Advanced Bayesian Computation Week 8

Rajarshi Guhaniyogi
Winter 2018

March 6, 2018

Random Projected Gaussian Process: Banerjee et al., 2013

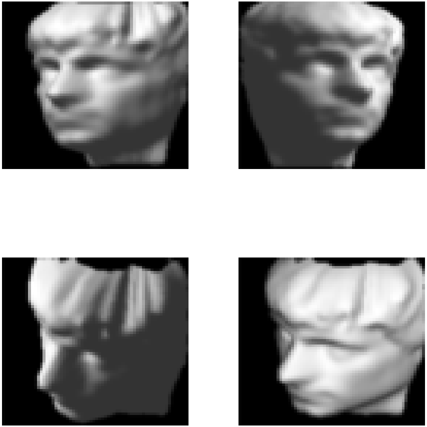
- We have already shown one technique to select knots.
- However it was computationally cumbersome.
- What if we completely avoid the choice of knots.
- $\boldsymbol{w} = (w(\boldsymbol{s}_1), ..., w(\boldsymbol{s}_n))'$, $\boldsymbol{\Phi}$ is an $n^* \times n$ random matrix.
- $\tilde{w}(\boldsymbol{s}) = \mathrm{E}[w(\boldsymbol{s})|\boldsymbol{\Phi w}] = \boldsymbol{c}(\boldsymbol{s})'\boldsymbol{\Phi}'(\boldsymbol{\Phi C_\theta \Phi}')^{-1}\boldsymbol{\Phi w}$.
- $\boldsymbol{c}(\boldsymbol{s}) = (C(\boldsymbol{s}, \boldsymbol{s}_1, \boldsymbol{\theta}), ..., C(\boldsymbol{s}, \boldsymbol{s}_n, \boldsymbol{\theta}))'$.
- $\tilde{\epsilon}(\boldsymbol{s}) \overset{ind.}{\sim} N(0, C(\boldsymbol{s}, \boldsymbol{s}, \boldsymbol{\theta}) - \boldsymbol{c}(\boldsymbol{s})'\boldsymbol{\Phi}'(\boldsymbol{\Phi C_\theta \Phi}')^{-1}\boldsymbol{\Phi c}(\boldsymbol{s}))$.

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

- 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 $\boldsymbol{\Phi}$ rather than randomly selecting entries of $\boldsymbol{\Phi}$.

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.



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)

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.

- Horizontal pose angle of each image is given.

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.

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`)

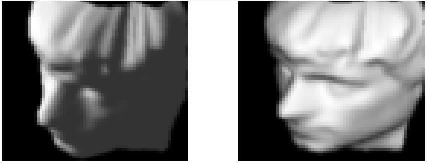
Scientific Question & Challenges

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 predictors

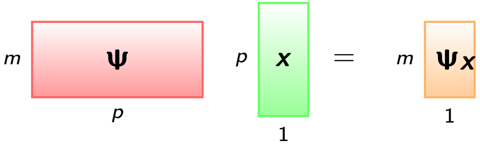
State-of-the-art approaches: unsatisfactory performance

Issues with existing approaches

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

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).

Compressed Gaussian Process



- $\boldsymbol{\Psi} = ((\Psi_{ij}))$, $\Psi_{ij} \sim N(0,1)$: Choice motivated by the popular compressed sensing literature (Ji et al., 20018).
- $\boldsymbol{x} = \boldsymbol{z} + \boldsymbol{\delta}$, $\boldsymbol{z} \in \mathscr{M}$, $\boldsymbol{\delta} \sim N(\boldsymbol{0}, \tau^2 \boldsymbol{I}_p)$.

Compressed GP model

$$
\begin{aligned}
y &= \mu(\boldsymbol{\Psi x}) + \epsilon,\ \epsilon \sim N(0, \sigma^2) \\
\mu(\cdot)|\sigma^2 &\sim GP(0, \sigma^2 K(\cdot,\cdot,\phi)) \\
K(\boldsymbol{x}_i, \boldsymbol{x}_j, \phi) &= \exp\left(-\phi||\boldsymbol{x}_i - \boldsymbol{x}_j||^2\right)
\end{aligned}
$$

Model fitting requires $n \times n$ matrix inversion at each MCMC.

