Predictive Distribution for Gaussian Process Models with Design-Based Subagging

Linglin He[a], Yufan Liu[b] and Ying Hung[a]

[a]Rutgers University, Department of Statistics and Biostatistics
[b] Dun & Bradstreet

Oct 13, 2017

Overview

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 $\boldsymbol{x} \in R^d$ and produces univariate output $y(\boldsymbol{x})$. To analyze the experiments, $y(\boldsymbol{x})$ is assumed to be a realization from a stochastic process model:

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

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

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

where θ is a vector of correlation parameter for the correlation function.

Gaussian Process Model (Estimation)

- Given n observed realizations $\boldsymbol{X}_n$ and $\boldsymbol{Y}_n$, the log-likelihood function, ignoring a constant, can be written as

$$\ell(\boldsymbol{X}_n, \boldsymbol{y}_n, \phi) = -\frac{1}{2\sigma^2}(\boldsymbol{y}_n - \boldsymbol{X}_n\beta)^T R_n^{-1}(\boldsymbol{\theta})(\boldsymbol{y}_n - \boldsymbol{X}_n\beta) - \frac{1}{2}\log|R_n(\boldsymbol{\theta})| - \frac{n}{2}\log(\sigma^2),$$

  where $R_n(\boldsymbol{\theta}) = [\psi(\boldsymbol{x}_i - \boldsymbol{x}_j); \boldsymbol{\theta}), i, j = 1, \ldots, n]$ is an $n \times n$ correlation matrix.

- The MLE can be obtained by

$$\hat{\boldsymbol{\beta}}_n = (\boldsymbol{X}_n^T R_n^{-1}(\boldsymbol{\theta})\boldsymbol{X}_n)^{-1}\boldsymbol{X}^T R_n^{-1}(\boldsymbol{\theta})\boldsymbol{y}_n,$$

$$\hat{\sigma}_n^2 = (\boldsymbol{y}_n - \boldsymbol{X}_n\hat{\boldsymbol{\beta}}_n)^T R_n^{-1}(\boldsymbol{\theta})(\boldsymbol{y}_n - \boldsymbol{X}_n\hat{\boldsymbol{\beta}}_n)/n,$$

$$\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}}\{n\log(\hat{\sigma}_n^2) + \log|R_n(\boldsymbol{\theta})|\}.$$

Gaussian Process Model (Prediction)

- Based on the MLEs, we are interested in predicting y_{n+1} at an untried new input $\boldsymbol{x}_{n+1}$ and quantifying the uncertainty. To achieve this, the conventional plug-in method predicts y_{n+1} by a distribution $g(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{Y}_n, \hat{\phi}_n)$ which is normally distributed with mean

$$\mu(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{y}_n, \hat{\phi}_n) = \boldsymbol{x}_{n+1}^T \hat{\boldsymbol{\beta}}_n + \gamma_n(\hat{\boldsymbol{\theta}}_n)^T R_n^{-1}(\hat{\boldsymbol{\theta}}_n)(\boldsymbol{y}_n - \boldsymbol{X}_n \hat{\boldsymbol{\beta}}_n)$$

and variance

$$\sigma^2(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{y}_n, \hat{\phi}_n) = \hat{\sigma}_n^2\{1 - \gamma_n(\hat{\boldsymbol{\theta}}_n)^T R_n^{-1}(\hat{\boldsymbol{\theta}}_n)\gamma_n(\hat{\boldsymbol{\theta}}_n)\}$$

