Highly Structured Models in High Energy Astrophysics

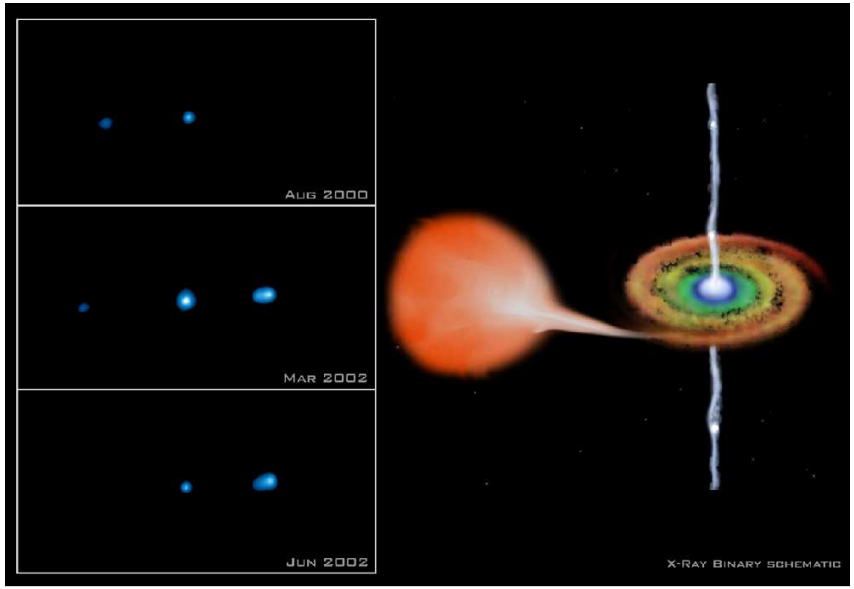
David A. van Dyk

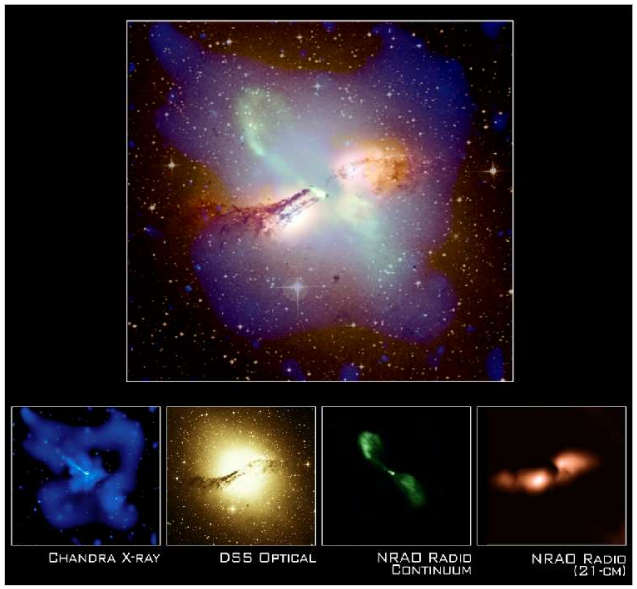
Department of Statistics
University of California, Irvine

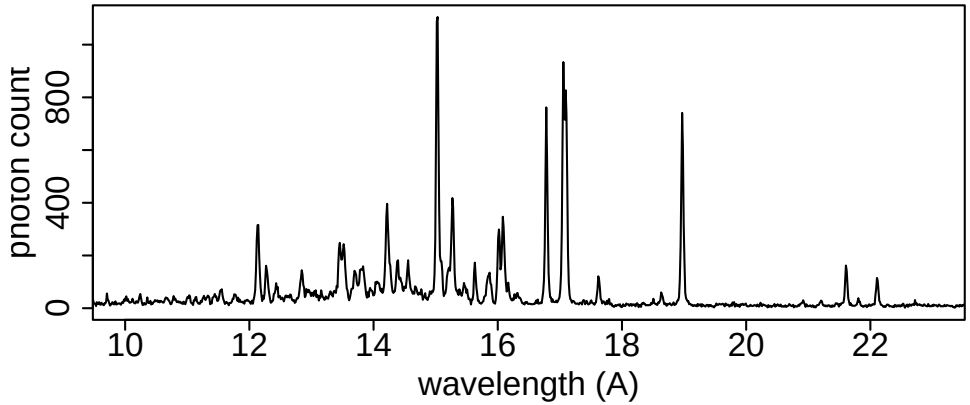
Joint work with
The California-Harvard Astrostatistics Collaboration

Reconstruction of the Physical Enviornment of a Stellar Corona
is joint with
Hosung Kang, Alanna Connors, and Vinay Kashyap

Complex Astonomical Sources

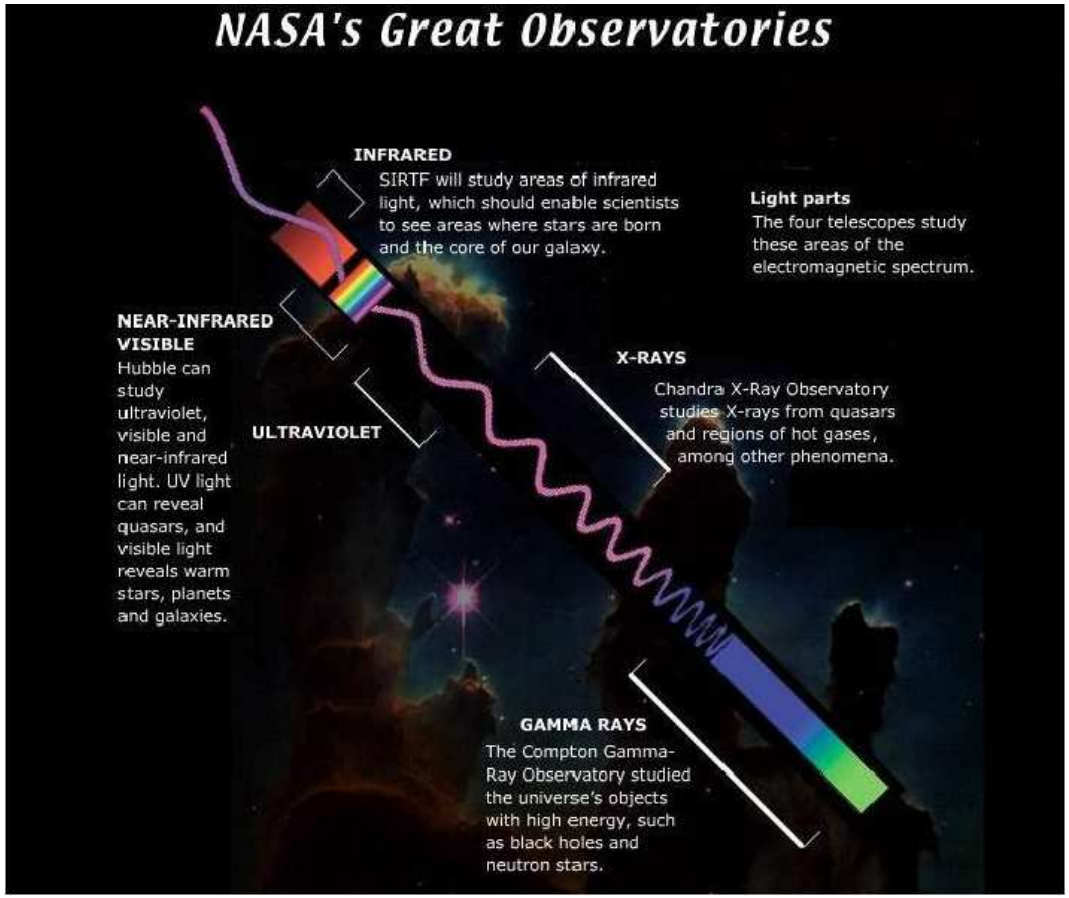






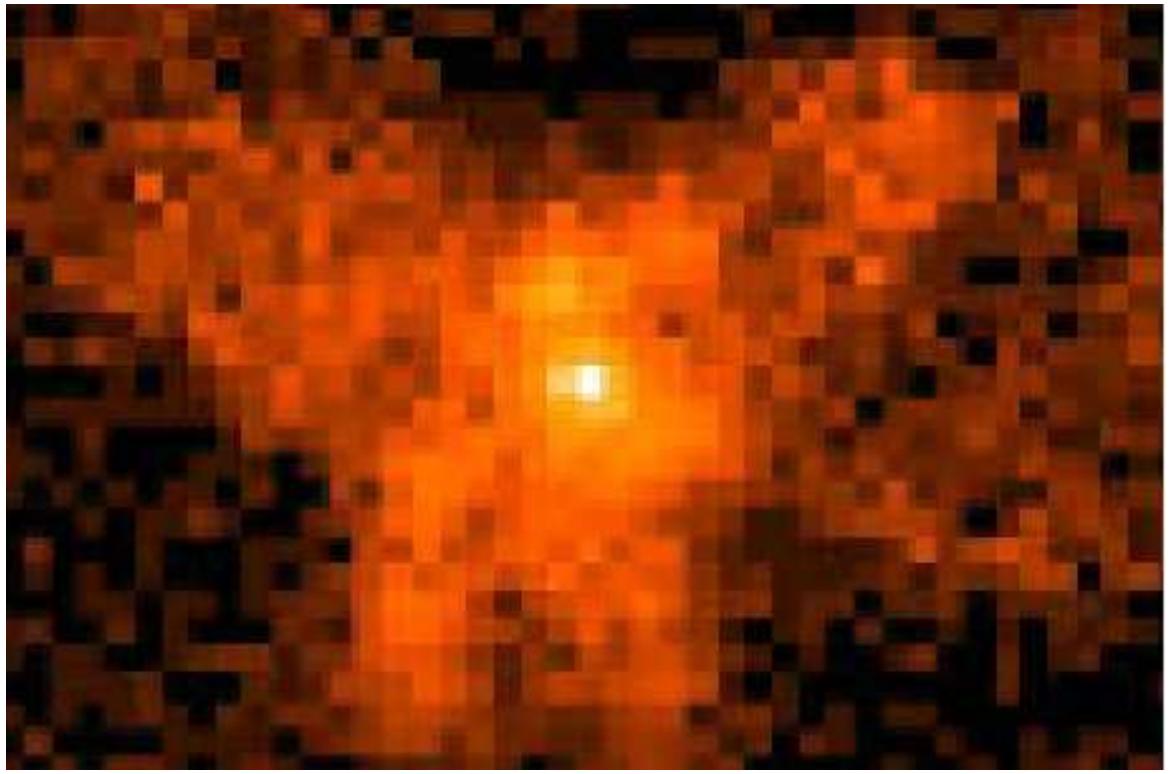
Images may exhibit Spectral, Temporal, and Spatial Characteristics.

