Multistage Anslysis on Solar Spectral Analyses with Uncertainties in Atomic Physical Models

Xixi Yu

Imperial College London

xixi.yu16@imperial.ac.uk

23 October 2018

Acknowledgments

Joint work with the International Space Science Institute (ISSI) team

"Improving the Analysis of Solar and Stellar Observations"

The Solar Corona

- The solar corona is a complex and dynamic system
- Measuring physical properties in any solar region is important for understanding the processes that lead to these events

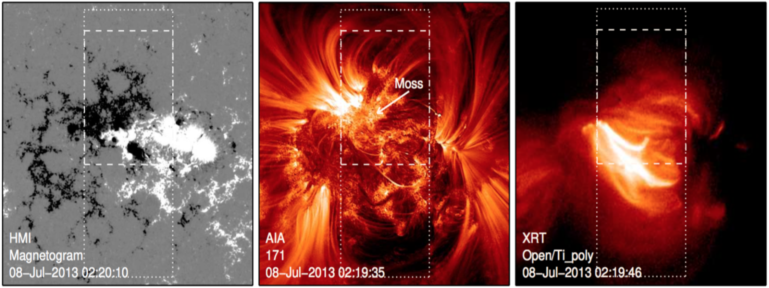


Figure: The photospheric magnetic field measured with HMI, million degree emission observed with the AIA Fe IX 171,Å channel, and high temperature loops observed with XRT

Aim

- We want to infer physical quantities of the solar atmosphere (density, temperature, path length, etc.), but we only observe intensity
- Inferences also rely on models for the underlying atomic physics
- How to address **uncertainty** in the atomic physics models?

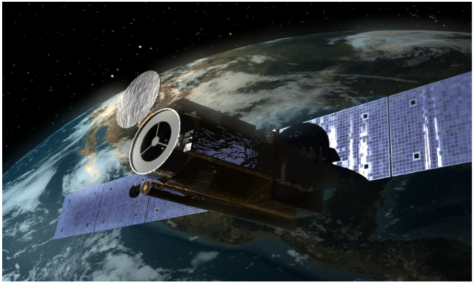


Figure: Hinode spacecraft.

An Astrophysical Perspective

- The importance of accounting for statistical errors is well established in astronomical analysis:

  a measurement is of little value
  without an estimate of its credible range

- However:
  - Uncertainty is often ignored entirely
  - If the second analysis is depending on the first analysis, use the best-fit value, e.g., by minimizing χ^2 results
- Lead to erroneous interpretation of the data!

A Statistical Perspective

- Problem: A two-stage analysis where
  primary: $X \mid \epsilon$
  secondary: $Y \mid \theta, \epsilon$
- Aim: To design a **robust principled** method
- Method: To estimate θ via the analysis of Y, but still depends on ϵ that can be estimated in the primary analysis
- Idea: The output from the primary analysis is required for the secondary analysis

Statistical Methods

- **Standard method**

$$p(\theta \mid Y, \hat{\epsilon}_0) \tag{1}$$

where $\hat{\epsilon}_0$ is the best-fit of $p(\epsilon \mid X)$ from the primary analysis

- **Multiple imputation**: multiple imputation combining rules with Gaussian approximation
- **Pragmatic Bayesian method**

$$p_{\mathrm{pB}}(\theta, \epsilon \mid Y) = p(\theta \mid Y, \epsilon) \cdot p(\epsilon) \tag{2}$$

- **Fully Bayesian method**

$$p_{\mathrm{fB}}(\theta, \epsilon \mid Y) = p(\theta \mid Y, \epsilon) \cdot p(\epsilon \mid Y) \tag{3}$$

Multiple imputation

- Sample M sets of independent parameter estimates from $p(\epsilon \mid X)$ along with their estimated variance-covariance matrices:

$$\epsilon^{(m)} \text{ and } \mathrm{Var}(\epsilon^{(m)}) \quad \text{for } m = 1, \ldots, M \tag{4}$$

- Make Gaussian assumption and use **multiple imputation combining rules** (a set of simple moment calculations)
- Parameter estimate:

$$\epsilon = \frac{1}{M} \sum_{m=1}^{M} \epsilon^{(m)} \tag{5}$$

Total uncertainty:

$$T = W + (1 + \frac{1}{M})B \tag{6}$$

$W = \frac{1}{M} \sum_{m=1}^{M} \mathrm{Var}(\epsilon^{(m)})$ and $B = \frac{1}{M-1} \sum_{m=1}^{M} (\epsilon^{(m)} - \epsilon)(\epsilon^{(m)} - \epsilon)^\top$ are the statistical and systematic uncertainties respectively

- However: M is typically **small**

  What if we have a large Monte Carlo (MC) sample?

