \sum_{i=1}^{n} I_{\{s_i=j\}}$. We can realistically model n_j as a Poisson variable with some mean m_j, for $j = 0,...,K$. Because these Poisson means vary with exposure time, however, the relative intensities, $w_j = m_j/\sum m_j$, are of more direct interest. Writing $w = (w_0,...,w_K)$, and given n, the Poisson model for $(n_0,...,n_K)$ yields a Multinomial model,[12]

$$(n_0,...,n_K)\big|w, n, K \sim \text{Multinomal}(n; w), \tag{6}$$

where $\sum_{j=0}^{K} w_j = 1$. Under this parameterization, the relative strengths of the sources and background can be succinctly expressed by the vector $w = (w_0,...,w_K)$ without further reference to n. Accordingly, all inference is performed given n, because its value tells us nothing about the number of sources or their parameters.

To complete our introduction of the model we derive the likelihood function, which is the probability of the data expressed as a function of the parameters. The likelihood tells us what values of the parameters are supported by the data and is a key component for principled statistical inference. Let $\mathcal{J}_j$ be the set of photons originating from source j (including $j = 0$) and let $\mathcal{J}$ be the entire collection of observed photons.[13] Also, denote the value of the PSF centered at μ and evaluated at (x, y) by $f_\mu(x, y)$. Lastly, here and throughout, we let $\theta_j = \{w_j, \mu_j, \alpha_j, \gamma_j\}$ denote the parameters associated with source j, for $j = 1,...,K$. Similarly, for the background, we let $\theta_0 = w_0$. We let Θ_K denote all the source (and background) specific parameters i.e., $\Theta_K = \{\theta_0,...,\theta_K\}$. The remaining parameters are K and s. As already discussed, we treat n as fixed, and impose the constraint that the likelihood is zero unless $\sum_{j=0}^{K} n_j = n$. Combining the different parts of the model yields the *full model* likelihood

$$\begin{aligned} L_n^{\text{full}}(\Theta_K, K) &\equiv p(x, y, E|\Theta_K, K, s, n) \\ &\propto \prod_{i\in\mathcal{J}/\mathcal{J}_0} f_{\mu_{s_i}}(x_i, y_i) g_{\alpha_{s_i},\gamma_{s_i}}(E_i), \end{aligned} \tag{7}$$

where

$$g_{\alpha_{s_i},\gamma_{s_i}}(E_i) = \frac{{\alpha_{s_i}}^{\alpha_{s_i}}}{\gamma_{s_i}^{\alpha_{s_i}}\Gamma(\alpha_{s_i})} E_i^{\alpha_{s_i}-1} e^{-\alpha_{s_i}E_i/\gamma_{s_i}}. \tag{8}$$

The maximum energy $E_{\max}$ and the image area are assumed to be known quantities, rather than parameters to be inferred. They are therefore omitted from the likelihood, as are all terms not involving the parameters. In later sections, we compare analyses under the full model to analyses under the *spatial-only model* that does not use the spectral information. The likelihood of the spatial-only model is

$$L_n^{\text{sp}}\big(\Theta_K^{\text{sp}}, K\big) \equiv p\big(x, y\big|\Theta_K^{\text{sp}}, K, s, n\big) \propto \prod_{i\in\mathcal{J}/\mathcal{J}_0} f_{\mu_{s_i}}(x_i, y_i). \tag{9}$$

The notation $\Theta_K^{\text{sp}} = \{w_0,...,w_K; \mu_1,...,\mu_K\}$ represents the set of spatial parameters. Note that, although w does not explicitly appear in either likelihood, the data does nevertheless constrain w in both cases. In particular, the likelihoods indicate probable values of s which in turn indicate probable values of w.

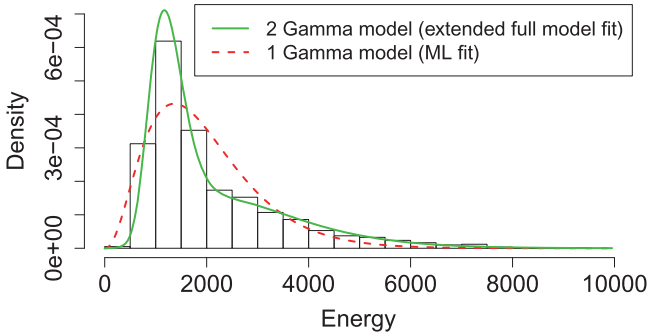

Figure 1. Fitting *gamma* distributions to a counts spectrum. The histogram shows the observed spectrum of the brightest of the *Chandra* sources in the Orion field in Section 7.2 (from one iteration of our algorithm; see Section 4.3), and the curves show *gamma* model fits. The solid line (green) is the extended full model fit of the two-*gamma* spectral model and the dashed line (red) is the maximum likelihood fit of the one-*gamma* model.

[11] I is an indicator function that is zero if its argument is false and one otherwise.

[12] The Multinomial distribution mass function assigns the probability $(n!/(n_0!\cdots n_K!))\prod w_0^{n0}\cdots w_K^{nK}$ to the allocation given by $(n_0,...,n_K)$ of $n = \sum_{i=0}^{K} n_i$ objects into $K + 1$ categories.

[13] Mathematically, $\mathcal{J}_j$ is the set of photon indices associated with source j, that is, $\mathcal{J}_j = \{i: s_i = j\}$, for $j = 0,...,K$, and $\mathcal{J} = \bigcup_{j=0}^{K} \mathcal{J}_j = \{1,...,n\}$.

Conceptually, our method is to apply Bayes rule, briefly reviewed in Section 4.1, to the likelihoods displayed in Equations (7) and (9) to yield a distribution summarizing our knowledge of the parameters given the data, i.e., the joint posterior distribution.

2.4. Extensions of the Spectral Model

In some situations the *gamma* spectral model given by Equation (4) is not sufficiently flexible to capture the spectral shape of the observed sources. For example, Figure 1 shows the observed spectrum of the brightest source in the *Chandra* observation analyzed in Section 7. In particular, the histogram shows the spectrum using one likely assignment of photons produced during the iterations of our algorithm (see Section 4.3). The dashed red curve shows the maximum likelihood fit of the *gamma* distribution to the observed spectrum. The *gamma* does not fit the distribution closely. This causes a problem because inference based on the (misspecified) *gamma* spectral model will suggest there are two sources instead of one in order to better capture the spectral distribution of the source.

To solve this problem, we use a mixture of two *gamma* distributions for a more general spectral model. That is,

$$E_i \Big| \left(\alpha_{j1}, \alpha_{j2}, \gamma_{j1}, \gamma_{j2}, \pi_{j1}, \pi_{j2}, s_i = j\right) \sim \pi_{j1} gamma\left(\alpha_{j1}, \frac{\alpha_{j1}}{\gamma_{j1}}\right) + \pi_{j2} gamma\left(\alpha_{j2}, \frac{\alpha_{j2}}{\gamma_{j2}}\right), \quad (10)$$

for $i = 1,...,n$, where the parameters π_j and $\pi_{j2} = 1 - \pi_{j1}$ are the weights of the two *gamma* components. When this two *gamma* mixture spectral model is substituted for the one *gamma* spectral model in Equation (7) we obtain the following *extended full model* likelihood

$$L_n^{\text{ext}}(\Theta_K^{\text{ext}}, K) \equiv p(x, y, E|\Theta_K^{\text{ext}}, K, s, n) \propto \prod_{i \in \mathcal{J}/\mathcal{J}_0} \left(f_{\mu_{s_i}}(x_i, y_i) \sum_{l=1}^{2} \pi_{s_i l} g_{\alpha_{s_i l}, \gamma_{s_i l}}(E_i) \right). \quad (11)$$

The notation Θ_K^{ext} denotes $\{\theta_0^{\text{ext}},...,\theta_K^{\text{ext}}\}$, where $\theta_j^{\text{ext}} = \{w_j, \mu_j, \alpha_{j1}, \alpha_{j2}, \gamma_{j1}, \gamma_{j2}, \pi_{j1}, \pi_{j2}\}$ gives the parameters associated with source j, for $j = 1,...,K$, and $\theta_0^{\text{ext}} = \theta_0$. The solid green curve in Figure 1 shows the extended full model fit of the *gamma* mixture spectral model. In this example, the mixture of *gammas* quite closely fits the observed spectrum and generally there did not appear to be unwarranted splitting of sources into two in our numerical studies using this model.

Even greater flexibility of the spectral model could be gained by considering a mixture of more than two *gammas*, but this was not necessary in our numerical studies. For the *XMM* data of Section 6, the one-*gamma* spectral model is sufficient in that, for both of the sources, the maximum likelihood fit of the one-*gamma* and the two-*gamma* models resulted in essentially identical fits when using a feasible allocation of photons. In the interest of simplicity, we only use the extended full model when necessary (i.e., in Section 7), and elsewhere use the full model given in Equation (7).

2.4.1. Detecting Spectral Model Inadequacy

A natural question is how one should decide if the source spectral model is inadequate for our purpose of allocating photons among the different sources (and background). There are two potential indications of spectral model misspecification. First, analysis may tend to divide bright sources into two. In particular, when the algorithm (see Section 4.3) finds many instances of sources very close together this indicates that the spectral model is probably not adequate.[14] A second indication of inadequacy of the spectral model comes from considering inference under the spatial-only model. We can inspect the empirical distribution of the photons assigned to a source in iterations of the spatial-only algorithm. If this empirical distribution differs substantially from a *gamma* distribution then it is unlikely that the one-*gamma* spectral model is sufficiently flexible. Clearly, looking at the empirical spectral distribution of a source under the spatial-only model is only reliable if we can accurately assign photons based on spatial data alone. Thus, when possible it is best to select a bright source which is relatively isolated. In the presence of uncertainty about the shape of the spectral distributions to expect then it is usually sensible to use a mixture of at least two *gammas* (or perform analysis several times using mixtures of different numbers of *gammas*). In the presence of uncertainty about the shape of the spectral distributions to expect then it is usually sensible to use a mixture of at least two *gammas* (or perform analysis several times using mixtures of different numbers of *gammas*). One should be cautious of using a spectral model that is too complicated[15] because overfitting may decrease the benefits of modeling the spectral data.

3. ILLUSTRATIVE EXAMPLE

To motivate our method we present a simple simulated data example that illustrates the potential gains made possible by using the full model instead of the spatial-only model. We emphasize that this is a walk-through, designed to clarify the conceptual foundations of the method. A detailed description of our method is in Section 4. The simulated data consist of the spatial and spectral details of photons detected from three weak sources contaminated with background. The spatial data and the spectral distributions used for simulation are shown in Figure 2. The background average is 10 photons per unit square, and the numbers of photons from each source are drawn from Poisson distributions with means 100, 50, and 25, respectively. Thus, the background is very strong and contributes about 85% of the photons over the entire image, and about 40%, 53%, and 66%, respectively, in the three source regions. The true source positions are (1.5, 0), (0, 1), and (−2, 0), and their source regions are approximately circles of radius 1. All three sources have the same PSF, the 2D profile density shown in Figure 15 in Appendix C. The source spectral data is drawn from a *gamma* distribution plotted in Figure 2 (mean parameter 600 and shape parameter 3). In this simple illustration, all the sources have the same theoretical spectral distribution; however, this is not assumed in the fitted model,

[14] Misspecification of the PSF, and specifically under-estimation of its width, could have a similar effect.

[15] We can avoid an overly complicated model by imposing parametric constraints or utilizing substantial prior information to be sure only scientifically plausible spectral shapes are allowed.

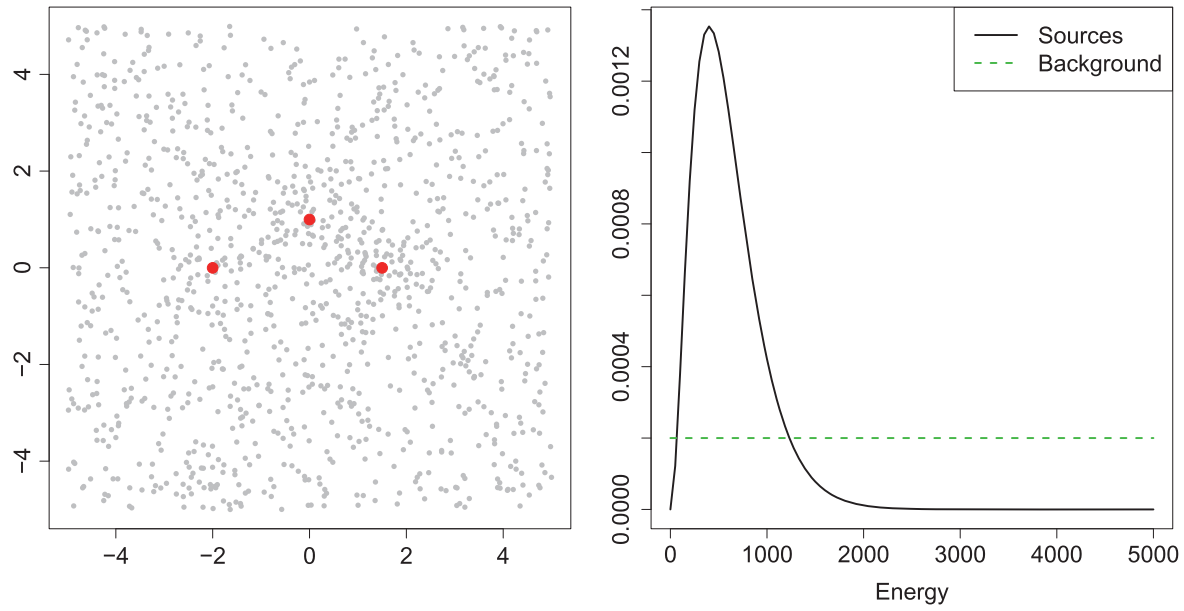

Figure 2. Illustrative simulation setup. Locations of three weak sources are shown as red dots over a scatter plot (left), as also are the adopted counts spectra of the sources and the background (right).

which is based on the likelihood in Equation (7). The theoretical background spectrum was uniform on (0, 5000).

We fit both the spatial-only and the full model to the simulated data. The resulting posterior probability distributions for K are shown in Figure 3. With the spatial data alone it is difficult to detect the faintest source, and consequently the most likely value of K is 2 rather than 3. The situation is much improved when we include the spectral data. The advantage of using the spectral information is due to a greater ability to distinguish the sources from the background, owing to the difference between the source spectra and the background spectrum.

Modeling the spectral data also improves estimation of the other parameters, even if we consider the fits based on $K = 3$. (This is the correct value of K and is identified by the full model but not the spatial-only model.) In Table 2, the first bold column and the last column (not bold) show a summary of the fitted parameters for $K = 3$ under the full model and spatial-only model, respectively. When we consider $K = 3$, the greatest gains of using the full model are in estimating the parameters of the faintest source because this source is the hardest to distinguish from the background when using only spatial data.

In practice, the advantage of using the spectral data for estimating the source parameters is greater than is apparent when we only consider $K = 3$. When confronted with the summary of the fit of K under the spatial-only model (given in the left panel of Figure 3), a researcher is likely to rely on the parameter fits assuming $K = 2$. Thus, it is fair to compare the $K = 3$ fit under the full model with the $K = 2$ fit under the spatial-only model (i.e., the bold columns in Table 2). The latter is clearly substantially worse than the former, because the faint source goes undetected and has no fitted parameters.