Complex Data Collection Mechanisms



- A very small sample of instruments
- Earth-based, survey, interferometry, etc.
- X-ray alone: at least four planned missions
- Instruments have different data-collection mechanisms

Complex Scientific Questions

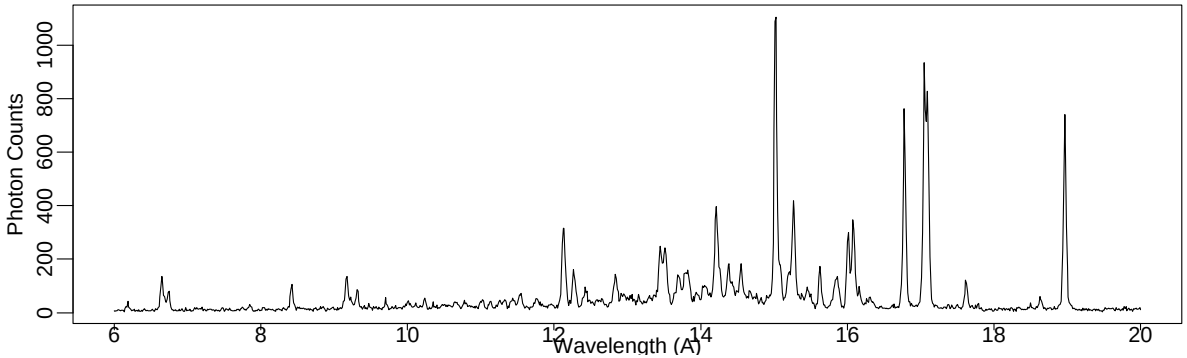


- Are the loops of hot gas *real*?

Outline of the Talk

I will examine a particular complex scientific question:

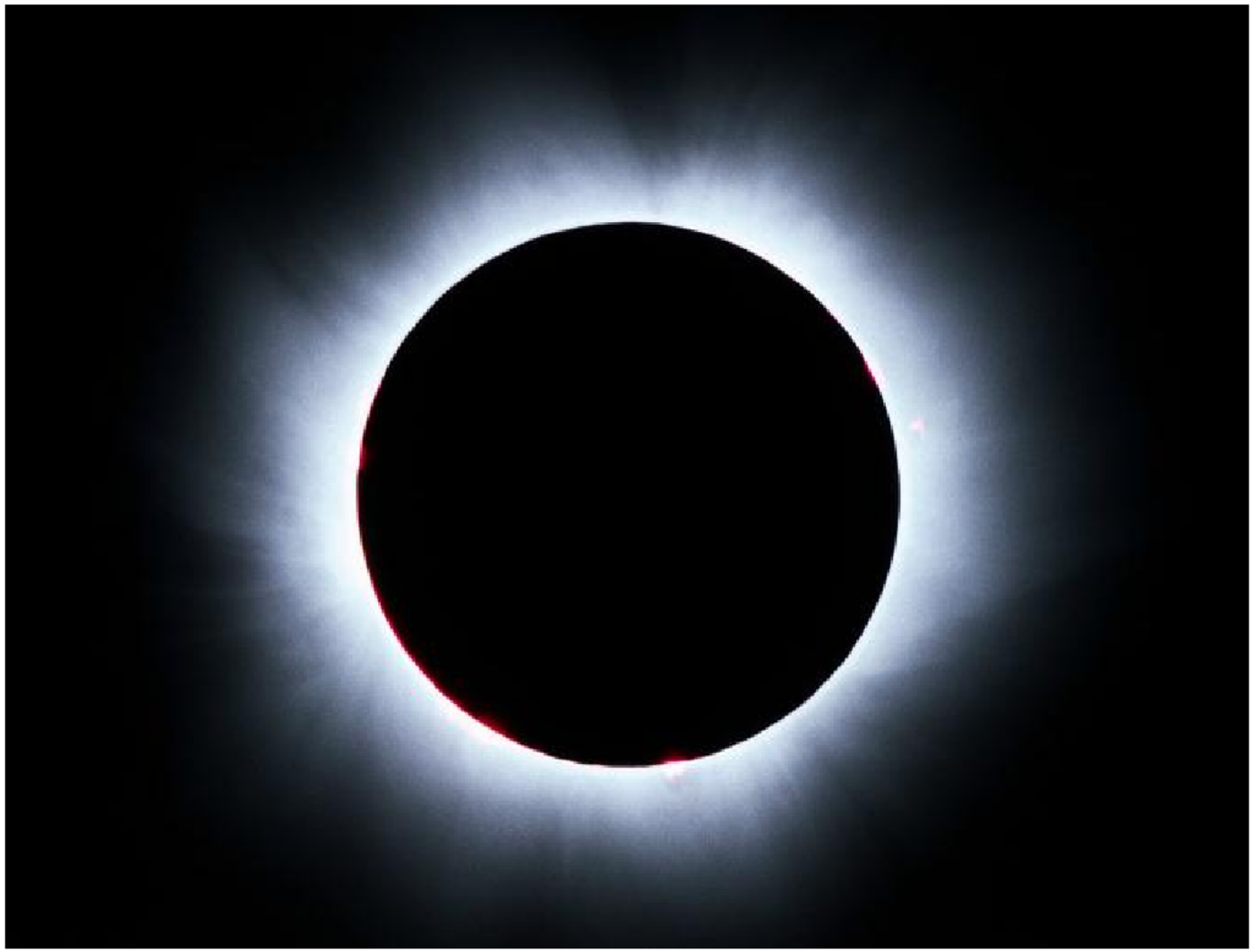
What does an ultra-high resolution spectrum tells us about the physical environment of a stellar source?



In particular, I will

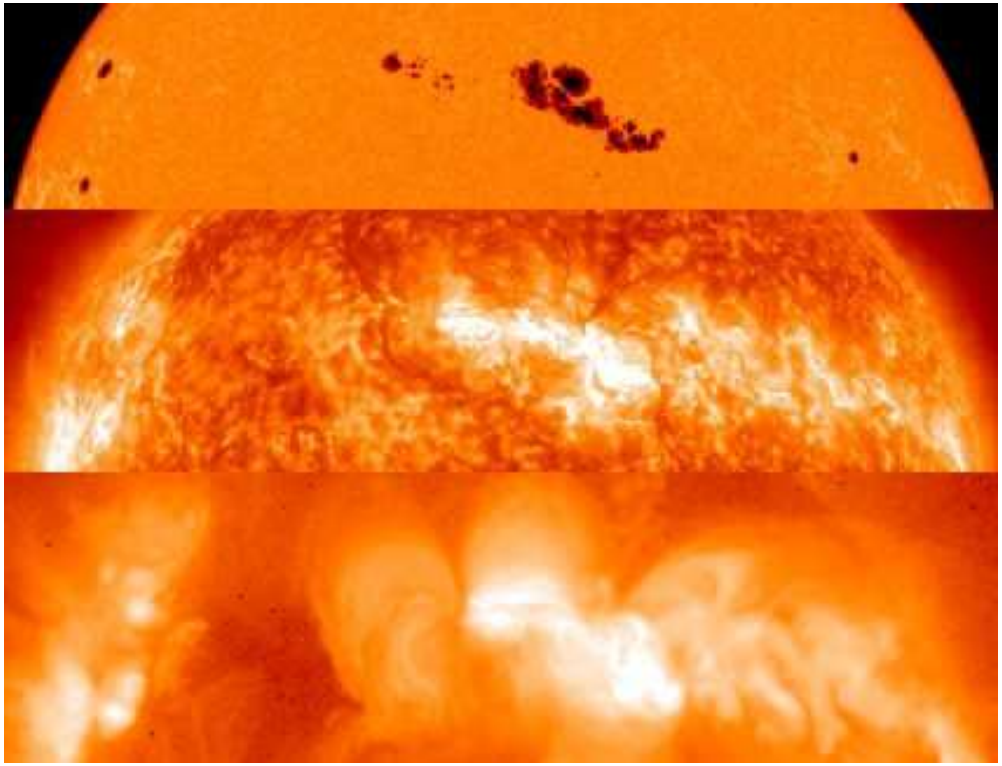
1. Build a statistical model the describes the physics underlying the spectrum,
2. Incorporate the data collection mechanism of the *Chandra X-ray Observ.*,
3. Discuss statistical inference under the model, and
4. Illustrate the method using simulation, data analysis, and model evaluation.

The Solar Corona During a Total Eclipse



The mostly X-ray emitting corona is fainter than the surface and is normally invisible.

The Solar Atmosphere

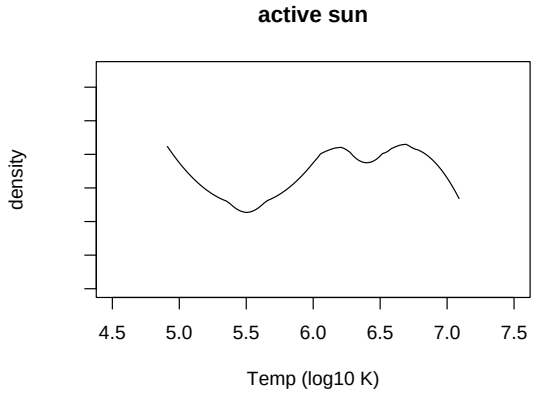


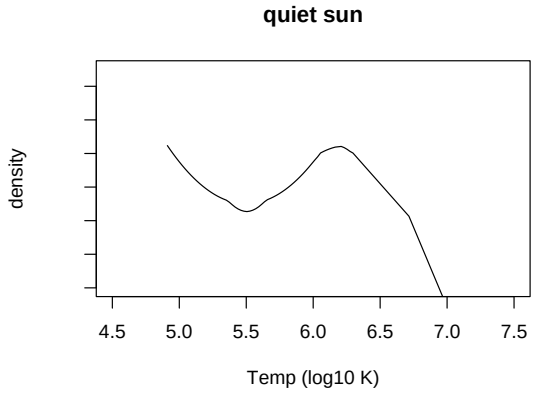
March 2001: Largest Sunspot Group in a Decade.