Strategy when sample size (n) is large

Large sample approximation of CGP

$$y = \tilde{\mu}(\boldsymbol{\Psi x}) + \epsilon,$$

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

Strategy when sample size (n) is large

Large sample approximation of CGP

$$y = \tilde{\mu}(\boldsymbol{\Psi x}) + \epsilon,$$

$\tilde{\mu}(\cdot) \rightarrow$ 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

$$y = \tilde{\mu}(\boldsymbol{\Psi x}) + \epsilon,$$

$\tilde{\mu}(\cdot) \to$ approximation of $\mu(\cdot)$.

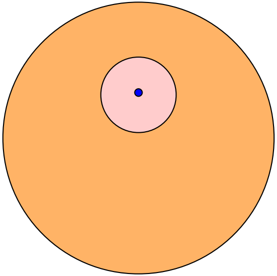
- $\tilde{\mu}$ can be chosen from the rich class of low rank Gaussian processes.
- Following Banerjee et al. (2013) we choose

$$\boxed{\tilde{\mu}(\boldsymbol{\Psi x}) = E(\mu(\boldsymbol{\Psi x}) | \boldsymbol{\Phi} \mu(\boldsymbol{\Psi X}))}$$

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

- Each MCMC iteration requires $n^* \times n^*$ matrix inversion.
- $n^* << n$ implies havoc computational gain.

General Theoretical Setup: Guhaniyogi et al., 2013



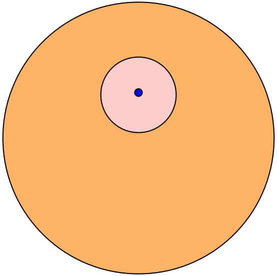
True regression function $\mu_0 \in \mathscr{C}^s$

Class of regression functions fitted to the data

ρ metric ball of radius ϵ_n around the truth

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

General Theoretical Setup: Guhaniyogi et al., 2013



True regression function $\mu_0 \in \mathscr{C}^s$

Class of regression functions fitted to the data

ρ metric ball of radius ϵ_n around the truth

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

Under what condition it shrinks fast enough?

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

Main Results

Theorem

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

Main Results

Theorem

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

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

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

$$(1-\kappa)\sqrt{\frac{m}{p}}||\boldsymbol{x}_i - \boldsymbol{x}_j|| < ||T(\boldsymbol{x}_i) - T(\boldsymbol{x}_j)|| < (1+\kappa)\sqrt{\frac{m}{p}}||\boldsymbol{x}_i - \boldsymbol{x}_j||.$$

- Additionally noise reduction is achieved through T.

Main Results

Theorem

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

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

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

$$(1-\kappa)\sqrt{\frac{m}{p}}||\boldsymbol{x}_i - \boldsymbol{x}_j|| < ||T(\boldsymbol{x}_i) - T(\boldsymbol{x}_j)|| < (1+\kappa)\sqrt{\frac{m}{p}}||\boldsymbol{x}_i - \boldsymbol{x}_j||.$$

- Additionally noise reduction is achieved through T.

Isomap face data analysis: Set up

- 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.

Isomap face data analysis: Competitors

Frequentist Competitors
Compressed Random Forest (CRF)
Distributed Supervised Learning (DSL)

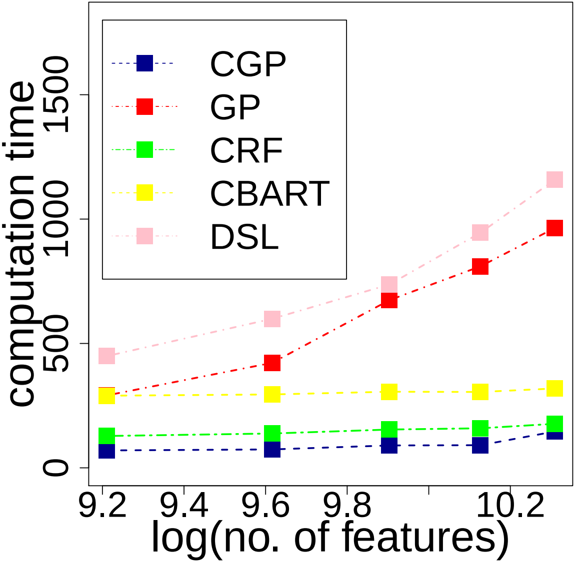
Isomap face data analysis: Competitors

Frequentist Competitors
Compressed Random Forest (CRF)
Distributed Supervised Learning (DSL)

Bayesian Competitors
GP
2GP
Compressed Bayesian Additive Regression Tree (CBART)

- Compress high dimensional predictors and apply RF and BART on compressed predictors.