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

Gaussian Process Model

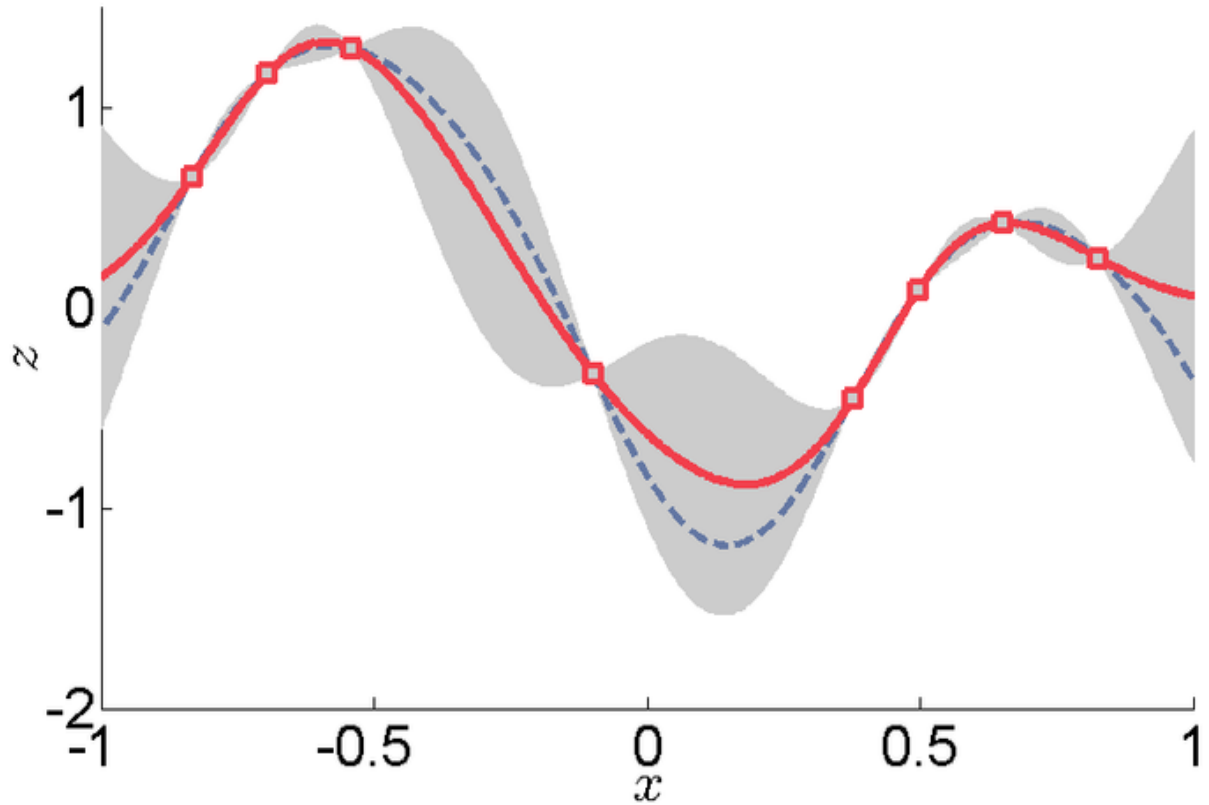


Figure: Example of Gaussian Process model in 1-D

Challenge and Motivation

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

