Accounting for missing lines in atomic emissivity databases using DEM analysis with high-resolution X-ray Spectra

Victoria Liublinska
Joint work with CHASC: V. Kashyap, X.-L. Meng, D. van Dyk

Harvard University

1 Dec, 2009

Main objective

- Using observed stellar emission spectrum estimate the shape of the DEM (differential emission measure)
- Use discrepancy between predicted and observed counts to identify lines that were omitted in the atomic emission table.

Main objective

- Using observed stellar emission spectrum estimate the shape of the DEM (differential emission measure)

- Use discrepancy between predicted and observed counts to identify lines that were omitted in the atomic emission table.

Observed data:

- Data source is Chandra X-ray Telescope that observes counts from active G-type binary Capella (LETG and MEG)

- Counts are recorded in a certain prespecified number of channels with varying width (*spectral resolution*)

Model: Parameters and latent variables

- Let $Y_i^{obs} \sim Pois(\xi_i)$, where ξ_i is photon intensity in channel $i, i = 1 \dots I$ and $\xi_i = \xi_i^{source} + \xi_i^{bkg}$
- Let λ_j be *true intensities* that correspond to each of J bins.

  After taking into account distortion effect from the instrument we get $\xi^{source} = Md\lambda$, where d is a vector of effective area or ARF (*censoring probability*) and M is a JxI probability matrix that represents RMF (*blurring effect*) with column sums =1.
- For Chandra LETGS the blurring effect can be described my scaled t_4 and vector d is known.

Model: Parameters and latent variables

- Let $G^{C,k}(T)$ be contribution function (or $Jx2^R$ matrix) coming from *continuum* from element k at temperature T.
- Let $G^{L,k}(T)$ be contribution function (or matrix) coming from all *lines* of element k at temperature T.
- γ - abundance ($Kx1$ vector), μ - DEM ($2^R x1$ vector)
- Each true bin intensity consists of $\lambda_j = \lambda_j^C + binned\{\lambda_l^L\}$, where $\lambda_j^C = \sum_k \lambda_j^{C,k}$ and $\lambda_l^L = \sum_k \lambda_l^{L,k}$
  $\lambda_j^{C,k} \propto \gamma_k G^{C,k} \mu$ and $\lambda_l^{L,k} \propto \gamma_k G^{L,k} \mu$.

Illustration of step-by-step data degradation

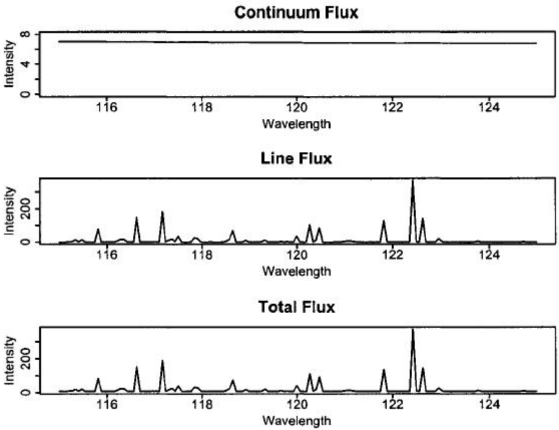


Figure: Illustration of the convolution of counts with continuum

Illustration of step-by-step data degradation

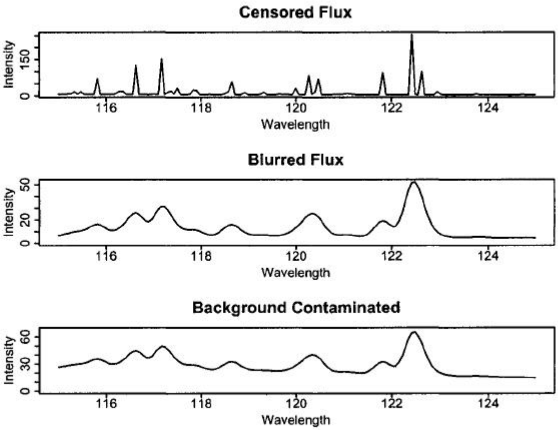


Figure: Exaggerated illustration of stochastic censoring, blurring and background contamination

Method: Data augmentation

- Final model for channel intensities is

$$\xi = Md\left(\sum_k \gamma_k \left\{G^{C,k} + binned\left[G^{L,k}\right]\right\}\right)\mu + \xi^{bkg}$$

- The following latent variables are defined:
  $Y_{1:I}$ - background-free channel counts
  $Z^{-}_{1:J}$ - censored bin counts
  $Z_{1:J}$ - bin counts
  $Z^{L}_{1:J}$ - counts generated by binned lines (vs. $Z^{C}_{1:J}$ corresponding to continuum)
  $Z^{L}_{1:L}$ - counts generated by each line separately
  $U^{-,k}_{1:T}$ - counts from continuum in each temperature bin but "censored" (for element k)
  $U^{k}_{1:T}$ - counts corresponding to continuum in temperature bins
  $V^{-,k}_{1:T}, V^{k}_{1:T}$ - counts coming from all lines of element k