A Particular Problem

- We have a **large** ensemble of MC sample that represents the variability from the primary analysis:

$$\mathcal{M} = \{\epsilon^{(m)}, m = 1, \ldots, M\}$$

- How to use this ensemble?
- Two methods:
  - **Discrete uniform**
  - **Gaussian approximation via Principal Component Analysis (PCA)**
- A Case Study in **FeXIII**

Physical Parameters

- n_k [1]: number of free electrons per unit volume in plasma
- T_k: electron temperature
- d_k: path length through the solar atmosphere
- $\theta_k = (\log \mathrm{n}_k, \log \mathrm{d}_k)$
- m: index of the emissivity curve
- Expected intensity of line with wavelength λ:

$$\epsilon_\lambda^{(m)}(\mathrm{n}_k, \mathrm{T}_k)\mathrm{n}_k^2\mathrm{d}_k$$

- $\epsilon_\lambda^{(m)}(\mathrm{n}_k, \mathrm{T}_k)$ is the plasma emissivity for the line with wavelength λ in pixel k

[1] Subscript k is the pixel index

Data: Observed Intensity

- Data from the Extreme-Ultraviolet Imaging Spectrometer (EIS) on *Hinode* spacecraft.

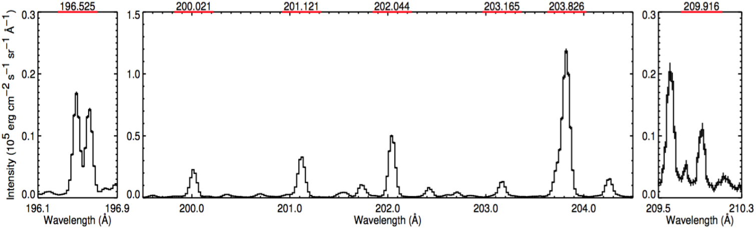


Figure: Example EIS spectrum of seven Fe XIII lines

- Spectral lines with wavelengths $\Lambda = \{\lambda_1, \ldots, \lambda_J\}$
- Observed intensities for K pixels and J wavelengths:

$$\hat{D} = \{D_k = (I_{k\lambda_1}, \ldots, I_{k\lambda_J}), k = 1, \ldots, K\}$$

- Standard deviation $\sigma_{k\lambda_j}$ are also measured

Uncertainty: Emissivity

- Emissivity: how strongly energy is radiated at a given wavelength
- Simulated from a model accounting for uncertainty in the atomic data
- Suppose a collection of M emissivity curves are known

$$\mathcal{M} = \{\epsilon_\lambda^{(m)}(\mathrm{n}_k, \mathrm{T}_k), \lambda \in \Lambda, m = 1, \ldots, M\}$$

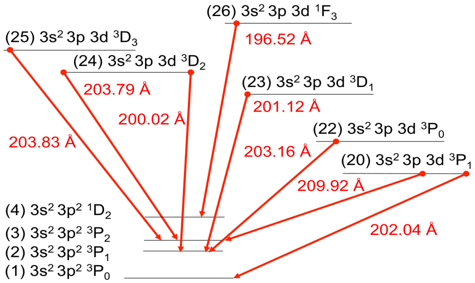


Figure: A simplified level diagram for the transitions relevant to the 7 lines considered here.

Statistical Model

Independent Prior distributions

$$p(m, \theta_k) = p(m)\ p(\log \mathrm{n}_k)\ p(\log \mathrm{d}_k) \tag{7}$$

$$m \sim p(m) \tag{8}$$

$$\log_{10} \mathrm{n}_k \sim \text{Uniform}(\text{min} = 7, \text{max} = 12) \tag{9}$$

$$\log_{10} \mathrm{d}_k \sim \text{Cauchy}(\text{center} = 9, \text{scale} = 5) \tag{10}$$

Likeihood $L(m, \theta_k \mid D_k)$

$$I_{k\lambda} \mid m, \mathrm{n}_k, \mathrm{d}_k \overset{\text{indep}}{\sim} \text{Normal}\left(\epsilon_\lambda^{(m)}(\mathrm{n}_k, \mathrm{T}_k)\mathrm{n}_k^2\mathrm{d}_k,\ \sigma_{k\lambda}^2\right), \quad \text{for } \lambda \in \Lambda \tag{11}$$

