Predictive Distribution for Gaussian Process Models with Design-Based Subagging

Linglin He\textsuperscript{a}, Yufan Liu\textsuperscript{b} and Ying Hung\textsuperscript{a}

\textsuperscript{a}Rutgers University, Department of Statistics and Biostatistics

\textsuperscript{b} Dun & Bradstreet

Oct 13, 2017
Overview

1. Introduction
2. LHD-Based Block Bootstrap
3. Consistency of the Bootstrap Estimators
4. Bootstrap Predictive Distribution
5. Numerical Studies
6. Future Work
7. Summary
Computer Experiments

- Computer experiments refer to the study of real systems using complex mathematical models.
- They have been widely used as alternatives to physical experiments, which sometimes are unethical, impossible, inconvenient or too expensive.
- They are nearly deterministic in the sense that a particular input will produce almost the same output.
- Therefore, it is desirable to build an interpolator for computer experiment outputs and use it as an emulator for the actual computer experiment.
Gaussian Process Model for Computer Experiments

- Consider a computer experiment which has $n$ inputs $\mathbf{x} \in \mathbb{R}^d$ and produces univariate output $y(\mathbf{x})$. To analyze the experiments, $y(\mathbf{x})$ is assumed to be a realization from a stochastic process model:

$$ Y(\mathbf{x}) = \mu(\mathbf{x}) + Z(\mathbf{x}), $$

- mean function: $\mu(\mathbf{x}) = \mathbf{x}^T \beta$
- $Z(\mathbf{x})$: stationary Gaussian process with mean 0 and covariance function $\sigma^2 \psi$
- covariance function:

$$ \text{Cov}\{Z(\mathbf{x}_i), Z(\mathbf{x}_j)\} = \sigma^2 \psi(\mathbf{x}_i - \mathbf{x}_j; \theta) $$

where $\theta$ is a vector of correlation parameter for the correlation function.
Given \( n \) observed realizations \( X_n \) and \( Y_n \), the log-likelihood function, ignoring a constant, can be written as

\[
\ell(X_n, y_n, \phi) = -\frac{1}{2\sigma^2} (y_n - X_n\beta)^T R_n^{-1}(\theta)(y_n - X_n\beta) - \frac{1}{2} \log |R_n(\theta)| - \frac{n}{2} \log(\sigma^2),
\]

where \( R_n(\theta) = [\psi(x_i - x_j); \theta), i, j = 1, \ldots, n] \) is an \( n \times n \) correlation matrix.

The MLE can be obtained by

\[
\hat{\beta}_n = (X_n^T R_n^{-1}(\theta) X_n)^{-1} X_n^T R_n^{-1}(\theta) y_n,
\]

\[
\hat{\sigma}_n^2 = (y_n - X_n\hat{\beta}_n)^T R_n^{-1}(\theta)(y_n - X_n\hat{\beta}_n)/n,
\]

\[
\hat{\theta}_n = \arg \min_{\theta} \left\{ n \log(\hat{\sigma}_n^2) + \log |R_n(\theta)| \right\}.
\]
Based on the MLEs, we are interested in predicting $y_{n+1}$ at an untried new input $x_{n+1}$ and quantifying the uncertainty. To achieve this, the conventional plug-in method predicts $y_{n+1}$ by a distribution $g(x_{n+1} | X_n, Y_n, \hat{\varphi}_n)$ which is normally distributed with mean

$$
\mu(x_{n+1} | X_n, y_n, \hat{\varphi}_n) = x_{n+1}^T \hat{\beta}_n + \gamma_n(\hat{\theta}_n)^T R_n^{-1}(\hat{\theta}_n)(y_n - X_n \hat{\beta}_n)
$$

and variance

$$
\sigma^2(x_{n+1} | X_n, y_n, \hat{\varphi}_n) = \hat{\sigma}_n^2 \{1 - \gamma_n(\hat{\theta}_n)^T R_n^{-1}(\hat{\theta}_n) \gamma_n(\hat{\theta}_n)\}
$$

where $\gamma_n(\hat{\theta}_n)$ is the correlation between the new observation and the existing data, i.e. $\gamma_n(\hat{\theta}_n) = [\psi(x_i - x_{n+1}; \hat{\theta}_n), i = 1, \ldots, n]$. 