The improvement in separation of the sources (and background) can be further understood from Table 3, which summarizes the probability that each photon originated from each source or the background, again under the optimistic assumption that $K = 3$ (see Section 4.3 for additional details). The rows of the table indicate the true photon origin, and the columns indicate the fitted origins. The table entries are the average probabilities, across photons, of the different fitted origins. Ideally the matrices would be identity matrices with "1"s along the diagonal and "0"s elsewhere, but because of the strength of the background many source photons are mixed up in the background. For example, for a photon originating from the leftmost source, the spatial-only model on average assigns probabilities of 0.095 and 0.800 that it originated from the correct source and the background, respectively, reflecting the difficulty in detecting the location of this faint source. Under the full model, the average probability of correct assignment is increased to 0.358, a substantial improvement. Indeed, for each of the three sources, nearly half as many photons are mixed up with the background under the full model. Our improved ability to correctly assign photons under the full model (relative to the spatial-only model) naturally leads to improved estimation of the parameters of the faint source, as illustrated in Table 2. There is a similar effect for the other sources though it is less pronounced because, being brighter, they are easier to detect from the spatial data alone.

4. BAYESIAN MODEL FITTING

4.1. Bayesian Inference

The Bayesian perspective provides a coherent approach for combining all available information to infer the unknown model parameters Θ_K, K, and s. First, our knowledge (or lack of knowledge) as to the likely values of the parameters before seeing the current data is quantified using a *prior distribution*. Once the data are observed, Bayes' Theorem allows us to combine the likelihood and the prior distribution to yield the *posterior distribution* of the parameters. Recall, the likelihood is the probability of the data given the parameters. The posterior distribution expresses our updated knowledge of the parameters after seeing the data. Bayes' Theorem states that, for generic data and parameter vector ψ, the posterior

Table 2
Fitted Parameters Under the Full and Spatial-only Models

	Truth	Full Model		Spatial-only Model			
k	3	**3**		**2**		3	
$P(K=k\|\text{data})$	⋯	**0.95**		**0.85**		0.14	
μ_{1x}	1.5	**1.51**	(1.41, 1.61)	**1.43**	(1.27, 1.58)	1.44	(1.29, 1.59)
μ_{1y}	0	**−0.01**	(−0.10, 0.09)	**0.04**	(−0.08, 0.17)	0.02	(−0.10, 0.14)
μ_{2x}	0	**−0.08**	(−0.20, 0.04)	**−0.09**	(−0.28, 0.12)	−0.03	(−0.22, 0.15)
μ_{2y}	1	**1.11**	(1.00, 1.23)	**0.96**	(0.80, 1.13)	0.99	(0.84, 1.15)
μ_{3x}	−2	**−1.96**	(−2.17, −1.76)	⋯	⋯	−1.37	(−2.40, 0.35)
μ_{3y}	0	**0.06**	(−0.15, 0.27)	⋯	⋯	−0.24	(−1.44, 0.75)
w_1	0.083	**0.068**	(0.057, 0.078)	**0.063**	(0.049, 0.076)	0.062	(0.049, 0.076)
w_2	0.058	**0.064**	(0.053, 0.076)	**0.055**	(0.041, 0.068)	0.052	(0.039, 0.066)
w_3	0.033	**0.028**	(0.019, 0.036)	⋯	⋯	0.017	(0.003, 0.030)
w_0	0.826	**0.841**	(0.826, 0.855)	**0.883**	(0.866, 0.900)	0.868	(0.848, 0.887)
γ_1	600	**536**	(478, 592)	⋯	⋯	⋯	⋯
γ_2	600	**735**	(646, 820)	⋯	⋯	⋯	⋯
γ_3	600	**634**	(397, 826)	⋯	⋯	⋯	⋯
α_1	3	**3.92**	(2.89, 4.97)	⋯	⋯	⋯	⋯
α_2	3	**2.94**	(2.18, 3.69)	⋯	⋯	⋯	⋯
α_3	3	**2.76**	(1.62, 3.82)	⋯	⋯	⋯	⋯

Note. The columns in bold give the fits that would likely be relied upon in practice for the two models. The intervals in parentheses indicate the 16% and 84% posterior quantiles, i.e., Bayesian 1σ equivalent intervals.

Table 3
Photon Allocation Proportions for the Spatial-only and Full Models

		Average Allocation Probabilities							
		Spatial-only Model				Full Model			
Source (True Intensity)	No. Photons in Simulation	Background	Right	Middle	Left	Background	Right	Middle	Left
Background (10/sq)	1001	0.917	0.037	0.033	0.013	0.940	0.022	0.026	0.012
Right (100)	84	0.566	0.354	0.068	0.012	0.318	0.557	0.113	0.012
Middle (70)	67	0.593	0.073	0.303	0.031	0.313	0.122	0.505	0.060
Left (40)	42	0.800	0.034	0.071	0.095	0.431	0.066	0.145	0.358

distribution is

$$p(\psi|\text{data}) = \frac{p(\text{data}|\psi)p(\psi)}{p(\text{data})}, \tag{12}$$

where $p(\text{data}|\psi)$ is the likelihood function and $p(\psi)$ is the prior distribution. The denominator $p(\text{data})$ is simply a normalizing constant which ensures the posterior integrates to one. In our case, the data is (x, y, E) and under the full model $\psi = \{\Theta_K, K, s\}$ so

$$p\big(\Theta_K, K, s\big|x, y, E\big) = \frac{p\big(x, y, E\big|\Theta_K, K, s\big)p\big(\Theta_K, K, s\big)}{p(x, y, E)}. \tag{13}$$

Here, all probabilities are conditional on n but this is suppressed. The likelihood $p(x, y, E|\Theta_K, K, s)$ is given in Equation (7), and the prior distribution $p(\Theta_K, K, s)$ is described in Section 4.2. Referring back to the illustrative example in Section 3, the marginal posterior distribution of K,

$$p(K|x, y, E) = \sum_s \int p\big(K, s, \Theta_K\big|x, y, E\big)d\Theta_K, \tag{14}$$

is displayed in Figure 3. Given the number of unknown parameters, it is not possible to plot their joint posterior distribution, but we can derive and plot the marginal posterior distribution of any one parameter, as in Equation (14) and Figure 3.

4.2. Completing the Model Formulation: Prior Distributions

Following the Bayesian approach, we specify prior distributions for each of the unknown parameters. First, the positions of the point sources are a priori assumed to be independently and uniformly distributed across the image. That is,

$$\mu_j \sim \text{Uniform} \tag{15}$$

for $j = 1,..., K$. In principle, informative priors can be used if prior information on source locations is available. For example, we might set $\mu_j \sim N(\mu_{j0}, \sigma^2_{j0})$, where (μ_{j0}, σ_{j0}), for $j = 1,..., K$, specifies knowledge of the source locations.[16]

[16] The notation $N(\mu_{j0}, \sigma^2_{j0})$ denotes a Gaussian distribution with mean μ_{j0} and variance σ^2_{j0}.

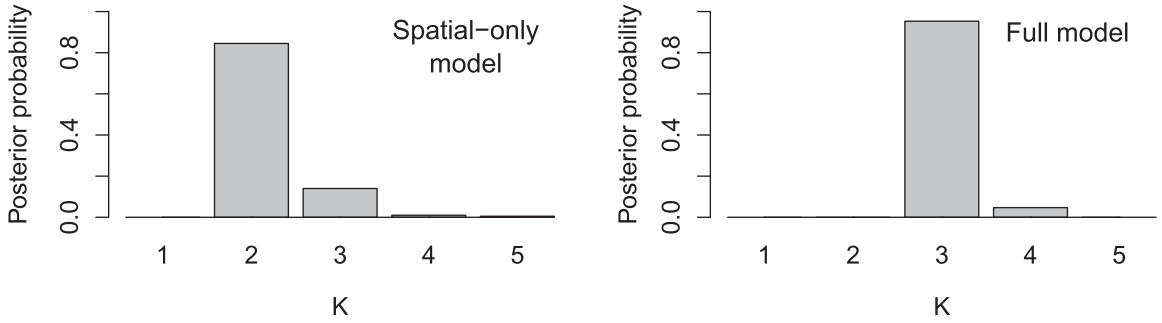

Figure 3. Probability distribution of the number of sources based on the spatial-only model (left) and the full model (right). In this simulation, the true value is $K = 3$.

Next, the vector w, that specifies relative intensities, is given a Dirichlet[17] prior distribution with parameter $(\lambda,...,\lambda)$. A Dirichlet random variable is a probability vector, i.e., it is a vector with non-negative entries that sum to one. We set $\lambda = 1$ throughout. This choice is uniform on the probability vector, but very slightly favors sources of equal size. Indeed, setting $\lambda = 1$ means the Dirichlet prior has as much information as a single photon count added to each source (including a single count added to the background).[18] Regarding the realized vector of source and background counts $(n_0,...,n_K)$, recall that Equation (6) specifies a Multinomial distribution for $(n_0,...,n_K)$, given w, n, and K. Since $(n_0,...,n_K)$ is a function of the parameter (or latent variable) s, Equation (6) is effectively a prior distribution for s.[19]

External information about the number of sources is amalgamated into a prior for K, which we assume to be Poisson with mean parameter κ.[20] Under the Poisson prior, the fitted value of K is relatively robust to the choice of κ because the PSF is completely specified.[21] Indeed, we show in Section 5.1 that the posterior mode for the number of sources may correctly identify the true value of K, even when κ is quite different from K. Therefore, in practice it is adequate to use the Poisson prior for K with κ set to any reasonable guess of the number of sources.

To complete the model specification, we must assign prior distributions for the source spectral distribution parameters α_j and γ_j, for $j = 1,...,K$. Typically there is sufficient data to overwhelm these prior distributions. Thus, we are not overly concerned with the exact form of these priors. For concreteness, however, we mention that one set of priors we use is $\alpha_j \sim gamma(2, 0.5)$ and $\gamma_j \sim \text{Uniform}(E_{\min}, E_{\max})$, for $j = 1,...,K$, where $E_{\min}$ is the minimum observed energy.[22]

To summarize, our prior distribution for the full model parameters Θ_K, K and s is

$$p(\Theta_K, K, s) = p(\mu, \alpha, \gamma, s|K, w)p(w|K)p(K) \propto \left(\prod_{j=0}^{K} \alpha_j e^{-0.5\alpha_j}\right) \prod_{j=0}^{K} w_j^{n_j} \left(\prod_{j=0}^{K} w_j\right)^{\lambda-1} \frac{\kappa^K}{K!}, \quad (16)$$

where μ, α and γ denote $(\mu_1,...,\mu_K)$, $(\alpha_1,...,\alpha_K)$, and $(\gamma_1,...,\gamma_K)$, respectively. The second term on the second line of Equation (16) comes from the Multinomial prior distribution for s. In the case of the extended full model given in Equation (11), the priors for α_{jl}, γ_{jl}, $l = 1, 2$, are the same as those for α_j, γ_j, and the prior for π_{j1} is a Beta(2, 2) distribution,[23] for $j = 1,...,K$. (No prior for π_{j2} is needed because this parameter is determined by π_{j1}, for $j = 1,...,K$.) The prior for the spatial model parameters is Equation (16) without the first term.

4.3. Statistical Computation and Model Fitting

Given the likelihood in Equation (7) and the prior distribution in Equation (16), we can apply Bayes' Theorem to obtain the posterior distribution of Θ_K, K and s (see Equation (13)). The resulting posterior distribution is a complicated function, which we summarize by the low-dimensional marginal distributions as described in Section 4.1 and their means and standard deviations. These summaries are used to estimate the model parameters and their error bars.

We accomplish the necessary numerical integration, e.g., as in Equation (14), using Monte Carlo methods, a cornerstone of statistical computing (Shao & Ibrahim 2000; Liu 2008; Christian & Casella 2011, p. 49–66). The idea of Monte Carlo

[17] The Dirichlet probability density function is $f(p_0,...,p_K) = \left(\Gamma(\sum_{i=0}^{K}\lambda_i)/\prod_{i=0}^{K}\Gamma(\lambda_i)\right)\prod_{i=0}^{K} p_i^{\lambda_i-1}$, for all p_i such that $\sum_{i=0}^{K} p_i = 1$ and $p_i \geqslant 0$ for $i = 0,...,K$, and is zero otherwise. Here, $(\lambda_0,...,\lambda_K)$ is a parameter, and Γ is the gamma function.

[18] Suppose the source counts are observed to be $(n_0,...,n_K)$ and follow a Multinomial distribution with probability vector w. Then, assuming a priori $w \sim \text{Dirichlet}(\lambda_0,...,\lambda_K)$, it can be shown that $w|(n_0,...,n_K) \sim \text{Dirichlet}(n_0 + \lambda_0,...,n_K + \lambda_K)$. Because λ_j is treated just like n_j in this posterior distribution, λ_j can be viewed as a "prior count" and we say the Dirichlet prior is as informative as λ_j counts added to source j, for $j = 0,...,K$.

[19] The parameter w is called a *hyper-parameter* because it appears in the prior distribution of s but is itself of interest and thus has its own prior distribution.

[20] While other priors for K are possible, the Poisson is simple and only moderately informative. Indeed, the equality of mean and variance captures the typical level of prior information we expect, e.g., if we suspect 10 sources, then an analysis yielding between 8 and 12 sources would seem quite reasonable, but we are unlikely to consider, say, 100 sources as a realistic possibility. Even less informative priors may sometimes be desirable, but it generally makes sense to use any reliable prior information that is available to guard against model misspecification. (Prior information about K also helps our algorithm to converge slightly more quickly.)

[21] If the PSF were not fully specified, it would be difficult to distinguish a few sources with a wide PSF from many sources with a narrow PSF. Thus, the fitted number of components of a general finite mixture model can be quite sensitive to the choice of prior on this parameter. Accounting for misspecified PSFs or uncertainties in their calibration is beyond the scope of this work (see Lee et al. 2011 and Xu et al. 2014 for possible strategies).

[22] More generally, if K is large and some of the sources are faint, it may be beneficial to model the distribution of the spectral parameters across the sources. This strategy is known as hierarchical modeling and is known to have statistical advantages in terms of the estimates of the individual spectral parameters. Such hierarchical spectral structures are left as a topic for future work.

[23] For $\alpha, \beta > 0$, the Beta(α, β) distribution density is $f(x) = (\Gamma(\alpha + \beta)/\Gamma(\alpha)\Gamma(\beta))x^{\alpha-1}(1 - x)^{\beta-1}$ for $x \in [0, 1]$, and is zero otherwise. Here, Γ is the Gamma function.

algorithms is to simulate values of the generic parameter ψ from the posterior distribution in Equation (12) to obtain a *Monte Carlo sample* $\{\psi^{(1)},..., \psi^{(T)}\}$. For example, in Figure 3, the height of the bin centered at k is the proportion of the Monte Carlo draws with $K^{(t)}$ equal to k, i.e.,

$$P(K = k|x, y, E) \approx \frac{1}{T}\sum_{t=1}^{T} I_{\{K^{(t)}=k\}}, \quad (17)$$

for $k = 1,..., K$.

