Fully Bayesian Analysis of Low-Count Astronomical Images

David A. van Dyk[1] Alanna Connors[2]

[1]Department of Statistics
University of California, Irvine

[2]Eurika Scientific

Thanks to James Chiang, Adam Roy, and
The California Harvard AstroStatistics Collaboration.

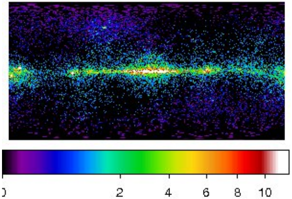
2007 Joint Statistics Meetings

Outline

Outline

Data Generation in High-Energy Astrophysics

- Low Counts
  - Imaging X-ray and γ-ray detectors typically count a *small number of photons* in each of a *large number of pixels*.
- Instrumentation
  - Point Spread Functions can vary with energy and location
  - Exposure Maps can vary across an image
  - Background Contamination



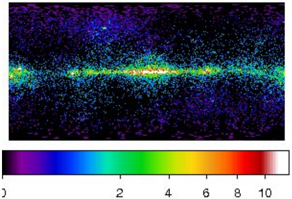
EGERT γ-ray counts >1GeV
(entire sky and mission life).

Sample Chandra psf's
(Karovska et al., ADASS X)

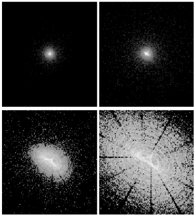
EGERT exposure map
(area $\times$ time)

Data Generation in High-Energy Astrophysics

- Low Counts
  - Imaging X-ray and γ-ray detectors typically count a *small number of photons* in each of a *large number of pixels*.
- Instrumentation
  - Point Spread Functions can vary with energy and location
  - Exposure Maps can vary across an image
  - Background Contamination



EGERT γ-ray counts >1GeV
(entire sky and mission life).

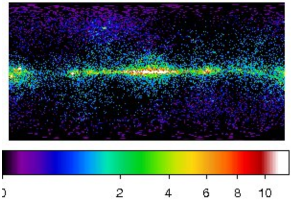


Sample Chandra psf's
(Karovska et al., ADASS X)

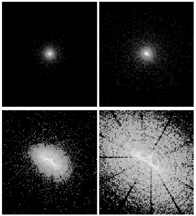
EGERT exposure map
(area $\times$ time)

Data Generation in High-Energy Astrophysics

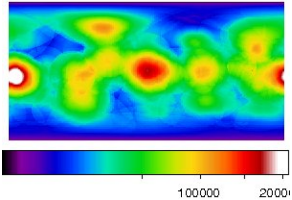
- Low Counts
  - Imaging X-ray and γ-ray detectors typically count a *small number of photons* in each of a *large number of pixels*.
- Instrumentation
  - Point Spread Functions can vary with energy and location
  - Exposure Maps can vary across an image
  - Background Contamination



EGERT γ-ray counts >1GeV (entire sky and mission life).



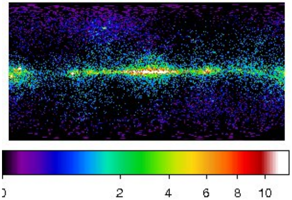
Sample Chandra psf's (Karovska et al., ADASS X)



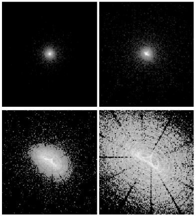
EGERT exposure map (area $\times$ time)

Data Generation in High-Energy Astrophysics

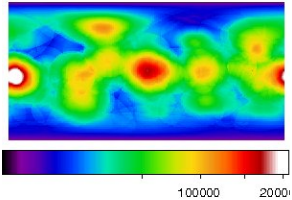
- Low Counts
  - Imaging X-ray and γ-ray detectors typically count a *small number of photons* in each of a *large number of pixels*.
- Instrumentation
  - Point Spread Functions can vary with energy and location
  - Exposure Maps can vary across an image
  - Background Contamination



EGERT γ-ray counts >1GeV
(entire sky and mission life).



