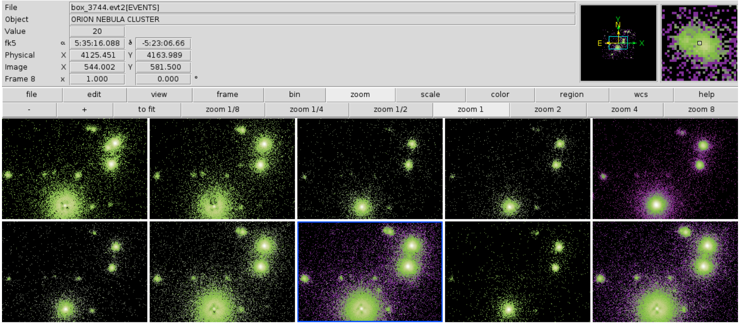
Detection: Overlapping Sources

David Jones
Harvard University Statistics Department

April 16, 2013

Introduction

- X-ray data: coordinates of photon detections
- PSFs of close sources overlap
- Aim: inference for number of sources and their intensities, positions and spectral distributions



Contamination approach (Kashyap et al. 1994)

- Circle sources and solve a set of linear equations describing the intensities and contamination of each source circle from background and other sources
- Issues
  - Not clear how the circles should be drawn
  - Gaussian PSFs
  - Only works with small overlap
  - Only works with few sources

There are also kernel approaches but these don't have the advantages of dealing with the allocation of photons exactly

Clustering Approach: Basic Model and Notation

Data $= y_{ij}$
$n_i = \#$ photons detected from source i
$\boldsymbol{\mu}_i =$ centre of source i
$k = \#$ sources (components)

$$
\begin{aligned}
y_{ij}|\boldsymbol{\mu}_i, n_i, k &\sim \text{PSF centred at } \boldsymbol{\mu}_i \;\; j = 1, \ldots, n_i, i = 0, \ldots, k \\
(n_0, n_1, \ldots, n_k)|w, k &\sim \text{Mult}(n; (w_0, w_1, \ldots, w_k)) \\
(w_0, w1, \ldots, w_k)|k &\sim \text{Dirichlet}(\alpha, \alpha, \ldots, \alpha) \\
\boldsymbol{\mu}_i|k &\sim \text{Uniform over the image} \;\; i = 1, 2, \ldots, k \\
k &\sim \text{Pois}(\theta)
\end{aligned}
$$

- Component with label 0 is background and its "PSF" is uniform over the image (so its "centre" is irrelevant)
- Reasonably insensitive to θ, the prior mean number of sources

3rd Dimension: Spectral Data

Can we distinguish the background and sources more accurately if we model the energy of the photons as well?

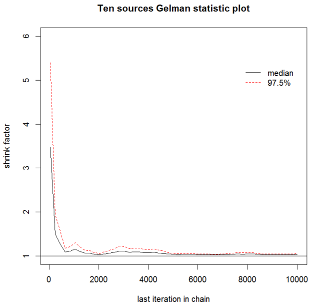
$$\begin{aligned} e_{ij}|\alpha, \beta &\sim \text{Gamma}(\alpha, \beta) \ \text{ for } i = 1, \ldots, k \\ e_{0j} &\sim \text{Uniform to some maximum} \\ \alpha &\sim \text{Gamma}(a_\alpha, b_\alpha) \\ \beta &\sim \text{Gamma}(a_\beta, b_\beta) \end{aligned}$$

Using a (correctly) "informative" prior on α and β versus a diffuse prior made very little difference to results.

