Rajarshi Guhaniyogi
Winter 2018

March 6, 2018
We have already shown one technique to select knots. However it was computationally cumbersome. What if we completely avoid the choice of knots. 

\[ w = (w(s_1), ..., w(s_n))', \Phi \text{ is an } n^* \times n \text{ random matrix.} \]

\[ \tilde{w}(s) = E[w(s)|\Phi w] = c(s)'\Phi'(\Phi C_\theta \Phi')^{-1}\Phi w. \]

\[ c(s) = (C(s, s_1, \theta), ..., C(s, s_n, \theta))'. \]

\[ \tilde{\epsilon}(s) \overset{ind.}{\sim} N(0, C(s, s, \theta) - c(s)'\Phi'(\Phi C_\theta \Phi')^{-1}\Phi c(s)). \]

\[ y(s) = x(s)'\beta + \tilde{w}(s) + \tilde{\epsilon}(s) + \epsilon(s), \epsilon(s) \sim N(0, \tau^2) \]

They showed that the covariance matrix is better conditioned under this idea than modified predictive process. They have also proposed some ideas to design the matrix \( \Phi \) rather than randomly selecting entries of \( \Phi \).
Data Motivation: Isomap Face Dataset (http://web.mit.edu/cocosci/isomap/datasets.html)

- 698 images of an artificial face.
- 2-dim projection of each image: $64 \times 64 = 4096$ pixels in size.
- Horizontal pose angle of each image is given.

Scientific Question & Challenges

- Predict horizontal pose angle of an image based on image pixels.
- Challenges:
  - Complex nonlinear relationship between the response (pose angle) and predictors.
  - Predictors are lying on a complex nonlinear manifold.
  - Large number of predictors and large sample size.
Data Motivation: Isomap Face Dataset (http://web.mit.edu/cocosci/isomap/datasets.html)

Scientific Question & Challenges

Predict horizontal pose angle of an image based on image pixels.

- Horizontal pose angle of each image is given.
Data Motivation: Isomap Face Dataset (http://web.mit.edu/cocosci/isomap/datasets.html)

698 images of an artificial face. 2-dim projection of each image: \(64 \times 64 = 4096\) pixels in size. Horizontal pose angle of each image is given.

Scientific Question & Challenges

Predict horizontal pose angle of an image based on image pixels.

Challenges:

- Complex nonlinear relationship between the response (pose angle) and predictors.
- Predictors are lying on a complex nonlinear manifold.
- Large number of predictors and large sample size.

Horizontal pose angle of each image is given.

**Scientific Question & Challenges**

Predict horizontal pose angle of an image based on image pixels.

**Challenges:**

- Complex nonlinear relationship between the response (pose angle) and predictors.
- Predictors are lying on a complex nonlinear manifold.

- Horizontal pose angle of each image is given.
Data Motivation: Isomap Face Dataset (http://web.mit.edu/cocosci/isomap/datasets.html)

698 images of an artificial face. 2-dim projection of each image: $64 \times 64 = 4096$ pixels in size. Horizontal pose angle of each image is given.

Scientific Question & Challenges

Predict horizontal pose angle of an image based on image pixels.

Challenges:

- Complex nonlinear relationship between the response (pose angle) and predictors.
- Predictors are lying on a complex nonlinear manifold.
- Large number of predictors and large sample size.

- Horizontal pose angle of each image is given.
State-of-the-art approaches: unsatisfactory performance

Issues with existing approaches

A. Unsatisfactory predictive uncertainty.
B. No theory justification.
C. Not scalable with large sample size and predictions.
State-of-the-art approaches: unsatisfactory performance

Issues with existing approaches

A Unsatisfactory predictive uncertainty.
B No theory justification.
C Not scalable with large sample size and predictors.

1. Tree based approaches: CART (Breiman, 1984), Random Forest (Breiman, 2001) (A, B, C), BART (Chipman et al., 2008) (B, C), Treed GP (Gramacy et al., 2007) (B, C).

2. Two stage approaches: clustering high dimensional predictors (Belkin et al., 2003) followed by independent model fitting in each cluster (A, B).

3. Model Based Full Bayesian approaches: GP latent variable models (Lawrence, 2005), PCA for mixture models (Chen et al., 2010) (C).
\(\Psi_{m \times p} = x_{p \times 1}\) where \(\Psi_{ij} \sim N(0, 1)\): Choice motivated by the popular compressed sensing literature (Ji et al., 2018).

\(x = z + \delta, \quad z \in \mathcal{M}, \quad \delta \sim N(0, \tau^2 I_p)\).