Method: Data augmentation

- Joint posterior for the augmented model is

$$
\begin{aligned}
&p(\gamma, \mu, V, V^-, U, U^-, Z, Z^L_{1:J}, Z^L_{1:L}, Z^-, Y | Y^{obs}) \propto \\
&\propto p(\gamma, \mu) p(V|\gamma) p(V^-|V) p(U|\mu) p(U^-|U) p(Z|U^-) \\
&p(Z^L_{1:J}|Z) p(Z^L_{1:L}|Z^L_{1:J}) p(Z^-|Z) p(Y|Z^-) p(Y^{obs}|Y) \qquad (1)
\end{aligned}
$$

- Flat conjugate prior distributions were assigned to $\gamma_k \sim Gamma(1, 0)$
- DEM μ is being smoothed using multiscale analysis, therefore the prior is
  $\mu_{0,0} \sim Gamma(1, 0)$ - parameter for the overall sum
  $\rho_{r,k} \sim Beta(\alpha_r, \alpha_r)$ - splitting factors

Gibbs steps

Updating latent variables:

1. $Y_i | Y_i^{obs}, \xi_i^{source}, \xi_i^{bkg} \sim Binomial\left(Y_i^{obs}, \frac{\xi_i^{source}}{\xi_i^{bkg} + \xi_i^{source}}\right), i = 1, \ldots, I$

2. $Z_j^- | Y, M, d, \lambda \sim \sum_i Multinomial\left(Y_i, \frac{(M_{1i} d_1 \lambda_1, \ldots, M_{Ji} d_J \lambda_J)}{\sum_j (M_{ji} d_j \lambda_j)}\right)$
   since $\xi^{source} = Md\lambda$, we get $Z_j^- \sim Pois(d_j \lambda_j)$

3. $Z_j | Z_j^-, d_j, \lambda_j \sim Z_j^- + Pois((1 - d_j)\lambda_j), j = 1, \ldots, J$
   so that $Z_j \sim Pois(\lambda_j)$

Gibbs steps

Updating latent variables:

4. $Z_j^L | Z_j, \lambda^L, \lambda^C \sim \mathit{Binomial}\left(Z_j, \frac{\lambda_j^L}{\lambda_j^L + \lambda_j^C}\right), j = 1, \ldots, J$

5. $(Z_{l1}^L, \ldots, Z_{lj}^L) | Z_j^L, \lambda_{1:L}^L, \lambda_{1:J}^L \sim \mathit{Multinomial}\left(Z_j^L, \frac{(\lambda_{l1}^L, \ldots, \lambda_{lj}^L)}{\sum_h (\lambda_{lh}^L)}\right), j = 1, \ldots, J$, line l is in bin j

(in EM these steps will be replaced by corresponding expected values)

Gibbs steps

Updating latent variables:

4. Define c_t^k to be column sums for $G^{k,C}$ with $c_*^k = \max_t\{c_t^k\}$ and $\tilde{G}^{k,C} = G^{k,C}/c_*^k$.
   Remember that $\lambda^{k,C} = \gamma_k G^{k,C}\mu$. Then

   $$U_t^{-,k}|Z_{1:J}^L, G_c^* \sim \sum_j Multinomial\left(Z_j^L, \frac{(\tilde{G}^{k,C}(j,1)\mu_1, \ldots, \tilde{G}^{k,C}(j,T)\mu_T)}{\sum_t \tilde{G}^{k,C}(j,t)\mu_t}\right)$$

   such that each $U_t^{-,k}|\mu_t, c^k \sim Pois(c_t^k\mu_t) \sim Pois(\tilde{c}_t^k c_*^k \mu_t)$

5. In order to bring all counts to the same scale we use the same idea as in "decensoring":
   $U_t^k|U_t^{-,k}, c^k, \mu_t \sim U_t^{-,k} + Pois((1-\tilde{c}_t^k)c_*^k\mu_t), t = 1, \ldots, T$ after that $U_t^k|\mu_t, c^k \sim Pois(c_*^k\mu_t)$

Gibbs steps

Updating latent variables:

4. The same method is applied to get $V_t^{-,k}|\mu_t \sim Pois(l_t^k \mu_t)$ and $V_t^k|\mu_t \sim Pois(l_*^k \mu_t)$, where l_t^k is column sums for $G^{k,L}$, etc.
5. Then we calculate $U_t = \sum_k \left(U_t^k + V_t^k\right) \sim Pois(g^* \mu_t)$ where $g^* = \sum_k \gamma_k \left(c_*^k + l_*^k\right)$
6. Counts $U_{1:T}$ go through multiscale smoothing.