A somewhat unusual feature of our model is that the number of parameters is determined by the value of K, the unknown number of sources. This necessarily conditional structure means that it only makes sense to consider the posterior distributions of the other parameters for a given inferred value of K (Park et al. 2008 discuss a somewhat similar conditional inference in the context of locating emission lines). For an illustration of why this is so, consider the intensity w_3 of the "third" source in an image. The parameter w_3 does not have the same interpretation when there are three sources versus four, because what is the "third" source in the first scenario may combine two sources from the latter scenario. In fact, for $K = 2$ the parameter w_3 does not even exist. In general, there is no clear relationship between the parameters under scenarios with different values of K. This prevents us from considering the unconditional posterior distribution of, say, w_3. Instead, we are interested in posterior summaries given a particular value of K, such as $p(w_3|K = k, x, y, E)$. For example, the second row of Table 2 provides an estimate of the posterior mean of w_2 conditional on $K = 3$, under the full model,[24]

$$\hat{w}_2^F(k) = \frac{\sum_{t=1}^{T} w_2^{(t)} I_{\{K^{(t)}=k\}}}{\sum_{t=1}^{T} I_{\{K^{(t)}=k\}}} = 0.080. \quad (18)$$

More generally, for each one-dimensional parameter τ, we calculate the Monte Carlo estimate

$$\hat{\tau}(k) = \frac{\sum_{t=1}^{T} \tau^{(t)} I_{\{K^{(t)}=k\}}}{\sum_{t=1}^{T} I_{\{K^{(t)}=k\}}}. \quad (19)$$

In practice, we choose a value of k at which K has relatively high posterior probability, such as the posterior mode, because otherwise the parameters estimated are unlikely to correspond to properties of real sources. (Indeed, our algorithm does not accurately estimate parameters under unlikely values of K.) We may decide to consider several different values if the posterior of K is not concentrated on one value. This can be useful despite the fact that, as we have mentioned, the number and interpretation of the parameters is not consistent across values of K.

The most popular method for obtaining the Monte Carlo samples needed for estimates such as that in Equation (18) is Markov chain Monte Carlo (MCMC). This an iterative algorithm in which we generate a new value of the parameters $\psi^{(t)}$ at each iteration by drawing from a distribution $\mathcal{G}$ that only depends on $\psi^{(t-1)}$ (and the data) and not earlier members of the Monte Carlo sample. Continuing for T iterations we obtain a sample $\{\psi^{(1)},..., \psi^{(T)}\}$ of correlated parameter values, which is usually called an MCMC chain. Appropriate choice of $\mathcal{G}$ ensures that the sample mimics the posterior distribution in the sense that as $T \to \infty$ the sample empirical distribution approaches the posterior distribution. In implementation, a draw from an appropriate $\mathcal{G}$ is typically achieved through two steps: first a new value of the parameters ψ^* is proposed, and then this value is either accepted or rejected with some probability.[25] The Metropolis–Hastings algorithm (Metropolis et al. 1953 and Hastings 1970) is an example of such an algorithm. The reader is referred to Gelman et al. (2013) for details, including discussion of efficient choices of $\mathcal{G}$ and monitoring of convergence to the posterior distribution (which is usually done by running multiple MCMC chains in parallel and checking that their behavior is sufficiently similar based on some criterion).

In standard MCMC algorithms the parameter space being explored is fixed throughout. In our context this means the number of sources would have to be known. We therefore turn to reversible jump Markov chain Monte Carlo (RJMCMC) algorithms (first introduced by Green 1995), which allow configurations with differing numbers of sources to be explored. There have been a number of uses of RJMCMC in other astronomy contexts, for example, Umstätter et al. (2005), Brewer & Stello (2009), Jasche & Wandelt (2013), and Walmswell et al. (2013). In RJMCMC algorithms, so called "jump" steps update the value of K, the name referring to a jump between parameter spaces (or "models"). These steps are performed by drawing $K^{(t)}$ from a distribution only depending on $\psi^{(t-1)} = (\Theta^{(t-1)}, K^{(t-1)}, s^{(t-1)})$, in the same spirit as ordinary MCMC iterations. Feasible values of the parameters $\Theta^{(t)}$ and $s^{(t)}$ must simultaneously be drawn because their dimension and interpretation change with K. It is this high dimensional sampling that makes RJMCMC challenging. In RJMCMC algorithms, $K^{(t)}$ is only allowed to differ from $K^{(t-1)}$ by at most one. This constraint facilitates the proposal of appropriate parameters $\Theta^{(t)}$ and $s^{(t)}$; RJMCMC moves between configurations by splitting, combining, creating or destroying sources in the model. The standard RJMCMC algorithm for Gaussian mixtures was introduced in Richardson & Green (1997), and Wiper et al. (2001) illustrated RJMCMC for *gamma* mixtures. Our BASCS software essentially combines these two algorithms. Additional details are given in Appendices A and B. For the analyses found in Sections 5 and 7 we specify the number of iterations for which our RJMCMC algorithm was run (which depended on the observed convergence rate and run time). A single iteration of our RJMCMC algorithm consists of one proposal to change K and ten MCMC updates of the other parameters, i.e., the number of MCMC iterations is ten times greater than the stated number of RJMCMC iterations. In Section 6 we fix K and use MCMC, and thus directly specify the number of MCMC iterations. Our standard approach is to run ten RJMCMC (or MCMC) chains to allow monitoring of convergence, but for simplicity the final results are always computed using a single chain.

As discussed in Section 4.2, having detailed information about the PSF means our estimates are insensitive to the prior on K (see also Section 5.1). Knowledge of the PSF also aids computation in that it limits the number of feasible

[24] The superscript F in Equation (17) indicates that the Monte Carlo samples were drawn from the posterior derived under the full model.

[25] An appropriate choice of $\mathcal{G}$ and the corresponding rejection probability to use, to ensure convergence of the sample empirical distribution to the posterior, can be calculated by appealing to the "reversibility condition" (see texts on the theory of Markov chain convergence e.g., Feller 1968).

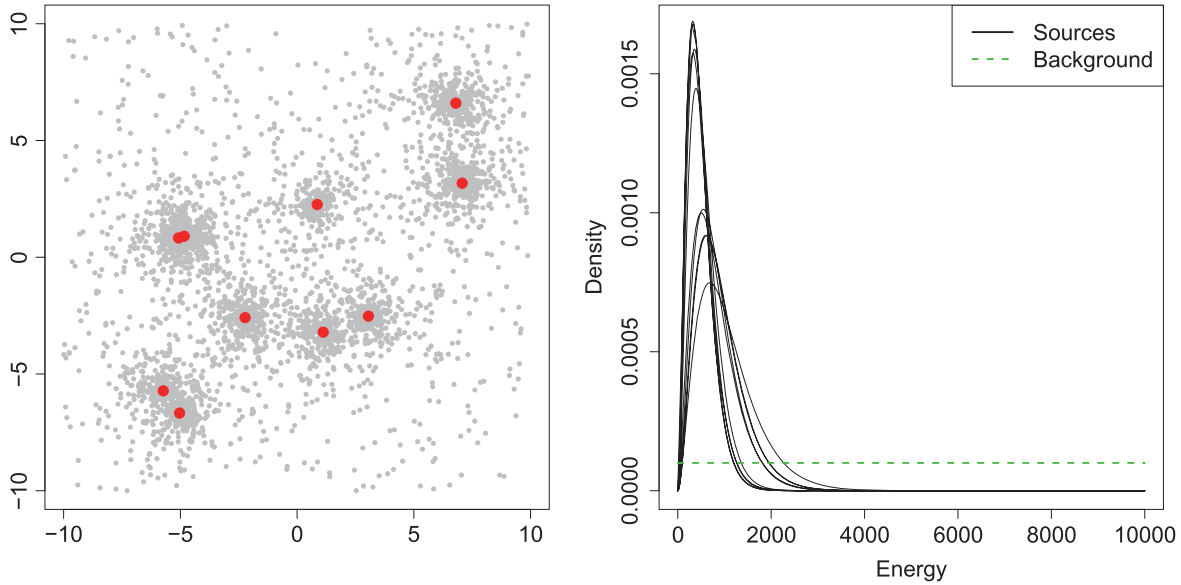

Figure 4. Simulated data set for the 10 source case. The simulated spatial counts distribution (left) and the adopted spectra for each source and the background (right) are shown. The true locations of the 10 sources are marked by large (red) dots in the left plot.

configurations, meaning the RJMCMC algorithm does not have to jump across many values of K. This keeps the number of iterations until approximate convergence comparatively low. Thus, knowledge of the PSF means that, despite the difficulties that are commonly thought to surround mixture models fit with RJMCMC algorithms, our proposed approach is relatively stable and robust. Nonetheless, when the number of sources is clear, MCMC algorithms should be used because they are computationally preferable to RJMCMC algorithms (see Section 6 for an analysis using an MCMC algorithm). In particular, MCMC algorithms are faster per iteration and fewer iterations are needed to obtain enough samples for a given K value of interest. One further challenge is moderate sensitivity to the spectral model, which is the reason why in some applications the *gamma* spectral model must be replaced by the *gamma* mixture spectral model introduced in Section 2.4.

5. SIMULATION STUDIES

Simulated data are used to assess two important aspects of our method: (i) the sensitivity of the fit for K on its prior distribution; and (ii) the performance of the method under a range of different source and background parameters. In the second case, of particular interest is the comparison of inference for the parameters under the spatial-only model and full models (given in Equations (7) and (9)).

5.1. Sensitivity to Prior Distribution on K

To illustrate robustness to the prior on K, we simulated data for a one-source ($K_{\text{true}} = 1$) and a ten-source ($K_{\text{true}} = 10$) reality and drew inference for the number of sources under three different settings of the prior mean κ (1, 3, and 10). Ten data sets were simulated under each reality, each consisting of images of 20 by 20 spatial units and spectral data (simulated under the single *gamma* spectral model). We randomly placed the sources in the central 18 by 18 region of the image, avoiding the edges so that source photons are largely contained within the image. The mean number of photons m_j from source j was chosen randomly from the interval 100 to 500, for $j = 1,...,K$. The mean total number of photons from the background in each data set, m_0, was set to 400, an average of 1 photon per unit square. The number of photons from source j (or the background) was then simulated from a Poisson distribution with mean m_j, for $j = 0,...,K$. Spatial coordinates for the photons were chosen by sampling from the PSF (or the Uniform distribution in the case of the background). We used the King profile density for the PSF; the same PSF is used for analysis of the data sets in Sections 6 and 7.

To complete the data sets, we simulated spectral data under the single *gamma* spectral model (and from a Uniform distribution in the case of the background). We drew the spectral distribution parameters α_j (shape) and α_j/γ_j (rate), $j = 1,...,K$, from the Gaussian distributions $N(3, 0.2^2)$ and $N(0.005, 0.001^2)$, respectively, truncating both distributions to be strictly positive. The resulting spectral parameters are similar to those fitted for the *XMM* data set in Section 6. An example simulated data set is shown in the left panel of Figure 4. The right panel shows the true spectral distributions for the same data set.

For each of the 20 simulated data sets, ten RJMCMC chains were run to assess convergence, but for simplicity only one chain per data set was used in the final analysis.[26] The chains were run for 200,000 RJMCMC iterations, the first 100,000 of which formed the convergence period (or burnin) and were discarded. For each data set, the posterior probability of being in state $K = k$ was calculated, using Equation (17), for all feasible values of k. Figure 5 summarizes the inference for K under the ten-source ($K_{\text{true}} = 10$, left panels) and one-source ($K_{\text{true}} = 1$, right panels) realities, for $\kappa = 1$, 3 and 10 (top, middle and bottom panels, respectively). Recall that κ is the prior mean number of sources. The 25% and 75% quantiles of the posterior probabilities across the ten data sets are indicated for each value of K.

[26] For the purposes of convergence diagnostics, we initialized each chain by randomly choosing between 1 and 20 sources and then deterministically spreading them out around the edge of the image space.

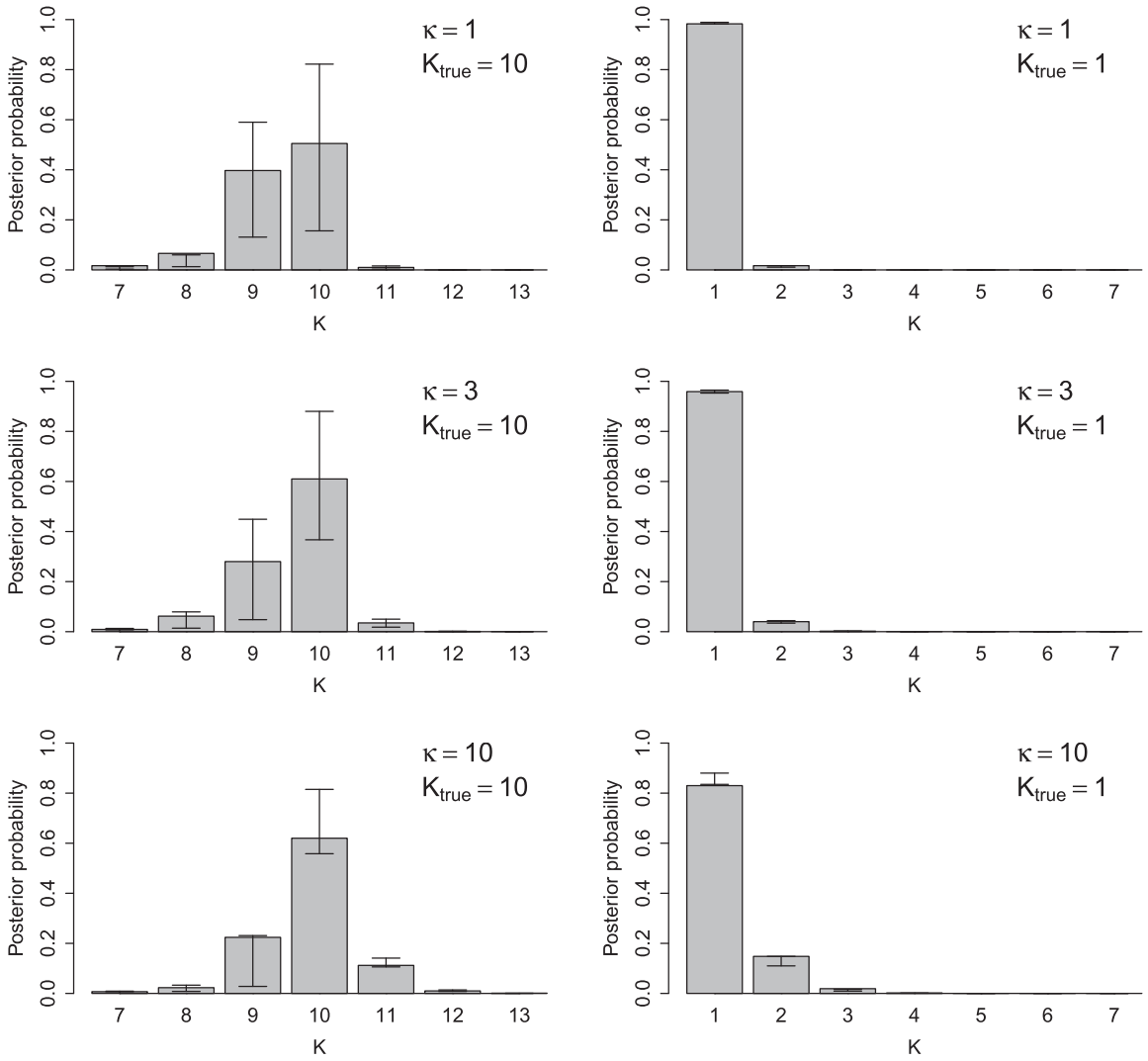

Figure 5. Average posterior probabilities of plausible values of K across ten data sets. Left plots show posteriors for the ten-source reality ($K_{\rm true} = 10$) with prior mean values of $\kappa = 1$, 3, 10 from top to bottom. Right plots show posteriors for the one-source reality ($K_{\rm true} = 1$) with $\kappa = 1$, 3, 10. In each plot, the 25% and 75% quantiles across the 10 data sets are indicated by the vertical error bars for each value of K.