Joint posterior distribution

$$p(m, \theta_k \mid D_k) \propto L(m, \theta_k \mid D_k)\ p(m, \theta_k), \tag{12}$$

Standard VS Pragmatic VS Fully Bayesian Models

Standard method

$$p(\theta_k \mid D_k, m = 1) \quad \text{with } m = 1 \text{ the default emissivity} \tag{13}$$

Pragmatic Bayesian posterior distribution

$$p(m, \theta_k \mid D_k) = p(\theta_k \mid D_k, m)\; p(m). \tag{14}$$

Fully Bayesian posterior distribution

$$p(m, \theta_k \mid D_k) = p(\theta_k \mid D_k, m)\; p(m \mid D_k). \tag{15}$$

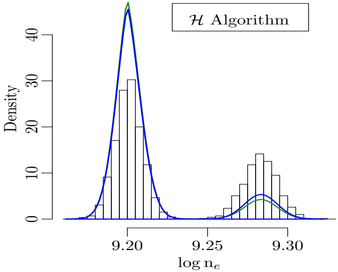
Aim: To compare the three methods, "standard, pragmatic and fully Bayesian", applied to a single-pixel and multiple pixels

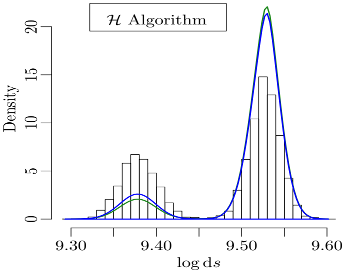
Prior I: Discrete Uniform (Done)

- Fully Bayesian Model I:
  - Use **Bayesian Methods** to incorporate information in the data for narrowing the uncertainty in the atomic physics calculation
  - Joint posterior distribution:
    $$p(m, \theta_k \mid D_k) \propto L(m, \theta_k \mid D_k)\ p(m)\ p(\theta_k)$$
  - $p(m) = \frac{1}{M}$, i.e., there are only $M = 1000$ **equally likely** emissivity curves as a priori
- Solution:
  - Obtain a sample of $\mathcal{M}$ that accounts for the uncertainty in the atomic data, m
  - m is treated as an unknown parameter
  - Obtain sample from $p(m, \theta \mid D)$ via **two-step Monte Carlo (MC) samplers** or **Hamiltonian Monte Carlo** (HMC)
- Conclusions:
  - We are able to incorporate uncertainties in atomic physics calculations into analyses of solar spectroscopic data

Prior I: Multimodal Posterior Distributions

- **Bimodal** posterior distributions occur
  - **Two modes** correspond to **two emissivity curves**
  - **Inaccurate relative size** of two modes in HMC
- Reason: **Not enough** emissivity curves
- Challenge: **Sparse** selection of emissivity curves



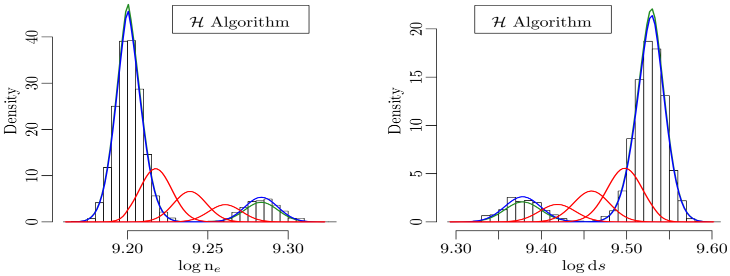


Prior I: Multimodal Posterior Distributions

- A computational issue: **Inaccurate relative size** of two modes
- A computational solution:
  Adding a few **synthetic** replicate emissivity curves with augmented set $\mathcal{M}^{\mathrm{aug}} \supset \mathcal{M}$:

$$\mathcal{M}^{\mathrm{aug}}/\mathcal{M} = \{w_1 * \mathsf{Emis}_{471} + w_2 * \mathsf{Emis}_{368}\},$$

  where $(w_1, w_2) = (0.75, 0.25), (0.50, 0.50), \&(0.25, 0.75)$



Aim of Prior II

Come up with a way to efficiently represent

the high dimensional joint distribution

of the uncertainty of the emissivity curves.