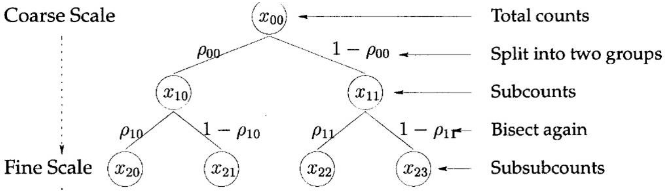


Figure: Binary Tree for Multiscale analysis (Nowak, Kolaczyk (2000))

Gibbs steps

Updating parameters:

- γ_k is updated using line counts $Z^L_{1:L}$ (after step 5):

$$\gamma_k | Z^L_{l_1}, \ldots, Z^L_{ln_k}, \mu \sim Gamma\left(\sum_h Z^L_{l_h} + 1\right) / \sum_h \lambda^L_{l_h}$$

  where $\lambda^L_l = \sum_t G^{k,L}_{lt} \mu_t$ and $Z^L_{l_1} \ldots Z^L_{l_{n_k}}$ are lines corresponding to element k

- DEM is updated after multiscale smoothing procedure $\mu_t \sim Gamma(U^{smoothed}_t + 1)/g^*, t = 1, \ldots, T.$

(in EM these steps will be replaced by corresponding posterior modes to get MAP values)

Normalized effective area (ARF) for simulated and Chandra LETGS data

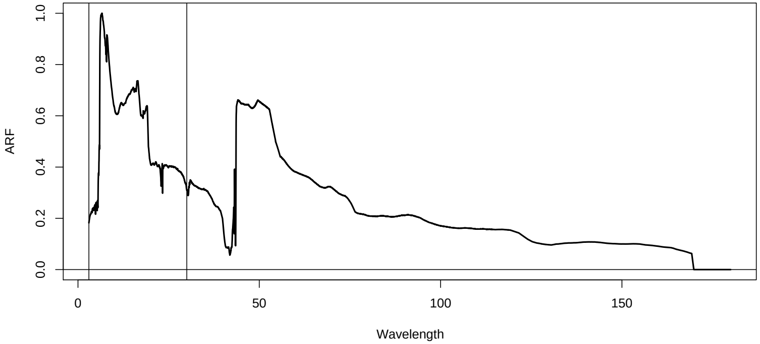


Figure: Effective area ("censoring probability") for Low Energy Transmission Spectrometer (LETGS) on Chandra

Normalized "censoring" probabilities $\tilde{l}^k_{1:T}$ (within $3\ldots30A$)

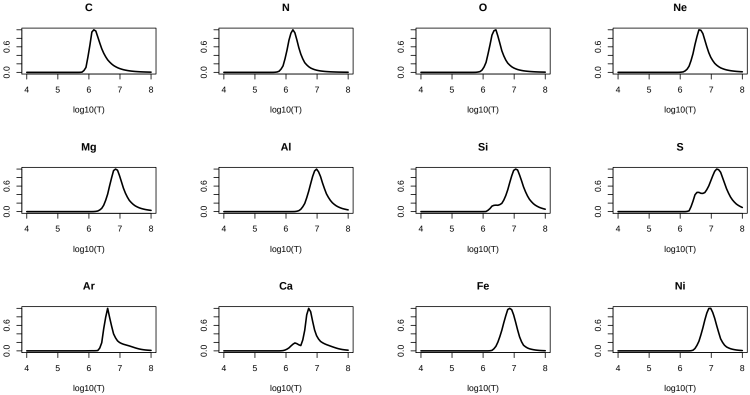


Figure: Counts at short wavelengths don't give us enough information about DEM below $10^6 K$

Three EM calculations for simulated data

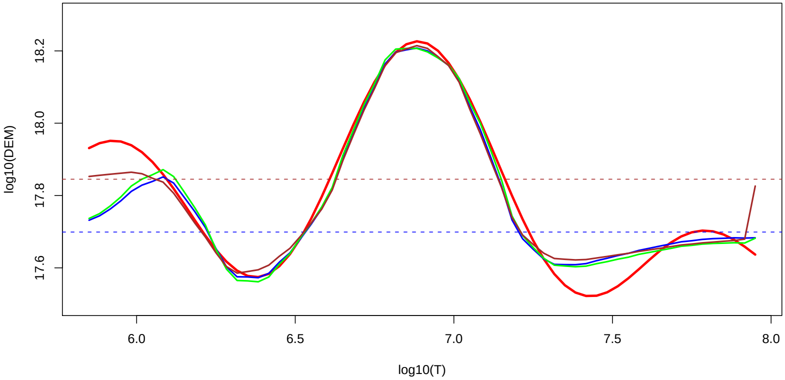


Figure: MAP values for μ with moderate smoothing $\alpha_r = 4$ and different starting values for γ and μ

Results for simulated data: Residuals

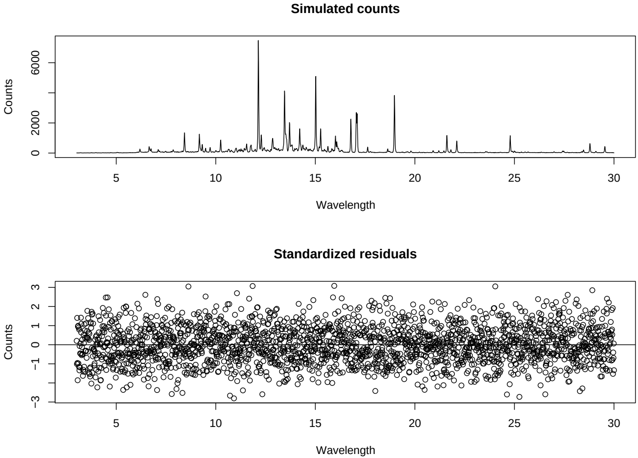


Figure: Standardized residuals $(Y_i^{obs} - \hat{\xi}_i)/\sqrt{\hat{\xi}_i}$, about 4.2% fall outside 1.96.

