Scalable Gaussian Processes for Characterizing Multidimensional Change Surfaces

April 18, 2016
William Herlands
Committee: Daniel Neill, Alex Smola, Wilbert Van Panhuis
Chair: Dave Choi

Outline

1. Motivation
2. Gaussian process introduction
3. Change surface model
4. Analysis of measles in the United States
Complex Changes

• In human systems changes are often complex
  – Policy interventions take time to trickle through government bureaucracy
  – Environmental hazards affect populations differentially

• Simple changepoint models are not sufficiently expressive

Why do we care?

• Understand past changes
  – Explore spatio-temporal heterogeneity
  – Model the rate of changes in different areas

• Enable more accurate or equitable policies

• Applications
  – Measles incidence in the U.S
  – Concerns about lead-tainted water in NYC
Our objectives

- Model complex changes in real world data
  - Multiple, flexible function regimes
  - Non-discrete changes
  - Non-monotonic changes
  - Heterogeneous changes over space, time, etc.

Gaussian Processes (GP)

- Non-parametric prior over smooth functions
  \[ f(x) \sim \mathcal{GP}(m(x), k(x, x')) \]
  \[ m(x) = \mathbb{E}[f(x)] \]
  \[ k(x, x') = \text{cov}(f(x), f(x')) \]

- Covariance function is a kernel. Defines the covariance of function values
Gaussian Processes (GP)

- Any finite set of $f(x)$ is Normally distributed
  \[ [f(x_1), \ldots, f(x_m)] \sim N(m(x), K) \]

- Observation model
  \[ y(x) = f(x) + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon) \]

- Marginal log likelihood optimization
  \[
  \log p(y | \theta) \propto - \log |K + \sigma_\varepsilon I| - y^T (K + \sigma_\varepsilon I)^{-1} y
  \]

Full Model

- Our model is a convex combination of $f_i$
  \[ y(x) = s_1(x) f_1(x) + \ldots + s_r(x) f_r(x) + \varepsilon_n \]

  - Switching functions
    \[ s_i(x) \in \Delta^r \]
  - Functional regimes
Model part 1: Functional Regimes

- GP prior for each functional regime
  - Use flexible stationary kernels
    \[ f_i \sim \text{GP}(0, K_i), \ i = 1, \ldots, r \]

Model part 2: Change Surfaces

- Changepoint
  \[ s_i = I(t < T_i) \]

- Non-discrete changepoint
  \[ s_i = \text{softmax}(t - T_i) \]

- Change surface
  \[ s_i = \text{softmax}(w_i(t)) \]
  \[ s_i = \sigma(w_i(t)) \]
Model part 2: Change Surfaces

- Random Kitchen Sink features for $w_i(x)$
  - Variable rate of change
  - Non-monotonic
  - Heterogeneous over input

\[ w_i(x) = \sum_{j=1}^{q} a_j \cos(\omega_j^T x + b_j) \]

Full Model

- Gaussian process change surface model

\[ y(x) = \sum_{i=1}^{r} \sigma(w_i(x)) f_i(x) + \epsilon_n \]

\[ f_i(x) \sim GP(0, K_i) \]

- Can depict this as a single Gaussian process with covariance function

\[ k_{\text{all}}(x, x') = \sum_{i=1}^{r} \sigma(w_i(x))k_i(x, x')\sigma(w_i(x')) \]
Scalable Inference

• Log likelihood naively $O(n^3)$
  \[ \log p(y \mid \theta) \propto -\log |K + \sigma I| - y^T (K + \sigma I)^{-1} y \]

• We develop scalable Kronecker inference using the Weyl bound, $O(Dn^{D+1/D})$

Measles in the United States

• Data
  – Monthly incidence rates 1935 – 2003
  – Continental United States and D.C.
  – $x \in \mathbb{R}^3$, 2D space and 1D time
  – Measles vaccine introduced in 1963
Measles in 3 states

CA

ME

MI

Measles in 3 states

CA

ME

MI
Measles in 3 states

- GP change surface
  - 2 functional regimes
  - $W_f(x)$ as RKS with 5 features
- **Not a causal model!**
Measles in 3 states

“Change slope” from $\sigma(w(x)) = 0.25 \rightarrow 0.75$

“Change date” per state $\sigma(w(x)) = 0.5$
Change date for measles in U.S.