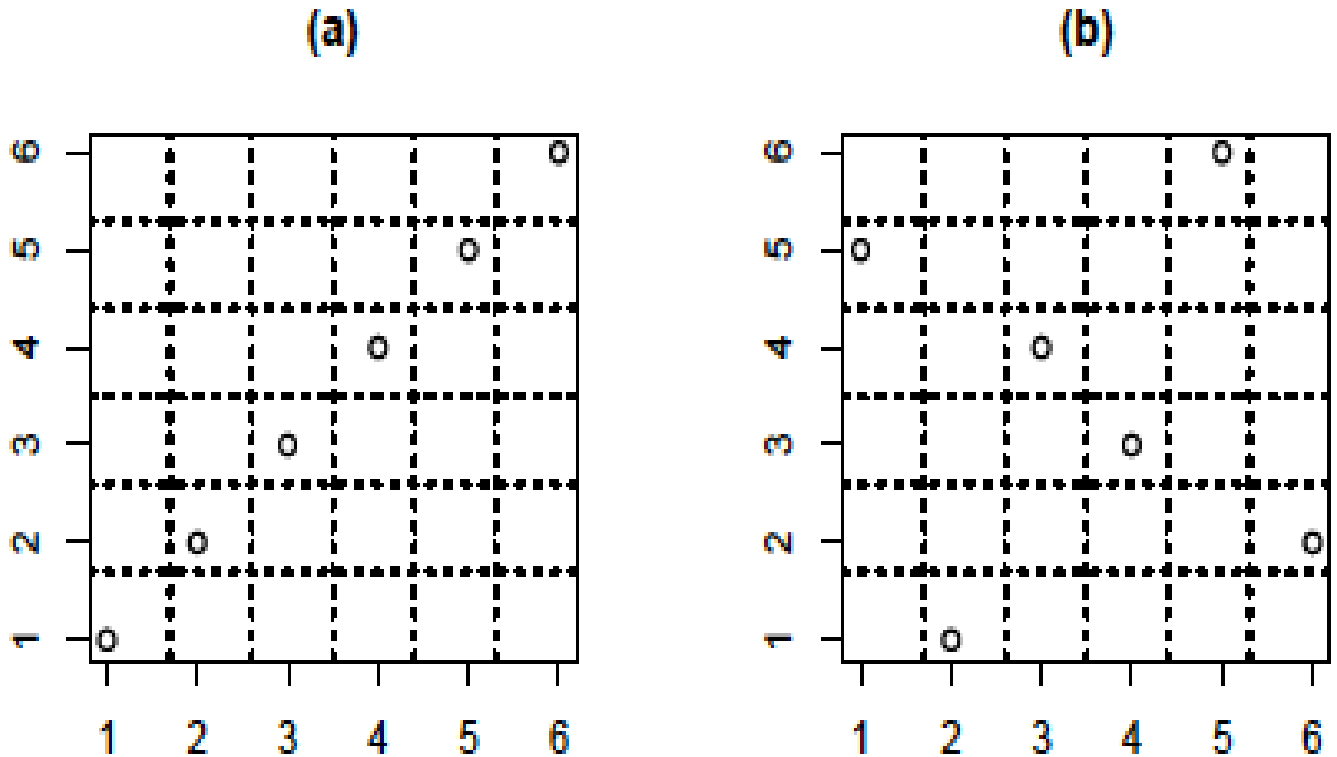


Figure: Example of Latin Hypercube Design in 2-D

Example of LHD-Based Block Bootstrap

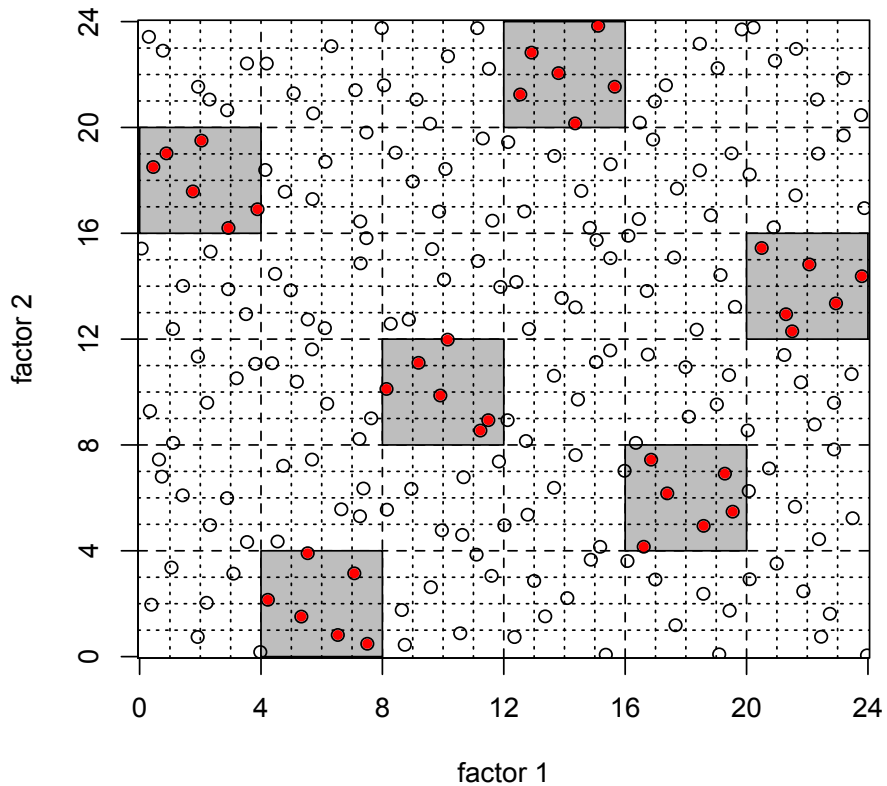


Figure: $d = 2,\ l = 24,\ b = 4,\ m = 6,\ |\mathcal{B}_n(\boldsymbol{i})| = 6,\ N = 36,\ n = 216$