Results for simulated data: Abundance

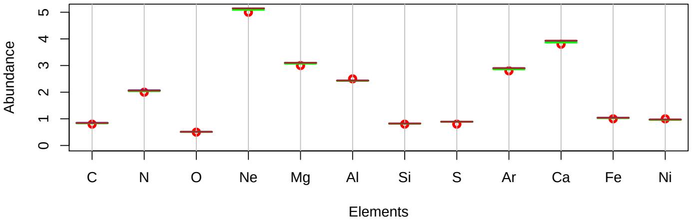


Figure: Red dots represent true values, lines (green, brown and blue) show results from three EM runs

Results for simulated data: Spectrum

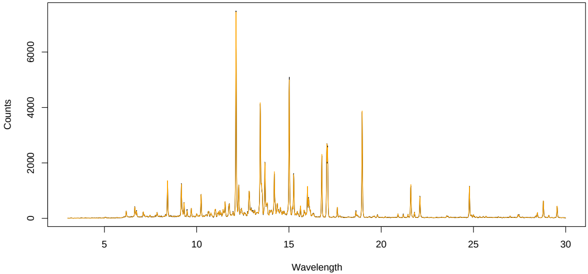


Figure: Estimated expected intensity $\xi_{1:J}$ (orange) superimposed on actual counts (black)

Data available for analysis

- Spectrum of Capella collected using Chandra's HRC-S (High Resolution Camera) with the LETGS diffraction grating (Low Energy Transmission Grating Spectrometer), wavelength range $3-160A$ and channel width $0.0125A$.
- High resolution spectrum of Capella collected using Chandra's ACIS-S (Advanced CCD Imaging spectrometer) with MEG diffraction grating (Medium Energy Grating), effective wavelength range $2-30A$ and bin width $0.005A$.

(First dataset was used by Hosung Kang (PhD thesis (2005)) and we seek to replicate results as well as compare them to the ones from the new dataset)

Observed data: LETGS

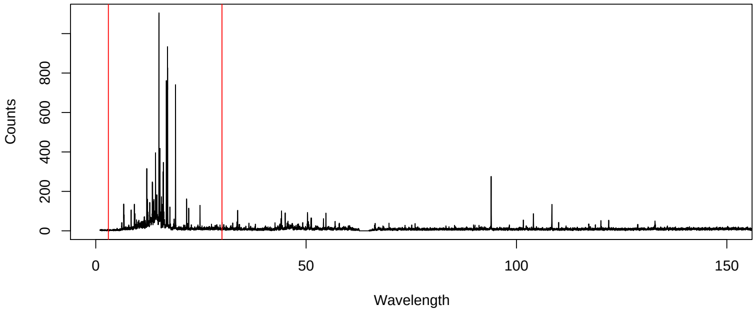


Figure: Observed counts obtained from Low Energy Transmission Spectrometer (LETGS) on Chandra

Normalized effective area (ARF) for Chandra LETGS data

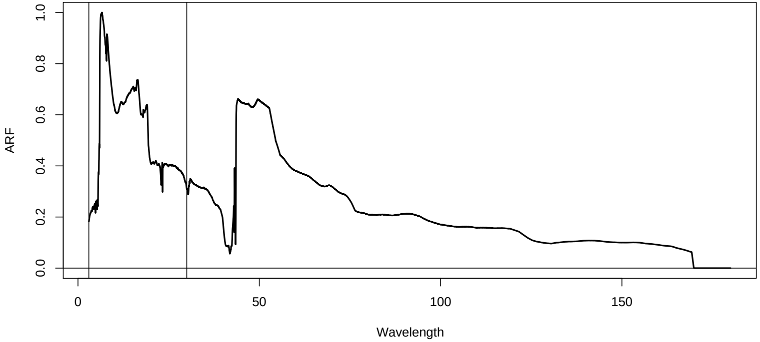


Figure: Effective area ("censoring probability") for Low Energy Transmission Spectrometer (LETGS) on Chandra