- Optical, Extreme UV, and X-ray Images
- Reveal different layers of atmosphere
- Higher Energy Emission
  → Hotter source
  → Extended atmosphere
- X-ray: Hot plasma arching high above the solar surface inside the loops of magnetic fields.

The complex structure in the X-ray emission across the solar corona is a tracer of the temperature and density of the plasma.

The Environment of the Solar Corona





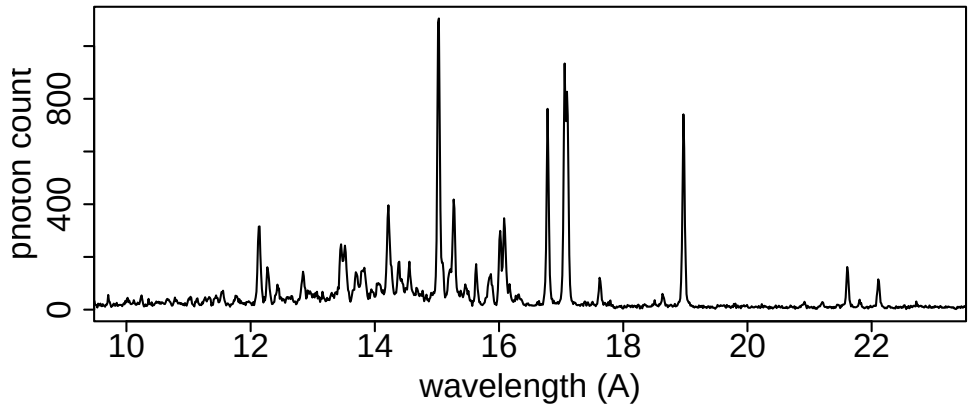
Temp density of coronal plasma
(DEM: Diffuse Emission Measure)

element	abundance (%/%H)
H	1.00000
He	0.07943
C	0.00039
N	0.00010
O	0.00077
Ne	0.00012
Mg	0.00014
Al	0.00001
Si	0.00013
S	0.00002
Fe	0.00013

There is MUCH less information available for stellar corona.

Data for a Stellar Corona

- No star except the sun can be imaged.
- Ultra-high resolution spectral data is available from the *Chandra X-ray Observatory*.



- *Chanda* counts photons in a large number of narrow spectral bins.

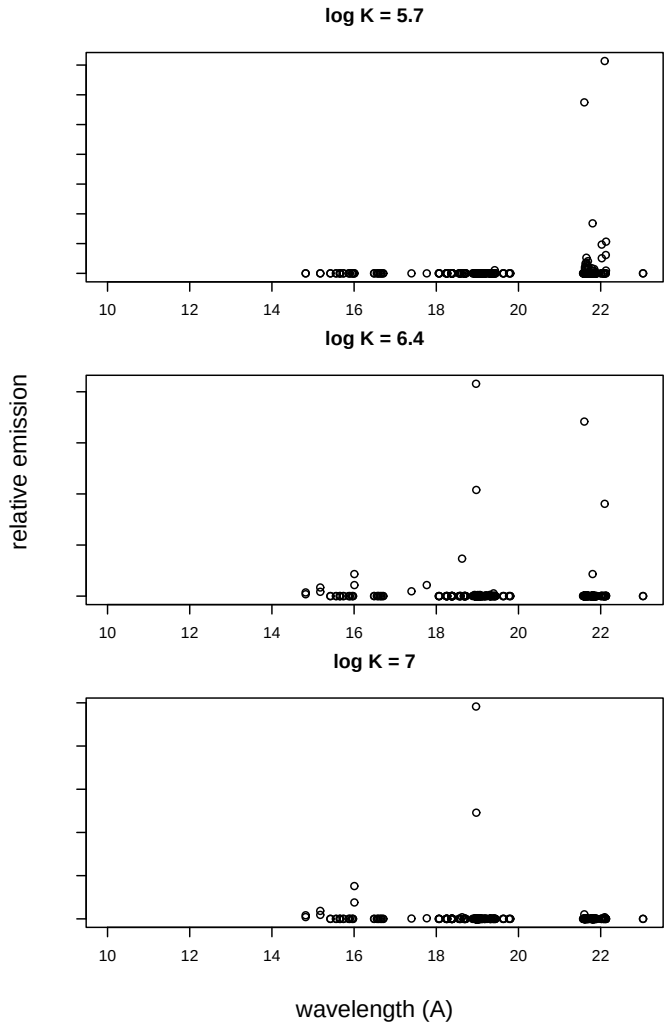
Unlocking the information in this forest of spectral lines requires subtle statistical analysis.

Physics of a Stellar Corona

- A stellar corona is made up of very hot plasma ($> 10^6$K).
- Ions are in an excited state: The electrons populate higher energy states.
- An (inelastic) collision of two ions:
  - The ions slow down;
  - Electrons jump to higher energy states;
  - Ions spontaneously decay to a lower more stable energy state; and
  - The difference in energy between the two states is emitted in the form of a photon.
- The energy difference is unique to the state transition of a particular ion.
- The frequency of a particular state transition is informative as to the temperature and density of the source.

Each line in the forest can be identified with a particular ion, and thus we obtain information on the environment in a stellar corona.

Identifying the Temperature Distribution



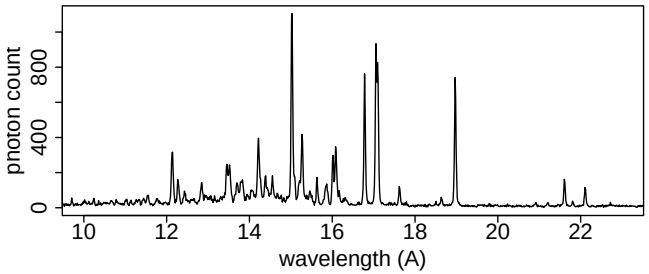
- The relative strength (*Emissivity*) of the Oxygen lines varies with the temperature of the plasma.
- If the corona is relatively hot, we expect the emission lines that correspond to more energetic quantum states to be relatively strong.
- The relative size of the Oxygen lines is informative as to the temperature of the plasma.

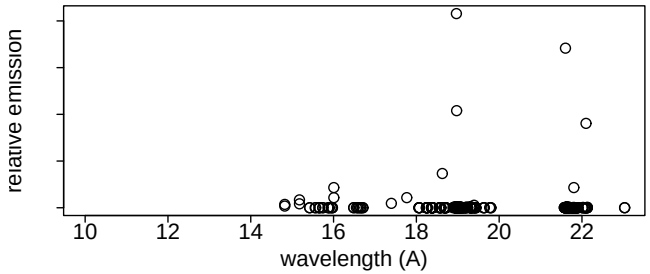
The data is a mixture of elements each at a mixture of temperatures. We aim to identify these two mixtures.

A Finite Mixture Distribution

- Photons are counted in a large number of narrow energy bins.
- We use a multinomial distribution with probability vector Π.
- We divide temperature into a small number of bins.

$$\Pi = \sum_{\text{temp}} \pi_t^{\text{T}} \left\{ \sum_{\text{elements}} \pi_e \left(\Upsilon_{et}^{\text{E}} + \Upsilon_{et}^{\text{C}} \right) \right\}$$





- Here we add a continuum term, Υ_{et}^{C}, which is the result of another physical process in a stellar corona.

The Complete Data

A complete-data table

	Temperature Bins			
Element	5.0	5.1	5.2	etc.
N	5	1	10	
O	6	7	8	
Ne	15	10	25	
Mg	5	3	10	
Al	9	14	10	
Si	0	1	1	
etc.				

For each photon, we would like to know:

1. The element that emitted the photon, and
2. The temperature of the plasma where that emitting element resided.

This *complete data* could be compiled into a two-way table of photon counts.

The cell probabilities for this table are of primary scientific interest.

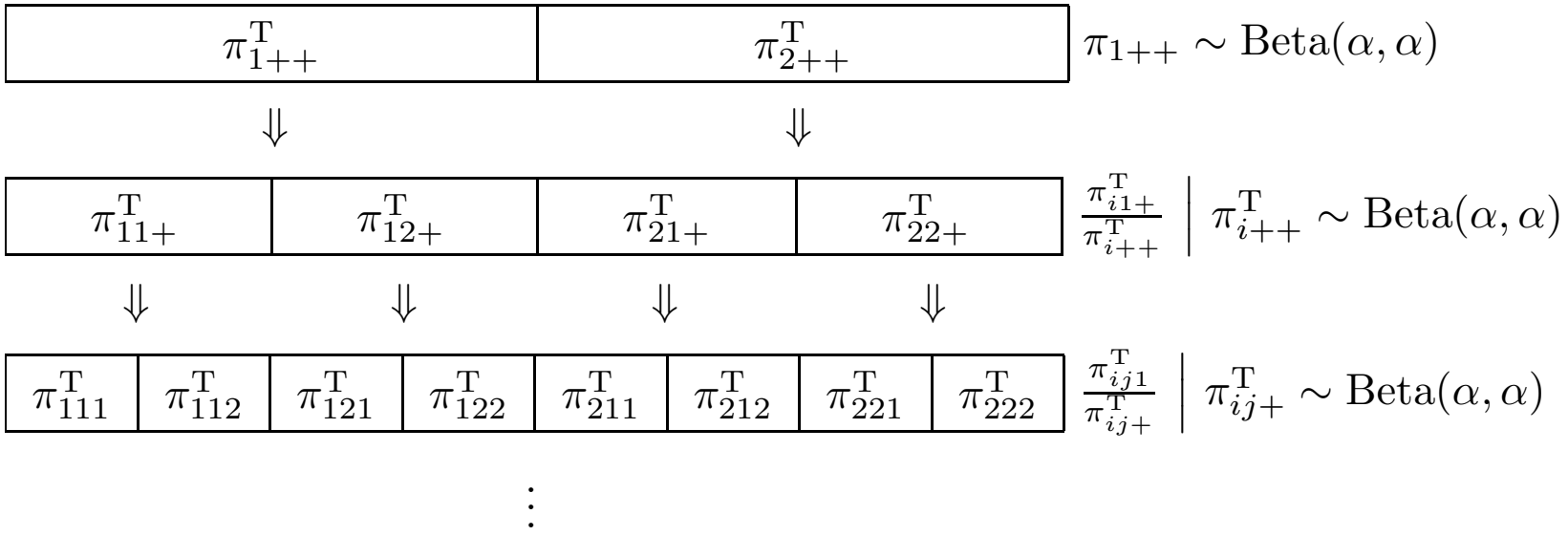
The Complete-Data Model

- We use an independence model on the complete-data two-way table: We assume the temperature distribution is the same for each element.
- Independent prior distributions are used on the probability vectors for the marginal tables of elements and temperature bins.
  - A Dirichlet prior could be used on the elemental probability vector to shrink toward a particular value. Currently we use a flat prior distribution.
  - A smoothing multiscale prior distribution is used on the temperature probability vector.