Sample Chandra psf's
(Karovska et al., ADASS X)



EGERT exposure map
(area $\times$ time)

Outline

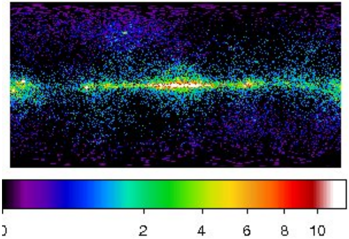
Scientific Goals

Given our blurry, low-count, inhomogeneous, contaminated data we would like to learn about the structure and unexpected features of an astronomical source.

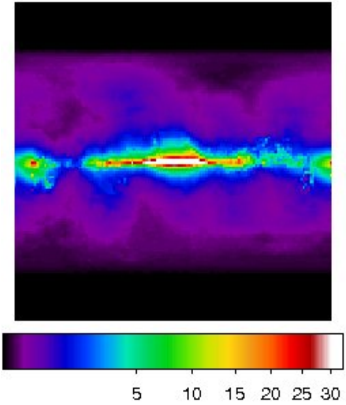
- What does the source *look* like?
- Are there interesting features?
- Are these features *statistically significant*?
- Are these features an indication of something beyond our current physical understanding of the source?
- *Is our physical model sufficient to explain the data?*

Example: Searching for the γ-ray halo.

data



physical model



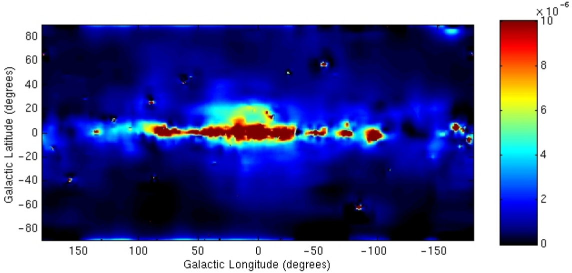
Is there excess emission/structure in the data?

Example: Residual Emission

- Dixon et al. fit a model of the form

  `Physical Model + Multi-Scale Residual`

  to the data, using Haar wavelets for the residual.
- Thresholding wavelet coefficient led to the following fit:



Example: But is it Real??

- Dixon et al. wondered....

  "The immediate question arises as to the statistical significance of this feature. Though we are able to make rigorous statements about the coefficient-wise and level-wise FDR, similar quantification of object-wise significance (e.g., 'this blob is significant at the n sigma level') are difficult."

Outline

Three Statistical Goals for Low-Count Image Analysis

Automate: We would like to automate

1. *model fitting* to avoid subjective stopping rules used to control reconstruction quality, and
2. *the search for structure* to avoid choosing parameters to enhance supposed structure.

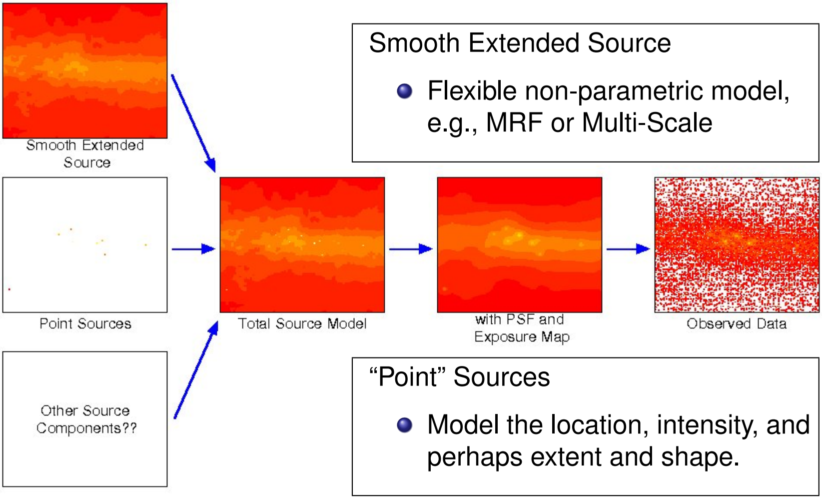
Formulate: We would like to formulate low-count image analysis in the terms of *statistical theory* to better understand the characteristics of the results.