Observed data: MEG

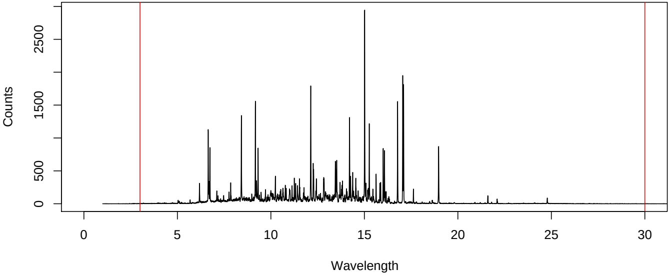


Figure: Observed counts obtained from Medium Energy Grating (MEG) on Chandra

Normalized effective area (ARF) for Chandra MEG data

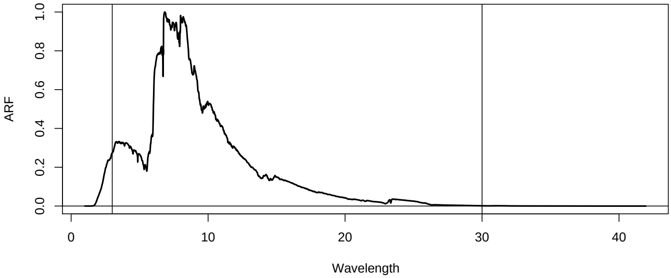
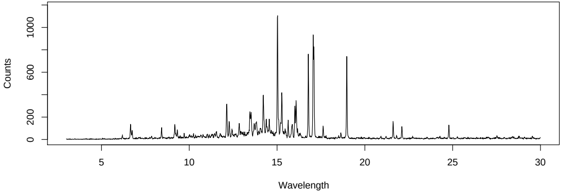
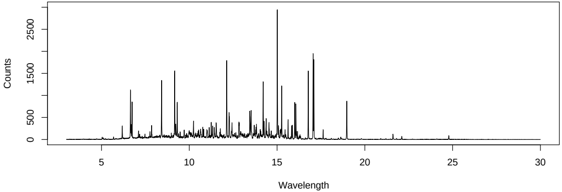


Figure: Effective area ("censoring probability")for Medium Energy Grating (MEG) on Chandra

Relevant ranges for LETGS and MEG together





Results for LETGS and MEG: DEM

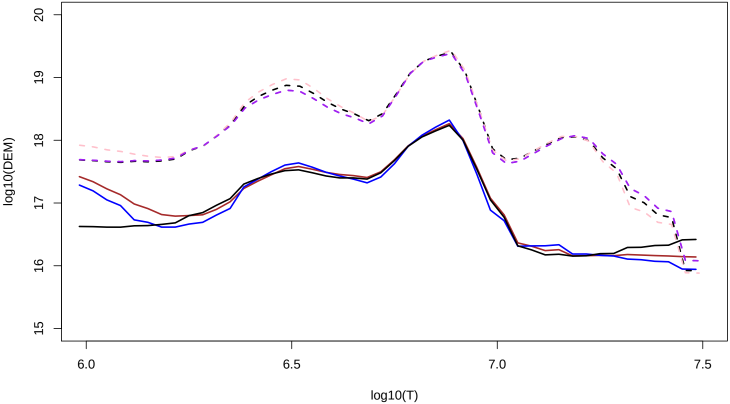


Figure: MAP values for μ. Three bottom lines correspond to three EM runs on LETGS data and top lines correspond to EM ran on MEG data

Results for LETGS and MEG: Abundance

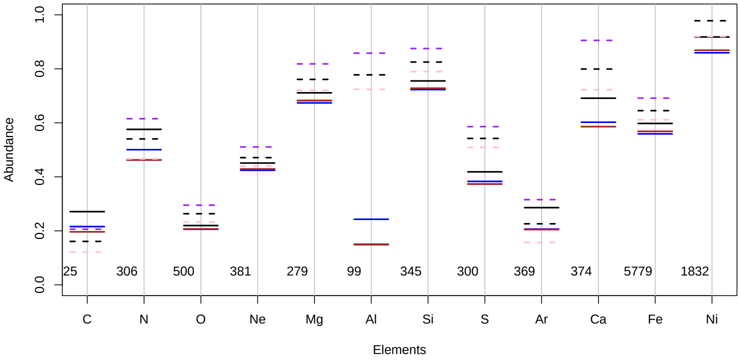


Figure: Results for abundance estimation for two data sources

Identifying missing lines using model output

- Atomic physicists continue discovering/calculating/identifying new lines for atoms of different chemical elements, but calculations are prone to errors. Many minor lines are not present in compiled databases but will show up in the observed spectrum.

- There were several attempts in the past to compare observed spectrum to the expected one (based on current theoretical models). For example, J.-U. Ness at al. (2003) studied the region around NeIX lines at 13.5A that is significantly blended by iron lines.

- The **idea** is to use DEM model output (residuals vs. estimated intensity) from a wider spectrum range to infer about possibly missing or misplaced sets of lines.

  We can start with the simulated model and see how it reacts to omitting strong or weak lines (of course, our inference is limited by the binned structure of the data).