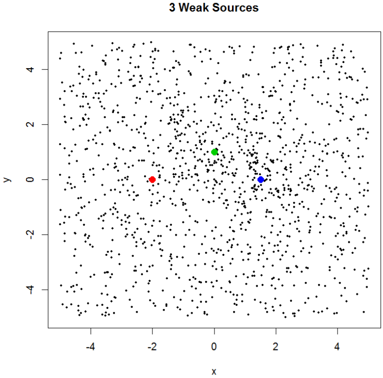
RJMCMC

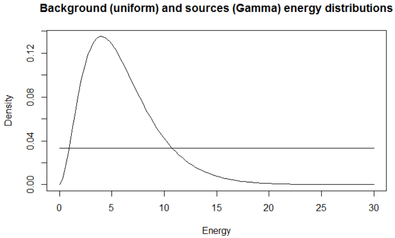
- Similar to Richardson & Green 1997
- Knowledge of the PSF makes things easier
- Insensitive to θ e.g. posterior for ten sources with $\theta = 3$:

	Number of Components						
	7	8	9	10	11	12	13
Mean	0.029	0.058	0.141	0.222	0.220	0.157	0.082
SD	0.018	0.019	0.022	0.029	0.027	0.021	0.014



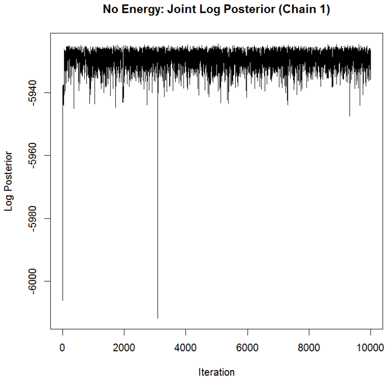
Simulated Data

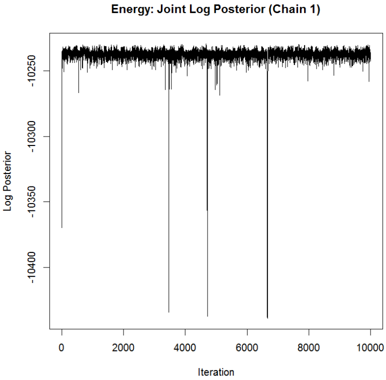




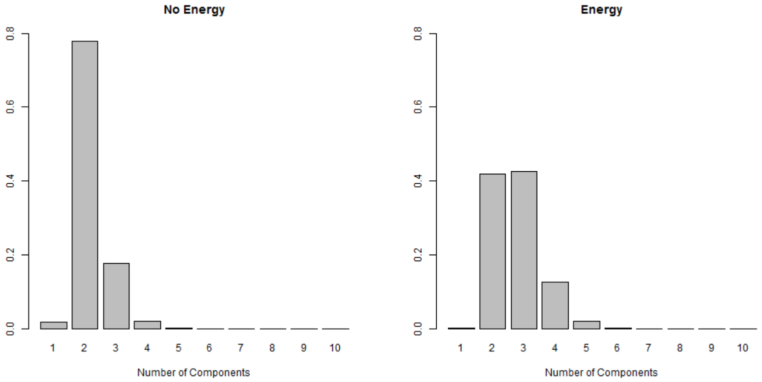
- Source region (2 SD) is about 28% of the area and contains about 41% of the observations
- Positions $(-2, 0), (0, 1), (1.5, 0)$ with intensities $50, 100, 150$ respectively