For each state, date where $\sigma(w(x)) = 0.5$

1961.5
1967.2

Change slope for measles in U.S.

For each state, slope of $\sigma(w(x)) = 0.75 \rightarrow 0.25$

0.156
0.297
Regression Analysis

• Explore factors that affect the change date
  – Birth and death rates
  – Population numbers per age segment
  – Income information
  – Government hospital and health workers
  – Slope of change surface
  – Average temperature

<table>
<thead>
<tr>
<th>Dependent variable: Change date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average date rate 0-4</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Average date rate 5-9</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Average birth rate</td>
</tr>
</tbody>
</table>

Gini of family income: 32.317*
(12.071)

Slope of change surface: 37.913**
(8.976)

<table>
<thead>
<tr>
<th>Population 0-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.00002</td>
</tr>
<tr>
<td>(0.00003)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Population 5-9</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00002</td>
</tr>
<tr>
<td>(0.00003)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Average temperature (°F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.025</td>
</tr>
<tr>
<td>(0.041)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Constant</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,946.783**</td>
</tr>
<tr>
<td>(7.614)</td>
</tr>
</tbody>
</table>

Observations: 46
R²: 0.618
Adjusted R²: 0.446

Note: *p<0.05; **p<0.01
Regression Analysis

• Gini of family income
  – Economically depressed communities
  – Rural regions

• Slope of change surface
  – Fewer cases nationwide enable more effective immunization later

Conclusions

• Introduced model for “change surfaces” in real world data
• Developed scalable inference for additive, non-stationary Gaussian processes
• Identified heterogeneity in first years of the measles vaccine
• Used the results of the change surface model for policy relevant conclusions
Acknowledgements

• Committee
  – Daniel Neill, Alex Smola, Wilbert van Panhuis
• Chair
  – Dave Choi
• Collaborators*
  – Andrew Wilson
  – Seth Flaxman
  – Hannes Nickisch

*Subset of paper accepted to AISTATS 2016

Questions?

Fin.
Backup slides

Conclusions

• Introduced model for “change surfaces” in real world data
• Developed scalable inference for additive, non-stationary Gaussian processes
• Identified heterogeneity in first years of the measles vaccine
• Used the results of the change surface model for policy relevant conclusions
Spectral Mixture Kernels

\[ \sum_{q=1}^{Q} \omega_q \cos(2\pi (\tilde{x} - \tilde{x'})^T m_q) \prod_{p=1}^{P} \exp(-2\pi^2 (\tilde{x}_p - \tilde{x'}_p)^2 v_q^{(p)}) \]

**Algorithm 2** Initialize spectral mixture kernels

1: for \( k_i : i = 1 : r \) do
2:     for \( d = 1 : D \) do
3:         Compute \( x^{(d)} \in \{ x : \sigma(w_t(x)) > 0.5 \} \)
4:         Sample \( s \sim |FFT(sort(y(x^{(d)})))|^2 \)
5:         Fit Q component 1D GMM to \( s \)
6:         Initialize \( \omega_q = \text{std}(y) \times \phi_q; m_q = \mu_q; v_q = \sigma_q \)
7:     end for
8: end for

Inference

- Compute log marginal likelihood
  \[ \log p(y | \theta) \propto -\log |K + \sigma_\varepsilon I| - y^T(K + \sigma_\varepsilon I)^{-1}y \]

- General Kronecker methods for scalability
  - Assume: \( x \in X = X^{(1)} \times \ldots \times X^{(D)} \)
  - Assume: multiplicative kernel across D
  - Then we can decompose kernel matrix,
    \[ K = K_1 \otimes \cdots \otimes K_D \]
Inference

• For additive kernels

\[ K_i = K_1 \otimes \ldots \otimes K_D \]

\[ K = \sum_{i=1}^{r} K_i = \sum_{i=1}^{r} K_{i,1} \otimes \ldots \otimes K_{i,D} \]

• \( K^{-1} \) can be computed efficiently using LCG*

• But how can we compute the \( \log|K| \) ?