Example of LHD-Based Block Bootstrap

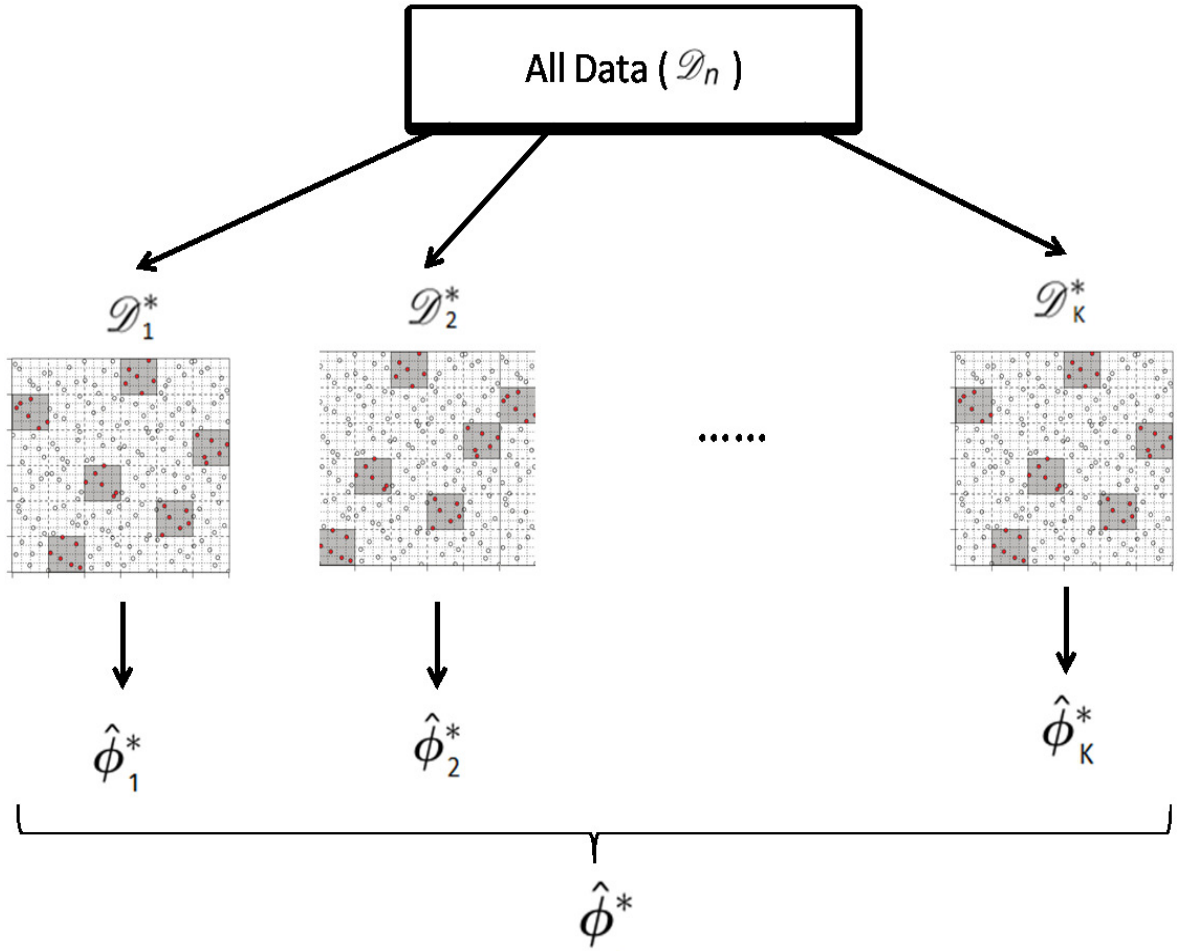


Figure: $d = 2,\ l = 24,\ b = 4,\ m = 6,\ |\mathcal{B}_n(\boldsymbol{i})| = 6,\ N = 36,\ n = 216$