Evaluate: We would like to evaluate

1. the *statistical error* in the fitted reconstruction under the assumed model,
2. the likelihood that supposed structures exist in the astronomical source, and
3. *the plausibility of the model assumptions.*

Outline

A Statistical Model for the Data Generation Process



Smooth Extended Source

- Flexible non-parametric model, e.g., MRF or Multi-Scale

“Point” Sources

- Model the location, intensity, and perhaps extent and shape.

Outline

Advantages of A Model-Based Formulation

1. The use of well defined statistical estimates such as ML estimates, MAP estimates, or posterior means, eliminates the need for ad-hoc stopping rules (Esch et al., ApJ, 2004).
2. Statistical theory allows computation of statistical errors with Bayesian / frequency properties (Esch et al., 2004).
3. Allows us to incorporate knowledge from other data.
4. Principled methods for comparing / evaluating models.
5. Quantify evidence for supposed structure under a flexible model.

Outline

Using Outside Information

Outside information can be critical with low-count data. Lucky, such information is often available as high-count high-resolution data from a different energy band (e.g., Optical or Radio).

Incorporating Information Through Model Components

- The number of and location of point sources.
- Smoothing parameters for extended source.
- Characterize spatial variation of smoothing parameters.

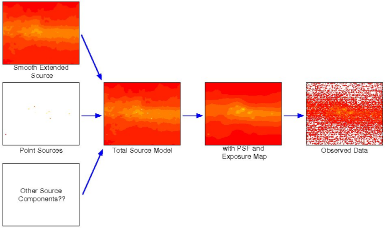
Incorporating Information Through Bayesian Prior Distributions

- Include a region where a point source is likely to exist.
- Encourage param values *similar* to those from better data.

Use of prior distributions offers a more flexible approach than setting parameters.

Outline

Is The Baseline Model Sufficient?



- Start with known parameterizied physical model (null).
- Residual is fit with a flexible multi-scale model.
- Is there structure in residual?

- We fit a *finite mixture distribution* with an unknown number of components:

  `Physical Model + Multi-Scale Residual`

- If we fit the two-component model, we can look for structure in the fitted residual.
- Tests are technically and computationally challenging.

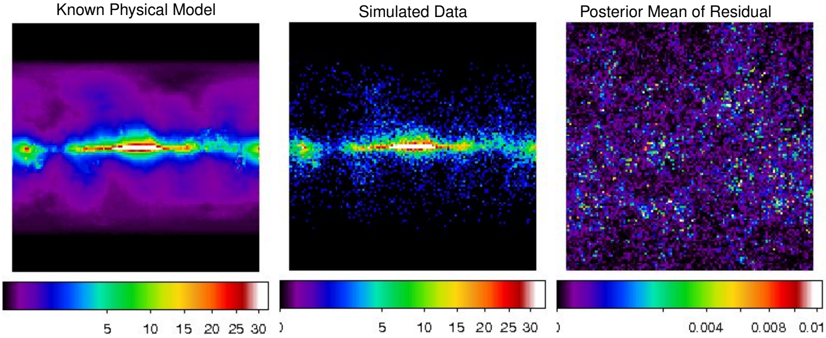
A Simulation Study

- We simulated data under the supposed physical model:

  `Physical Model`

- We fit the two component model:

  `Physical Model + Multi-Scale Residual`

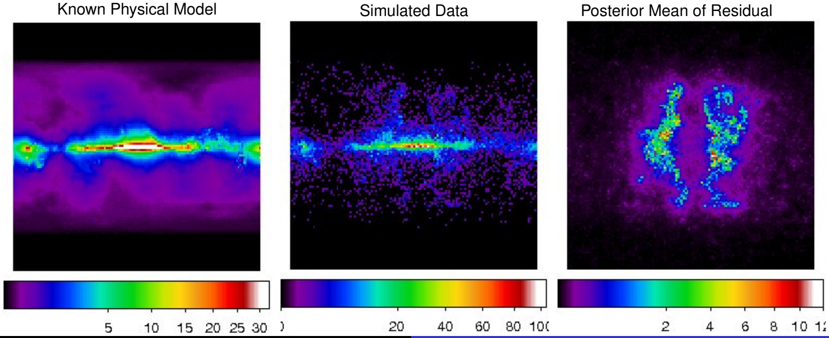


Simulation Under the Alternative

- We simulated data under a model with the supposed physical model plus a physically possible feature:

  `Physical Model + Multi-Scale Residual`

- We fit the same two component model.

