Multiple datasets of different sizes
Hierarchical Gaussian Process with Haar wavelet mean process

Shihao Yang

01/23/2017 STAT310 Astrostatistics
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

<table>
<thead>
<tr>
<th>energy</th>
<th>ea1346</th>
<th>err1346</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.40</td>
<td>411.64</td>
<td>0.10</td>
</tr>
<tr>
<td>0.50</td>
<td>1044.55</td>
<td>0.02</td>
</tr>
<tr>
<td>0.60</td>
<td>1030.93</td>
<td>0.01</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>10.30</td>
<td>17.98</td>
<td>0.04</td>
</tr>
<tr>
<td>10.40</td>
<td>13.11</td>
<td>0.04</td>
</tr>
<tr>
<td>10.50</td>
<td>10.40</td>
<td>0.05</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>X_Ray_energy</th>
<th>A_eff</th>
<th>A_err</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.93</td>
<td>760.00</td>
<td>7.16</td>
</tr>
<tr>
<td>4.51</td>
<td>362.43</td>
<td>5.69</td>
</tr>
<tr>
<td>5.41</td>
<td>307.15</td>
<td>2.70</td>
</tr>
<tr>
<td>8.03</td>
<td>76.33</td>
<td>3.52</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>X_Ray_energy</th>
<th>A_eff</th>
<th>A_err</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.17</td>
<td>352.45</td>
<td>5.70</td>
</tr>
<tr>
<td>2.98</td>
<td>410.27</td>
<td>10.07</td>
</tr>
<tr>
<td>3.44</td>
<td>402.52</td>
<td>8.03</td>
</tr>
</tbody>
</table>

Multiple datasets of different sizes
XRCF Correction Factor

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

Shihao Yang

Multiple datasets of different sizes
The fear of imbalanced dataset

- if datasets are of same quality, then larger datasets should dominate small datasets
- discount large datasets $\iff$ 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
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: $y_{ij}|f_i \overset{iid}{\sim} \mathcal{N}(f_i(x_{ij}), \sigma_{ij}^2)$
- 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**
  - \( f_i | m \sim \mathcal{GP}(m, k_i), (f_1, f_2, f_3) \perp \perp m \)
  - \( y_{ij} | f_i \sim \mathcal{N}(f_i(x_{ij}), \sigma_{ij}^2), (y_{i1}, y_{i2}, \ldots) \perp \perp f_i \)
  - \( \Rightarrow y_i | m \sim \mathcal{N}(m(x_i), k_i(x_i, x_i) + \Sigma_i) \)

- **Prior**
  - \( m \sim \mathcal{GP}(0, k_m) \)

- **Posterior**
  - for new point \( x_* \) and \( m_* = m(x_*) \):

\[
\begin{pmatrix}
  y_1 \\
  y_2 \\
  y_3 \\
  m_*
\end{pmatrix}
\sim
\mathcal{N}
\begin{pmatrix}
  0 \\
  (k_m(x_1, x_1) + k_1(x_1, x_1) + \Sigma_1) \\
  k_m(x_2, x_1) \\
  k_m(x_3, x_1) \\
  k_m(x_*, x_1)
\end{pmatrix}
\begin{pmatrix}
  k_m(x_1, x_2) \\
  k_m(x_2, x_2) + k_2(x_2, x_2) + \Sigma_2 \\
  k_m(x_3, x_2) \\
  k_m(x_3, x_2) + k_3(x_3, x_3) + \Sigma_3 \\
  k_m(x_*, x_2) \\
  k_m(x_*, x_3) \\
  k_m(x_*, x_3)
\end{pmatrix}
\begin{pmatrix}
  k_m(x_1, x_3) \\
  k_m(x_2, x_3) \\
  k_m(x_3, x_3) + k_3(x_3, x_3) + \Sigma_3 \\
  k_m(x_*, x_3) \\
  k_m(x_*, x_3)
\end{pmatrix}
\]
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:

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

- $l_i$ controls the smoothness (variability/wiggling) along the curve
- $\gamma_i$ 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”:
  $$\gamma_1 \geq \gamma_2 = \gamma_3$$
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...