Computation Time (in seconds): CGP is the fastest



Mean Squared Prediction error (MSPE): Compressed methods perform best

$y_1, \ldots, y_k \rightarrow$ **observed**, $y_1^*, \ldots, y_k^* \rightarrow$ **predicted**

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

τ	CGP	GP	CBART	CRF	DSL	2GP
0.03	$0.14_{0.059}$	$0.92_{0.074}$	$0.06_{0.005}$	$0.05_{0.007}$	$0.68_{0.023}$	$0.95_{0.062}$
0.06	$0.09_{0.006}$	$0.79_{0.056}$	$0.09_{0.007}$	$0.09_{0.008}$	$0.75_{0.015}$	$0.94_{0.041}$
0.10	$0.12_{0.008}$	$0.83_{0.077}$	$0.12_{0.005}$	$0.13_{0.011}$	$0.54_{0.014}$	$0.92_{0.013}$

Table: MSPE and standard error (computed using 20 samples) for all the competitors over 50 replications

Coverage and Length of 95% Predictive Intervals

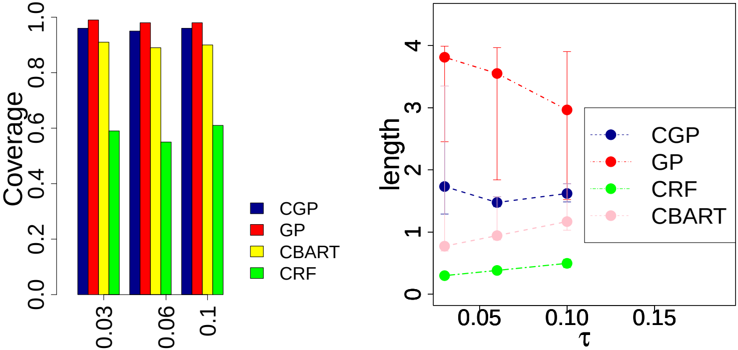


Figure: coverage and length of 95% PI's for CGP, GP, CBART, CRF. 95% CI's are shown at each point

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 $C_\nu(\boldsymbol{s}, \boldsymbol{s}')$ which is 0 when $||\boldsymbol{s} - \boldsymbol{s}'|| > \nu$.
- These are known as tapered correlation kernels.
- Wendland (1995) proposed tapered correlation kernels and later Gneting (2002) formalized the concept.

Gaussian process with compactly supported correlation functions

- Kaufman et al. (2009) proposed

$$y(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})'\boldsymbol{\beta} + w(\boldsymbol{s})\eta(\boldsymbol{s}) + \epsilon(\boldsymbol{s}), \;\; \epsilon(\boldsymbol{s}) \sim N(0, \tau^2).$$

- $w(\cdot) \sim GP(0, C_{\boldsymbol{\theta}}(\cdot,\cdot))$, $\eta(\cdot) \sim GP(0, C_{\nu}(\cdot,\cdot))$.
- The covariance matrix of $\boldsymbol{y} = (y(\boldsymbol{s}_1), ..., y(\boldsymbol{s}_n))'$ 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 ν is chosen properly.
- In practice, we do not know how to choose ν.
- ν acts as a tuning parameter that is adjusted based on the available computational resources.