Comparison of Prior I and Prior II

Prior I (Done)

- Joint posterior distribution:
$$p(m, \theta_k \mid D_k) \propto L(m, \theta_k \mid D_k)\ p(m)\ p(\theta_k)$$
- $p(m) = \frac{1}{M}$
- A computational trick: adding a few **synthetic** replicate emissivity curves

Prior II

- Joint posterior distribution:
$$p(\epsilon(r_k), \theta_k \mid D_k) \propto L(\epsilon(r_k), \theta_k \mid D_k)\ p(r_k, \theta_k)$$
- r_k is the PCA transformation of emissivity curve, ϵ
- $p(r)$ is a high dimensional **distribution**
- An algorithm: summarizing the distribution with multivariate standard Normal distribution via **PCA**

Prior II: Gaussian approximation via PCA

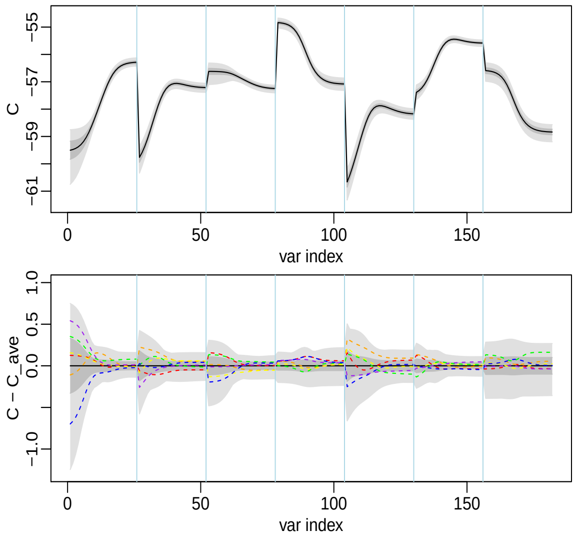
- In $\sqrt{}$ space
- $J = 16$ PCs capture 99% of total variation
- **PCA generated** emissivity curve replicate based on **the first J PCs**

$$\epsilon^{\text{rep}} = \bar{\epsilon} + \sum_{j=i}^{J} r_j \beta_j v_j,$$

where

- $\bar{\epsilon}$: average of all 1000 emissivity curves
- r_j: random variate generated from the standard Normal distribution
- β_j^2, v_j: eigenvalue and eigenvector of component j in the PCA representation

Prior II: Plot of Original Emissivity Curves



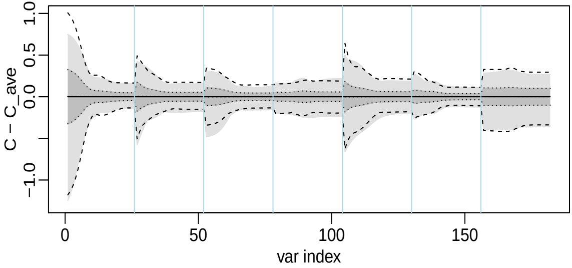
Top panel:

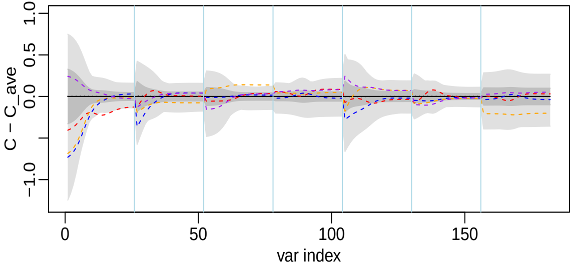
- Each part corresponds one of the seven lines
- Light gray area: all 1000 emissivities
- Dark gray area: middle 68% of emissivities
- Solid black curve: $\bar{\epsilon}$

Bottom panel:

- Same as above, but using $\epsilon - \bar{\epsilon}$
- Colored dashed curves: six randomly selected curves

Prior II: Plot of PCA Generated Emissivity Curves





Light & dark gray areas are same as above

Top panel:

- Dashed lines: all 1500 PCA generated emissivities
- Dotted lines: the middle 68% of PCA emissivities

Bottom panel:

- Colored dashed curves: selection of PCA curves

Prior II: Pragmatic VS Fully Bayesian Models

Pragmatic Bayesian model

$$p(r, \theta \mid D) = p(\theta \mid D, r)\ p(r \mid D) = p(\theta \mid D, r)\ p(r) \tag{16}$$