Compressed GP model

\[
\begin{align*}
y &= \mu(\Psi x) + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \\
\mu(\cdot) | \sigma^2 &\sim GP(0, \sigma^2 K(\cdot, \cdot, \phi)) \\
K(x_i, x_j, \phi) &= \exp(-\phi \| x_i - x_j \|^2)
\end{align*}
\]

Model fitting requires \(n \times n\) matrix inversion at each MCMC.
Strategy when sample size \((n)\) is large

Large sample approximation of CGP

\[ y = \tilde{\mu}(\Psi x) + \epsilon, \]

\(\tilde{\mu}(\cdot) \rightarrow \text{approximation of } \mu(\cdot).\)
Strategy when sample size \((n)\) is large

Large sample approximation of CGP

\[ y = \tilde{\mu}(\Psi x) + \epsilon, \]

\(\tilde{\mu}(\cdot) \rightarrow\) approximation of \(\mu(\cdot)\).

- \(\tilde{\mu}\) can be chosen from the rich class of low rank Gaussian processes.
Strategy when sample size \((n)\) is large

Large sample approximation of CGP

\[ y = \tilde{\mu}(\Psi x) + \epsilon, \]

\(\tilde{\mu}(\cdot) \rightarrow \) approximation of \(\mu(\cdot)\).

- \(\tilde{\mu}\) can be chosen from the rich class of low rank Gaussian processes.
- Following Banerjee et al. (2013) we choose

\[
\tilde{\mu}(\Psi x) = E(\mu(\Psi x) | \Phi \mu(\Psi X))
\]

\(\Phi\) is an \(n^* \times n\) matrix, \(\Phi_{ij} \sim N(0, 1)\).

- Each MCMC iteration requires \(n^* \times n^*\) matrix inversion.
- \(n^* \ll n\) implies havoc computational gain.
General Theoretical Setup: Guhaniyogi et al., 2013

True regression function $\mu_0 \in C^s$

Class of regression functions fitted to the data

$\rho$ metric ball of radius $\epsilon_n$ around the truth

$$\rho(\mu, \mu_0)^2 = \frac{1}{n} \sum_{i=1}^{n} (\mu(x_i) - \mu_0(x_i))^2$$
True regression function $\mu_0 \in \mathcal{C}^s$

Class of regression functions fitted to the data

$\rho$ metric ball of radius $\epsilon_n$ around the truth

$$\rho(\mu, \mu_0)^2 = \frac{1}{n} \sum_{i=1}^{n} (\mu(x_i) - \mu_0(x_i))^2$$

Under what condition it shrinks fast enough?

Ordinary GP regression shrinks at the rate $n^{-s/(2s+p)}$. 
Main Results

Theorem

1. $\mathcal{M}$ is a $d$ dimensional $C^{r_1}$ compact sub-manifold of $\mathbb{R}^p$. 

Winter 2018
Main Results

**Theorem**

1. $\mathcal{M}$ is a $d$ dimensional $C^{r_1}$ compact sub-manifold of $\mathbb{R}^p$.
2. $T : \mathbb{R}^p \to \mathbb{R}^m$, $m << p$ s.t. restriction of $T$ in $\mathcal{M}$ is a $C^{r_2}$ diffeomorphism onto its image.
3. $s < \min\{2, r_1 - 1, r_2 - 1\}$.

Then $\epsilon_n = n^{-s/(2s+d)} \log(n)^{d+1}$.

- $T(x) = \Psi x$ is both dimension reducing map and a diffeomorphism onto its image as w.p. $1 - \phi_n$

$$
(1 - \kappa) \sqrt{\frac{m}{p}} \|x_i - x_j\| < \|T(x_i) - T(x_j)\| < (1 + \kappa) \sqrt{\frac{m}{p}} \|x_i - x_j\|.
$$

- Additionally noise reduction is achieved through $T$. 

Theorem

1. $\mathcal{M}$ is a $d$ dimensional $C^{r_1}$ compact sub-manifold of $\mathbb{R}^p$.
2. $T : \mathbb{R}^p \to \mathbb{R}^m$, $m << p$ s.t. restriction of $T$ in $\mathcal{M}$ is a $C^{r_2}$ diffeomorphism onto its image.
3. $s < \min\{2, r_1 - 1, r_2 - 1\}$.

Then $\epsilon_n = n^{-s/(2s+d)} \log(n)^{d+1}$.

- $T(x) = \Psi x$ is both dimension reducing map and a diffeomorphism onto its image as w.p. $1 - \phi_n$

$$(1 - \kappa) \sqrt{\frac{m}{p}} \|x_i - x_j\| < \|T(x_i) - T(x_j)\| < (1 + \kappa) \sqrt{\frac{m}{p}} \|x_i - x_j\|.$$  

- Additionally noise reduction is achieved through $T$. 