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*

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

Note: $\hat{T}_N^* \to 0$ $prob - P_{N,\omega}^*, prob - P$ if for any $\epsilon > 0$ 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*

Objectives of Research

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 $\{\boldsymbol{X}_n, \boldsymbol{y}_n\}$, *let* $\{\boldsymbol{X}_N^*, \boldsymbol{y}_N^*\}$ *be a bootstrap sample, a bootstrap predictive distribution is defined by*

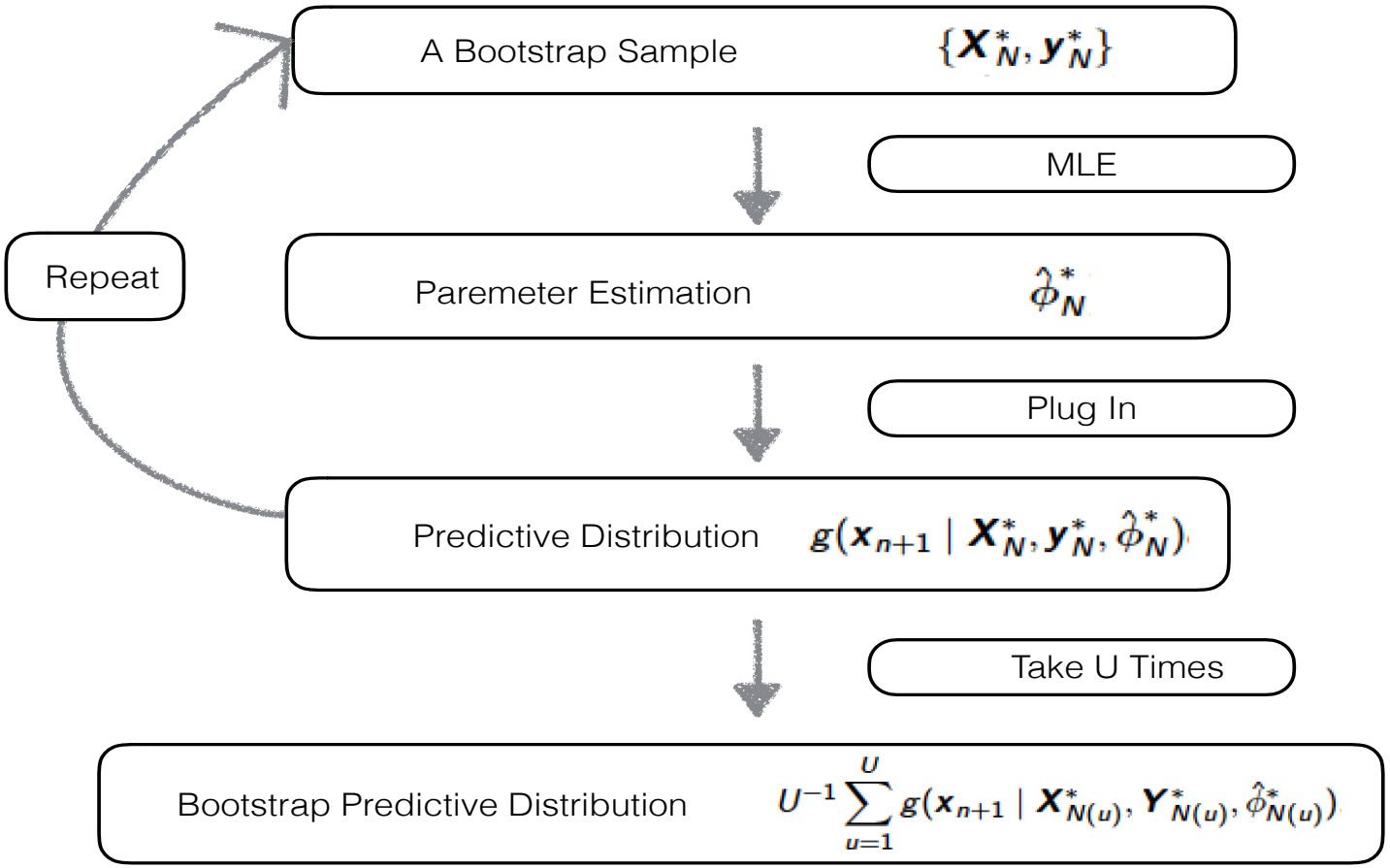
$$g^*(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{y}_n) = \int g(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_N^*, \boldsymbol{y}_N^*, \hat{\phi}_N^*) dP^*(\boldsymbol{X}_N^*, \boldsymbol{y}_N^* \mid \boldsymbol{X}_n, \boldsymbol{y}_n),$$