Joint Log Posterior



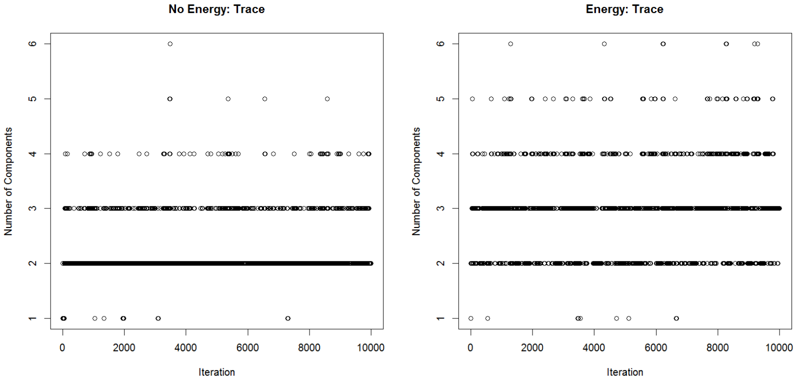


Posterior of k

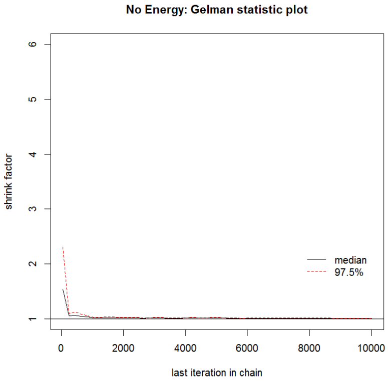


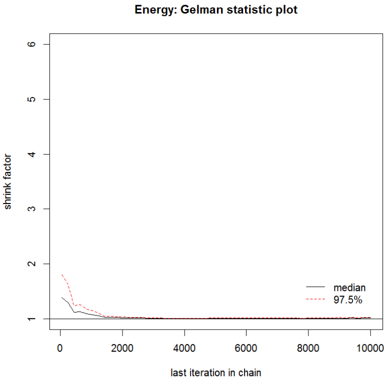
- Aggregation over 10 chains of the posterior probabilities (for each k the SD over the 10 chains is small)
- When not using the energy information we usually can't find the faintest source

Chain 1: Posterior of k Trace



Gelman-Rubin: Posterior of k





- Gelman-Rubin statistics were 1.00 (C.I. 1.01) and 1.01 (C.I. 1.01) respectively

Allocation of Photons

Table: Allocation breakdown: (a) ignoring energy information

Source (intensity)	**Average No. Photons**	**Average Allocation Breakdown**			
		Background	**Left**	**Middle**	**Right**
Background (10/sq)	1015	0.876	0.035	0.040	0.049
Left (50)	38	0.798	0.121	0.067	0.014
Middle (100)	97	0.502	0.168	0.189	0.141
Right (150)	152	0.481	0.043	0.159	0.317

Table: Allocation breakdown: (b) using energy information

Source (intensity)	**Average No. Photons**	**Average Allocation Breakdown**			
		Background	**Left**	**Middle**	**Right**
Background (10/sq)	1015	0.894	0.024	0.038	0.045
Left (50)	38	0.531	0.278	0.165	0.026
Middle (100)	97	0.293	0.122	0.346	0.239
Right (150)	152	0.305	0.028	0.141	0.526

- Background is more easily distinguished from the sources when we include the energy information

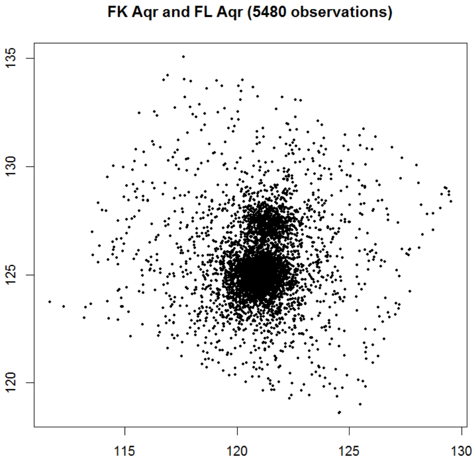
Parameter Inference

Table: Parameter estimation (a) no energy information (b) with energy information

	μ_{11}	μ_{12}	μ_{21}	μ_{22}	μ_{31}	μ_{32}	w_1	w_2	w_3	w_b	α	β
Mean	-1.266	0.839	0.401	0.549	1.798	-0.054	0.049	0.067	0.086	0.798	NA	NA
SD	0.069	0.125	0.067	0.068	0.030	0.046	0.002	0.002	0.003	0.001	NA	NA
MSE	0.543	0.718	0.165	0.207	0.090	0.005					NA	NA
SD/Mean							0.050	0.027	0.032	0.001	NA	NA
Mean	-1.790	-0.101	-0.234	1.042	1.584	-0.044	0.040	0.077	0.115	0.768	2.827	0.459
SD	0.037	0.064	0.033	0.026	0.019	0.022	0.001	0.001	0.002	0.000	0.013	0.003
MSE	0.045	0.014	0.056	0.002	0.007	0.002					0.030	0.002
SD/Mean							0.036	0.018	0.014	0.000	0.004	0.006

- The effects are obviously less pronounced when the sources are more easily distinguished from the background

Real Data



- Additional question: can we distinguish the spectral distributions of the sources?