Simulated counts for 13..14A region

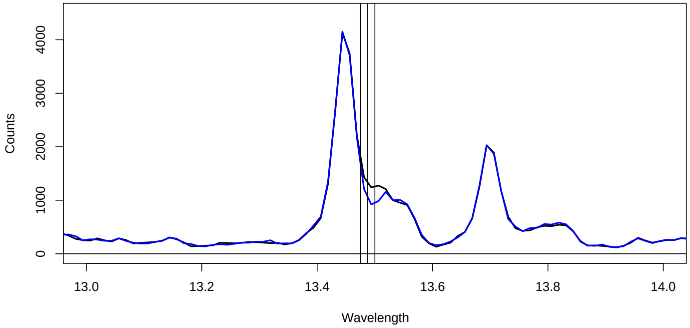


Figure: Black spectrum corresponds to simulated counts using all lines and blue spectrum corresponds to counts generated from the model with 14 missing Fe lines around 13.48A (all lines within two chosen bins)

Simulated counts for 13..14A region

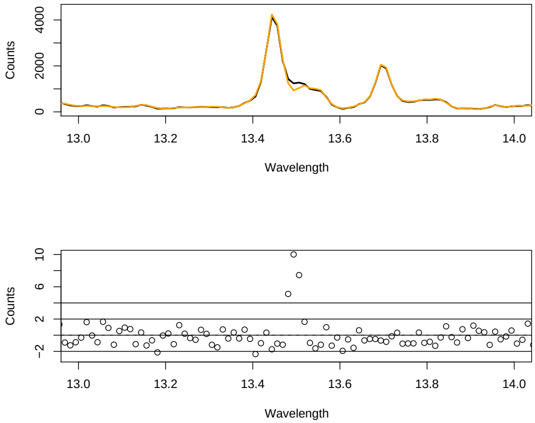


Figure: (a)Black spectrum is the same and orange spectrum corresponds to intensities $\xi_{1:J}$ estimated using the model with 14 missing Fe lines around 13.48A. (b)Bottom graphs shows standardized residuals, out of 2160 channels, only these 3 were beyond 4σ (and only 0.14 are expected to be beyond 4σ).

Results for MEG data

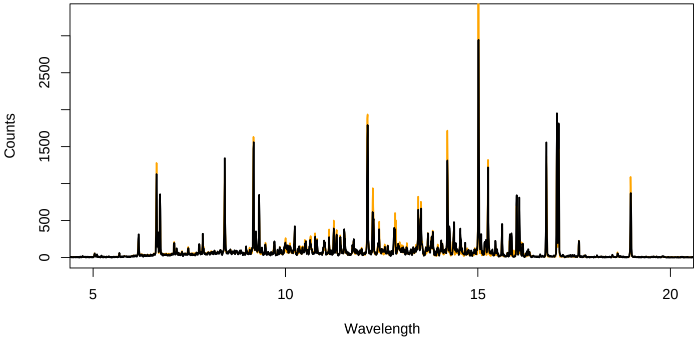


Figure: Estimated expected intensity $\xi_{1:J}$ (orange) superimposed on actual counts (black)

Results for MEG data

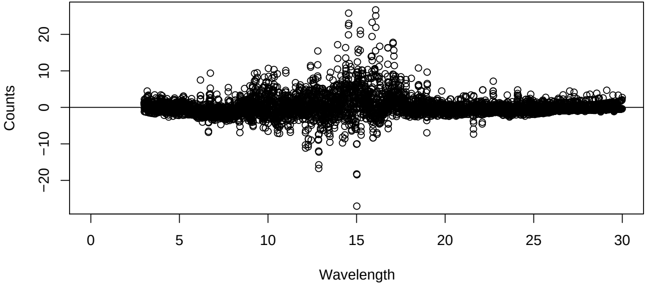
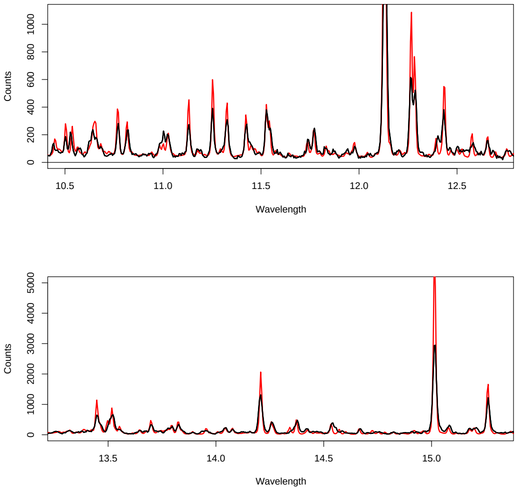
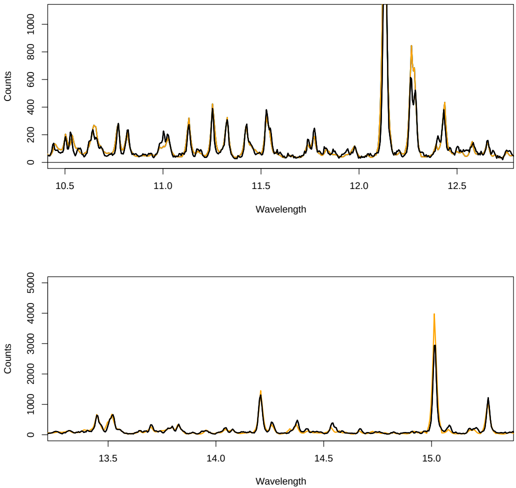


