Statistica Sinica Preprint No: SS-2019-0376	
Title	Gaussian Process Prediction using Design-Based Subsampling
Manuscript ID	SS-2019-0376
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202019.0376
Complete List of Authors	Linglin He and Ying Hung
Corresponding Author	Ying Hung
E-mail	yhung@stat.rutgers.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

Gaussian Process Prediction using Design-Based Subsampling

Linglin He and Ying Hung

Rutgers University

Abstract: Gaussian process (GP) models are widely used in the analysis of computer experiments. However, two issues have not been solved satisfactorily. The first is a computational issue that prevents GP models from being more widely applied, especially for massive data with high-dimensional inputs. The second is the underestimation of the prediction uncertainty in GP modeling. To tackle these problems simultaneously, we propose two methods for constructing GP predictive distributions based on a new version of bootstrap subsampling. The new subsampling procedure borrows the strength of space-filling designs to provide an efficient subsample, and thus reduce the computational complexity. Compared with the plug-in approach, this procedure provides unbiased predictors and offers an efficient analogue of conventional bootstrap predictive distributions with empirical coverage probabilities closer to their nominal levels. We illustrate the proposed methods using two complex computer experiments with high-dimensional inputs and tens of thousands of simulation outputs.

Key words and phrases: Computer experiment; Experimental design; Kriging; Sub-bagging; Space-filling design; Uncertainty quantification.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

1. Introduction

Computer experiments examine real systems using complex mathematical models. They are widely used as alternatives to physical experiments, especially when studying complex systems. In many situations, a physical experiment is infeasible because it is unethical, impossible, inconvenient, or too expensive. A mathematical model of a system can often be developed and input/output pairs can be produced with the help of computers. Computer experiments are widely used in science and engineering. Typically, such experiments require a great deal of time and computing. Furthermore, they are nearly deterministic, in the sense that a particular input will produce almost the same output if given to the computer experiment on another occasion. Therefore, it is desirable to build an interpolator for computer experiment outputs, and to use this as an emulator for the actual experiment. Additional discussions of the design and analysis of computer experiments can be found in Santner, Williams and Notz (2003) and Fang, Li and Sudjianto (2006).

A Gaussian process (GP) model (or kriging) is a flexible and widely used method in the analysis of computer experiments; however, there are two critical issues in GP modeling. The first is a computational issue that prevents GP models from being more widely used, especially with high-

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

dimensional inputs and massive outputs. This is because the modeling and prediction of a GP involve significant manipulations of an $N \times N$ correlation matrix, where N is the sample size, requiring $O(N^3)$ computations and often resulting in a singularity. This problem is even more critical when analyzing complex computer experiments, because the estimation of high-dimensional correlation parameters often leads to numerical instability in the estimation and prediction. The second issue is how to accurately quantify the uncertainty based on a GP. It is well known that the GP predictive interval constructed by substituting the true parameters by the estimators, often called the plug-in predictor, underestimates the uncertainty (Santner, Williams and Notz (2003), p.98). Although numerous works examine each of these issues, to the best of our knowledge, there is no systematic approach that addresses both simultaneously, which is the main focus of this study.

The computational issue is well recognized in the literature and a number of methods have been proposed. Some methods address this problem by changing the model to one that is computationally convenient. Here, examples include the works of Rue and Held (2005), Cressie and Johannesson (2008), Banerjee et al. (2008), Gramacy and Lee (2008), Wikle (2010), Chang et al. (2014), Castrillon et al. (2015), Mak and Joseph (2018), and

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

Wang, Yang and Stufken (2019). Another approach is to approximate the likelihood for the original data. Here, examples include the works of Nychka (2000), Stein, Chi and Welty (2004), Furrer, Genton and Nychka (2006), Snelson and Ghahramani (2006), Fuentes (2007), Kaufman, Schervish and Nychka (2008), Gramacy and Apley (2015), and Nychka et al. (2015). Nevertheless, most existing methods are developed for data sets collected from a regular grid under a low-dimensional geostatistical setting. However, these assumptions are often violated in computer experiments because high-dimensional inputs are common, and the computational expense often prohibits running such experiments over a dense grid of input configurations. A commonly used approach for high-dimensional computer experiments is to impose a sparsity constraint on the correlation matrix (Kaufman, Schervish and Nychka (2008), Kaufman et al. (2011)). However, it has been shown that this method does not work well for parameter estimation (Stein (2013), Liang et al. (2013)), which is crucial for the GP predictor. In addition, the connection between the degree of sparsity and the computation time is nontrivial.