Figure 5 shows that, for the ten-source reality, the posterior probability is concentrated around $K=$ 9–11, regardless of which of the three values of κ is used. Indeed, the prior probability of ten sources specified by the prior with $\kappa = 10$ is nearly 1.25 million times that of the probability specified by the prior with $\kappa = 1$. Despite the difference in the prior probability as a function of κ, the posterior probabilities of $K = 10$ are quite consistent; the average (across simulations) differs by only about 0.1 (comparing $\kappa = 1$ with $\kappa = 10$, see Figure 5). In other words, there is about a 1.25 multiplicative increase in the posterior probability of ten sources when the prior mean is changed from $\kappa = 1$ to $\kappa = 10$. This modest difference in posterior probability is acceptable as it is unlikely that prior information would allocate the truth 1.25 million to one odds.

There are appreciable differences among the simulated data sets as indicated by the quantiles in Figure 5. This is to be expected because the source positions and intensities are chosen randomly. Some of the simulated data sets have two sources very close to each other, making it hard to determine that they are distinct. In some cases, it is possible to separate these very close sources based on the spectral data (using the full model), i.e., if the spectral data appear to come from two *gamma* distributions rather than one. However, in other cases it is difficult to separate such nearby sources, even with the spectral data. Indeed, checks confirmed that data sets with sizeable posterior probability at $K = 9$ under the full model include overlapping sources that cannot be separated by eye and have similar spectral distributions. Posterior probability at K values of 11 and above appear because chance clusters of photons are sometimes mistaken for separate sources. The precise location of these "ghost" sources, however, is highly erratic across RJMCMC iterations. There is limited evidence for them in the data and thus wide error bars for their "locations" in the posterior distribution.

Inference is also robust to the choice of κ under the one-source reality ($K_{\rm true} = 1$). The posterior mode is clearly $K = 1$ for all three values of κ. Owing to the skewness of the Poisson density, the difference in prior probability of $K = 1$ across the different κ values is less dramatic than that for $K = 10$. When $\kappa = 1$ the a priori probability of $K = 1$ is around 800 times that when $\kappa = 10$. Consequently, the difference in posterior probabilities is also less noticeable. Indeed, the qualitative difference in the posteriors under $\kappa = 1$ and $\kappa = 10$ is marginal, see Figure 5.

Our key conclusion is that the posterior probability of the true number of sources K seems insensitive to the prior probability assigned to K, at least when using the Poisson prior. Consequently, the value of κ only needs to be in the region of the true number of sources in order for the fit for K to be reasonable. These conclusions match our intuition that knowing the precise PSF statistically constrains the mixture model sufficiently for the data to drive the fitted values of the parameters. Our simulations are representative of typical data sets, but establishing similar conclusions for smaller datsets may require more studies. A data set could also be larger than those in our simulations, but as κ (and K) increases, greater Poisson variance means that the absolute deviation of κ from the true number of sources has progressively less influence on posterior inferences. (Intuitively, it is more reasonable to a priori suspect 101 sources when there are 110, than to suspect 1 when there are 10.) In our context, prior information typically consists of previous observations, possibly from a different wavelength band. Therefore, it can be assumed that the information is quite reliable and gross prior "misspecification" is unlikely. Clearly, priors other than the Poisson distribution can be considered if a more diffuse prior distribution is desired.

5.2. Utility of the Spectral Model

Here we investigate the performance of our model and methods for a range of background intensities, source separations and relative source intensities. We compare the performance of the spatial-only and full models. For simplicity, we simulated data for a two-source ($K_{\rm true} = 2$) reality. In each simulation, the number of photons from the background and the number from each source were drawn from Poisson distributions with respective means m_0, m_1, m_2. We set $m_2 = 1000$ and $m_1 = m_2/r$, for $r = 1, 2, 5, 10, 50$. We refer to r as the relative intensity of the two sources. To set m_0 and quantify the strength of the simulated background in an astronomically meaningful way we define a source region in terms of the PSF. Specifically, we again use the King profile PSF and define the source region as the region with PSF greater than 10% of its maximum. (The King profile density has no finite moments). We next define q to be the probability that a photon from a source falls within its source region and set the background per source region to be $m_0 = bqm_2$, for $b = 0.001, 0.01, 0.1, 1$. That is, the mean number of background photons in the faint source region was varied between 1/1000 and 1 times the mean number of photons from the faint source falling in the same region. As we shall discuss and unsurprisingly, the faint source was difficult to locate in data sets that were simulated with $b = 1$ and less so for those simulated with $b = 0.001$. Finally, the separation of the two sources was set to be 0.5, 1, 1.5, or 2 distance units. These separations can be interpreted using the fact that our source regions are approximately circles of radius 1.

Spectral data was also simulated for source and background photons. An aim of this simulation study aims is to investigate how much using the spectral data improves the fitted parameters. Since sources can only be distinguished by their spectra if their spectra are different, we used different spectra for the two simulated sources; specifically we set $\alpha_1 = 3$, $\gamma_1 = 600$, $\alpha_2 = 6$, and $\gamma_2 = 1500$.

In summary, our simulation study consists of a $5 \times 4 \times 4$ grid of configuration settings ($r = 1, 2, 3, 5, 10, 50$; $b = 0.001, 0.01, 0.1, 1$; and source separations of 0.5, 1, 1.5, 2). One hundred data sets were simulated for each of the resulting 80 configurations, and analyzed using first the spatial-only model and then the full model. In particular, for each data set our algorithm was run for 20,000 RJMCMC iterations, the first 10,000 of which formed the convergence period (or burnin) and were discarded.[27] The median posterior probability of two sources is shown in Figure 6 for each of the different simulation settings. The left and the right panels correspond to the spatial-only and full models, respectively. We use the median posterior probability across the 100 simulated data sets because in a few simulations the faint source is unusually bright or unusually faint, which noticeably effects the mean posterior probability of two sources. Nevertheless, summaries based on the mean posterior probability are qualitatively very similar, albeit with slightly more noise. We have organized the results by background intensity because in practical applications background is often well determined.

In images simulated with relative intensity 50 the posterior probability of two sources tends to be low. This is because $r = 50$ corresponds to a faint source intensity of $m_2 = 20$, while the brighter source has intensity $m_1 = 1000$. Thus, the faint source is typically not bright enough to be distinguished from noise; its photons can be adequately explained as a random cluster formed of photons from the brighter source or the background. In this case the posterior probability peaks sharply at $K = 1$. The spatial-only model is more likely to mistake a cluster of background photons for *a* faint source and therefore, in the case of $r = 50$ and small source separation, typically gives slightly higher posterior probabilities of two sources than the full model (but the probabilities are still very small). For less extreme relative intensities, using the full model increases the posterior probability of two sources. The improvement is particularly noticeable for relative intensities 5 and 10, regardless of the background strength. The spectral distribution of source counts reduces the plausibility that the faint source is just a cluster of photons from the background or the bright source. When both sources are bright and reasonably separated both the spatial-only and full models give high posterior probability at $K = 2$.[28]

To fit the source parameters, we fix K at its posterior mode value and use Equation (19). Although the fitted parameters of the bright source are always accurate, those for the faint source may be poor, especially if the posterior mode of K is at 1 or if "ghost" sources have appreciable posterior probability. The accuracy of the faint source's fitted parameters essentially follows the pattern seen in Figure 6. When the real faint source is very weak or located too close to the bright source, then a fitted second source (when the posterior mode of K is greater than 1) is likely to be a "ghost" consisting mainly of a cluster of photons from the background or the bright source. In which

[27] This is a relatively small number of RJMCMC iterations, but since our simulated data sets were quite small images each including only two sources, we found it to be sufficient.

[28] One curiosity, present in the left panels of Figure 6 (spatial-only model), is that when both of the sources are reasonably bright, greater median posterior probability of two sources is obtained when the background is *stronger*. This phenomenon occurs because, in the presence of strong background, deviations between the PSF and the observed counts are difficult to detect, whereas, with weak background, such deviations may be attributed to spurious additional sources. (Indeed, the posterior probability of $K = 3$ is typically greater at low background levels than at high background levels.) When the full model is used this effect is diminished. The curiosity is not qualitatively important because the bright sources are well identified in all cases. Clearly weaker background is preferred as it improves the chance of detecting (real) faint sources.

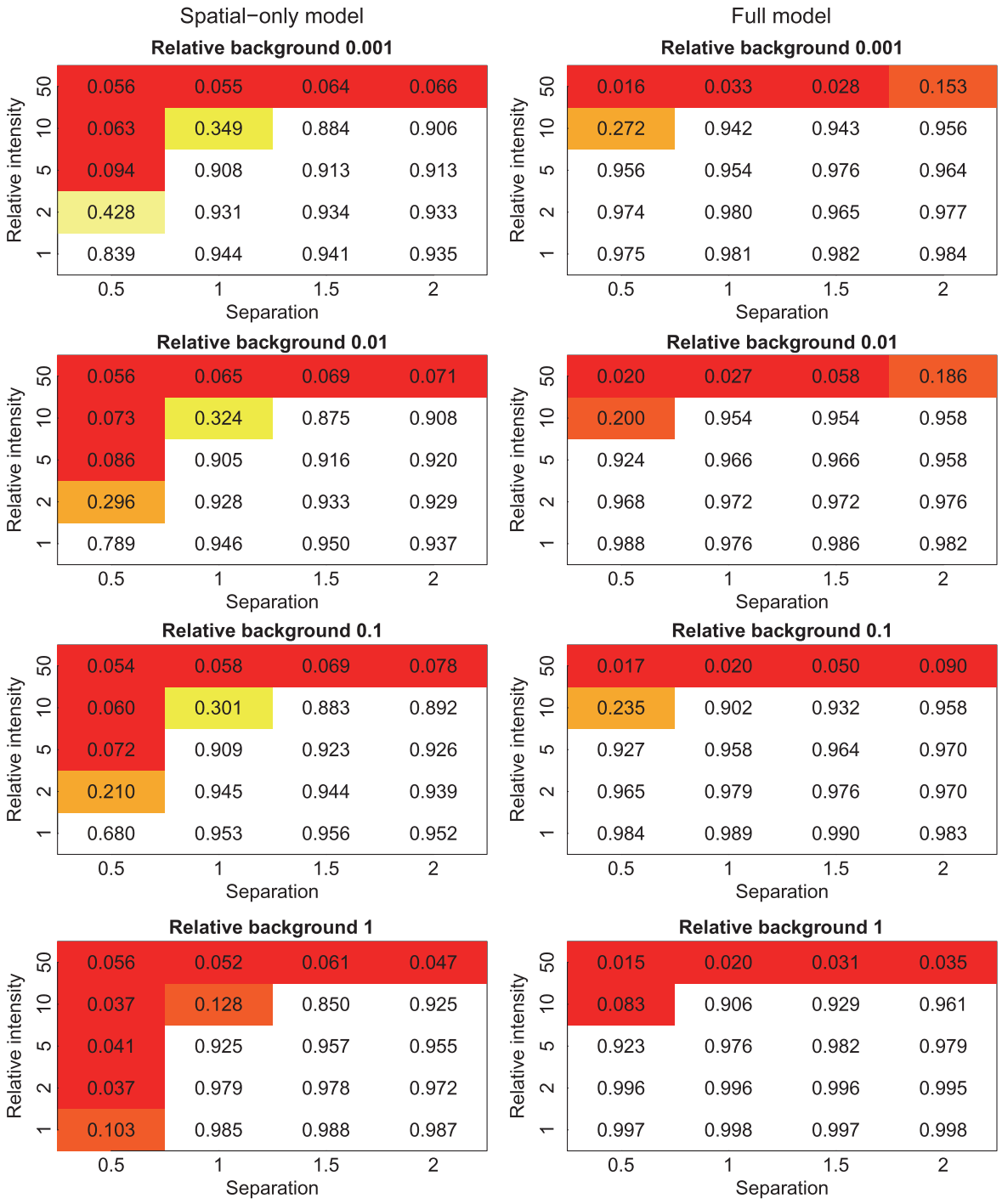

Figure 6. Exploring the sensitivity of our algorithm to source separation, relative strengths, and background level. The median posterior probability of $K = 2$ across the 100 simulations is shown; $K_{\rm true} = 2$ in all cases. The results from the spatial-only model (left column) and the full model (right column) are both shown. Red indicates probabilities less than 0.1, and white indicates probabilities greater than 0.5. (Intermediate colors indicate probabilities between 0.1 and 0.5.)

case, its fitted parameters bear little resemblance to those of the true faint source. This is illustrated in Figure 7, which shows the mean (conditional on $K = 2$) posterior locations of the two sources for all 100 data sets under each configuration of simulation settings. Crosses indicate the true locations of the sources. The mean posterior locations of the bright source (red dots) are not always visible in the plots because they are often in the middle of the red crosses. The location of the bright source becomes slightly harder to fit as the intensity of the faint source increases. (This is at least partly because the background intensity is proportional to the faint source intensity.) The size of the dots indicate the posterior probability of two sources.