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

$$\tilde{g}^*(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \mathbf{y}_n) = U^{-1} \sum_{u=1}^{U} g(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_{N(u)}^*, \boldsymbol{Y}_{N(u)}^*, \hat{\phi}_{N(u)}^*),$$

When $U \to \infty$, $\tilde{g}^*(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \mathbf{y}_n)$ *converges to* $g^*(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{y}_n)$.

Direct Density Prediction



Construction Methods

Definition 2 (Normal approximation)

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

$$\tilde{\mu}^*(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{y}_n) = U^{-1} \sum_{u=1}^{U} \mu(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_{N(u)}, \boldsymbol{y}_{N(u)}, \hat{\phi}^*_{N(u)})$$

$$\tilde{\sigma}^{2*}(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{y}_n) = U^{-1} \sum_{u=1}^{U} \sigma^2(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_{N(u)}, \boldsymbol{y}_{N(u)}, \hat{\phi}^*_{N(u)}).$$

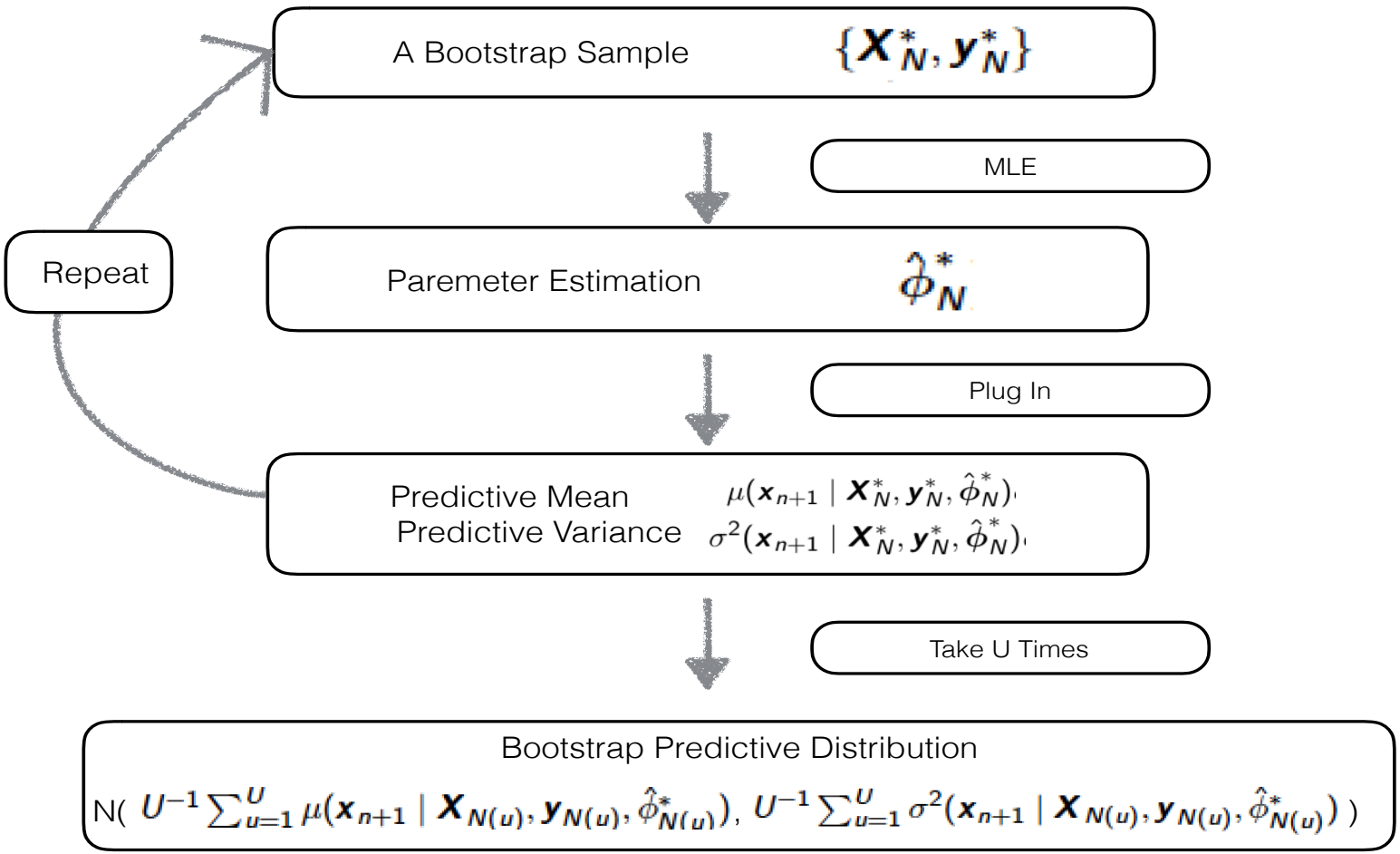
When $U \to \infty$, $\tilde{\mu}^*(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{y}_n)$ *converges to*

