Fitting Narrow Spectral Lines in High Energy Astrophysics Using Incompatible Gibbs Samplers

David A. van Dyk (Department of Statistics, University of California, Irvine, USA, dvd@ics.uci.edu),
Taeyoung Park (Department of Statistics, Harvard University, USA, tpark@stat.harvard.edu), and
Aneta Siemiginowska (Harvard-Smithsonian Center for Astrophysics, USA, aneta@head-cfa.harvard.edu).

Abstract

Spectral analysis in high-energy astrophysics explores the the distribution of the energies of photons emitted from an astronomical source. The shape and structure of this distribution gives clues as to the composition, density, temperature, relative velocity, and distance of the source. Thus, spectral analysis is key to our understanding of the physical environment and structures of astronomical sources, the processes and laws which govern the births and death of planets, stars, and galaxies, and ultimately the structure and evolution of the universe. From a statistical point of view a typical stellar spectra can be formulated as a finite mixture distribution composed of one (or more) continuum terms and a set of emission line terms. While the continuum describes the general shape of a spectrum, each emission line represents a positive aberration from the continuum in a narrow band of energies. Emission lines are used to model the emission resulting from electrons falling to a lower energy shell in a particular ion. Thus, emission lines are important in the investigation of the composition of a source. The Doppler shift of the location of a known spectral line (such as a particular hydrogen line) can also be used to determine the relative velocity of a source. Thus, determining the precise location of emission lines is a critical task.

Here, we focus on a single narrow emission line that can be modeled with a Gaussian distribution or a delta function. Spectral data are typically contaminated by several non-trivial physical processes including non-homogeneous stochastic censoring, blurring of photon energies, and background contamination. To account for these processes, we construct a highly structured model that is formulated in terms of several layers of missing data. In this poster we discuss this model along with several specially designed MCMC samplers that are constructed as Gibbs samplers but use incompatible conditional distributions. We verify that the resulting chains have the target posterior distribution as their stationary distribution but are much faster to converge than more standard Gibbs samplers in this case. As an illustration, we apply our models, methods, and computational strategies to the X-ray spectrum of the high redshift quasar, PG1634+706.

Modeling the X-ray Spectrum

X-ray Spectra:

- The Gaussian assumptions that are inherent in traditional χ^2 fitting are inappropriate for low count data in each bin of the X-ray spectrum.
- Instead, we explicitly model photon arrivals as an inhomogeneous Poisson process (van Dyk et al., 2001).

The Basic Spectral Model: The X-ray spectrum can be separated into (1) a continuum term and (2) a set of emission lines.

- The continuum is described by a parametric form (e.g., a power law).
- The emission lines are statistically described by adding a narrow Gaussian, a narrow Lorentzian, or a delta function to the continuum.

Data Degradation:

- The photons are subject to various physical processes which significantly degrade the source model. Namely,
  1. absorption,
  2. the effective area,
  3. blurring of photon energies (stochastic redistribution), and
  4. background contamination.
- We design a highly structured multilevel spectral model with components for the data degradation processes (van Dyk et al., 2001).

Efficient X-ray Spectral Fitting

Hierarchical Structure of Missing Data:

- $\mathbf{Y}_{\text{mis}1}$: variables that describe the data contamination processes.
- $\mathbf{Y}_{\text{mis}2}$: a mixture component indicator variable that indicates which photons originated from the line and which from the continuum.

Partitions of Model Parameters:

- ψ : parameters for the continuum.
- μ : a parameter for the line location.
- ν : a parameter for the line width; for a delta function, ν is set to 0.

Difficulty with Identifying Narrow Emission Lines:

- When an emission line is modeled with a delta function, a Gibbs sampler used to fit the finite mixture distribution breaks down because the line location cannot move from its starting value.
- We suggest fitting the location of a narrow line without conditioning upon the mixture components indicator variables.

Incompatible Gibbs Samplers:

- Park and van Dyk (2006) devised several incompatible Gibbs samplers for a spectral model with a delta function or Gaussian line.
- Our goal is to construct quickly converging samplers with the target posterior distribution as their stationary distribution.

Figure 1: Illustration of deriving incompatible Gibbs samplers by partially marginalizing over the missing data. The top figure corresponds to the delta function line profile, and the bottom figure the Gaussian line profile.

Simulation Study

Four Different Cases: Data are simulated under the following cases.

- **Case 1**: No emission line.
- **Case 2**: A **narrow** and weak line at 2.85 keV with width 0.04 keV.
- **Case 3**: A broad and **strong** line at 3.4 keV with width 0.207 keV.
- **Case 4**: A **narrow** and strong line at 3.4 keV with width 0.04 keV.

Posterior Inferences:

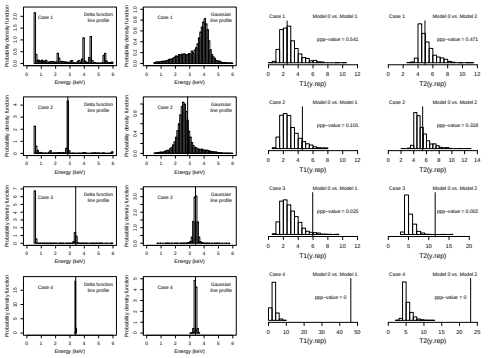

Figure 2: The left column corresponds to the posterior distribution of the **delta function line location** and the right column that of the **Gaussian line location**. The vertical solid lines represent the true line locations.