The full model again yields more accurate fits. The fitted locations of the faint source (blue dots) center around its true location (blue crosses) for $r \leqslant 10$, even when the source separation is small. For the spatial-only model there is more scatter. Under both models, when $r = 50$ we can see that many of the fitted faint source locations correspond to spurious clusters of photons surrounding the bright source. As the

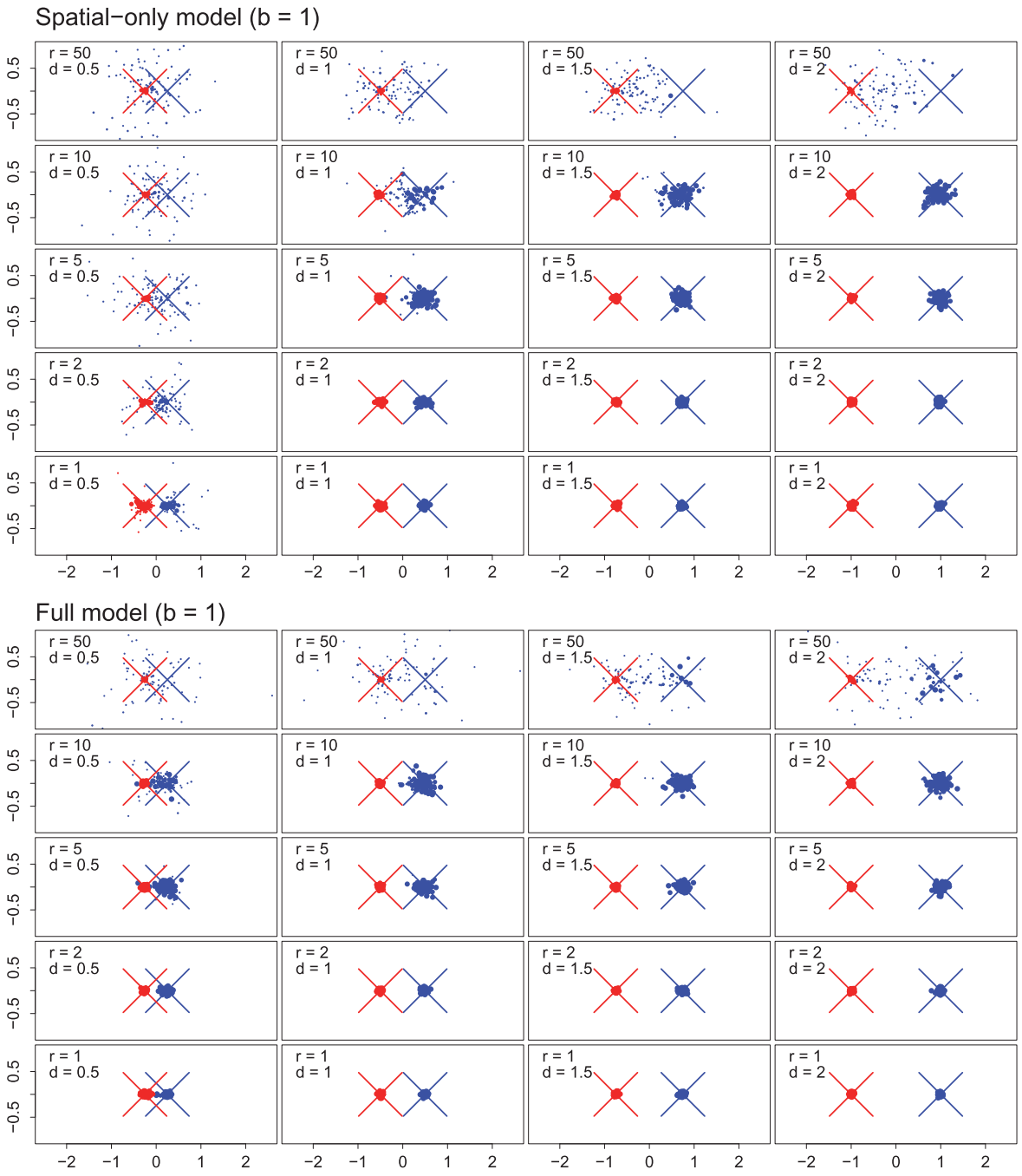

Figure 7. Sensitivity of location determination as a function of source separation, relative strength, and background level. The simulation is the same as that in Figure 6. Mean posterior locations of two sources for each of 100 simulations, under the spatial-only model (top 20 plots) and the full model (bottom 20 plots). Red and blue dots give the mean posterior locations for each simulation of the bright and faint sources, respectively. The large "X"s of corresponding color indicate the true locations. The diameters of the dots are proportional to the posterior probabilities of two sources. The relative background, relative source intensity, and source separation are indicated by b, r, and d, respectively.

separation increases some of the fitted faint source locations are halfway between the true locations of the two sources. This occurs when the posterior distribution of the faint source x-coordinate is bimodal, a spurious cluster of photons and the real faint source both being supported as possible second sources. In an actual analysis this bimodal behavior would be apparent from inspection of the posterior draws of the source location. For $r = 50$ and large separation, the full model sometimes accurately fits the faint source location, but the spatial-only model never does. The behavior of the other fitted parameters follows the same pattern illustrated in Figure 7 because the fitted source locations indicate how well photons are allocated to the correct source. This is confirmed by inspecting tables of the mean (or median) squared error of each parameter (not shown).

The number of Monte Carlo samples used in estimating the mean posterior locations (conditional on $K = 2$) is determined by the posterior probability of two sources, and thus is indicated by the size of the dots. Very small dots may have non-negligible Monte Carlo error, i.e., the true posterior mean

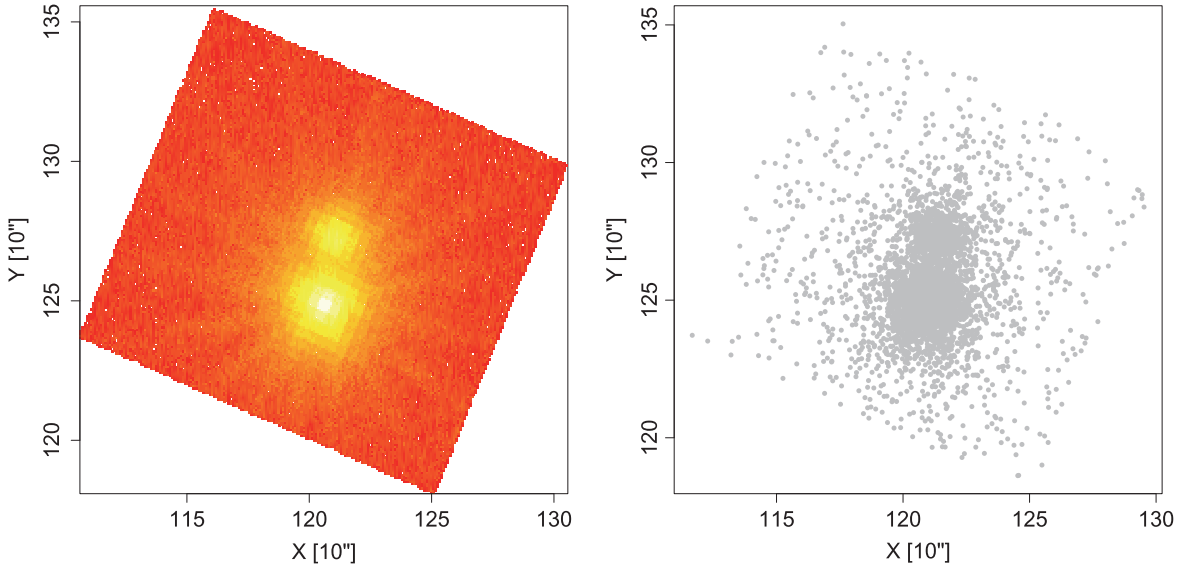

Figure 8. Visual binary FK and FL Aqr observed with *XMM-Newton* (FK is the brighter source at bottom). The XMM obs_id is 0151450101. Shown is a counts image with 10″ bins and arbitrary origin (left), and a scatter plot of a subset of 6000 events over a 5 ks subexposure (right).

location (conditional on $K = 2$) may be somewhat inaccurately approximated. This is because applying Equation (18) for each parameter does not accurately compute the mean of $p(\Theta_K|K, x, y, E)$ for values of K that have low posterior probability.[29] However, in practice, when the number of sources is unknown, it makes sense to only consider values of K with relatively high posterior probability. Furthermore, one typically checks the level of Monte Carlo error for the values of K of interest, by running multiple chains. Large variation in the parameter estimates across the chains indicates high Monte Carlo error. In which case, one should run the chains longer in order to obtain a larger Monte Carlo sample.

6. APPLICATION I: *XMM* DATA SET

We now apply the spatial-only and full models to an XMM observation (obs_id 0151450101) of the apparent visual binary FK Aqr and FL Aqr. The data consist of the spatial and spectral information of around 540,000 photons detected during a 47 ks exposure. The spatial data is displayed in Figure 8 as both an image (left) and a scatter plot (right), and the spectrum is plotted in Figure 9. The moderate overlap of the sources and high counts make this a good test of our model. In particular, we expect that the spatial-only model and full model analyses to be similar (for the spatial parameters) because of the large amount of spatial information. Furthermore, since the data clearly indicate two sources, we can concentrate on verifying that our model yields sensible posterior inference using standard MCMC. (This gives draws from the joint posterior for a fixed number of sources and therefore results in inference that is simpler to interpret than inference resulting from RJMCMC.) Use of the more complicated RJMCMC analysis is reserved for the *Chandra* data set in Section 7 because there is non-negligible uncertainty in K for that data set.

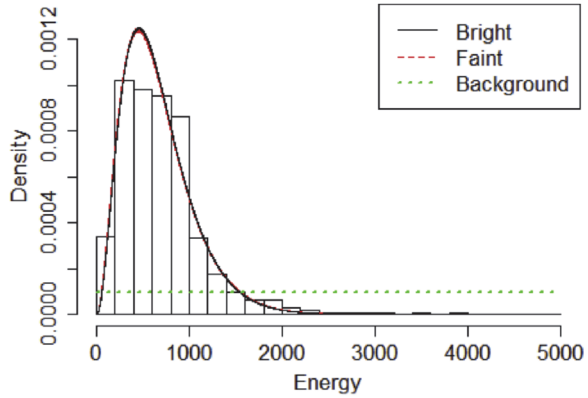

Figure 9. Histogram of the spectral data in the *XMM* observation of FK Aqr and FL Aqr. Plotted are 1000 spectra for the bright (solid black lines) and faint (dashed red lines) sources, each corresponds to a posterior sample of the spectral parameters. (The posterior variance is small on this scale.) The background spectra is shown by the dotted green line.

In the image shown on the left of Figure 8 the sources seem to have faint "spokes." Approaches for modeling these features are suggested in Read et al. (2010) and Read & Saxton (2012)[30], but we use the unaltered King profile PSF for simplicity. As mentioned in Section 2.1, the spatial data are binned when recorded on the observatory LCD screen. However, the bins are small in comparison to the *XMM* PSF so our use of a model that treats the data as unbinned is reasonable. (See Section 8 for further discussion.)

For the spatial-only model and the full model, ten MCMC chains (with K fixed at 2) were run for 20,000 MCMC iterations, the first 10,000 of which formed the convergence period (or burnin) and were discarded.[31] The large amount of data means that the source locations are precisely fit by both models, as can be seen in Table 4. However, the posterior mean

[29] We could instead fix $K = 2$ and run a standard MCMC algorithm to obtain a large enough posterior sample to accurately fit the mean posterior locations. We do not pursue this strategy because the fitted parameters that are conditional on unlikely values of K are of little practical use.

[30] http://xmm2.esac.esa.int/external/xmm sw cal/calib/rel notes/

[31] Note that, since we used standard MCMC and there are only two bright sources, the number of MCMC iterations until convergence was relatively small.

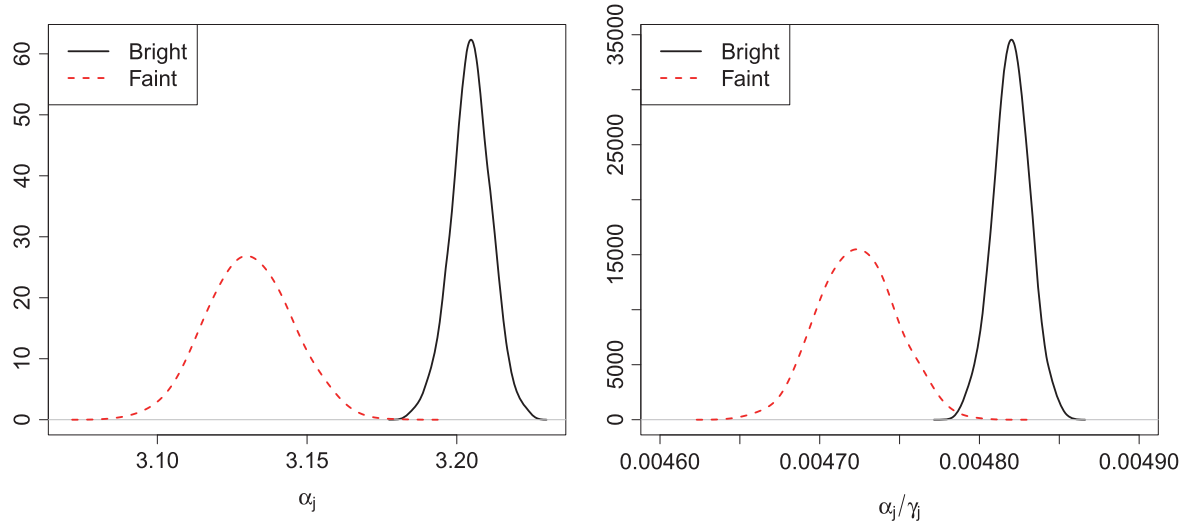

Figure 10. Posterior distributions of the parameters of the *gamma* distributions used to model the spectra of FK Aqr and FL Aqr. The posterior distributions of the shape and rate parameters are shown in the left and right panels, respectively.

Table 4
Posterior Means Under the Spatial-only Model and the Full Model

	Spatial-only Model		Full Model	
μ_{1x}	120.974	(120.973, 120.975)	120.973	(120.973, 120.974)
μ_{1y}	124.873	(124.873, 124.874)	124.873	(124.872, 124.874)
μ_{2x}	121.396	(121.394, 121.398)	121.397	(121.395, 121.399)
μ_{2y}	127.319	(127.317, 127.321)	127.326	(127.324, 127.328)
w_1	0.717	(0.716, 0.718)	0.732	(0.731, 0.732)
w_2	0.182	(0.181, 0.182)	0.189	(0.189, 0.190)
w_0	0.102	(0.101, 0.102)	0.079	(0.079, 0.079)
γ_1	⋯	⋯	664.86	(664.43, 665.30)
γ_2	⋯	⋯	662.78	(661.78, 663.87)
α_1	⋯	⋯	3.205	(3.199, 3.211)
α_2	⋯	⋯	3.131	(3.118, 3.144)

Note. The parenthetic intervals are 1σ error bars computed using 16% and 84% posterior quantiles.

of the relative intensity of the background is about 20% lower for the full model. This is presumably due to a greater ability to separate source and background counts with the additional information given by the spectral data. In particular, photons from the sources can be found across the entire image so there is a tendency to over-estimate the background intensity without some non-spatial way of distinguishing its photons from those of the sources.

Until now, it has not been possible to distinguish the spectral distributions of these two sources. Conventional fitting of the spectra extracted from non-overlapping source regions give statistically indistinguishable results, with identical column density $N_{\rm H} \approx 1.0-1.6$ (10^{20} cm^{-2}), double temperature components $kT_1 \approx 0.25-0.26$ (keV), $kT_2 \approx 0.78-0.82$ (keV), and metallicities $Z \approx 0.12-0.14$. This remarkable coincidence could be attributed to strong contamination of FL Aqr by photons from FK Aqr. Our algorithm, which eliminates such contamination, can answer the question of how similar the two sources are. Of course, a comparison of the source spectra shapes is only possible using the full model. Figure 9 shows 1000 spectra sampled from the posterior distribution[32] for the bright (black solid lines) and faint (red dashed lines) sources; for each source, all 1000 spectra are very similar and so appear as a single curve. We observe that the bright source spectra are very similar to the faint source spectra, which is consistent with the difficulty in distinguishing the spectral distributions of the two sources in previous analyses.

Although the overall shapes of the two spectral are similar (Figure 9), we can distinguish them by examining the parameters of their underlying *gamma* distribution. Figure 10 plots the posterior distributions of these parameters for the two sources and shows that they clearly differ. We have plotted the shape and rate parameters, because the shape and variance differ more than the mean. The posterior distributions in Figure 10 indicate that there is very little uncertainty in the spectral parameters; the intervals in Table 4 convey a similar message. This precision is obtained because of the large amount of data combined with the fact that our method properly accounts for uncertainty in photon origins and jointly fits spectral and spatial parameters. Although our analysis is only physically accurate to the extent that the source spectra can reasonably be modeled with *gamma* distributions, it nevertheless provides evidence that the spectra do differ in some way. More detailed conclusions would be possible with a physics-based spectral model that accounts for emission lines and other spectral features. A possible extension of this work is to replace the *gamma* spectral model with a more complete model. A computationally less intensive approach is described in Section 7.2.