The multi-scale prior distribution smooths toward a smooth distribution on the temperature of the coronal plasma.

A Multiscale Prior Distribution

Low Resolution

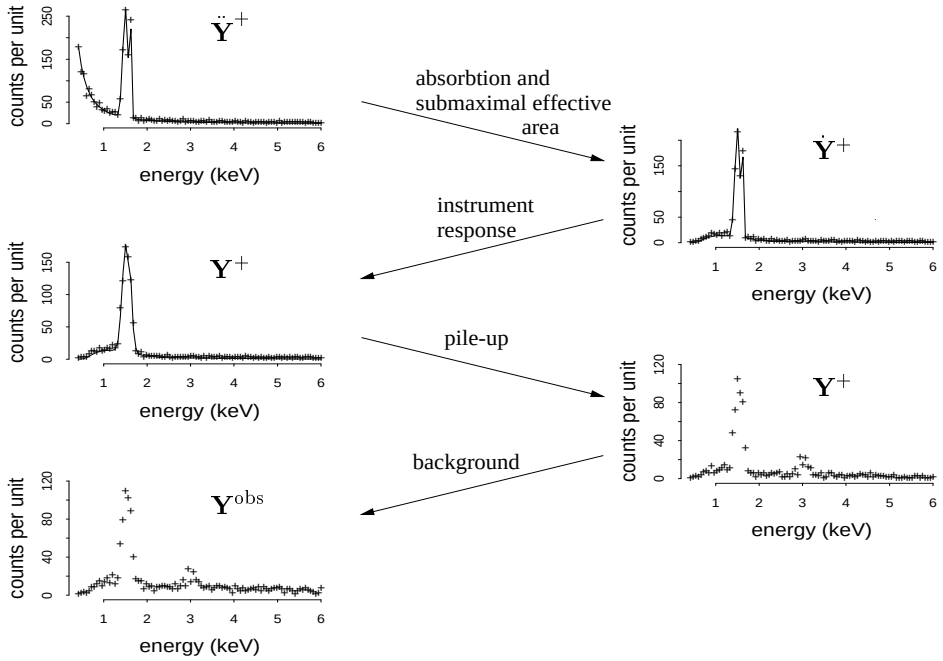


High Resolution

Unfortunately, the data collection process is further complicated by the missing data mechanism.

Modeling the Missing Data Mechanism

Modelling the *Chandra* data collection mechanism.



- The method of Data Augmentation: EM algorithms and Gibbs samplers.
- We can separate a complex problem into a sequence of problems, each of which is easy to solve.

We wish to directly model the sources and data collection mechanism and use statistical procedures to fit the resulting highly-structured models and address the substantive scientific questions.

A Model-Based Statistical Paradigm

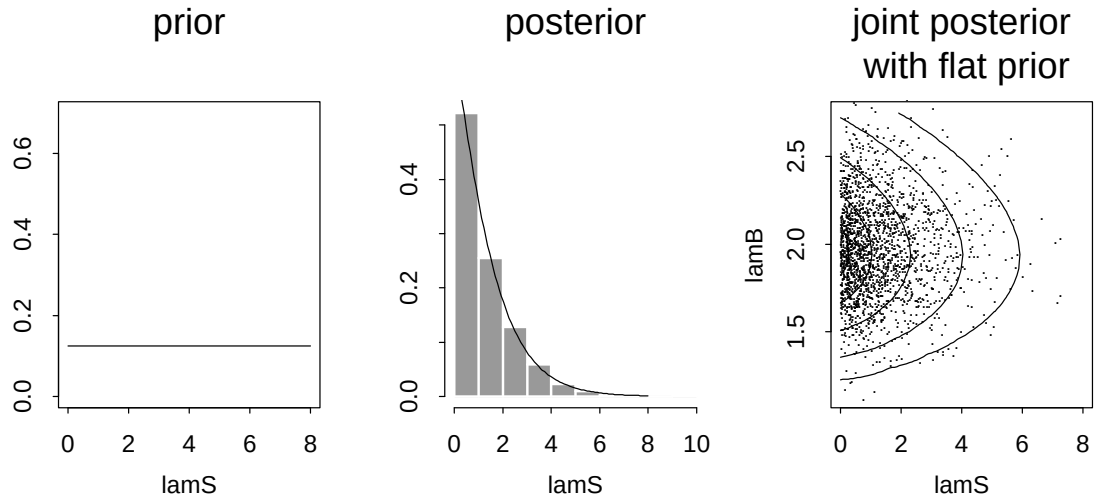
1. Model Building
   - Model source spectra, image, and/or time series
   - Model the data collection process
     - background contamination
     - instrument response
     - effective area and absorption
     - pile up
   - Results in a highly structured hierarchical model
2. (Model Based) Mode of Statistical Inference
   - Bayesian posterior distribution
   - Maximum likelihood estimation
   - Asymptotic approximations (e.g., χ^2 fitting)
3. Sophisticated Statistical Computation Methods Are Required
   - Goals: stability and easy implementation
   - Emphasize natural link with models: *The Method of Data Augmentation*

Bayesian Inference Using Monte Carlo

The Building Block of Bayesian Analysis

1. The sampling distribution: $p(Y|\psi)$.
2. The prior distribution: $p(\psi)$.
3. Bayes theorem and the posterior distribution: $p(\psi|Y) \propto p(Y|\psi)p(\psi)$

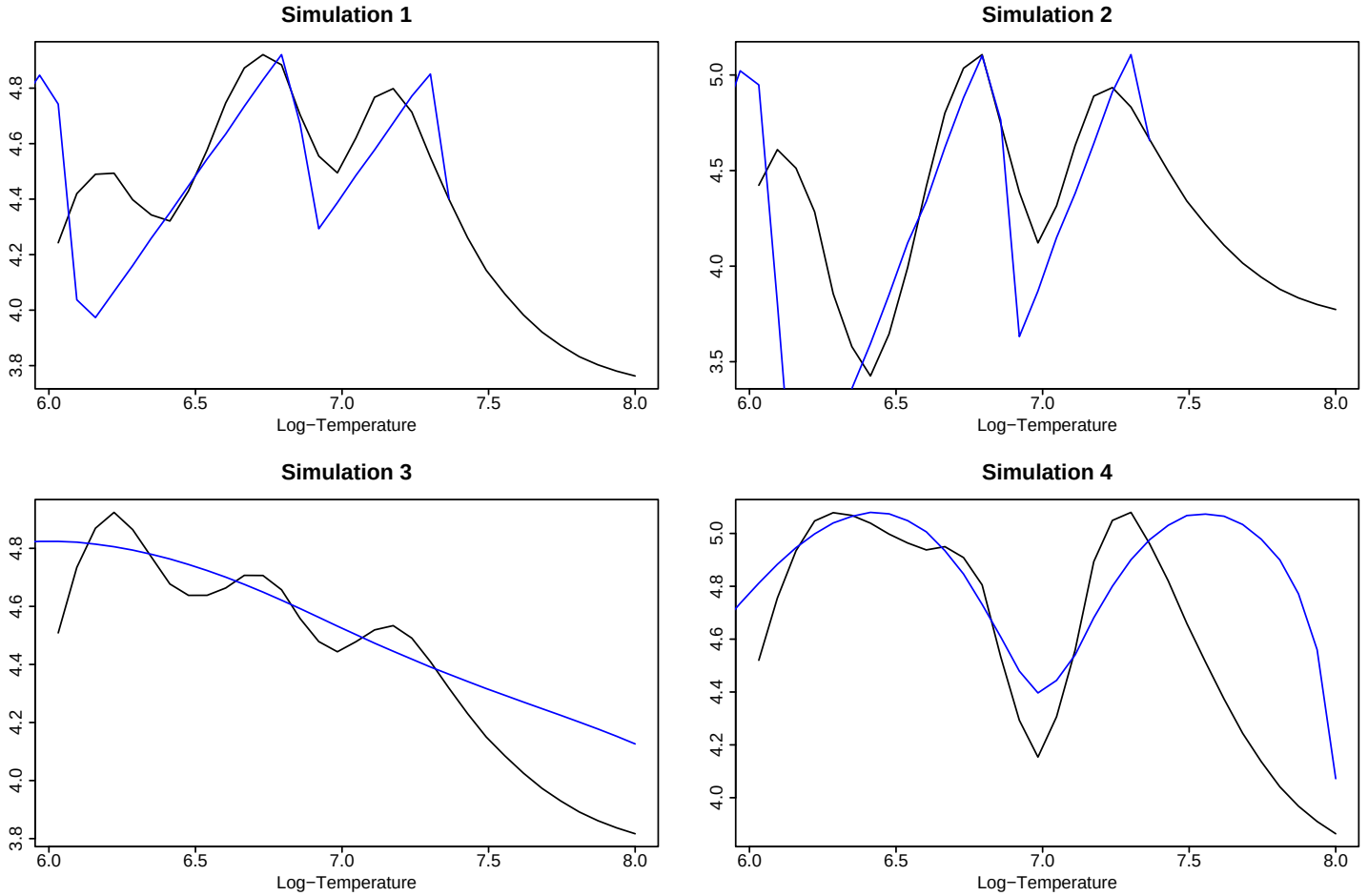
Inference Using a Monte Carlo Sample:



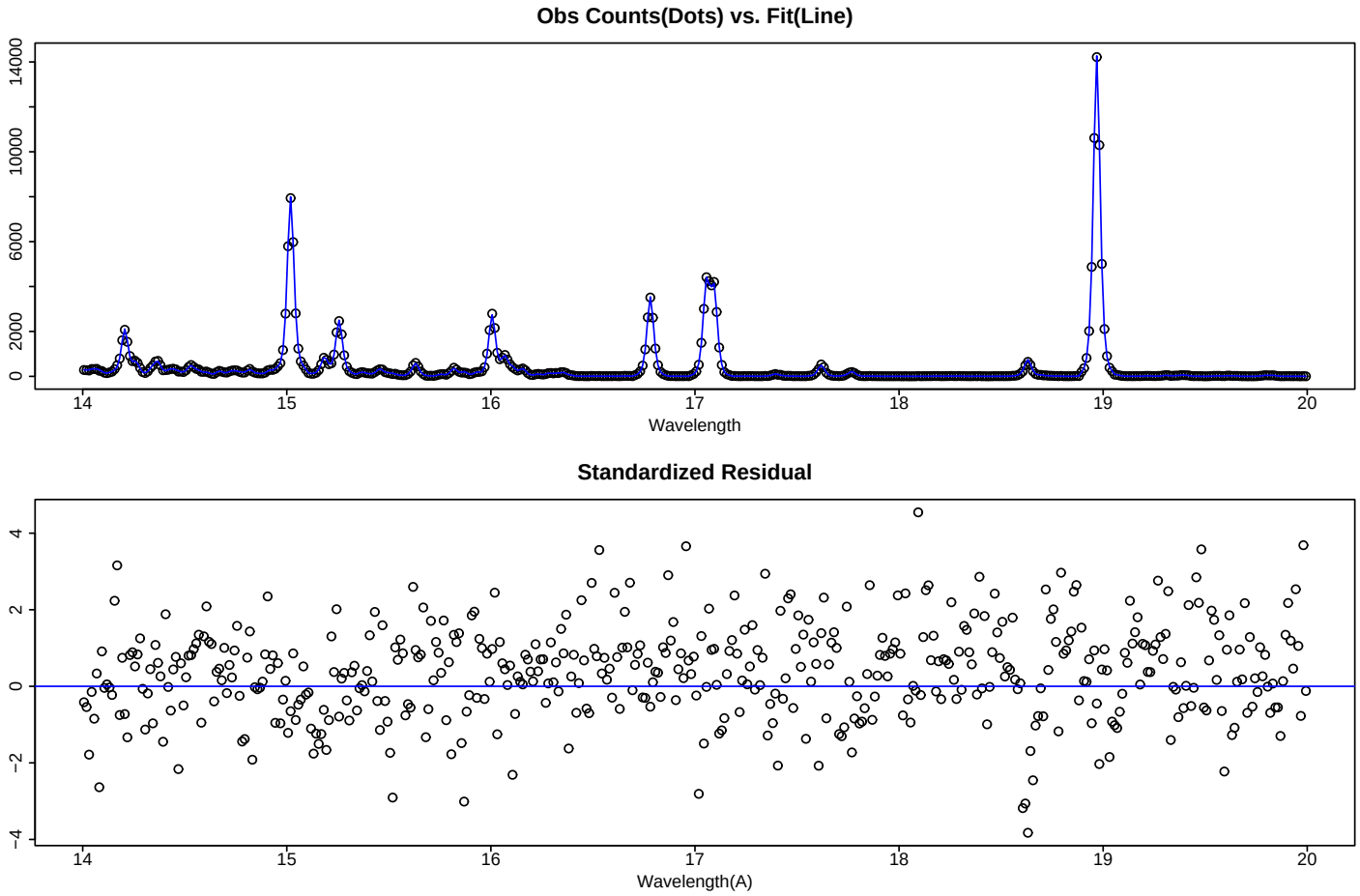
Obtain a Monte Carlo sample via the Gibbs Sampler:

How it All Works to Deconvolve a Stellar Spectrum

A Simulation Study:

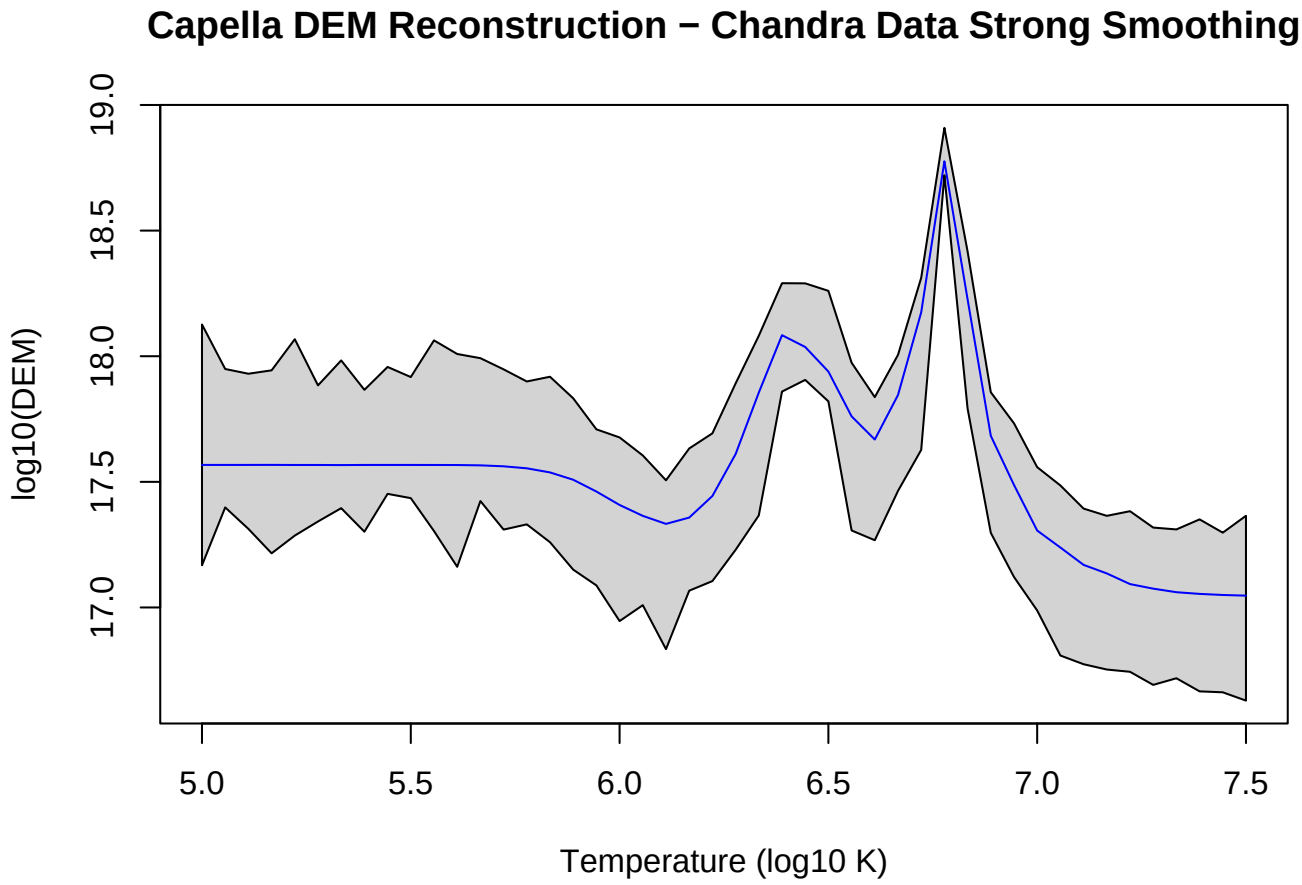


Evaluating the Fit in Simulation 4



std residual: (fit - count)/sqrt(fit)

Reconstruction of Capella's DEM



Posterior means (blue line) and 95% pointwise posterior intervals (grey area).

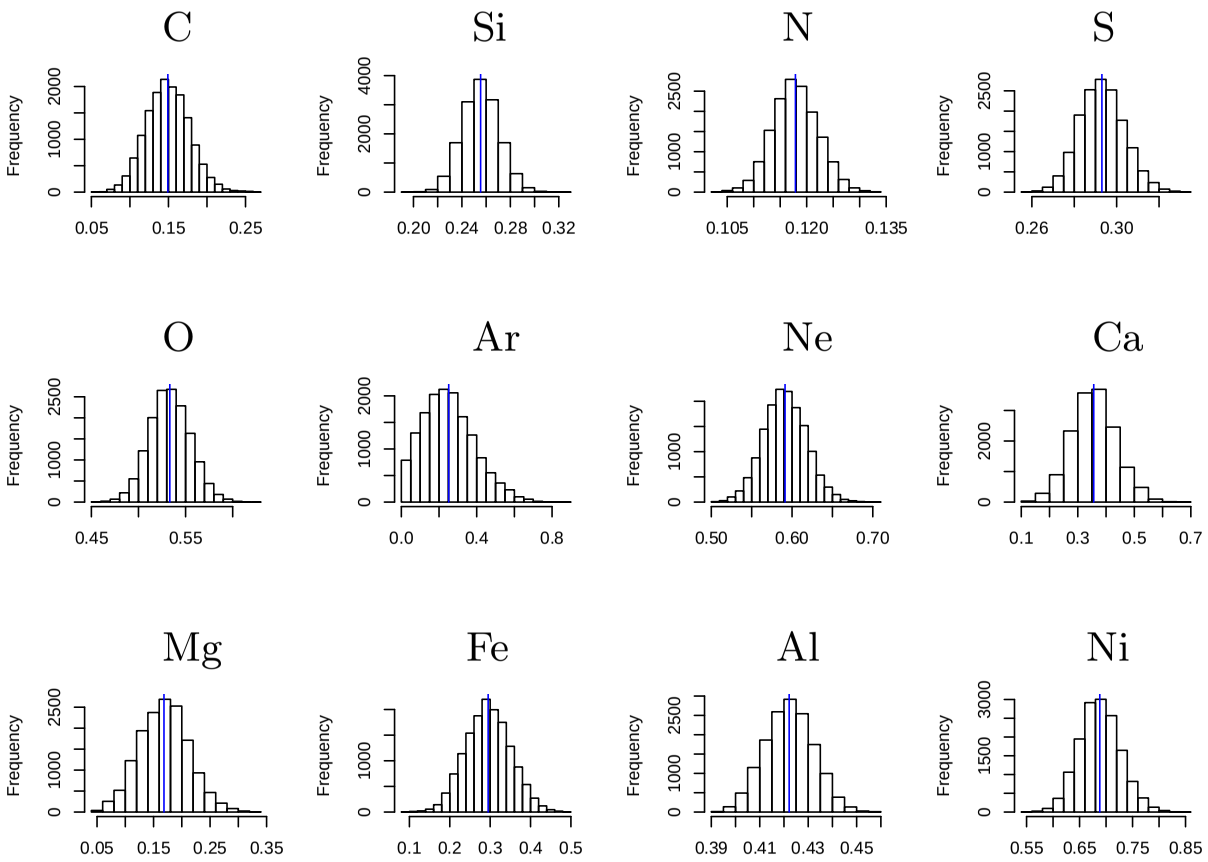
DEM : the distribution of the temperature of the coronal plasma.