Gaussian Process Model

Figure: Example of Gaussian Process model in 1-D
Introduction

Challenge and Motivation

- Computational issue that hinders GP from broader application
  - Modelling and making inference involves manipulations of a $n \times n$ correlation matrix $R_n(\theta)$, such as the calculation of $R^{-1}_n(\theta)$ and $|R_n(\theta)|$. The computational order is $O(n^3)$.

- The underestimation of GP predictor uncertainty
  - The resulting plug-in predictors tend to underestimate the uncertainty because variance is obtained by substituting the true parameters with their estimators.
LHD Example

Figure: Example of Latin Hypercube Design in 2-D
Example of LHD-Based Block Bootstrap

Figure: $d = 2$, $l = 24$, $b = 4$, $m = 6$, $|B_n(i)| = 6$, $N = 36$, $n = 216$
Example of LHD-Based Block Bootstrap

Figure: $d = 2$, $l = 24$, $b = 4$, $m = 6$, $|B_n(i)| = 6$, $N = 36$, $n = 216$
Consistency of the Bootstrap Estimators: Converge in Probability

**Theorem 1**

Assume regularity conditions are satisfied. If $m = o(n^{1/d})$ and $m \to \infty$, then

$$\hat{\phi}_N - \hat{\phi}_n \to 0 \quad \text{prob} - P_{N,\omega}, \text{prob} - P. $$

Note:

$$\hat{T}_N^* \to 0 \quad \text{prob} - P_{N,\omega}, \text{prob} - P \text{ if for any } \epsilon > 0 \text{ and any } \delta > 0, $$

$$\lim_{n \to \infty} P\{P_{N,\omega}^*(|\hat{T}_N^*| > \epsilon > \delta)\} = 0. $$

* Yibo Zhao, Yasuo Amemiya and Ying Hung

Efficient Gaussian Process Modelling Using Experimental Design-based Subagging

*Statistica Sinaca*
Objective of this research is to construct a predictive distribution that is

- 1) easy to compute (due to the subsampling);
- 2) allow a better quantification of predictive uncertainty
Construction Methods

Definition 1 (Direct density prediction)

Given the realization \( \{X_n, y_n\} \), let \( \{X^*_N, y^*_N\} \) be a bootstrap sample, a bootstrap predictive distribution is defined by

\[
g^*_n(x_{n+1} \mid X_n, y_n) = \int g(x_{n+1} \mid X^*_N, y^*_N, \hat{\phi}^*_N) dP^*(X^*_N, y^*_N \mid X_n, y_n),
\]

Based on the subsamples obtained from LHD-based bootstrap, a Monte Carlo estimate of the bootstrap predictive distribution can be obtained by

\[
\tilde{g}^*_n(x_{n+1} \mid X_n, y_n) = U^{-1} \sum_{u=1}^{U} g(x_{n+1} \mid X^*_N(u), Y^*_N(u), \hat{\phi}^*_N(u)),
\]

When \( U \to \infty \), \( \tilde{g}^*_n(x_{n+1} \mid X_n, y_n) \) converges to \( g^*_n(x_{n+1} \mid X_n, y_n) \).
Direct Density Prediction

A Bootstrap Sample \( \{X^*_N, y^*_N\} \)

Parameter Estimation \( \hat{\phi}^*_N \)

Predictive Distribution \( g(x_{n+1} | X^*_N, y^*_N, \hat{\phi}^*_N) \)

Take U Times

Bootstrap Predictive Distribution

\[
U^{-1} \sum_{u=1}^{U} g(x_{n+1} | X^*_N(u), y^*_N(u), \hat{\phi}^*_N(u))
\]
Definition 2 (Normal approximation)