7. APPLICATION II: CHANDRA DATA SET

We analyze a *Chandra* observation of the Orion Nebula Cluster using the spatial-only model and the extended full model given in Equation (10). The extended full model is used because the full model is not sufficiently flexible to capture the shape of the source spectra, as explained in Section 2.4. The specific data set we analyze is a subset of ObsID 1522 that omits the central source, a region where the PSF is distorted

[32] To reduce correlation, every 10th sample of the original 10,000 stored MCMC samples of the spectral parameters was used.

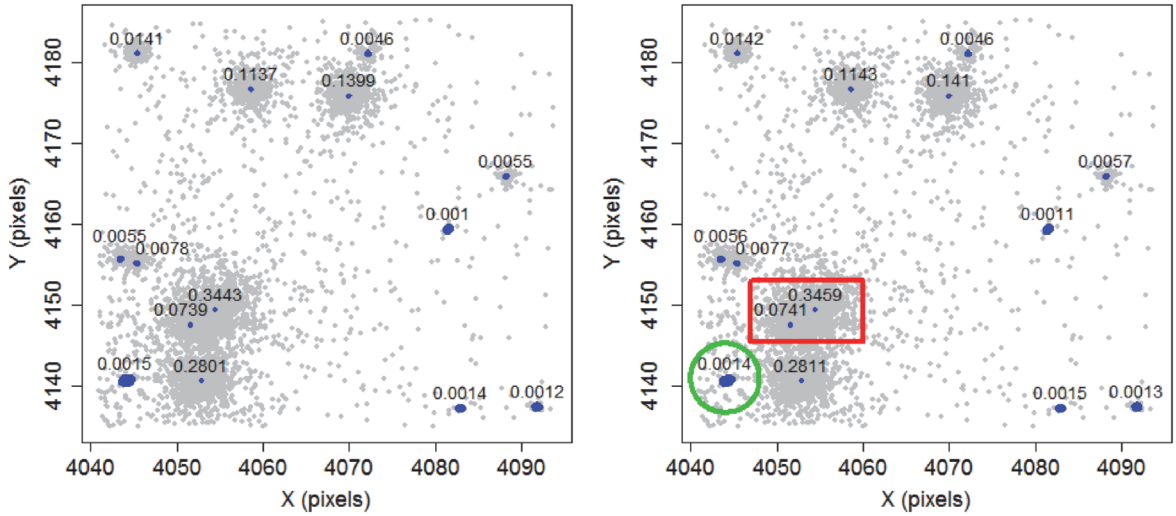

Figure 11. *Chandra* observation of a crowded field near the center of the Orion Nebula Cluster. This field is approximately 25″ × 25″ in size, and is centered at (R.A., decl.)=(5:35:15.4,-05:23:04.68). Shown in blue are approximate 90% posterior credible regions for source locations, under the spatial-only model (left), and the extended full model (right). The figures next to the regions indicate the estimated relative intensities. The credible region of the source with the largest location uncertainty is circled in green (right panel). The red rectangular box encloses two overlapping sources (right panel) for which we carry out a detailed follow-up spectral analysis (Section 7.2).

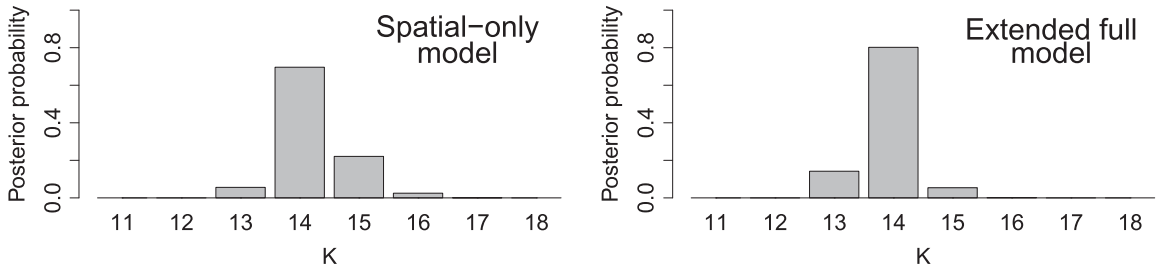

Figure 12. Number of sources detected in the analysis of the *Chandra* observation in Figure 11. Posterior of K based on the spatial-only model (left) and the extended full model (right).

due to strong pile-up (Figure 11). The data include events that occurred within the first 20 ks of the observation, of which there are ≈14,000.

7.1. Analysis Using the Spatial-only and Extended Full Models

For both models, ten RJMCMC chains were run for 150,000 RJMCMC iterations, the first 100,000 of which formed the convergence period (or burnin) and were discarded. The posterior distribution of the number of sources, under the spatial and extended full models, is displayed in Figure 12. The mode of both posteriors is at 14. However, the spatial-only model shows slightly more uncertainty, and some support for 15 sources.

As mentioned in Section 4, K determines the number and meaning of the other model parameters and therefore we must condition on a value of K to draw meaningful inferences for them from the RJMCMC output. Figure 11 shows 90% posterior credible regions (blue) for the locations of the sources under the two models, given $K = 14$. Each credible region shows an area which has 0.9 posterior probability (given $K = 14$) of containing the location of the relevant source, i.e., an integral of the posterior distribution of the source location (given $K = 14$) over this area would evaluate to 0.9. The credible regions look to be similar under the two models. The estimated relative intensities also appear in Figure 11 and are also similar, but are slightly lower under the spatial-only model for most sources. This is due to a higher estimate of the relative background intensity under the spatial-only model (0.0053 versus 0.0006 under the extended full model[33]). Table 5 gives the the posterior mean fit of the source locations and relative intensities under the extended full model for $K = 14$. The detected sources are also matched to the source catalog from the *Chandra* Orion Ultradeep Project (COUP; Getman et al. 2005).

Other observations of Orion suggest that the source circled (in green) in the right panel of Figure 11 is a genuine source. Its location is more uncertain than other sources because it is more difficult to detect. Indeed, with an estimated intensity

[33] The background is likely inaccurately estimated by both models because the King profile PSF that we use is an approximation to the *Chandra* PSF; the latter is more concentrated at its center. Thus, in our analysis, too many photons are allocated to the wings of the sources, deflating the background. That our analysis has still found genuine sources illustrates that it is nottoo sensitive to the PSF, at least in the case of specifying overly heavy wings. If instead the raytraced PSF (ChaRT: http://cxc.cfa.harvard.edu/chart/) is used, then the estimate of the background is higher because this PSF has lighter wings than the King profile. The lighter wings also lead to the detection of four additional faint sources: one has an optical counterpart, one does not, and two cannot be confirmed optically because they are close to a bright source. Further investigation of these sources and modeling possible variations in the PSF are topics for future work. For example, temporal information can potentially be used as a diagnostic to assess whether any of the detected weak sources are in fact due to fluctuations in the PSFs of the bright sources.

Table 5
Extended Full Model Fit for the *Chandra* Observation in Figure 11

COUP #	μ_{jx}		μ_{jy}		Relative Intensity (%)	
732	4054.42	(4054.41, 4054.43)	4149.45	(4149.44, 4149.46)	34.59	(34.16, 35.03)
745	4052.83	(4052.81, 4052.84)	4140.67	(4140.66, 4140.68)	28.11	(27.71, 28.51)
689	4069.93	(4069.91, 4069.94)	4175.93	(4175.91, 4175.94)	14.10	(13.79, 14.40)
724	4058.57	(4058.56, 4058.59)	4176.73	(4176.71, 4176.74)	11.43	(11.16, 11.71)
744	4051.53	(4051.50, 4051.55)	4147.57	(4147.55, 4147.60)	7.41	(7.14, 7.68)
765	4045.40	(4045.35, 4045.46)	4181.20	(4181.15, 4181.25)	1.42	(1.32, 1.53)
649	4088.16	(4088.08, 4088.24)	4165.95	(4165.87, 4166.03)	0.57	(0.50, 0.63)
766	4045.36	(4045.27, 4045.45)	4155.18	(4155.10, 4155.25)	0.77	(0.68, 0.87)
788	4043.48	(4043.36, 4043.61)	4155.74	(4155.64, 4155.84)	0.56	(0.47, 0.64)
682	4072.11	(4072.01, 4072.21)	4181.12	(4181.03, 4181.22)	0.46	(0.39, 0.52)
640	4091.73	(4091.53, 4091.92)	4137.42	(4137.26, 4137.59)	0.13	(0.10, 0.16)
664	4081.43	(4081.22, 4081.63)	4159.41	(4159.21, 4159.61)	0.11	(0.08, 0.14)
665	4082.84	(4082.67, 4083.02)	4137.28	(4137.14, 4137.43)	0.15	(0.12, 0.19)
779	4044.39	(4043.86, 4044.60)	4140.72	(4140.43, 4140.90)	0.14	(0.09, 0.18)
Background	⋯	⋯	⋯	⋯	0.06	(0.01, 0.10)

Note. Posterior mean locations and relative intensities (as percentages), with 68% intervals indicated.

between 13 and 25 counts, this source is at the edge of detectability of local detection methods, particularly since the estimate of the local background in such methods would be high due to contamination from nearby bright sources. Thus, we expect that more basic approaches would either have failed to find this source, or would only find it by rendering their detection threshold to a point where spurious detections became problematic. Indeed, the reason the spatial-only model gives non-negligible weight to 15 sources (see Figure 12) is that it tends to split sources into two. The problem is that a single empirical PSF may exhibit chance variations that appear to be evidence for multiple PSFs. The spatial-only model also mistakes clusters of background photons for sources. The locations and spectra of these spurious sources show considerable posterior variability. Although any particular instance has low probability, there are multiple instances that together create erroneous support for an additional source. The main advantage of using the spectral information, in this example, is that it mitigates these issues, leading to a greater certainty that there are really 14 sources. Additionally, under the extended full model, the standard deviations of the parameters are almost invariably slightly smaller.

7.2. Spectral Analysis of the Disentangled Sources

The extended full model only captures the basic shape of the source spectra and we now illustrate how detailed follow-up spectral analysis can incorporate probabilistic event allocations. We perform this analysis for the two overlapping sources that are enclosed in the red box in the right panel of Figure 11 (COUP sources #732 and #744; Getman et al. 2005). Their estimated relative intensities are 0.3459 and 0.0741 under the extended full model. This is a good example to test the probabilistic event allocations, since the sources are close together (separation $\approx 1\farcs7$), each have sufficient counts for a useful spectral fit ($\approx$4350 and $\approx$910 counts between 0.5−7 keV for the bright and faint sources, respectively), and one source is substantially weaker than the other.

As described in Section 2.3, s_i indicates the source (or background) number associated with photon i. These are unknown parameters (or latent variables) that are updated at each iteration of the RJMCMC sampler. The variability in s_i indicates the uncertainty in the source of photon i (due to the PSF and uncertainty in the source parameters). We can account for this uncertainty by conducting many spectral analyses, each according to a sampled photon allocation (i.e., sampled values of s_i), and combining the results. We focus on photons with spatial location in the red box in Figure 11 (right panel) and to values of s_i sampled conditional on $K = 14$. Since we are only interested in COUP sources #732 and #744 we ignore any photons that are attributed to one of the other sources (in a given allocation). (The photons in the red box in Figure 11 are attributed to one of the other sources only rarely.)

Based on the photon allocations, we construct a sample of 1000 simulated spectral data sets for both sources, constructed from photon allocations based on every 10th iteration of the RJMCMC algorithm that sets $K = 14$ (up to the 10,000th RJMCMC iteration that sets $K = 14$). The variability in the source counts across the 1000 iterations is ± 17 for both the bright and faint sources. The specific photons that are allocated to each source also varies, even when the total source counts do not. Each individual spectrum is fit with an absorbed single temperature thermal model (`xsphabs*xsapec` in CIAO/*Sherpa* v4.6) fitting the absorption column ($N_{\rm H}$), temperature (kT), metallicity (Z), and normalization. A pile-up correction is needed for all spectra for the bright source since the measured count rate of 0.7 counts frame^{-1} is higher than the threshold at which pile-up becomes significant ($\approx$0.3 counts frame^{-1}). We use the `jdpileup` model in *Sherpa*, fitting the grade migration parameter α and the pile-up strength parameter f (Davis 2001). We call the entire collection of spectral fits the disentangled analysis.

For comparison, we also carry out a spectral analysis of the sources based on a naïve allocation of photons that collects events from within $1''$ of the fitted location of each source and assumes that there is no contamination from the other source. The only difference in the spectral model for the naïve and disentangled analyses is in how the effective areas are defined. In the case of the naïve analysis, a correction is made post-facto to the normalization based on how much of the source is expected to be included within the $1''$ source photons extraction radius. In the disentangled analysis, the assumed extraction

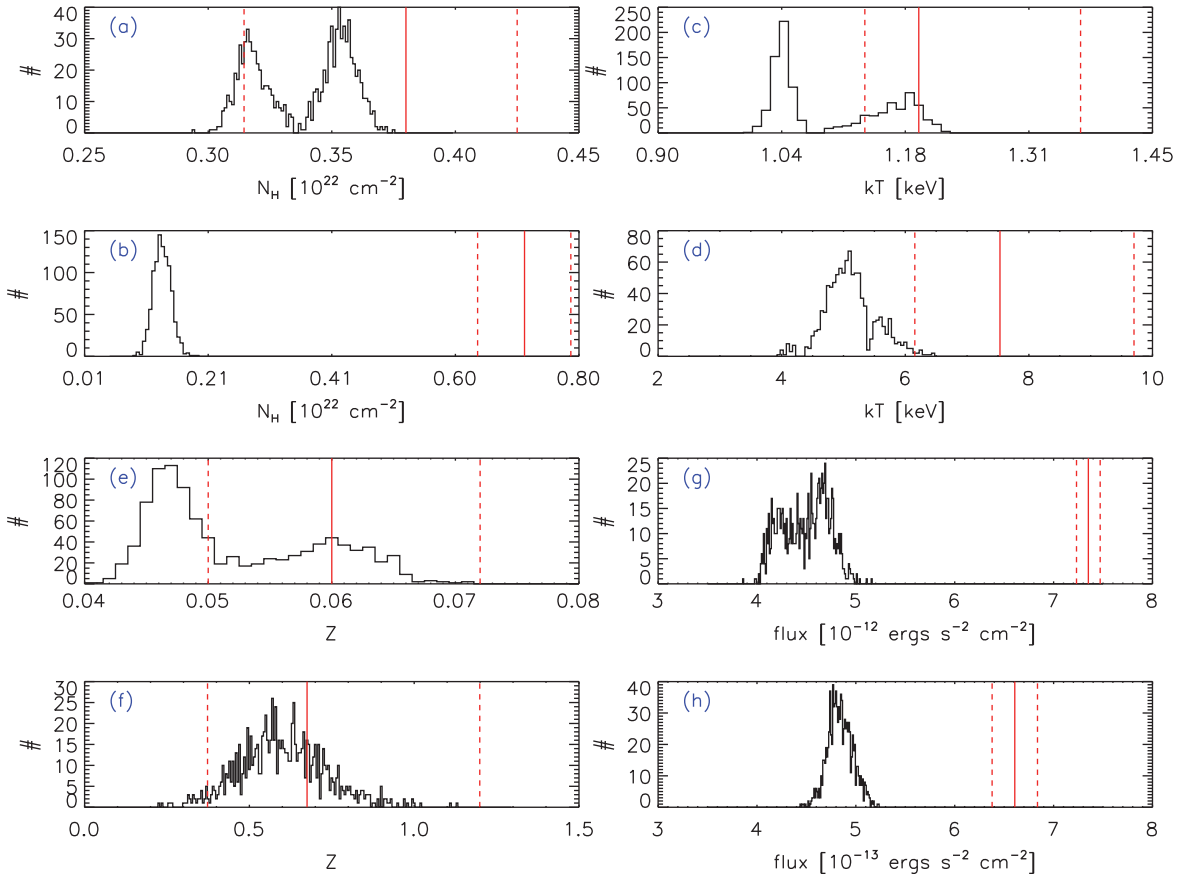

Figure 13. Detailed spectral analysis of overlapping COUP sources #732 and #744. Best-fit values of absorption column ((a), (b)), temperature ((c), (d)), metallicity ((e), (f)), and flux ((g), (h)) for the disentangled analysis, for each of 1000 allocations of the photons are shown as histograms. Panels (a), (c), (e), and (g) correspond to the bright source and panels (b), (d), (f), and (h) correspond to the fainter source. The naïve analysis best-fit values and their 68% intervals are shown by the solid and dashed red vertical lines, respectively. The width of the histograms only account for uncertainty due to the allocation of photons, and not additional statistical error, which is well described by the intervals shown for the naïve analysis.

radius for the spectra with allocated events is set to be $2\farcs5$ and the subsequent correction is negligible.