Figure: Standardized residuals. Given the number of bins (5400), we expect 0.34 of them to be greater then 4σ, but we get 486 (9%)

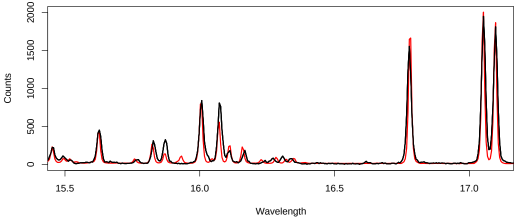
Poorly fitted regions with small bluring SD = 0.005

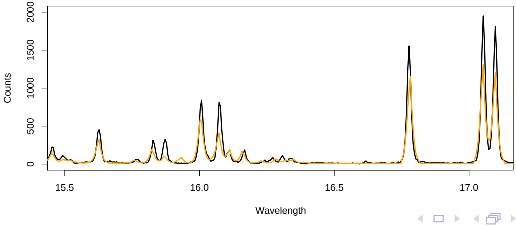


Same regions using bigger bluring SD = 0.01



Another region that is affected by SD increase in an opposite way





Possible model improvements before providing suggestions regarding missing lines

- The issue with overdispersion may occure due to
  - varying RMF (by flux size)
  - line shift caused by the telescope
  - errors of omission and commission in emissivity databases and by the source (Doppler effect)

- After obtaining a reasonable fit we can look at $N\sigma$ outliers and get a range for possibly missing lines.

Challenges in the Gibbs sampler

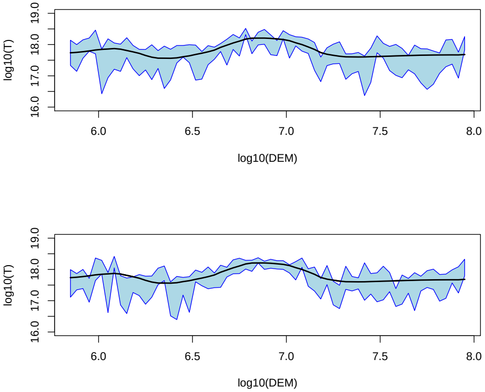


Figure: 95% posterior intervals for DEM from two simulations with 10000 iterations (and 5000 droped)

Two chains combined

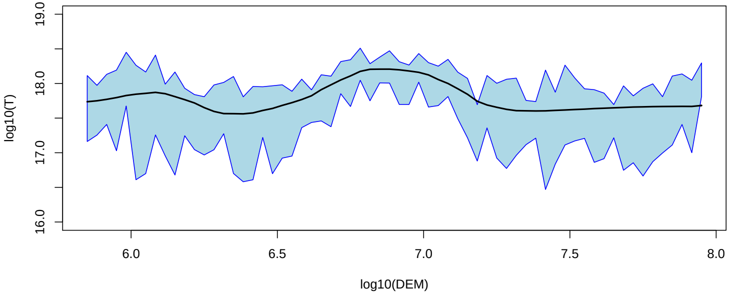


Figure: Combined chains cover almost all MAP values

Results for abundance for simulated data

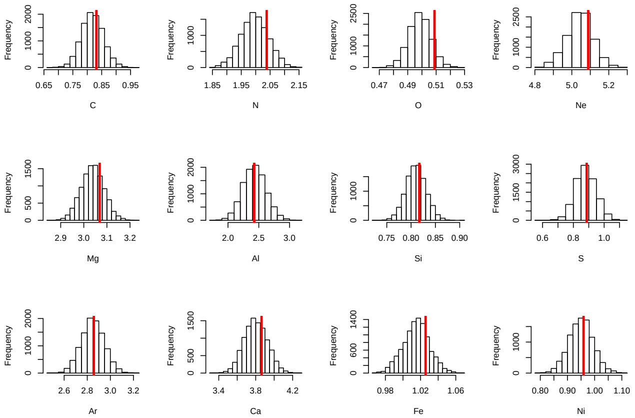


Figure: Histograms corresponding to the first chain (second chain gave similar results)