Utilizing LHD-based bootstrap approach, the Monte Carlo estimate of the predictive mean and variance are:

\[
\begin{align*}
\tilde{\mu}^* (x_{n+1} \mid X_n, y_n) &= U^{-1} \sum_{u=1}^{U} \mu(x_{n+1} \mid X_N(u), y_N(u), \hat{\phi}^*_N(u)) \\
\tilde{\sigma}^2 (x_{n+1} \mid X_n, y_n) &= U^{-1} \sum_{u=1}^{U} \sigma^2(x_{n+1} \mid X_N(u), y_N(u), \hat{\phi}^*_N(u)).
\end{align*}
\]

When \( U \to \infty \), \( \tilde{\mu}^* (x_{n+1} \mid X_n, y_n) \) converges to

\[
\mu^* (x_{n+1} \mid X_n, y_n) = \int \mu(x_{n+1} \mid X_N^*, y_N^*, \hat{\phi}^*_N) dP^*(X_N^*, y_N^* \mid X_n, y_n)
\]

and \( \tilde{\sigma}^2 (x_{n+1} \mid X_n, y_n) \) converges to

\[
\sigma^2 (x_{n+1} \mid X_n, y_n) = \int \sigma^2(x_{n+1} \mid X_N^*, y_N^*, \hat{\phi}^*_N) dP^*(X_N^*, y_N^* \mid X_n, y_n).
\]
Normal Approximation Prediction

A Bootstrap Sample

\( \{X^*_N, y^*_N\} \)

Parameter Estimation

\( \hat{\phi}^*_N \)

Predictive Mean

\[ \mu(x_{n+1} \mid X^*_N, y^*_N, \hat{\phi}^*_N) \]

Predictive Variance

\[ \sigma^2(x_{n+1} \mid X^*_N, y^*_N, \hat{\phi}^*_N) \]

Repeat

Take U Times

Bootstrap Predictive Distribution

\[ N( U^{-1} \sum_{u=1}^U \mu(x_{n+1} \mid X_{N(u)}, y_{N(u)}, \hat{\phi}^*_{N(u)}), U^{-1} \sum_{u=1}^U \sigma^2(x_{n+1} \mid X_{N(u)}, y_{N(u)}, \hat{\phi}^*_{N(u)})) \)
Theoretical Comparison

Theorem 2

Let $\sum_i$ be the summation of all $m^d$ blocks and $\sum_{\pi_1, \ldots, \pi_d}$ be the summation of independent permutation over $\{0, 1, \ldots, m-1\}$. Under assumption (A.3), we have

- **Direct density and normal approximation both have unbiased predictive mean.** i.e.,

$$E(\mu(x_{n+1} | x_n, y_n, \hat{\phi}_n) - \mu_1^*) = E(\mu(x_{n+1} | x_n, y_n, \hat{\phi}_n) - \mu_2^*)$$

$$= E\left(\frac{m^d - 1}{m^d} \sum_{i} \gamma_i(\hat{\theta}_n)^T R_{i,i}(\hat{\theta}_n)(y_i - x_i\hat{\beta}_n) + O(N^{-1/2})\right)$$

$$= 0$$
Simulation Setting

- $Y(x) = \mu(x) + Z(x)$
- $\mu(x) = x^T \beta$
- $\psi(x_1 - x_2) = \exp(-\sum_{i=1}^{2} |x_{1i} - x_{2i}|/\theta_i)$
- $\beta_1 = \beta_2 = \theta_1 = \theta_2 = \sigma^2 = 1$
- Generate $n=400$, $2500$ realizations on regular grid $[0, 1]^2$
- For each choice of sample size, a total of $100$ data sets are simulated
- $20$ LHD-based block bootstrap samples are collected
**Simulation studies**

**Table:** Comparisons of prediction in three untried settings with 100 replications (standard deviation in parenthesis).