The results of the spectral fits to the disentangled spectra are shown as histograms of best-fit values for $N_{\rm H}$, kT, Z, and model flux computed for each of the 1000 spectra, see Figure 13. In several cases, a bimodal distribution is apparent. This suggests that a multi-temperature component spectrum would be a better fit. The separation of the modes, however, is generally too small to be picked up by typical multi-temperature model fits. Not shown are the pile-up parameters for the bright source, which are consistent between the naïve and disentangled analyses ($(\alpha, f) = (0.6, 0.93)$ for naïve, and $(0.53, 0.89)$ for the disentangled spectra), though the former indicates that the pile-up strength is slightly higher. This is to be expected, since the naïve analysis is carried out for photons in the core of the PSF, where naturally pile-up is most significant. The disentangled spectra include photons from the wings, thus reducing the strength of pile-up effects and decreasing the correction needed to the source flux by about 60%.

The spread in the histograms in Figure 13 indicates the uncertainty in the best-fit values due to uncertainty in the allocation of photons. The best-fit values from the naïve calculation are shown as solid red vertical lines. The dashed red vertical lines give 68% intervals indicating the statistical errors, due to randomness in the photons emitted and detected, under the naïve analysis. These statistical errors do not account for uncertainty in the photon allocations. The histograms, on the other hand, represent only errors due to uncertainty in the photon allocations, but do not account for statistical errors (due to randomness in photon emission and detection). Because the two sources of error are independent, and because we expect the statistical errors for the disentangled analyses to be similar to those for the naïve analysis, the total errors could be represented by a perturbation of the histograms with σ equal to the statistical errors from the naïve analysis. For these data, with the exception of flux (panels (g) and (h) of Figure 13), the statistical errors dominate the errors due to uncertainty in the photon allocation. Despite this, the disentangled analysis provides reasonable evidence that the absorption column of the faint source (panels (b) of Figure 13) and the flux of the two sources (panels (g) and (h) of Figure 13) are different from the best-fit values under the naïve analysis.

The variability of the true parameters around each of the best fit values recorded in the histograms is expected to be similar to that indicated for the naïve fit. However, we did not calculate these uncertainties because of the large computational cost. For these data, with the exception of flux (panels (g) and (h) of Figure 13), the variability in the true spectral parameters around the best fit values is likely larger than the uncertainty in the best fit values (due to the uncertainty in the allocation of photons). Despite this, the disentangled analysis provides reasonable evidence that the absorption column of the faint source (panel (b) of Figure 13) and the flux of the two sources

(panels (g) and (h) of Figure 13) are different to the naïve analysis best-fit values.

Overall, the naïve analysis best-fit values for the fainter source are in greater disagreement with the disentangled analysis than those for the bright source. This is to be expected, since in the naïve analysis, the contamination of the fainter source by the brighter source is larger. Our algorithm effectively removes this contamination. This causes the spectral fit parameter values to change and the measured source flux of the fainter source to decrease. In summary, the observed changes to the spectral model parameters are as would be expected when contamination is reduced and the data quality is improved.

8. SUMMARY

We have developed a Bayesian statistical method that models spatial and spectral information from overlapping sources and the background, and jointly estimates all individual source parameters. The key contributions of our approach are the use of spectral information to improve spatial separation, coherent quantification of uncertainty, including that of the number of sources, and the probabilistic assignment of photons to the different sources. Our simulation studies show that using spectral information improves the detection of both faint and closely overlapping sources and increases the accuracy with which source parameters are inferred.

We have analyzed data from two sets of overlapping sources observed with *XMM* and *Chandra*. Traditional analysis of *XMM* observations of FK and FL Aqr, thought to be a visual binary, show that their spectra are not distinguishable. Our analysis confirms that the spectra are indeed similar, but nonetheless shows that they are separable. We have also carried out detailed spectral analysis on disentangled photons from a pair of close sources from near the center of the Orion Nebula Cluster observed with *Chandra*. We find that the spectral parameters change significantly after contamination is removed.

The data we have considered consists of event-level observations. In the more usual case of spatially binned data, the PSF could be updated to take account of the binning. If the spatial pixels are larger, the importance of spectral data is greater, because it is harder to spatially distinguish sources from each other and the background. Clearly however, unbinned data is preferred when available, and our method has the ability to use all the information in such data. Similar comments apply when the spectral data are grouped.

As with other detection procedures, an important question is how to combine information from multiple observations. Since our approach gives the posterior distribution of all the parameters, this can be used as the prior distribution in subsequent analyses. Thus, under the Bayesian framework it is straightforward to analyze the available observations sequentially, which is convenient in that different PSFs, for example, can be used for each analysis. This is critical if the observations are recorded by different observatories.

Another advantage of the Bayesian framework is that more complex models can straightforwardly be built in. For example, using a location or spectral dependent PSF would require only minimal changes to the method and code. Another extension is to include the different temporal signatures of overlapping sources to further separate them. Future work will focus on these and related issues as well as computational scalability.

This work was supported by the Smithsonian Competitive Grants Program for Science Fund 40488100HH0043 and was conducted under the auspices of the CHASC International Astrostatistics Center. CHASC is supported by NSF grants DMS 1208791 and DMS 1209232. D.E.J. acknowledges support from the Harvard Statistics Department, V.L.K. from a NASA contract to the *Chandra* X-Ray Center NAS8-03060 and from *Chandra* grant AR0-11001X, and D.v.D. from a Wolfson Research Merit Award provided by the British Royal Society and from a Marie-Curie Career Integration Grant provided by the European Commission. In addition, we thank CHASC members for many helpful discussions, especially Xiao-Li Meng, Andreas Zezas, Aneta Siemiginowska, Lazhi Wang, and Alex Blocker.

APPENDIX A
SPLIT AND COMBINE PROPOSALS IN RJMCMC

The purpose of this appendix is to detail our implementation of split-combine moves in the BASCS code. We assume the reader is familiar with MCMC and RJCMC algorithms. Those unfamiliar with MCMC we refer to Gelman et al. (2013) and the appendix of Xu et al. (2014). Those unfamiliar with RJMCMC we refer to Richardson & Green (1997) and Green (1995). The basic properties of the algorithm follow from the reversibility condition and the theory of Markov chain convergence dealt with in many probability and stochastic processes books, for example Feller (1968).

We concentrate on the split proposals used in BASCS because they are more complex than the combine proposals. In particular, we detail the steps of a split proposal in BASCS for the extended full model (the most complex case considered). The corresponding combine proposals are straightforwardly obtained by solving the equations appearing in our split proposal scheme for the parameters of the combined source (i.e., the parameters of the yet to split source). Conditions that are required of newly split sources must also be satisfied when sources are combined. Following the algorithm is a short description of the reasons that its novel features are necessary in the current context.

Let $\mu_j = (\mu_{jx}, \mu_{jy})$ be the location of the source the algorithm is attempting to split. Throughout this appendix, the parameters for the two newly proposed sources formed by a split will be subscripted as in the main parts of the paper except that a 1 will appear after the subscript j to indicate the first newly proposed source, and similarly a 2 will indicate the second newly proposed source, e.g., μ_{j1x} will denote the x-coordinate of the first newly proposed source formed by a split. The newly proposed sources are ordered so that $\min(\gamma_{j11}, \gamma_{j12}) \leqslant \min(\gamma_{j21}, \gamma_{j22})$, i.e., the smallest *gamma* distribution mean of the spectral model for the first newly proposed source is smaller than that of the second newly proposed source. For the full model the ordering used is $\gamma_{j1} \leqslant \gamma_{j2}$, and for the spatial-only model it is $\mu_{j1x} \leqslant \mu_{j2x}$. These orderings are solely for the purposes of proposals; the label switching problem is discussed separately in Appendix B. A split proposal is performed as follows:

Step 1: Spectral parameters proposal: draw $u \sim \text{Uniform}(0, 1)$.

(a) If $u > 0.5$, simulate $u_1 \sim \text{Beta}(2, 2)$, $t, v_2, v_3 \sim \text{Uniform}(0, 1)$ and $v_4, v_5 \sim \textit{gamma}(5, 5)$. For

$a = \pi_j/u_1$ and $b = (\pi_j + u_1 - 1)/u_1$ define

$$f(u_1, \pi_j) = \begin{cases} a & \text{if } a < 1 \\ 1 + e^{\frac{10}{a} - 10} \log(a) & \text{otherwise,} \end{cases} \quad (20)$$

$$g(u_1, \pi_j) = \begin{cases} b & \text{if } b > 0 \\ be^{10b} \log(b) & \text{otherwise.} \end{cases} \quad (21)$$

Then set

$$\pi_{j1} = tg(u_1, \pi_j) + (1 - t)h(u_1, \pi_j) \quad (22)$$

$$\pi_{j2} = \frac{\pi_j - u_1\pi_{j1}}{1 - u_1} \quad (23)$$

$$\gamma_{j11} = v_2\gamma_{j1} \quad (24)$$

$$\gamma_{j21} = \frac{1 - v_{11}v_2}{1 - v_{11}}\gamma_{j1} \quad (25)$$

$$\gamma_{j12} = \gamma_{j11} + \frac{v_3}{v_{12}}(\gamma_{j2} - \gamma_{j11}) \quad (26)$$

$$\gamma_{j22} = \gamma_{j11} + \frac{1 - v_3}{1 - v_{12}}(\gamma_{j2} - \gamma_{j11}) \quad (27)$$

$$\alpha_{j11} = v_4\alpha_{j1} \quad (28)$$

$$\alpha_{j12} = v_5\alpha_{j2} \quad (29)$$

$$\alpha_{j2l} = \frac{\rho_{j2l}\gamma_{j2l}^2}{A_{jl}} \quad \text{for } l = 1, 2, \quad (30)$$

where

$$A_{jl} = \rho_{j1}\gamma_{jl}^2\left(1 + \frac{1}{\alpha_{jl}}\right) - \rho_{j1l}\gamma_{j1l}^2\left(1 + \frac{1}{\alpha_{j1l}}\right) - \rho_{j2l}\gamma_{j2l}^2, \quad (31)$$

and $\rho_{j1} = w_j\pi_j$, $\rho_{j11} = w_{j1}\pi_{j1}$, $\rho_{j12} = w_{j2}\pi_{j2}$, $\rho_{j2} = w_j(1 - \pi_j)$, $\rho_{j12} = w_{j1}(1 - \pi_{j1})$, and $\rho_{j22} = w_{j2}(1 - \pi_{j2})$. The split proposal is immediately rejected if π_{j1} is not between $\min\left(1, \frac{\pi_j}{u_1}\right)$ and $\max\left(0, \frac{\pi_j + u_1 - 1}{u_1}\right)$, or $\gamma_{j21} > \gamma_{j22}$, or any of γ_{j11}, γ_{j21}, γ_{j12}, γ_{j22} are outside the range of the spectral data E.

(b) If $u \leqslant 0.5$, simulate $\pi_{j1}, \pi_{j2} \sim \text{Beta}(10, 1)$, $v_2, v_3 \sim \text{Uniform}(0, 1)$, and $v_4, v_5 \sim \text{Beta}(1, 5)$. Then set $u_1 = \pi_j$, $\gamma_{j11} = \gamma_{j1}$, $\gamma_{j21} = \gamma_{j2}$, $\alpha_{j11} = \alpha_{j1}$, $\alpha_{j21} = \alpha_{j2}$, and

$$\gamma_{j12} = \gamma_{j1} + v_2\left(E_{\max} - \gamma_{j1}\right) \quad (32)$$

$$\gamma_{j22} = \gamma_{j2} + v_3\left(E_{\max} - \gamma_{j2}\right) \quad (33)$$

$$\alpha_{j12} = 20v_4 \quad (34)$$

$$\alpha_{j22} = 20v_5. \quad (35)$$

Step 2: Spatial parameters proposal: simulate $u_1 \sim \text{Beta}(2, 2)$, $u_2 \sim S_2\text{Beta}(2, 2)$, and $u_3 \sim S_3\text{Beta}(2, 2)$ (where S_2 and S_3 are independent random signs) and set

$$w_{j1} = w_j u_1 \quad (36)$$

$$w_{j2} = w_j(1 - u_1) \quad (37)$$

$$\mu_{j11} = \mu_{jx} - u_2\sigma\sqrt{\frac{w_{j2}}{w_{j1}}} \quad (38)$$

$$\mu_{j21} = \mu_{jx} + u_2\sigma\sqrt{\frac{w_{j1}}{w_{j2}}} \quad (39)$$

$$\mu_{j12} = \mu_{jy} - u_3\sigma\sqrt{\frac{w_{j2}}{w_{j1}}} \quad (40)$$

$$\mu_{j22} = \mu_{jy} + u_3\sigma\sqrt{\frac{w_{j1}}{w_{j2}}}. \quad (41)$$

In our algorithm $\sigma = 1$ (tuning parameter).

Step 3: If $(\mu_{j1x} - \mu_{j'x})^2 + (\mu_{j1y} - \mu_{j'y})^2 < (\mu_{j1x} - \mu_{j2x})^2 + (\mu_{j1y} - \mu_{j2y})^2$ for some $j' \in \{1, ..., K\}/\{j\}$, then the split proposal is rejected. We also reject the split proposal if the proposed source locations are outside the convex hull of the spatial data (x, y).

Step 4: To update s to s' randomly assign photon i to the first newly proposed source with probability $p_i = p_{i1}/(p_{i1} + p_{i2})$, and otherwise to the second newly proposed source, for each $i \in \mathcal{J}_j$. Here

$$p_{il} = w_{jl}f_{(\mu_{jl1}, \mu_{jl2})}(x_i, y_i)\sum_{r=1}^{2}\pi_{jlr}g_{\alpha_{sjlr}, \gamma_{sjlr}}(E_i), \quad (42)$$

for $l = 1, 2$. We denote the probability of the particular allocation realized by P_{alloc}.