What is the PSF?

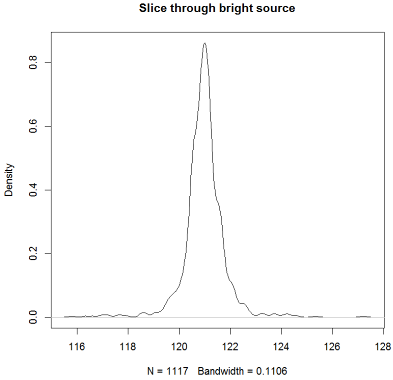
- Ideally a fairly accurate PSF can be obtained by training on non-overlapping sources
- In the absence of an accurate PSF:
  1. Approximate the number of sources (2 in this case)
  2. Obtain an EM estimate of the covariance of the PSF
- The presence of some clearly separated sources will obviously improve the accuracy of step 2 and generally reduce sensitivity to step 1

EM Estimate of the Covariance

- We obtained

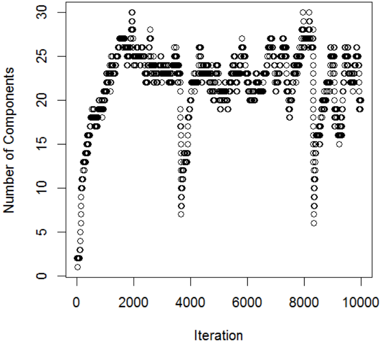
$$\hat{\Sigma}_{EM} = \begin{pmatrix} 0.562 & -0.020 \\ -0.020 & 0.479 \end{pmatrix}$$

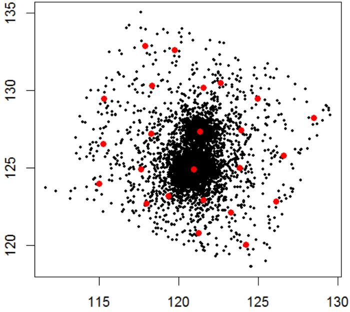
- A slice through the middle of the brighter source suggests the diagonal terms are not unreasonable



Problem!

- Behaves badly possibly because the background is not uniform





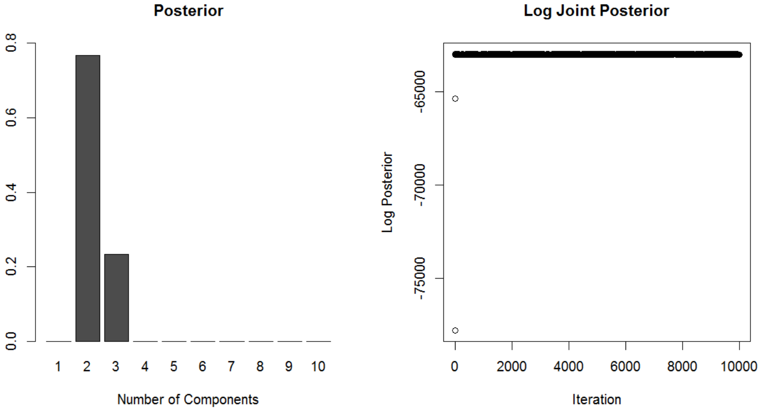
Solutions?

- The covariance matrix doesn't seem to be the issue. Scaling the EM estimate by a range of values made very little difference
- Ignoring the energy information also doesn't help
- **Current solution**:

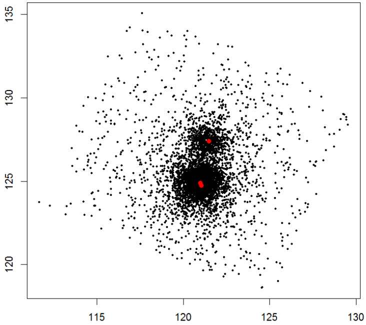
$$(w_0, w1, \ldots, w_k) | k \sim \text{Dirichlet}(\alpha, \alpha, \ldots, \alpha)$$

  previously $\alpha = 1$ but now we set $\alpha = 50$ to eliminate very weak sources
- Other ideas?