Figure 3: The left column corresponds to the posterior predictive checks for the **delta function emission line** and the right column for the **Gaussian emission line**. Small ppp-values (e.g., <0.1) give evidence for the line.

The Posterior Distribution of the Line Location:

- **Case 2** : With narrow lines, delta functions are better suited for fitting the data.
- **Case 3** : With broad lines, Gaussian lines are better suited.
- **Case 4** : With narrow strong lines, both line models work well.

Posterior Predictive Checks:

- We compare two models to quantify the evidence in the data for the emission line via $T(y_{\text{rep}}^{(\ell)}) = \log\{\sup_{\theta\in\Theta_1} L(\theta|y_{\text{rep}}^{(\ell)})/\sup_{\theta\in\Theta_0} L(\theta|y_{\text{rep}}^{(\ell)})\}$ for $\ell = 1,2,\ldots,1000$, where Θ_0 and Θ_1 represent the parameter spaces under MODELS 0 and 1, respectively, and $y_{\text{rep}}^{(\ell)}$ denotes replicated data sets under MODEL 0:
  - MODEL 0 : There is no emission line in the spectrum.
  - MODEL 1 : There is an emission line in the spectrum.
- Small posterior predictive p-values indicate MODEL 1 is preferable to MODEL 0.

Spectral Analysis of PG1634+706

The Quasar, PG1634+706:

- PG1634+706 is a redshift z=1.334 radio quiet and optically bright quasar. The source was observed with the *Chandra X-ray Observatory* as a calibration target six times on March 23 and 24, 2000.
- The fluorescent Fe-K-α emission line has been observed in the quasar rest frame of near **6.4 keV**, which corresponds to **2.74 keV** in the observed frame of PG1634+706.
- The **location** of the line indicates the ionization state of iron in the emitting plasma; the **width** of the line the velocity of the plasma.

Fitting a spectral model to the *Chandra* data of PG1634+706:

- For each of the six *Chandra* data sets, we fit the spectral model with a delta function or Gaussian emission line.
- After the fitting, we compare the Fe-K-α emission line (**2.74 keV**) observed in the other similar sources with the fitted line location in the X-ray spectrum of PG1634+706.

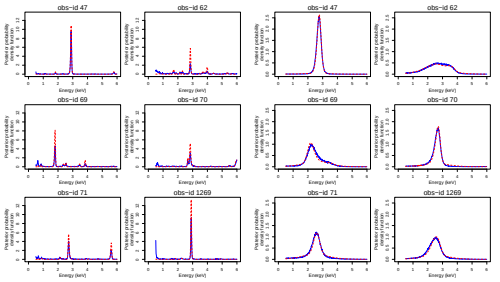

Figure 2: Posterior distributions of the **delta function line location** resulting from the spectral analysis of PG1634+706 data. Blue solid lines represent the posterior distributions, and red dotted lines the profile posterior distributions.

Figure 3: Posterior distributions of the **Gaussian line location** resulting from the spectral analysis of PG1634+706 data. Blue solid lines represent the posterior distributions, and red dotted lines the profile posterior distributions.

Posterior Inferences Given All Six Observations of PG1634+706:

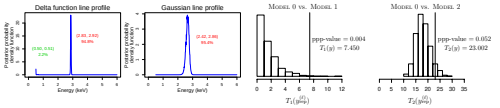

Figure 4: Posterior distributions of the delta function and Gaussian line locations given all six observations give evidence for the line near the Fe-K-α emission line (**2.74 keV**) observed in the other similar sources.

Figure 5: In each histogram, the observed test statistic (the vertical line) is compared with the test statistics from 1000 posterior predictive simulated data sets. Small ppp-values indicate stronger evidence of the emission line.

- The most probable line is located at $2.865^{+0.055}_{-0.035}$ **keV** for the delta function line profile; the posterior mean of the line location is 2.650 ± 0.222 **keV** for the Gaussian line profile.
- We compare three models for the delta function emission line:
  - MODEL 0 : There is no emission line in the spectrum.
  - MODEL 1 : There is an emission line at **2.74 keV** but unknown intensity.
  - MODEL 2 : There is an emission line with unknown location and intensity.
- We base our comparisons on the test statistic that is the sum of the loglikelihood ratio statistics for comparing MODEL m and MODEL 0, i.e.,

  $$T_m(y_{\text{rep}}^{(\ell)}) = \sum_{i=1}^{6} \log\left\{\frac{\sup_{\theta\in\Theta_m} L(\theta|y_{\text{rep}\,i}^{(\ell)})}{\sup_{\theta\in\Theta_0} L(\theta|y_{\text{rep}\,i}^{(\ell)})}\right\} \text{ for } m = 1,2 \text{ and } \ell = 1,2,\ldots,1000,$$

  where Θ_0, Θ_1, and Θ_2 represent the parameter spaces under MODELS 0, 1, and 2, respectively, and $y_{\text{rep}}^{(\ell)}$ denotes the collection of six replicated data sets under MODEL 0.
- The small ppp-values indicate MODELS 1 and 2 are preferable to MODEL 0.

Acknowledgments

We gratefully acknowledge funding for this partially provided by NSF Grants DMS-01-04129, DMS-04-38240, and DMS-04-06085 and by NASA Contracts NAS8-39073 and NAS8-03060(CXC). This work is a product of joint work with the California-Harvard astrostatistics collaboration (CHASC, www.ics.uci.edu/~dvd/astrostat.html) whose members include J. Chiang, A. Connors, D.A. van Dyk, V.L. Kashyap, X.-L. Meng, T. Park, Y. Yu, and A. Zezas.