Step 5: Simulate $u_{\text{split}} \sim \text{Uniform}(0, 1)$ and accept the proposed split if $u_{\text{split}} < \min\{1, A\}$ where

$$A = \begin{cases} \dfrac{p\left(\Theta'_{K+1}, K + 1, s'\middle|x, y, E\right)}{p(\Theta_K, K, s|x, y, E)}\dfrac{d_{K+1}}{b_K P_{\text{alloc}}} \times \dfrac{1}{\frac{1}{4}b_{2,2}(u_1)b_{2,2}(|u_2|)b_{2,2}(|u_3|)} \times \dfrac{1}{g_{5,5}\left(v_4\right)g_{5,5}\left(v_5\right)}|J_a| & \text{if } u > 0.5 \text{ (Step 1)} \\ \dfrac{p(\Theta'_K, K + 1, s'|x, y, E)}{p(\Theta_K, K, s|x, y, E)}\dfrac{d_{K+1}}{b_K P_{\text{alloc}}} \times \dfrac{1}{\frac{1}{4}b_{2,2}(|u_2|)b_{2,2}(|u_3|)b_{10,1}(\pi_{j1})b_{10,1}(\pi_{j2})} \times \dfrac{1}{b_{1,5}\left(v_4\right)b_{1,5}\left(v_5\right)}|J_b| & \text{otherwise.} \end{cases} \quad (43)$$

Here, the notation $b_{S,R}$ and $g_{S,R}$ denotes the $\text{Beta}(S, R)$ and $gamma(S, R)$ densities, respectively, and

$$b_K = \begin{cases} \dfrac{1}{K} & \text{if } K = 1 \\ \dfrac{1}{2}\dfrac{1}{K} & \text{otherwise,} \end{cases} \quad (44)$$

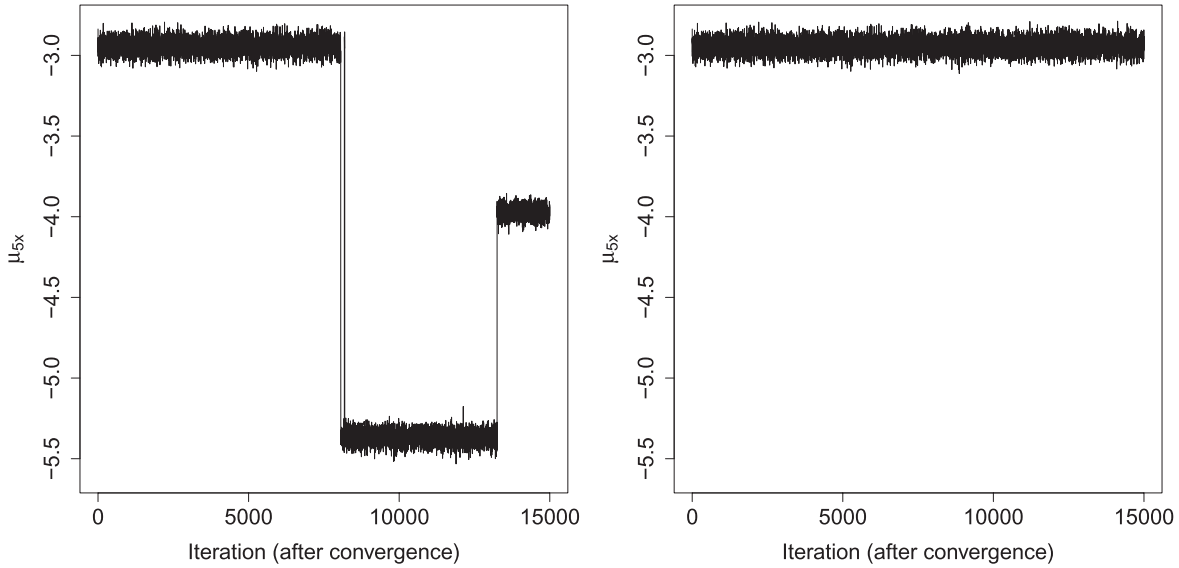

Figure 14. Trace plot of the parameter μ_{5x} from a simulation with ten sources (Section 5.1) before (left) and after (right) relabelling.

$$d_{K+1} = \begin{cases} \dfrac{1}{K+1} & \text{if } D_{j1,j2} \leqslant D_{j1,j'} \\ & \forall\, j' \in \{1,...,K\}/\{j\} \\ \dfrac{1}{2}\dfrac{1}{K+1} & \text{otherwise,} \end{cases} \tag{45}$$

where $D_{t_1,t_2} = ||(\mu_{t_1x}, \mu_{t_1y}) - (\mu_{t_2x}, \mu_{t_2y})||_2$.

The Jacobian $|J_a|$ is the determinant of a 16×16 block matrix. The determinant of the upper-left 6×6 block is $w_j\sigma^2/(u_1(1-u_1))$, and this is multiplied by the determinant of the lower-right block which is calculated numerically. The Jacobian $|J_b|$ is $20^2(E_{\max} - \gamma_{j1})(E_{\max} - \gamma_{j2})w_j\sigma^2/(u_1(1-u_1))$.

There are two features of BASCS that are not explicitly dealt with in standard approaches. The first is that the distributions we split and combine are themselves mixture distributions. The second is that BASCS randomly chooses from two proposal schemes for the spectral parameters in Step 1 because a single approach does not address all the possibilities. The approach in Step 1(a) splits each *gamma* distribution in the current source's spectral model into two, thus forming two new spectral models for the newly proposed sources. The key aspect of this approach is that the new spectral models are designed to both be similar to the original. This makes sense in a situation where two similar sources have been mistaken for one. The approach in Step 1(b) is designed to split one true source into two, with each newly proposed source accounting for one *gamma* component of the true spectral model. Thus, the two new source spectral models each typically have nearly all their weight on a single *gamma*, which is almost invariably the first component in the extended full spectral model (we sort the *gammas* by their means, in increasing order). Of course, we do not want to split a true source, but this split proposal is necessary in order to allow the reverse combine proposal, because the reversibility condition must be satisfied.

APPENDIX B
LABEL SWITCHING

A computational challenge is that the enumeration, or labelling, of individual sources changes stochastically during the iterations of an RJMCMC algorithm (and even during the iterations of an MCMC algorithm for a mixture model with a known number of components). For example, Figure 14 shows the value of μ_{5x} at each iteration of our algorithm (after convergence) before and after the labelling has been corrected (the data are from the simulation study involving ten sources described in Section 5.1, and in particular, μ_{5x} is the x-coordinate of the fifth source). Clearly, some such correction will be necessary in order for estimates such as that in Equation (17) to be meaningful.

We implemented two approaches to relabelling and, in our real data analyses, they gave essentially identical results. The first method was to impose a hard constraint. In one dimension a hard constraint typically involves ordering the component locations, but it is not clear how best to impose such a constraint in two dimensions. As most of the source positions were precisely fitted, we simply ran the RJMCMC algorithm until convergence and then selected a posterior draw of the positions and weights to use as a reference. Running the algorithm again (or continuing the initial run), at each iteration we labelled the current source closest to the brightest reference source as source one, then we looked for the source closest to the second brightest reference source, and so on. As in the one-dimensional case, this approach has the limitation that artificial "boundary" effects may be introduced when the posteriors of two source positions overlap. These effects indicate that the real posterior uncertainty has not been correctly recovered (unless there is some real information to support a hard constraint in our prior). However, in our real data analyses there was no evidence of such boundaries because, for probable values of K, all the source positions were precisely fitted and there was little overlap between the posteriors of source positions. In the case of the *Chandra* observation and $K = 14$, the fact that the posteriors of the source locations are non-overlapping can be seen from Figure 11.

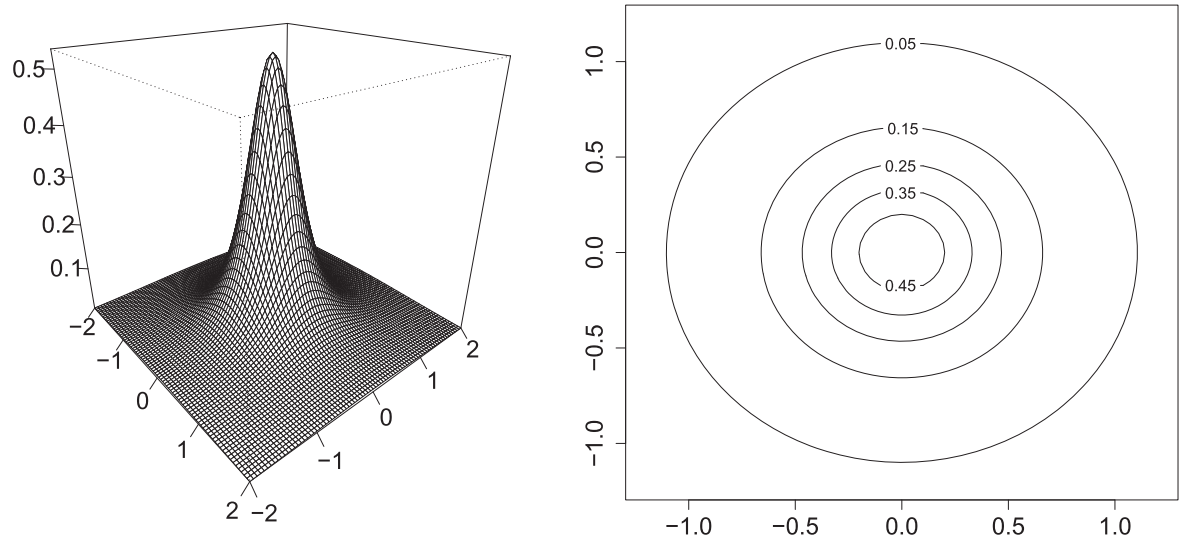

Figure 15. 2D King profile density (left), and its contours (right).

We also implemented the approach suggested by Cron & West (2011), by modifying their publicly available code to work for our model. This method also uses a reference and is based on a loss function. At iteration t, the most likely assignment of each photon is computed treating the current parameter values as the true parameters, and then again treating the reference parameters as the true parameters. If, assuming the current parameter values, photon i is most likely to have originated from component two, but another origin is most likely when assuming the reference parameter values, then we say there is a mismatch in allocation of photon i. The method used by Cron & West (2011) is to choose the relabeling that minimizes the number of mismatches at iteration t, and then proceed to the next iteration. This second approach is substantially more computationally expensive than the first. Therefore we use the first approach online and apply the second only if there are potential "boundary" effects (neither method is affected by the initial labels and therefore no problems are caused by applying both).

APPENDIX C
KING PROFILE

The functional form of the 2D King profile is

$$f(d) = \frac{C}{\left(1 + \left(d/d_0\right)^2\right)^{\eta}}$$

where

$$d(x, y, \omega) = \sqrt{(x\cos\omega + y\sin\omega)^2 + (y\cos\omega - x\sin\omega)^2/(1-\epsilon)^2}.$$

The constant C is determined numerically. The particular parameters we use for the 2D King profile are as follows; off-axis angle $\theta = 0$ arcmin, core radius $d_0 = 0.6$ arcsec, power-law slope $\eta = 1.5$, ellipticity $\epsilon = 0.00574$. The resulting probability density is displayed in Figure 15.

REFERENCES

Brewer, B. J., Foreman-Mackey, D., & Hogg, D. W. 2013, AJ, 146, 7
Brewer, B. J., & Stello, D. 2009, MNRAS, 395, 2226
Broos, P. S., Townsley, L. K., Feigelson, E. D., et al. 2010, ApJ, 714, 1582
Christian, R., & Casella, G. 2011, in Handbook of Markov Chain Monte Carlo, ed. S. Brooks, A. Gelman, G. Jones, & X.-L. Meng (Boca Raton, FL: CRC)
Cron, A. J., & West, M. 2011, The American Statistician, 65, 16
Davis, J. E. 2001, ApJ, 562, 575
Feller, W. 1968, An Introduction to Probability Theory and its Applications, Vol. I (New York, NY: Wiley)
Gelman, A., Carlin, J. B., Stern, H. S., et al. 2013, Bayesian Data Analysis (3rd ed., Boca Raton, FL: CRC)
Getman, K., Flaccomio, E., Broos, P., et al. 2005, ApJS, 160, 319
Green, P. J. 1995, Biometrika, 82, 711
Hastings, W. K. 1970, Biometrika, 57, 97
Jasche, J., & Wandelt, B. D. 2013, ApJ, 779, 15
Kashyap, V., Micela, G., Sciortino, S., et al. 1994, in AIP Conf. Proc. 313, The Soft X-Ray Cosmos, ed. E. M. Schlegel & R. Petre (Melville, NY: AIP), 239
King, I. 1962, AJ, 67, 471
Knoetig, M. L. 2014, ApJ, 790, 106
Kraft, R. P., Burrows, D. N., & Nousek, J. A. 1991, ApJ, 374, 344
Laird, E., Nandra, K., Georgakakis, A., et al. 2009, ApJS, 180, 102
Lee, H., Kashyap, V., van Dyk, D., et al. 2011, ApJ, 731, 126
Liu, J. S. 2008, Monte Carlo Strategies in Scientific Computing (Berlin: Springer)
Loredo, T. J. 1992, Statistical Challenges in Modern Astronomy, ed. E. D. Feigelson & G. J. Babu (Berlin: Springer)
McLachlan, G., & Peel, D. 2004, Finite Mixture Models (New York, NY: Wiley)
Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. 1953, JChPh, 21, 1087
Mukherjee, S., Feigelson, E. D., Babu, G. J., et al. 1998, ApJ, 508, 314
Park, T., Kashyap, V. L., Siemiginowska, A., et al. 2006, ApJ, 652, 610
Park, T., van Dyk, D. A., & Siemiginowska, A. 2008, ApJ, 688, 807
Primini, F., & Kashyap, V. 2014, ApJ, 796, 24
Read, A., & Saxton, R. 2012, XMM-Newton Current Calibration File Release Notes, Document XMM-CFF-REL-280 (Paris: ESA)
Read, A., Saxton, R., Guainazzi, M., Rosen, S., & Stuhlinger, M. 2010, XMM-Newton Current Calibration File Release Notes, Document XMM-CFF-REL-263 (Paris: ESA)
Richardson, S., & Green, P. J. 1997, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 59, 731
Safarzadeh, M., Ferguson, H. C., Lu, Y., Inami, H., & Somerville, R. S. 2015, ApJ, 798, 91
Shao, Q.-M., & Ibrahim, J. G. 2000, Monte Carlo Methods in Bayesian Computation (New York: Springer)

Titterington, D. M., Smith, A. F., Makov, U. E., et al. 1985, Statistical Analysis of Finite Mixture Distributions, Vol. 7 (New York: Wiley)
Umstätter, R., Christensen, N., Hendry, M., et al. 2005, PhRvD, 72, 022001
van Dyk, D. A., Connors, A., Kashyap, V., & Siemiginowska, A. 2001, ApJ, 548, 224
Walmswell, J., Eldridge, J., Brewer, B., & Tout, C. 2013, MNRAS, 435, 2171
Weisskopf, M. C., Wu, K., Trimble, V., et al. 2007, ApJ, 657, 1026
Wiper, M., Insua, D. R., & Ruggeri, F. 2001, Journal of Computational and Graphical Statistics, 10, 440
Xu, J., van Dyk, D. A., Kashyap, V. L., et al. 2014, ApJ, 794, 97