Multiple datasets of different sizes

Hierarchical Gaussian Process with Haar wavelet mean process

Shihao Yang

01/23/2017 STAT310 Astrostatistics

Background

- Statistics: internet-based big data & traditional survey data
- Astronomy: SED (spectral energy distribution) problem where OIR photometry must be fit simultaneously with X-ray spectra. Or in calibration studies, when measurements of the same quantity from different sources must be combined

Motivating Example - XRCF Correction Factor

- curve fitting from three different sources, with different quantity and quality
- the true curve has jumps

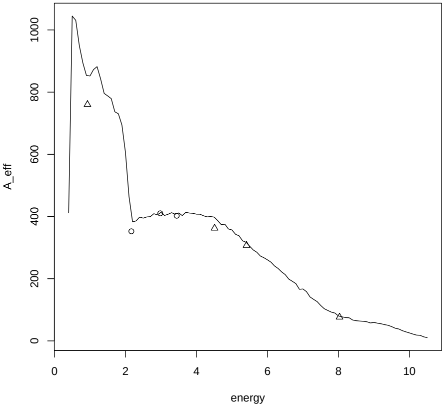
energy	ea1346	err1346
0.40	411.64	0.10
0.50	1044.55	0.02
0.60	1030.93	0.01
...	...	...
10.30	17.98	0.04
10.40	13.11	0.04
10.50	10.40	0.05

X_Ray_energy	A_eff	A_err
0.93	760.00	7.16
4.51	362.43	5.69
5.41	307.15	2.70
8.03	76.33	3.52

X_Ray_energy	A_eff	A_err
2.17	352.45	5.70
2.98	410.27	10.07
3.44	402.52	8.03

XRCF Correction Factor

Standard Errors are not consistent from one dataset to another...



The fear of imbalanced dataset

- if datasets are of same quality, then larger datasets should dominate small datasets
- discount large datasets $\Leftrightarrow$ large datasets has "worse" quality
- two possibilities (paradigms) for "worse" quality:
  - the large dataset is biased (e.g. internet-based data)
  - the large dataset has strong correlation (e.g. multi-level data or clustered data)
- both the two above could be loosely interpreted as "bias", but subtle difference in repeated sampling interpretation
- unknown systematic bias could be thought of as correlation in samples
- for XRCF Correction Factor, it is hard to believe physical instrument has systematic bias, so the correlation perspective is more suitable here

Vague intuitions about the model

- estimates in each dataset are strongly correlated with $\rho \propto L$
- between dataset independence
- hierarchical Gaussian process with random shift from common mean curve
- the standard error is conditional on the random shift, thus unconditionally the error is much larger compare to the true mean curve
- true curve has jumps $\Rightarrow$ wavelet transformation

the minimum non-trivial example

- the jumps in the curve are orthogonal to the problem of sizing issue of multiple datasets
- assume no jumps for now to focus on the primary problem
- once the primary problem is solve, we can add back jumps by working on the wavelet transformed domain

mathematical model

- Gaussian Process seems to be a nature choice for correlated error
- Multiple datasets $\Rightarrow$ hierarchical Bayesian model
- Naturally incorporates SE as conditional standard deviation

Hierarchical Gaussian Process

- Denote true curve as $m : x \mapsto m(x)$
- Each measurement instrument i has its own curve $f_i | m \sim \mathcal{GP}(m, k_i)$, where $k_i : (x, x') \mapsto k_i(x, x')$ is the kernel function
- Observations by each instrument has error conditional on instrument's inherited curve: $y_{ij} | f_i \overset{iid}{\sim} N(f_i(x_{ij}), \sigma_{ij}^2)$
- intuition for hierarchical structure: even if we can have infinite observation from each instrument, we still cannot recover true curve m, but rather we will have three instrument-specific curve f_1, f_2, f_3 that are around m. This is because in addition to observation error, each instrument has another layer of built-in error that is specific to that particular machine.

Hierarchical Gaussian Process - Formal Setup

- Likelihood
  - $f_i | m \sim \mathcal{GP}(m, k_i)$, $(f_1, f_2, f_3)_{\perp\!\!\!\perp} | m$
  - $y_{ij} | f_i \sim N(f_i(x_{ij}), \sigma_{ij}^2)$, $(y_{i1}, y_{i2}, \ldots)_{\perp\!\!\!\perp} | f_i$
  - $\Rightarrow \boldsymbol{y}_i | m \sim N(m(\boldsymbol{x}_i), k_i(\boldsymbol{x}_i, \boldsymbol{x}_i) + \Sigma_i)$
- Prior
  - $m \sim \mathcal{GP}(0, k_m)$
- Posterior
  - for new point $\boldsymbol{x}_*$ and $\boldsymbol{m}_* = m(\boldsymbol{x}_*)$:

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \boldsymbol{y}_3 \\ \boldsymbol{m}_* \end{pmatrix} \sim N\left(\boldsymbol{0}, \begin{pmatrix} k_m(\boldsymbol{x}_1,\boldsymbol{x}_1) + k_1(\boldsymbol{x}_1,\boldsymbol{x}_1) + \Sigma_1 & k_m(\boldsymbol{x}_1,\boldsymbol{x}_2) & k_m(\boldsymbol{x}_1,\boldsymbol{x}_3) & k_m(\boldsymbol{x}_1,\boldsymbol{x}_*) \\ k_m(\boldsymbol{x}_2,\boldsymbol{x}_1) & k_m(\boldsymbol{x}_2,\boldsymbol{x}_2) + k_2(\boldsymbol{x}_2,\boldsymbol{x}_2) + \Sigma_2 & k_m(\boldsymbol{x}_2,\boldsymbol{x}_3) & k_m(\boldsymbol{x}_2,\boldsymbol{x}_*) \\ k_m(\boldsymbol{x}_3,\boldsymbol{x}_1) & k_m(\boldsymbol{x}_3,\boldsymbol{x}_2) & k_m(\boldsymbol{x}_3,\boldsymbol{x}_3) + k_3(\boldsymbol{x}_3,\boldsymbol{x}_3) + \Sigma_3 & k_m(\boldsymbol{x}_3,\boldsymbol{x}_*) \\ k_m(\boldsymbol{x}_*,\boldsymbol{x}_1) & k_m(\boldsymbol{x}_*,\boldsymbol{x}_2) & k_m(\boldsymbol{x}_*,\boldsymbol{x}_3) & k_m(\boldsymbol{x}_*,\boldsymbol{x}_*) \end{pmatrix}\right)$$

Kernels and hyper-parameters

- even if the true curve m has jumps, the instrument-specific errors on top of m should be smooth (?)
- use Gaussian (radial basis function) kernel:

$$k_i(x, x') = \gamma_i \exp(-\frac{1}{2l_i^2}(x - x')^2)$$

- l_i controls the smoothness (variability/wiggling) along the curve
- γ_i controls the severity of random instrument-specific "bias"
- Assumptions:
  - the smoothness (degree of variability/wiggling along the curve) is the same across instrument $\Rightarrow l_1 = l_2 = l_3$
  - the large dataset may have bigger random "bias": $\gamma_1 \geq \gamma_2 = \gamma_3$

m curve and revisit of discontinuity

how about k_m for mean curve?
Now is the time to incorporate jumps:

- Discontinuity can be modeled by Haar wavelet under Gaussian Process umbrella
- m as Haar wavelet linear combination, where coefficients are independent Gaussian random variable
- m defined above is indeed a Gaussian Process with some induced kernel (needs further work)

Simulation for data generating process

- working on it now...