The second issue of quantifying the uncertainty in GP predictions is important, but has been overlooked in the literature. For example, most of the aforementioned methods address the computational issue, but adopt plug-in

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

predictors for inference, and therefore underestimate the prediction uncertainty. Moreover, with different approximation techniques, these methods bring in additional uncertainty that is difficult to quantify. Although methods such as Bayesian approaches (Handcock and Stein (1993), Kennedy and O'Hagan (2001), Schmidt and O'Hagan (2003)) and the regular bootstrap (Santner, Williams and Notz (2003), Luna and Young (2003)) have been proposed to provide a better quantification of the prediction uncertainty by incorporating the estimation uncertainty, they are computationally intensive and often intractable for massive data.

In this paper, a new framework is proposed to for constructing GP predictors and their predictive distributions, in which we combine the bootstrap predictive distribution with an experimental design-based stratified subsampling plan. Bootstrapping is an increasingly popular method for obtaining accurate confidence intervals and performing statistical inference (DiCiccio and Efron (1992), DiCiccio and Efron (1996), Efron and Tibshirani (1993). A direct application of bootstrap methods to construct predictive distributions for GP is conceptually attractive, but computationally prohibitive, especially for massive data. Therefore, we introduce a new bootstrap method using design-based subsampling, and propose two methods for constructing of bootstrap predictive distributions. We show that,

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

compared with the plug-in approach, this procedure not only provides unbiased predictors, but also offers an efficient analogue of conventional bootstrap predictive distributions with empirical coverage probabilities closer to their nominal levels. Moreover, theoretical comparisons with commonly used predictors are provided.

The remainder of the paper is organized as follows. In Section 2, we introduce the idea of a Latin hypercube design (LHD)-based block bootstrap and propose two methods for constructing predictive distributions. In Section 3, the proposed predictors are shown to be unbiased. Theoretical comparisons with the regular bootstrap and the plug-in approach are developed. In Section 4, the finite-sample performance of the proposed methods is investigated using simulation studies. Applications to two real examples of computer experiments are given in Section 5. Section 6 concludes the paper.

2.1 Gaussian process models for computer experiments

Consider a computer experiment that has n inputs $\boldsymbol{x} \in R^d$ and produces output $y(\boldsymbol{x})$. To analyze the experiment, $y(\boldsymbol{x})$ is assumed to be a realization from a stochastic process

$$Y(\boldsymbol{x}) = \mu(\boldsymbol{x}) + Z(\boldsymbol{x}), \tag{2.1}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

where the mean function is defined as $\mu(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$, and $Z(\boldsymbol{x})$ is a stationary Gaussian process with mean zero and covariance function $\sigma^2\psi$. The covariance function is defined as $cov\{Y(\boldsymbol{x}+\boldsymbol{h}), Y(\boldsymbol{x})\} = \sigma^2\psi(\boldsymbol{h};\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of correlation parameters for the correlation function $\psi(\boldsymbol{h};\boldsymbol{\theta})$, and $\psi(\boldsymbol{h};\boldsymbol{\theta})$ is a positive semidefinite function with $\psi(\mathbf{0};\boldsymbol{\theta}) = 1$ and $\psi(\boldsymbol{h};\boldsymbol{\theta}) = \psi(-\boldsymbol{h};\boldsymbol{\theta})$. Note that we assume the variables in the mean function are known, such a model is also known as universal kriging. However, the proposed framework is not limited to this assumption. It can be extended to incorporate various variable selection methods for GP models (Li and Sudjianto (2005)).

Suppose n realizations are observed and denoted by

$$\mathscr{D}_n = \{\big(\boldsymbol{x}_{t_1}, y(\boldsymbol{x}_{t_1})\big), \ldots, \big(\boldsymbol{x}_{t_n}, y(\boldsymbol{x}_{t_n})\big)\} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}.$$

Let $\boldsymbol{y}_n = (y_1, \ldots, y_n)^T$, $\boldsymbol{X}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$, and $\boldsymbol{\phi} = (\theta^T, \boldsymbol{\beta}^T, \sigma^2)^T$ be vectors of the parameters, and let Θ be the parameter space. Based on (2.1), the likelihood function can be written as

$$f(\boldsymbol{y}_n, \boldsymbol{X}_n; \boldsymbol{\phi}) = \frac{|R_n(\boldsymbol{\theta})|^{-1/2}}{(2\pi\sigma^2)^{n/2}} \exp\{-\frac{1}{2\sigma^2}(\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta})^T R_n^{-1}(\boldsymbol{\theta})(\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta})\},$$

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

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

matrix. Thus, the log-likehood function, ignoring a constant, is

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

Here, the parameters $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and σ are unknown. They are estimated using likelihood-based methods such as the maximum likelihood or restricted maximum likelihood (REML) (Irvine, Gitelman and Hoeting (2007)). Here, we focus on maximum likelihood estimators (MLEs); the results can be extended to the REML.