<table>
<thead>
<tr>
<th>Method</th>
<th>Summary statistics</th>
<th>( x_{n+1} )</th>
<th>( x_{n+2} )</th>
<th>( x_{n+3} )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>( n = 400 )</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plug-in</td>
<td>Mean</td>
<td>0.59 (0.9965)</td>
<td>1.05 (1.0447)</td>
<td>1.10 (1.0558)</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>0.01 (0.0030)</td>
<td>0.03 (0.0061)</td>
<td>0.02 (0.0040)</td>
</tr>
<tr>
<td>Density ( m = 4 )</td>
<td>Mean</td>
<td>0.63 (0.8666)</td>
<td>1.05 (0.9383)</td>
<td>1.09 (0.9533)</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>0.22 (0.0559)</td>
<td>0.15 (0.0298)</td>
<td>0.14 (0.0291)</td>
</tr>
<tr>
<td>Density ( m = 6 )</td>
<td>Mean</td>
<td>0.62 (0.8416)</td>
<td>1.08 (0.9604)</td>
<td>1.12 (0.9683)</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>0.25 (0.0741)</td>
<td>0.15 (0.0388)</td>
<td>0.13 (0.0380)</td>
</tr>
<tr>
<td>Normal ( m = 4 )</td>
<td>Mean</td>
<td>0.63 (0.8666)</td>
<td>1.05 (0.9383)</td>
<td>1.09 (0.9533)</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>0.17 (0.0287)</td>
<td>0.11 (0.0173)</td>
<td>0.09 (0.0150)</td>
</tr>
<tr>
<td>Normal ( m = 6 )</td>
<td>Mean</td>
<td>0.62 (0.8416)</td>
<td>1.08 (0.9604)</td>
<td>1.12 (0.9683)</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>0.14 (0.0249)</td>
<td>0.11 (0.0165)</td>
<td>0.10 (0.0148)</td>
</tr>
<tr>
<td><strong>( n = 2500 )</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plug-in</td>
<td>Mean</td>
<td>0.67 (0.9785)</td>
<td>1.10 (1.1682)</td>
<td>1.14 (1.1731)</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>0.02 (0.0017)</td>
<td>0.02 (0.0016)</td>
<td>0.01 (0.0013)</td>
</tr>
<tr>
<td>Density ( m = 4 )</td>
<td>Mean</td>
<td>0.66 (0.9183)</td>
<td>1.12 (1.0952)</td>
<td>1.14 (1.0899)</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>0.21 (0.0517)</td>
<td>0.11 (0.0272)</td>
<td>0.10 (0.0254)</td>
</tr>
<tr>
<td>Density ( m = 6 )</td>
<td>Mean</td>
<td>0.62 (0.8931)</td>
<td>1.10 (1.0751)</td>
<td>1.13 (1.0710)</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>0.19 (0.0401)</td>
<td>0.11 (0.0204)</td>
<td>0.11 (0.0216)</td>
</tr>
<tr>
<td>Normal ( m = 4 )</td>
<td>Mean</td>
<td>0.66 (0.9183)</td>
<td>1.12 (1.0952)</td>
<td>1.14 (1.0899)</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>0.15 (0.0129)</td>
<td>0.07 (0.0057)</td>
<td>0.06 (0.0050)</td>
</tr>
<tr>
<td>Normal ( m = 6 )</td>
<td>Mean</td>
<td>0.62 (0.8931)</td>
<td>1.10 (1.0751)</td>
<td>1.13 (1.0710)</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>0.13 (0.0115)</td>
<td>0.08 (0.0065)</td>
<td>0.08 (0.0061)</td>
</tr>
</tbody>
</table>
Future Work

Ongoing Work

Compare the predictive variance of direct density prediction and normal approximation with the case when full data is used.

Quantify the gain in predictive variance of LHD-based block bootstrap with SRS block bootstrap.
LHD-based block bootstrap borrows the strength of space-filling designs to provide an efficient subsampling plan and reduce computational complexity.

Two methods are proposed to construct bootstrap predictive distributions.

We show the unbiasedness of the predictive mean under both direct density prediction and normal approximation prediction.
Thank you!