20 random splitting of the data into 648 training and 50 test samples.

response is standardized to have unit variance.

To deal with a more challenging case, \( N(0, \tau^2) \) noise is added to each of 4096 pixels to form noisy predictors.

CGP model for large \( n \) is fitted to the data.

Predictive inference is carried out with summary measures mean squared prediction error (MSPE), coverage and length of 95% predictive interval.
### Frequentist Competitors

<table>
<thead>
<tr>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compressed Random Forest (CRF)</td>
</tr>
<tr>
<td>Distributed Supervised Learning (DSL)</td>
</tr>
</tbody>
</table>
Isomap face data analysis: Competitors

<table>
<thead>
<tr>
<th>Frequentist Competitors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compressed Random Forest (CRF)</td>
</tr>
<tr>
<td>Distributed Supervised Learning (DSL)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bayesian Competitors</th>
</tr>
</thead>
<tbody>
<tr>
<td>GP</td>
</tr>
<tr>
<td>2GP</td>
</tr>
<tr>
<td>Compressed Bayesian Additive Regression Tree (CBART)</td>
</tr>
</tbody>
</table>

- Compress high dimensional predictors and apply RF and BART on compressed predictors.
Computation Time (in seconds): CGP is the fastest

![Graph showing computation time vs log(number of features) for different models: CGP, GP, CRF, CBART, DSL. The X-axis represents the log of the number of features, and the Y-axis represents computation time. CGP is the fastest among all models.]
Mean Squared Prediction error (MSPE): Compressed methods perform best

\[ y_1, \ldots, y_k \rightarrow \text{observed}, \quad y_1^*, \ldots, y_k^* \rightarrow \text{predicted} \]

\[ MSPE = \frac{1}{k} \sum_{i=1}^{k} (y_i - y_i^*)^2 \]

<table>
<thead>
<tr>
<th>( \tau )</th>
<th>CGP</th>
<th>GP</th>
<th>CBART</th>
<th>CRF</th>
<th>DSL</th>
<th>2GP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.03</td>
<td>0.14_{0.059}</td>
<td>0.92_{0.074}</td>
<td>0.06_{0.005}</td>
<td>0.05_{0.007}</td>
<td>0.68_{0.023}</td>
<td>0.95_{0.062}</td>
</tr>
<tr>
<td>0.06</td>
<td>0.09_{0.006}</td>
<td>0.79_{0.056}</td>
<td>0.09_{0.007}</td>
<td>0.09_{0.008}</td>
<td>0.75_{0.015}</td>
<td>0.94_{0.041}</td>
</tr>
<tr>
<td>0.10</td>
<td>0.12_{0.008}</td>
<td>0.83_{0.077}</td>
<td>0.12_{0.005}</td>
<td>0.13_{0.011}</td>
<td>0.54_{0.014}</td>
<td>0.92_{0.013}</td>
</tr>
</tbody>
</table>

Table: MSPE and standard error (computed using 20 samples) for all the competitors over 50 replications
Figure: coverage and length of 95% PI's for CGP, GP, CBART, CRF. 95% CI's are shown at each point.
Gaussian process with compactly supported correlation functions

- Under Matern correlation kernel, the correlation between two points is positive even when they are sufficiently far apart.
- In practice, one may safely assume that two observations are not correlated to each other if they are sufficiently far apart.
- How to impose that restriction?
- What if we define a correlation kernel $C_\nu(s, s')$ which is 0 when $||s - s'|| > \nu$.
- These are known as tapered correlation kernels.
Kaufman et al. (2009) proposed

\[ y(s) = x(s)'\beta + w(s)\eta(s) + \epsilon(s), \quad \epsilon(s) \sim N(0, \tau^2). \]

- \( w(\cdot) \sim GP(0, C_\theta(\cdot, \cdot)), \eta(\cdot) \sim GP(0, C_\nu(\cdot, \cdot)). \)
- The covariance matrix of \( y = (y(s_1), ..., y(s_n))' \) becomes sparse.
- Use sparse matrix solvers to efficiently compute inverse.
- It was proved theoretically that this model will asymptotically provide the same inference as the full Gaussian process model without tapering if the tapering range \( \nu \) is chosen properly.
- In practice, we do not know how to choose \( \nu \).
- \( \nu \) acts as a tuning parameter that is adjusted based on the available computational resources.
Recall the model for modified predictive process

\[ y(s) = x(s)'\beta + \tilde{w}(s) + \tilde{\epsilon}(s) + \epsilon(s) \]