Tapered Predictive Process

▸ MPP slide Recall the model for modified predictive process

$$y(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})'\boldsymbol{\beta} + \tilde{w}(\boldsymbol{s}) + \tilde{\epsilon}(\boldsymbol{s}) + \epsilon(\boldsymbol{s})$$

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

$$\begin{aligned} \tilde{\epsilon}(\cdot) &\sim GP(0, C_{tap}(\boldsymbol{s}_1, \boldsymbol{s}_2)) \\ C_{tap}(\boldsymbol{s}_1, \boldsymbol{s}_2; \boldsymbol{\theta}) &= C_{\tilde{\epsilon}}(\boldsymbol{s}_1, \boldsymbol{s}_2; \boldsymbol{\theta}) C_\nu(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|) , \end{aligned}$$

- $C_\nu(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|)$ is a compactly supported correlation function on $[0, \nu]$.

Tapered Predictive Process

▸ MPP slide Recall the model for modified predictive process

$$y(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})'\boldsymbol{\beta} + \tilde{w}(\boldsymbol{s}) + \tilde{\epsilon}(\boldsymbol{s}) + \epsilon(\boldsymbol{s})$$

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

$$\begin{aligned} \tilde{\epsilon}(\cdot) &\sim GP(0, C_{tap}(\boldsymbol{s}_1, \boldsymbol{s}_2)) \\ C_{tap}(\boldsymbol{s}_1, \boldsymbol{s}_2; \boldsymbol{\theta}) &= C_{\tilde{\epsilon}}(\boldsymbol{s}_1, \boldsymbol{s}_2; \boldsymbol{\theta}) C_{\nu}(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|) \ , \end{aligned}$$

- $C_{\nu}(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|)$ is a compactly supported correlation function on $[0, \nu]$.

$$\begin{aligned} \nu &= 0 \Rightarrow MPP \\ \nu &= \infty \Rightarrow GSP \end{aligned}$$

Low rank models:Do they oversmooth?

- Mean square continuity and differentiability at $\boldsymbol{s}_0$ of a process $w(\cdot)$ requires existence of some vector $\nabla w(\boldsymbol{s}_0)$ with,

$$\lim_{\boldsymbol{s}\to\boldsymbol{s}_0} E\left(w(\boldsymbol{s}) - w(\boldsymbol{s}_0)\right)^2 = 0$$

$$\lim_{h\to 0} E\left(\frac{w(\boldsymbol{s}_0 + h\boldsymbol{u}) - w(\boldsymbol{s}_0)}{h} - \langle \nabla w(\boldsymbol{s}_0), \boldsymbol{u}\rangle\right)^2 = 0$$

Low rank models:Do they oversmooth?

- Mean square continuity and differentiability at $\boldsymbol{s}_0$ of a process $w(\cdot)$ requires existence of some vector $\nabla w(\boldsymbol{s}_0)$ with,

$$\lim_{\boldsymbol{s}\to\boldsymbol{s}_0} E\left(w(\boldsymbol{s}) - w(\boldsymbol{s}_0)\right)^2 = 0$$

$$\lim_{h\to 0} E\left(\frac{w(\boldsymbol{s}_0 + h\boldsymbol{u}) - w(\boldsymbol{s}_0)}{h} - \langle \nabla w(\boldsymbol{s}_0), \boldsymbol{u}\rangle\right)^2 = 0$$

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 $\mathscr{S}^*$.
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 $\mathscr{S}^*$, where $C_\nu(\cdot)$ is k-times differentiable.

Results

	True	Non-spatial	PP	Modified PP	Tapered PP
β_0	8.25	8.26 (8.15 , 8.27)	10.83 (9.29 , 12.60)	9.21 (7.83 , 10.97)	8.43 (7.20 , 9.64)
σ^2	6	–	8.95 (2.68 , 15.81)	5.07 (3.44 , 7.32)	4.06 (3.12 , 5.91)
τ^2	0.5	3.59 (3.30 , 3.88)	2.20 (2.02 , 2.40)	.73 (.39 , 1.17)	0.43 (0.34 , 0.55)
ϕ	4	–	2.78 (2.32 , 3.62)	2.73 (2.23 , 5.38)	4.09 (2.61 , 5.77)
G	–	3959.95	2397.21	347.16	146.72
P	–	3943.83	2502.70	1471.05	858.04
D	–	7903.79	4899.91	1818.22	1004.76
p_D	–	1.95	31.79	731.42	1010.30
DIC	–	2509.32	2000.50	1628.88	1370.06