*See Flaxman et al. (2015)

Inference

(Weyl, 1912) which states that for \( n \times n \) Hermitian matrices, \( M = A + B \), with sorted eigenvalues \( \mu_1, \ldots, \mu_n, \alpha_1, \ldots, \alpha_n, \) and \( \beta_1, \ldots, \beta_n \), respectively,

\[ \mu_{i+j-1} \leq \alpha_i + \beta_j \]

\[ \log(|A + B|) = \log(|M|) = \sum_{i=1}^{n} \log(\mu_i) \leq \sum_{i+j-1=1}^{n} \log(\alpha_i + \beta_j) \]
Inference

• Choosing indices $i, j \sum_{i+j-1=1}^{n} \log(\alpha_i + \beta_j)$

<table>
<thead>
<tr>
<th>Method</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimization for best pair</td>
<td>$O(n^2)$</td>
</tr>
<tr>
<td>“Middle” heuristic $i=j$ OR $i=j+1$</td>
<td>$O(n)$</td>
</tr>
<tr>
<td>Greedy search of $s$ pairs below and above previous pair</td>
<td>$O(2sn)$</td>
</tr>
</tbody>
</table>

Inference

• Scaling functions, $\sigma(w(x))$

$$K = S_1 K_1 S_1' + \cdots + S_r K_r S_r'$$  \hspace{1cm} (22)

where $S_i = \text{diag}(\sigma(w_i(x)))$ and $S_i' = \text{diag}(\sigma(w_i(x')))$. Employing the bound on eigenvalues of matrix products (Bhatia, 2013),

$$\text{sort}(\text{eig}(A \ast B)) \leq \text{sort}(\text{eig}(A)) \ast \text{sort}(\text{eig}(B))$$  \hspace{1cm} (23)
Inference – so what?!

- Linear complexity for additive kernels
  - $O(Dn^{D+1/D})$
- Scalable inference for non-separable kernels in space and time
- Scalable inference for non-stationary kernels
Numerical Experiments

• 2500 points of synthetic data
• 2 functional regimes defined by squared exponential kernels
• Change surface defined by $\sigma(w_{poly}(x))$

$$w_{poly}(x) = \sum_{i=0}^{3} \beta_i^T x^i, \beta_i \sim \mathcal{N}(0, 3I_D)$$

Results - Numerical
### Demographic Analysis

**Dependent variable:** Change slope

| Dependent variable | Estimate | Std. Error | t value | Pr(>|t|) |
|--------------------|----------|------------|---------|----------|
| Average date rate 0-4 | 2.974 | (10.959) |        |          |
| Average date rate 5-9 | -142.811 | (98.068) |        |          |
| Average birth rate | 0.001 | (0.003) |        |          |
| Gini of family income | -0.531** | (0.191) |        |          |
| Per capita income | 0.00000 | (0.00000) |        |          |
| Change date of change surface | 0.010** | (0.002) |        |          |
| Gov’t health and hospitals employees per population | -5.007 | (2.644) |        |          |
| Population 0-4 | 0.00000 | (0.00000) |        |          |
| Population 5-9 | -0.00000* | (0.00000) |        |          |
| Average temperature (°F) | -0.002** | (0.001) |        |          |
| Constant | -18.456** | (4.485) |        |          |

**Observations:** 46  
**R^2:** 0.904  
**Adjusted R^2:** 0.900  
**Residual Std. Error:** 0.014 (df = 31)  
**F Statistic:** 20.784** (df = 14; 31)

**Note:** *p<0.05; **p<0.01