Fully Bayesian model

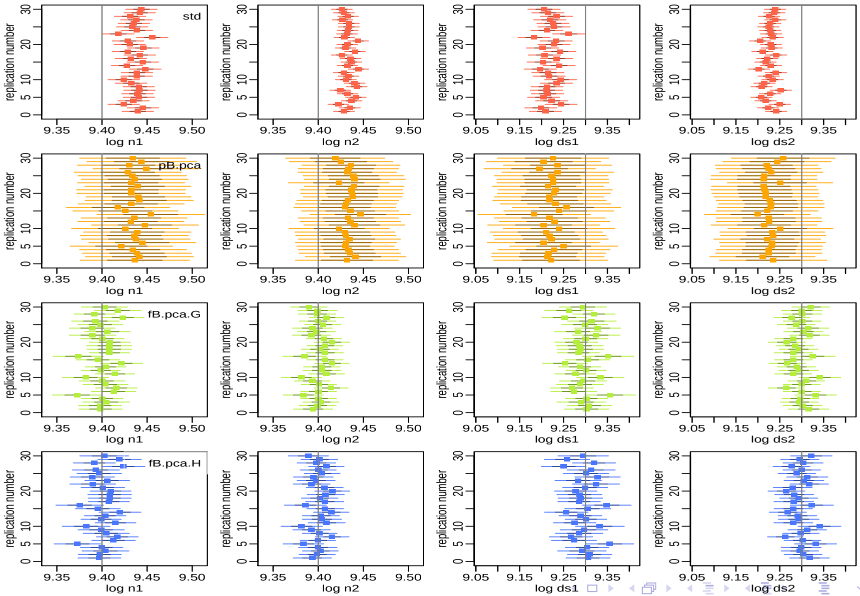
$$p(r, \theta | D) \propto L(r, \theta \mid D)\ p(r)\ p(\theta) \propto \prod_{k=1}^{K} L(r, \theta_k \mid D_k)\ p(\theta_k) \cdot p(r) \tag{17}$$

Two different computing algorithms are used: *two-step Gibbs sampler* and *Hamiltonian Monte Carlo (HMC)*

Prior II: Data Simulation, $J = 3$ and $K = 2$

- Pick the first two plixels: $\#1$ and $\#2$, i.e., here $K = 2$
- Fix $\theta_k = (\log \mathrm{n}_k, \log \mathrm{d}_k) = (9.4, 9.3)$ for $k = 1, 2$ (posterior mean for Pix1 in paper).
  Emissivity curve $= \epsilon^*_{471}$, which is computed from the first $J = 3$ PCs accounting for 42.23% of the total variance. (To make sure we are simulating and fitting under the *same* model.)
- Simulate 200 replicates for each pixel.
  Here, each of pixels is simulated with same values of parameters, but different replicates have different simulated data.

Prior II: Output summary, $J = 3$ and $K = 2$



Prior II: Output summary, $J = 3$ and $K = 2$

		Bias	RMSE	Coverage of 68% interval	Coverage of 95% interval	Average length of 95% interval
std	$\log n_1$	0.0359	0.1273	0	0.02	0.0346
	$\log n_2$	0.0336	0.0868	0	0	0.0226
	$\log d_1$	−0.0781	0.2690	0	0.01	0.0728
	$\log d_2$	−0.0733	0.1849	0	0	0.0480
prag Bayes	$\log n_1$	0.0366	0.0376	0.225	1	0.1187
	$\log n_2$	0.0345	0.0350	0.155	1	0.1114
	$\log d_1$	−0.0780	0.0817	0.205	1	0.2556
	$\log d_2$	−0.0752	0.0763	0.115	1	0.2405
fully Bayes Gibbs	$\log n_1$	0.0023	0.0121	0.68	0.95	0.0483
	$\log n_2$	0.0019	0.0096	0.73	0.97	0.0405
	$\log d_1$	−0.0053	0.0260	0.67	0.94	0.1026
	$\log d_2$	−0.0045	0.0207	0.72	0.975	0.0869
fully Bayes HMC	$\log n_1$	0.0024	0.0120	0.675	0.95	0.0481
	$\log n_2$	0.0019	0.0096	0.72	0.975	0.0404
	$\log d_1$	−0.0054	0.0259	0.65	0.945	0.1023
	$\log d_2$	−0.0046	0.0207	0.72	0.965	0.0867

Prior II: Output summary, $J = 3$ and $K = 2$

- Pragmatic Bayesian:
  - biased
  - the 68% intervals are significantly under coverage, the 95% intervals are significantly over coverage
- Fully Bayesian:
  - the 68% intervals are a little bit over coverage
  - smaller bias and RMSE
  - the 95% coverage is better, the smaller coverage length
  - results from the two different algorithms are almost the same and applying HMC is significantly faster, so keep using it

Next step, try to increase J and K!