Reconstructing Capella's Elemental Abundances

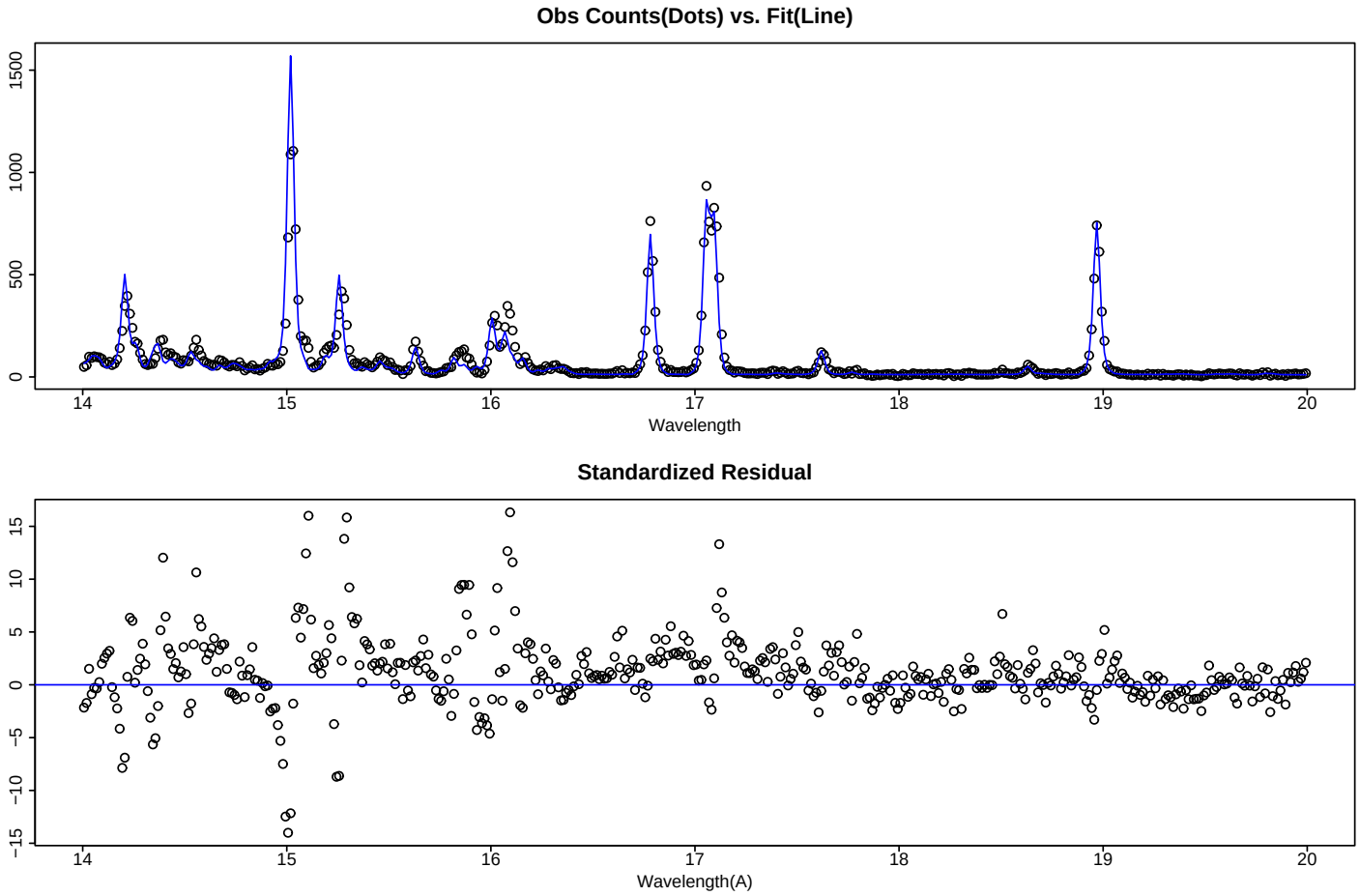
	Chandra (3-30A)			
Element	no. of lines	Mode	Mean	95% Interval
C	25	0.155	0.149	(0.097, 0.205)
Si	306	0.266	0.255	(0.227, 0.286)
N	500	0.122	0.118	(0.110, 0.126)
S	381	0.300	0.293	(0.273, 0.315)
O	279	0.542	0.533	(0.492, 0.577)
Ar	99	0.235	0.251	(0.025, 0.555)
Ne	345	0.599	0.591	(0.540, 0.644)
Ca	300	0.362	0.356	(0.206, 0.517)
Mg	369	0.177	0.168	(0.085, 0.256)
Fe	374	0.303	0.295	(0.190, 0.405)
Al	5779	0.428	0.422	(0.403, 0.442)
Ni	1832	0.707	0.688	(0.616, 0.767)

Relative Abundances (abundance over solar abundance).

Capella Relative Abundances



Evaluating the Fit



std residual: (fit - count)/sqrt(fit)

Improving the Fit

Sources of Errors

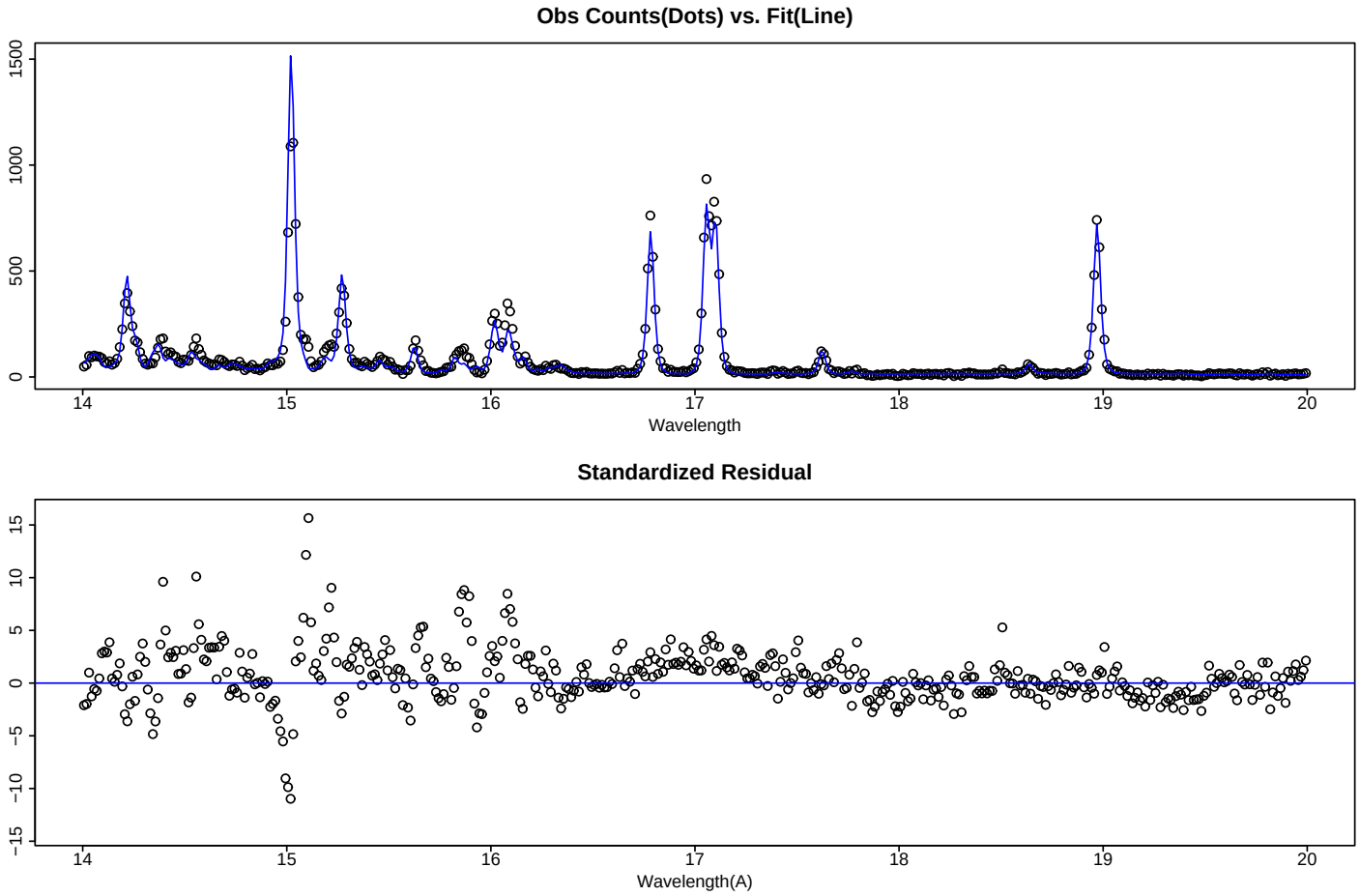
- Elements of the Emissivity Matrix are measured with error.
- There are many many weak lines, some of which are not accounted for.

Possible Solutions

- *Multiple Imputation*
  - Impute Emissivity matrices according to their measured errors.
  - Redo the analysis with each matrix.
  - Average the results.
  - DIFFICULTY: Errors are recorded, but correlations are not.
- Fit particular structures in the Emissivity Matrix.
  - The residual plots indicate that some of the strong lines should be shifted.