$$\mu^*(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{y}_n) = \int \mu(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}^*_N, \boldsymbol{y}^*_N, \hat{\phi}^*_N) dP^*(\boldsymbol{X}^*_N, \boldsymbol{y}^*_N \mid \boldsymbol{X}_n, \boldsymbol{y}_n)$$

and $\tilde{\sigma}^{2*}(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{y}_n)$ *converges to*

$$\sigma^{2*}(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{y}_n) = \int \sigma^2(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}^*_N, \boldsymbol{y}^*_N, \hat{\phi}^*_N) dP^*(\boldsymbol{X}^*_N, \boldsymbol{y}^*_N \mid \boldsymbol{X}_n, \boldsymbol{y}_n).$$

Normal Approximation Prediction



Theoretical Comparison

Theorem 2

Let $\sum_{\boldsymbol{i}}$ 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.,*

$$
\begin{aligned}
& E(\mu(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{y}_n, \hat{\phi}_n) - \mu_1^*) = E(\mu(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{y}_n, \hat{\phi}_n) - \mu_2^*) \\
= \quad & E(\frac{m^{d-1}-1}{m^{d-1}} \sum_{\boldsymbol{i}} \gamma_{\boldsymbol{i}}(\hat{\boldsymbol{\theta}}_n)^T R_{\boldsymbol{i},\boldsymbol{i}}^{-1}(\hat{\boldsymbol{\theta}}_n)(\boldsymbol{y}_{\boldsymbol{i}} - \boldsymbol{X}_{\boldsymbol{i}}\hat{\boldsymbol{\beta}}_n) + O(N^{-1/2})) \\
= \quad & 0
\end{aligned}
$$

Simulation Setting

- $Y(\boldsymbol{x}) = \mu(\boldsymbol{x}) + Z(\boldsymbol{x})$
- $\mu(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$
- $\psi(\boldsymbol{x}_1 - \boldsymbol{x}_2) = \exp(-\sum_{i=1}^{2} |x_{1i} - x_{2i}|/\theta_i)$
- $\beta_1 = \beta_2 = \theta_1 = \theta_2 = \sigma^2 = 1$
- 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).

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

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.

Summary

- **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!