Posterior of k



Three? Potential Binaries?



- Probably just an artifact of making the sources more similar in brightness through α (but could be useful with prior knowledge) - moderate choice of α needed
- More careful treatment of label switching is needed for inference for the parameters of potential binaries

Parameter Inference

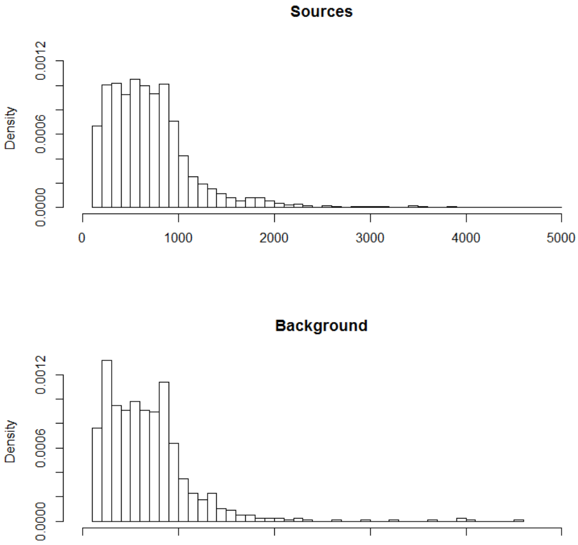
Table: Parameter estimation for FK Aqr and FL Aqr

	μ_{11}	μ_{12}	μ_{21}	μ_{22}	w_1	w_2	w_b	α	β
Mean	120.980	124.846	121.415	127.400	0.673	0.181	0.146	3.112	0.005
SD	0.017	0.017	0.036	0.036	0.007	0.005	0.005	0.062	0.000
MSE	0.000	0.000	0.001	0.001				0.004	0.000
SD/Mean					0.010	0.030	0.034	0.020	0.023

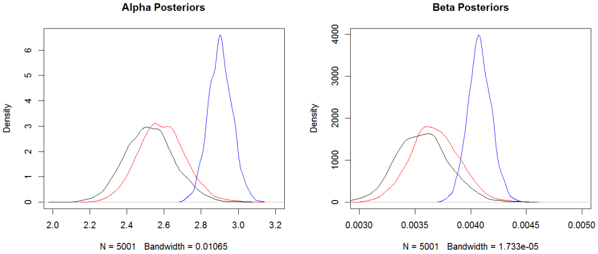
Extensions to Spectral Modeling

- The background spectral distribution doesn't appear to be uniform at all
- Model the spectral distributions of background and sources to all be different Gammas
- Will allow us to look at the question of whether the two sources have different spectral distributions

Background is Not Uniform

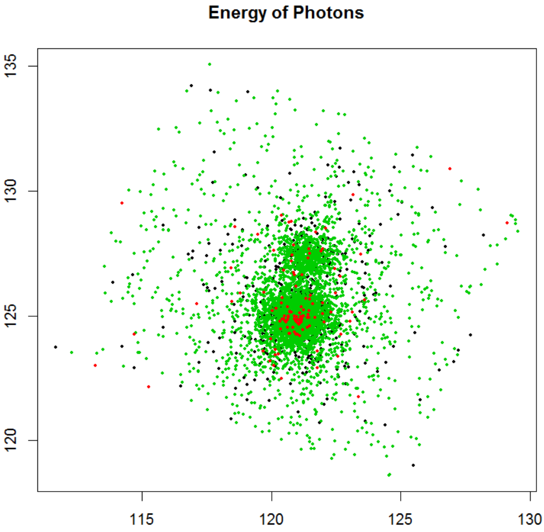


Comparing Spectral Distribution Parameters



- 95% posterior intervals for α_1 and α_2 are nearly disjoint

Should the dim source be similar to background?



Summary

- Works very well for simulated data
- Spectral model and possibly the background spatial model need some revisions to be realistic
- Need to investigate exactly why saturation occurs for the real data but not the simulated data
- Potential to separate spectral distributions of different sources