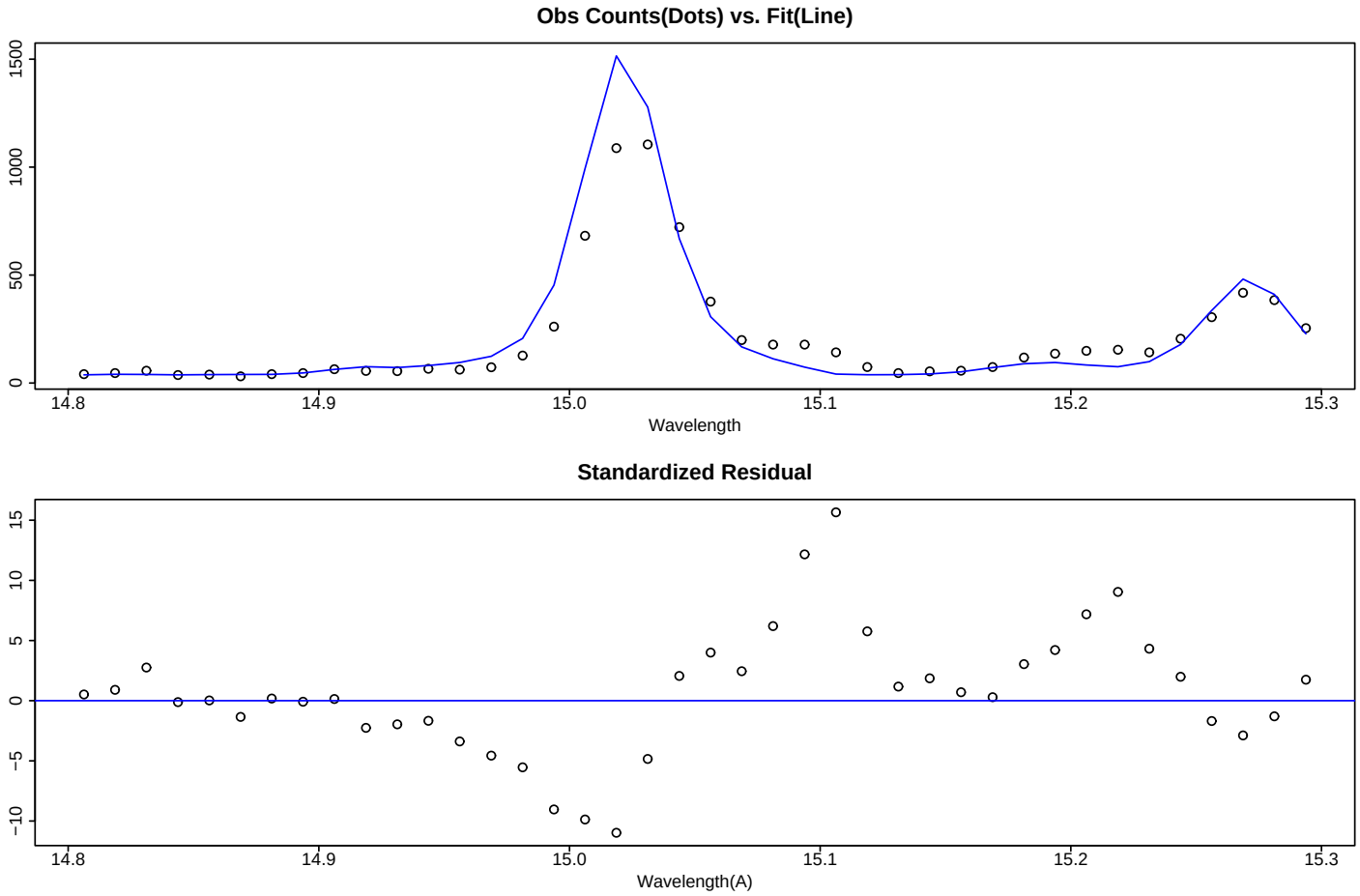
These are areas of current research.

Accounting for Errors in the Emissivity Matrix



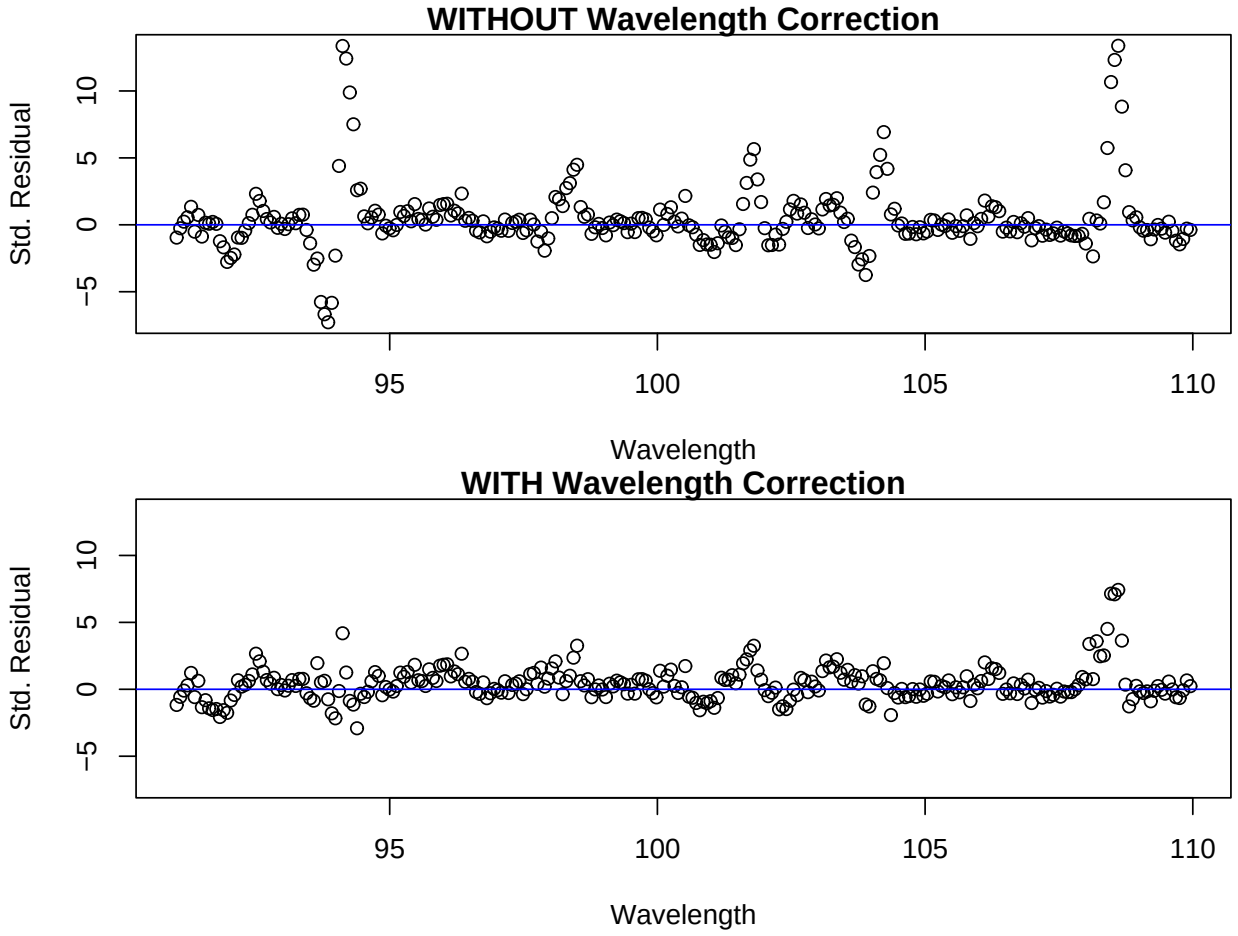
std residual: (fit - count)/sqrt(fit)

Magnification of Counts near 15 A

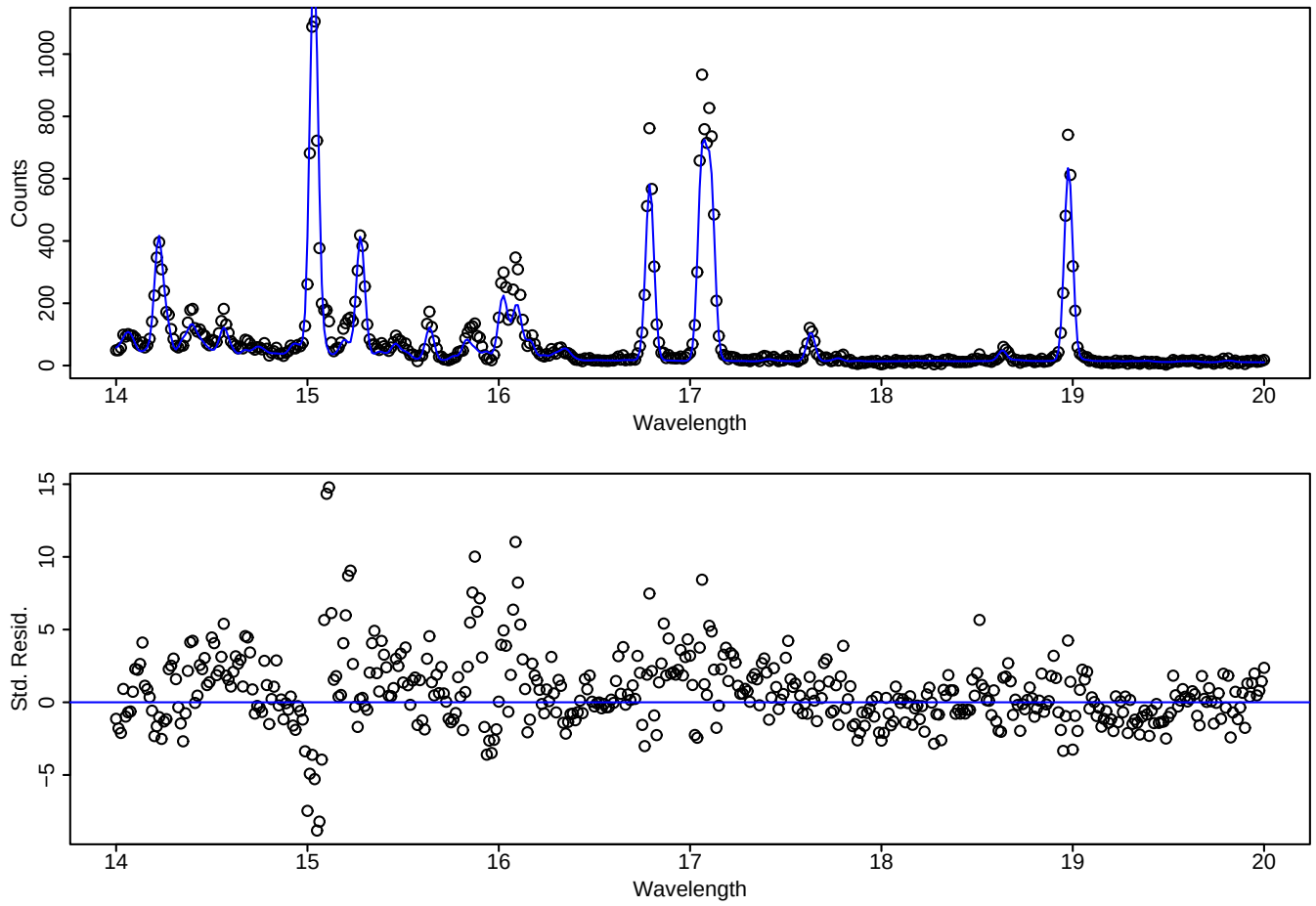


std residual: (fit - count)/sqrt(fit)