**Tapered adjustment** (Guhaniyogi et al., 2012; Sang et al., 2012)

\[ \tilde{\epsilon}(\cdot) \sim GP(0, C_{tap}(s_1, s_2)) \]

\[ C_{tap}(s_1, s_2; \theta) = C_{\tilde{\epsilon}}(s_1, s_2; \theta) C_{\nu}(|s_1 - s_2|) , \]

- \( C_{\nu}(|s_1 - s_2|) \) is a compactly supported correlation function on \([0, \nu]\).
Recall the model for modified predictive process

\[ y(s) = x(s)'\beta + \tilde{\nu}(s) + \tilde{\epsilon}(s) + \epsilon(s) \]

Tapered adjustment (Guhaniyogi et al., 2012; Sang et al., 2012)

\[ \tilde{\epsilon}(\cdot) \sim GP(0, C_{tap}(s_1, s_2)) \]
\[ C_{tap}(s_1, s_2; \theta) = C_{\tilde{\epsilon}}(s_1, s_2; \theta)C_{\nu}(||s_1 - s_2||) , \]

- \( C_{\nu}(||s_1 - s_2||) \) is a compactly supported correlation function on \([0, \nu]\).

\[ \nu = 0 \Rightarrow MPP \]
\[ \nu = \infty \Rightarrow GSP \]
Mean square continuity and differentiability at $s_0$ of a process $w(\cdot)$ requires existence of some vector $\nabla w(s_0)$ with,

$$\lim_{s \to s_0} E (w(s) - w(s_0))^2 = 0$$

$$\lim_{h \to 0} E \left( \frac{w(s_0 + hu) - w(s_0)}{h} - \langle \nabla w(s_0), u \rangle \right)^2 = 0$$
Mean square continuity and differentiability at $s_0$ of a process $w(\cdot)$ requires existence of some vector $\nabla w(s_0)$ with,

$$\lim_{s \to s_0} E (w(s) - w(s_0))^2 = 0$$

$$\lim_{h \to 0} E \left( \frac{w(s_0 + hu) - w(s_0)}{h} - \langle \nabla w(s_0), u \rangle \right)^2 = 0$$

**Theorem on Smoothness (Guhaniyogi et al., 2012)**

With matern correlation function having smoothness $m$,

1. Predictive Process model is infinitely mean square differentiable except at the set of knot points $\mathcal{S}^*$.

2. Modified Predictive Process is not mean square continuous at any point.

3. Tapered Predictive Process is $\min(m, k)$-times mean square differentiable except at $\mathcal{S}^*$, where $C_\nu(\cdot)$ is $k$-times differentiable.

Winter 2018
<table>
<thead>
<tr>
<th></th>
<th>True</th>
<th>Non-spatial</th>
<th>PP</th>
<th>Modified PP</th>
<th>Tapered PP</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>8.25</td>
<td>8.26 (8.15, 8.27)</td>
<td>10.83 (9.29, 12.60)</td>
<td>9.21 (7.83, 10.97)</td>
<td>8.43 (7.20, 9.64)</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>6</td>
<td>–</td>
<td>8.95 (2.68, 15.81)</td>
<td>5.07 (3.44, 7.32)</td>
<td>4.06 (3.12, 5.91)</td>
</tr>
<tr>
<td>$\tau^2$</td>
<td>0.5</td>
<td>3.59 (3.30, 3.88)</td>
<td>2.20 (2.02, 2.40)</td>
<td>.73 (.39, 1.17)</td>
<td>0.43 (0.34, 0.55)</td>
</tr>
<tr>
<td>$\phi$</td>
<td>4</td>
<td>–</td>
<td>2.78 (2.32, 3.62)</td>
<td>2.73 (2.23, 5.38)</td>
<td>4.09 (2.61, 5.77)</td>
</tr>
<tr>
<td>$G$</td>
<td>–</td>
<td>3959.95</td>
<td>2397.21</td>
<td>347.16</td>
<td>146.72</td>
</tr>
<tr>
<td>$P$</td>
<td>–</td>
<td>3943.83</td>
<td>2502.70</td>
<td>1471.05</td>
<td>858.04</td>
</tr>
<tr>
<td>$D$</td>
<td>–</td>
<td>7903.79</td>
<td>4899.91</td>
<td>1818.22</td>
<td>1004.76</td>
</tr>
<tr>
<td>$p_D$</td>
<td>–</td>
<td>1.95</td>
<td>31.79</td>
<td>731.42</td>
<td>1010.30</td>
</tr>
<tr>
<td>DIC</td>
<td>–</td>
<td>2509.32</td>
<td>2000.50</td>
<td>1628.88</td>
<td>1370.06</td>
</tr>
</tbody>
</table>