For a GP model, the MLEs can be obtained by

$$\hat{\boldsymbol{\beta}}_n = (\boldsymbol{X}_n^T R_n^{-1}(\boldsymbol{\theta})\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T R_n^{-1}(\boldsymbol{\theta})\boldsymbol{y}_n, \tag{2.2}$$

$$\hat{\sigma}_n^2 = (\boldsymbol{y}_n - \boldsymbol{X}_n\hat{\boldsymbol{\beta}}_n)^T R_n^{-1}(\boldsymbol{\theta})(\boldsymbol{y}_n - \boldsymbol{X}_n\hat{\boldsymbol{\beta}}_n)/n, \tag{2.3}$$

and

$$\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}}\{n\log(\hat{\sigma}_n^2) + \log|R_n(\boldsymbol{\theta})|\}, \tag{2.4}$$

where $|R_n(\boldsymbol{\theta})|$ is the determinant of the matrix $R_n(\boldsymbol{\theta})$.

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} using the distribution $g(\boldsymbol{x}_{n+1} \mid$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

$\boldsymbol{X}_n, \boldsymbol{Y}_n, \hat{\phi}_n)$, which is normally distributed with mean

$$\mu(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{y}_n, \hat{\phi}_n) = \boldsymbol{x}_{n+1}^T \hat{\boldsymbol{\beta}}_n + \gamma_n(\hat{\boldsymbol{\theta}}_n)^T R_n^{-1}(\hat{\boldsymbol{\theta}}_n)(\boldsymbol{y}_n - \boldsymbol{X}_n \hat{\boldsymbol{\beta}}_n) \quad (2.5)$$

and variance

$$\sigma^2(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{y}_n, \hat{\phi}_n) = \hat{\sigma}_n^2\{1 - \gamma_n(\hat{\boldsymbol{\theta}}_n)^T R_n^{-1}(\hat{\boldsymbol{\theta}}_n)\gamma_n(\hat{\boldsymbol{\theta}}_n)\}, \quad (2.6)$$

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

Such a predictor is often computationally infeasible for massive data because it requires manipulations of an $n \times n$ correlation matrix $R_n(\hat{\boldsymbol{\theta}}_n)$, such as the calculations of $R_n^{-1}(\boldsymbol{\theta})$ and $|R_n(\boldsymbol{\theta})|$, which are computationally intensive and often intractable owing to numerical issues. This is particularly difficult for massive data (i.e., large n) collected on nonregular grids, such as the space-filling designs commonly used in computer experiments, because Kronecker product techniques cannot be used to simplify the computation (Rougier (2008), Santner, Williams and Notz (2003)). Alternatives, such as Bayesian methods, suffer from the same difficulty. Furthermore, the resulting plug-in predictors tend to underestimate the uncertainty, because the variance in (2.6) is obtained by substituting the true parameters by their estimators.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

2.2 LHD-based block subsampling

The main way of achieving an efficient computational reduction in GP estimation and prediction is to incorporate a new version of bootstrap subsampling called LHD-based block subsampling. The idea and some empirical performance was first discussed by Liu and Hung (2015). The asymptotic properties for estimation and variable selection using LHD-based block subsampling are studied in Zhao, Amemiya and Hung (2018). Although the empirical results for this approach in different applications have shown promising performance (Liu and Hung (2015), Sun et al. (2019)), there is a lack of a systematic framework in which to construct the predictive distributions, and the corresponding theoretical justifications are not available in the literature.

The idea of bootstrap subsampling is attractive for achieving computational reductions, but direct applications with random subsamples are not efficient in GP estimation and prediction for two reasons. First, it is known in the experimental design literature that the estimation efficiency of simple random sampling can be improved by certain stratification, such as LHDs (McKay, Beckman and Conover (1979)). Second, it is shown by Zhu and Stein (2006) that including clusters of points is important for capturing the local behavior of the process, especially when the parameters are unknown

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

in a GP.