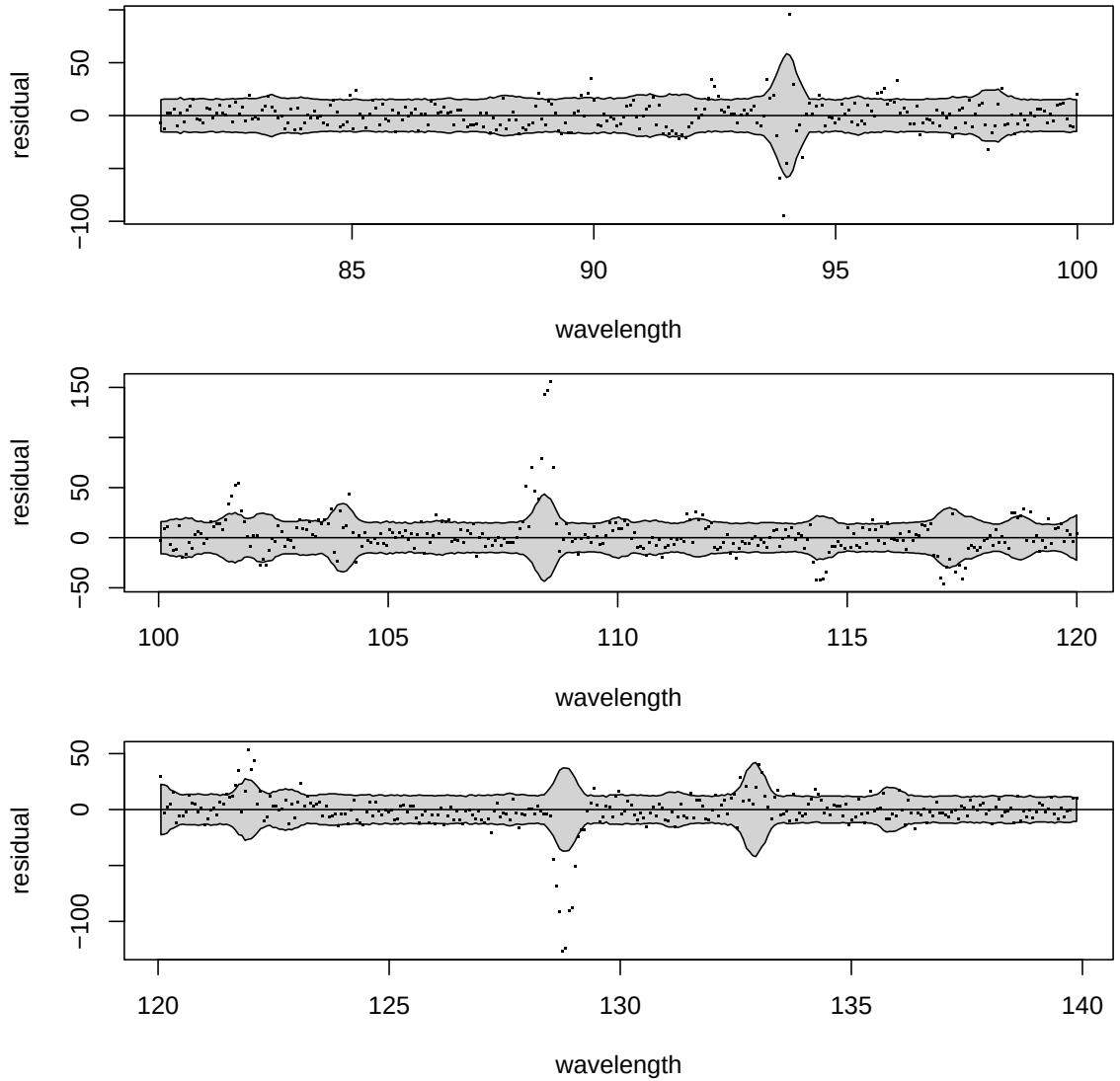
Fitting Line Locations to Improve the EUVE Fit



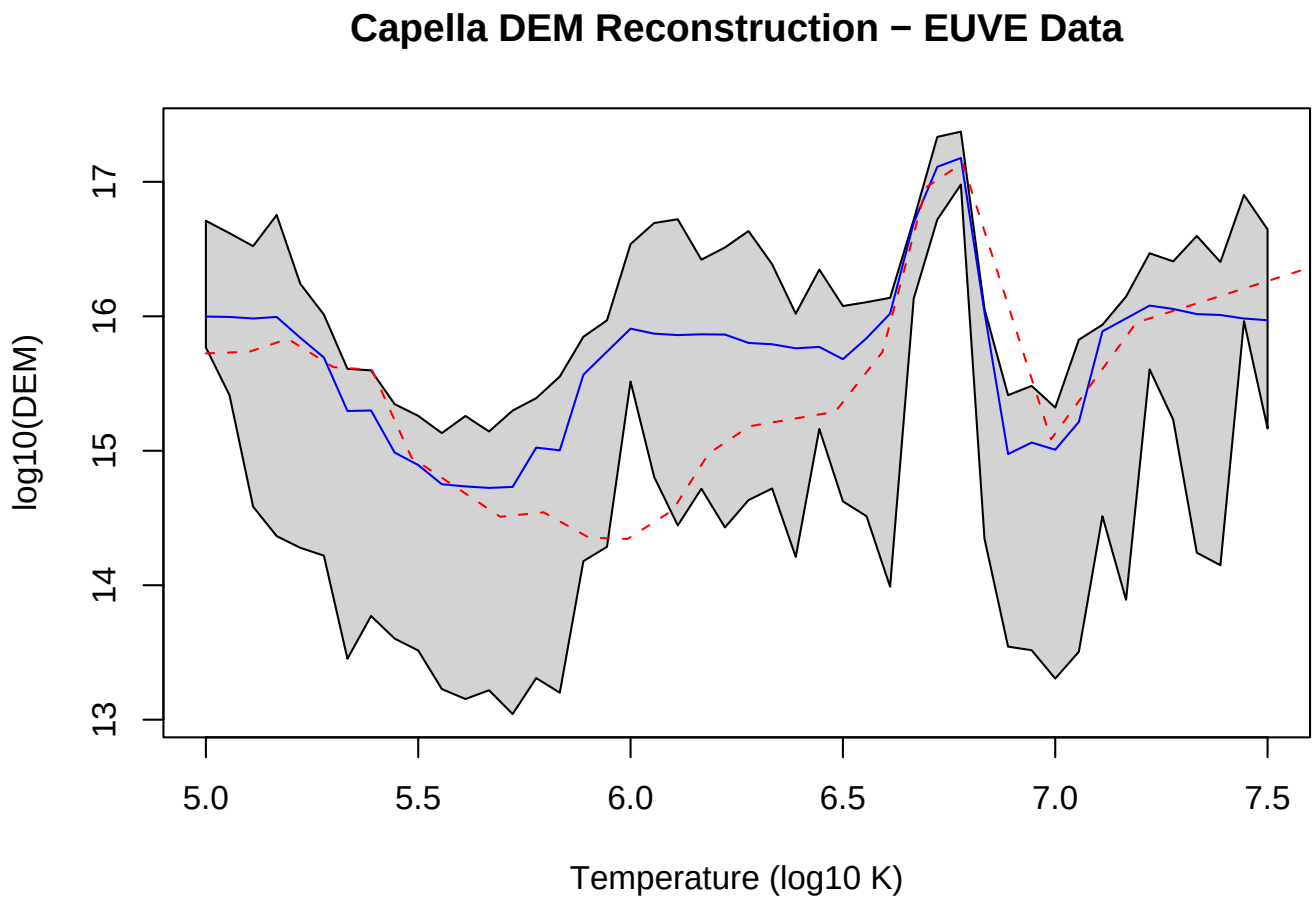
Fitting Line Locations to Improve the Chandra Fit



Evaluating the EUVE Fit



Fitting Capella's DEM using EUVE Data



Data Analysis in The New Millennium!

1. Application specific model are becoming ever more prevalent.
2. Model can be designed that explicitly account for complex systems and complex data collection mechanisms—*including missing data mechanisms*.
3. Ideally such models aim to directly answer complex substantive questions.
4. A Bayesian framework allows us to incorporate information from a variety of sources (e.g., instrumental calibration and quantum physical calculations).
5. Using such models requires sophisticated statistical inference and computational techniques.

References

van Dyk, D. A., Connors, A., Kashyap, V. L., & Siemiginowska, A. (2001). Analysis of Energy Spectrum with Low Photon Counts, *The Astrophysical Journal*, vol. 548, 224–243.

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

van Dyk, D. A. and Kang, H. (2003). Highly Structured Hierarchical Models for Spectral Analysis in High Energy Astrophysics. *Statistical Science*, vol. 19, 275–293.

Esch, D. N., Connors, A., Karovska, M., and van Dyk, D. A. (2004). A Image Restoration Technique with Error Estimates. *The Astrophysical Journal,* vol. 610, 1213–1227.

van Dyk, D. A., Connors, A., Esch, D. N., Freeman, P., Kang, H., Karovska, M., Kashyap, V. (2004). Deconvolution in High-Energy Astrophysics: Science, Instrumentation, and Methods (with discussion). Submitted to *Bayesian Analysis*.