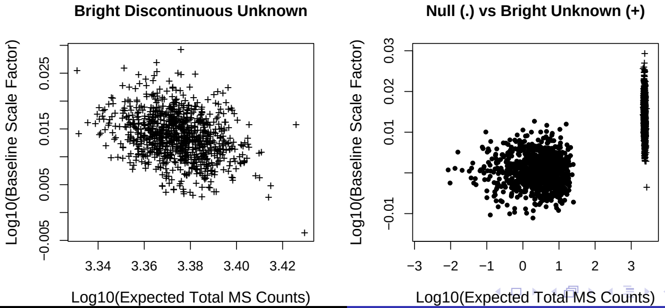


Evidence For The Added Component

We examine the joint posterior distribution of

1. Baseline Scale Factor: α
2. Expected Total MS Counts: β

in $\alpha\ \text{Physical Model} + \beta\ \dfrac{\text{Multi-Scale Residual}}{\sum \text{Multi-Scale Residual}}$.



Using a Bayesian prior to formulate frequentist test

A procedure:

1. Construct a prior distribution that favors a null hypothesis

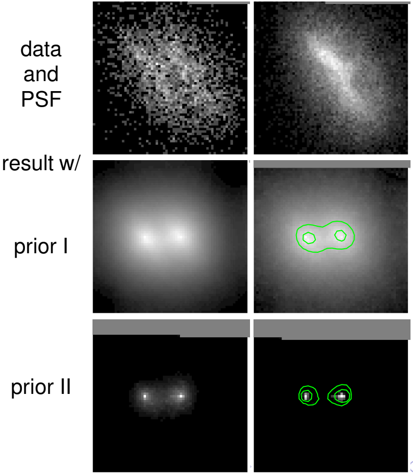
   H_0: object is a point source

2. Compute the posterior and evaluate the propensity of the alternative hypothesis

   H_A: an extended source

3. Using a test statistic, prior parameters can be used to set level (and power).

Simulation Result



Outline

Posterior Predictive P-values

- Is the deviation form the baseline model significant?
- Is the difference in the previous slide typical?
- For data generated under the null model, what is the sampling distribution of $\hat{\beta}$, the expected residual count under the two-component model?

We can answer these questions computationally:

- Sample replication datasets under the null model.
- Sample unknown parameters from their null posterior.
- Fit the two-component model to each replicate dataset.
- Compare the resulting distribution of $\hat{\beta}$ with the value fit to the actual data.

This strategy is computationally demanding!

Summary

- The search for highly irregular and unexpected structure in astronomical images posses many statistical challenges.
- Model-based methods allow us to make progress on formalizing an answering scientific questions.
- More Sophisticated computational methods and methods for summarizing high dimensional posterior distributions are yet to be explored.

For Further Reading I

- Connors, A. and van Dyk, D. A..
  How To Win With Non-Gaussian Data: Poisson Goodness-of-Fit.
  In *Statistical Challenges in Modern Astronomy IV*. to appear.

- van Dyk, D. A., Connors, A., Esch, D. N., Freeman, P., Kang, H., Karovska, M., and Kashyap, V.
  Deconvolution in High Energy Astrophysics: Science, Instrumentation, and Methods (with discussion).
  Bayesian Analysis, **1**, 189–236, 2006.

- Esch, D. N., Connors, A., Karovska, M., and van Dyk, D. A.
  An Image Reconstruction Technique with Error Estimates.
  The Astrophysical Journal, **610**, 1213–1227, 2004.

- Protassov, R., van Dyk, D. A., Connors, A., Kashyap, V. L. and Siemiginowska, A.
  Statistics: Handle with Care, Detecting Multiple Model Components with the Likelihood Ratio Test.
  The Astrophysical Journal, **571**, 545–559, 2002.