Convergence asesement for abundance: Chains

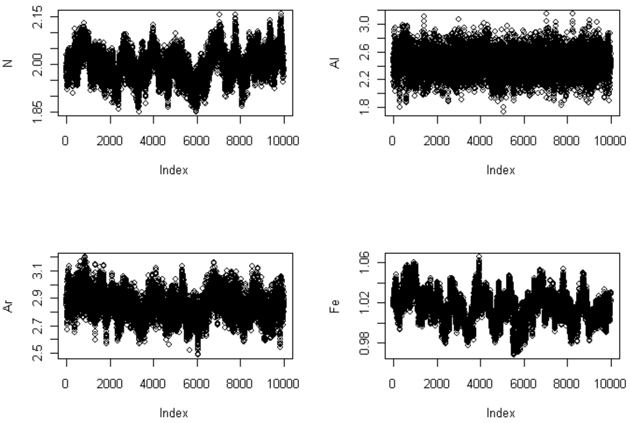


Figure: Chains for four chosen elements

Convergence result for abundance: ACF

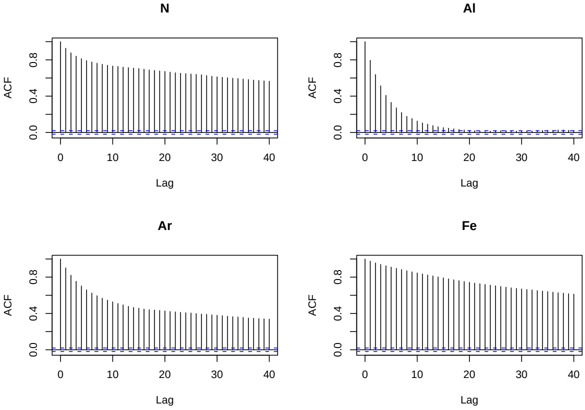


Figure: ACF for four chosen elements

Convergence issues with DEM: Chains

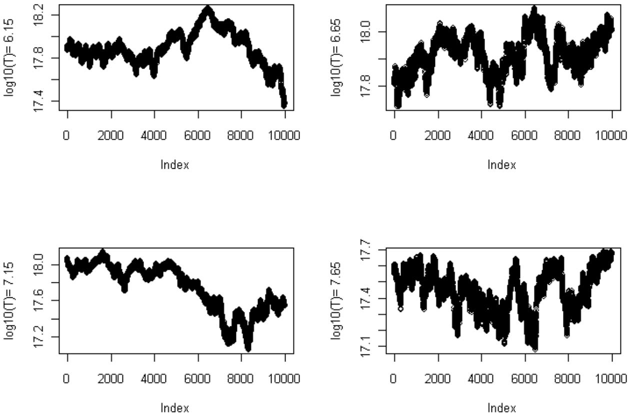


Figure: Chains for four chosen temperature points

Convergence issues with DEM: ACF

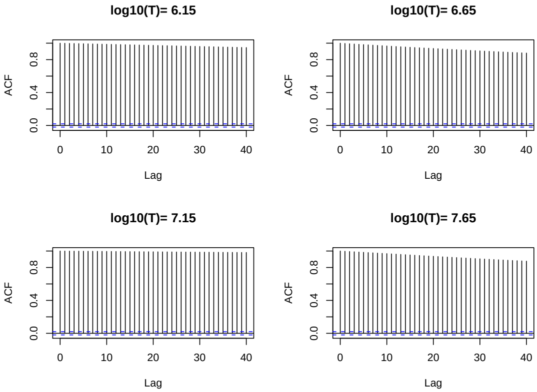


Figure: ACF for four chosen temperature points

Possible problematic steps in Gibbs sampler

"Desencoring" steps

$$U_t^k | U_t^{-,k}, c^k, \mu_t \sim U_t^{-,k} + Pois((1-\tilde{c}_t^k) c_*^k \mu_t), t = 1, \ldots, T$$

$$V_t^k | V_t^{-,k}, l^k, \mu_t \sim U_t^{-,l} + Pois((1-\tilde{l}_t^k) l_*^k \mu_t), t = 1, \ldots, T$$

such that $U_t^k | \mu_t, c^k \sim Pois(c_*^k \mu_t) \; \forall t$ and $V_t^k | \mu_t, l^k \sim Pois(l_*^k \mu_t) \; \forall t$

Low $\tilde{c}_t^k$ and $\tilde{l}_t^k$ values cause high autocorrelation between iterations.

Normalized "censoring" probabilities $\tilde{l}^k_{1:T}$ (within $3 \ldots 30A$)

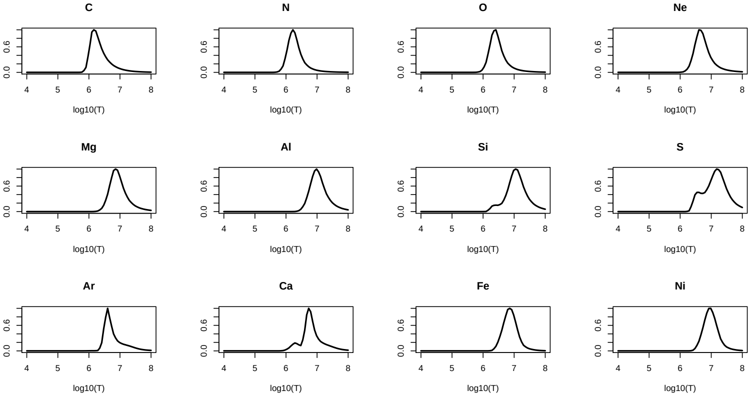


Figure: Counts at short wavelengths don't give us enough information about DEM below $10^6 K$