LHD-based block subsampling has the following advantages. First, because of the one-dimensional balance property inherited from LHDs, the subsamples can spread out uniformly over the complete data and, therefore, the resulting subsamples are more representative. Second, estimations and predictions calculated from LHD-based subsamples are expected to outperform those from simple random samples because of the well-developed understanding of variance reduction in LHD compared with that in simple random sampling (McKay, Beckman and Conover (1979)). Third, the clusters of points within the blocks capture the local behavior of the process, and therefore improves the estimation accuracy for correlation parameters, which is essential for GP prediction.

LHD-based subsampling follows three steps.

Step 1: *Denote the d-dimensional input space by* $\Gamma \in [0, l]^d$. *Divide each dimension into* m *equally spaced intervals so that* Γ *consists of* m^d *disjoint hypercubes/blocks. Define each block by mapping* $\boldsymbol{i}$ *to a d-dimensional hypercube*

$$\mathcal{B}_n(\boldsymbol{i}) = \{\boldsymbol{x} \in R^d : bi_j \leq x_j \leq b(i_j + 1) \ \ and \ \ j = 1, ..., d\},$$

where $\boldsymbol{i} = (i_1, ...i_d)$, *for* $i_j \in (0, ..., m-1)$, *represents the index of*

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

each hypercube/block, and $b = l/m$ *is the edge length of the hypercube. Let* $|\mathcal{B}_n(\boldsymbol{i})|$ *be the number of observations in the* $\boldsymbol{i}$*th block. To simplify the notation in the proof, we assume the data points are equally distributed over the blocks and* $|\mathcal{B}_n(\boldsymbol{i})| = n/m^d$*. Theoretically, this assumption can be relaxed to situations where the number of observations in each block is in the same order, that is,* $|\mathcal{B}_n(\boldsymbol{i}_i^*)| = O(n/m^d)$*, and the asymptotic properties developed in Section 3 remain valid. In practice, based on empirical experience, this procedure provides an efficient representation of the original data, as long as each bootstrap subsample does not contain empty hypercubes/blocks.*

Step 2: *Select* m *hypercubes according to a randomly generated* m*-run LHD, in which each column of the design matrix is a random permutation of* $\{0, \ldots, m-1\}$*. Denote the design points by* d*-dimensional vectors* $\boldsymbol{i}_1^*, \ldots, \boldsymbol{i}_m^*$ *and the corresponding selected blocks by* $\mathcal{B}_n(\boldsymbol{i}_1^*), \ldots, \mathcal{B}_n(\boldsymbol{i}_m^*)$*. The bootstrapped subsamples, denoted by* $y_1^*(\boldsymbol{x}_1^*), \ldots, y_N^*(\boldsymbol{x}_N^*)$*, are the observations in the selected blocks, where* $N = \sum_{i=1}^m |\mathcal{B}_n(\boldsymbol{i}_i^*)|$*. Based on the subsamples, the MLEs* $\hat{\boldsymbol{\phi}}_N^*$ *can be obtained from (2.2)–(2.4).*

Step 3: *Repeat the second step* U *times to obtain the bootstrapped MLEs* $\hat{\boldsymbol{\phi}}_{N(1)}^*, \cdots, \hat{\boldsymbol{\phi}}_{N(U)}^*$*. Based on these estimators, the bootstrap pre-*

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

dictive distributions can be constructed using the methods described in Section 2.3.

To illustrate the subsampling idea, we consider a simple example of a six-run two-dimensional LHD on the first panel of Figure 1 in the Supplementary Material. The design points are denoted by $\boldsymbol{i}_1^* = (0, 4)$, $\boldsymbol{i}_2^* = (1, 0)$, $\boldsymbol{i}_3^* = (2, 2)$, $\boldsymbol{i}_4^* = (3, 5)$, $\boldsymbol{i}_5^* = (4, 1)$, and $\boldsymbol{i}_6^* = (5, 3)$. On the second panel, consider $\Gamma \in [0, 24]^2$ with $d = 2$ and $l = 24$. The circles represent the settings in which computer experiments are performed, and the total sample size is $n = 216$. According to the LHD on the left, we have $m = 6$, $b = 4$, and $|\mathcal{B}_n(\boldsymbol{i})| = 6$. The corresponding LHD-based blocks are the six gray boxes in the right panel, and the red dots are the resulting subsamples with size $N = 36$.

Note that applying LHD-based block subsampling reduces the complexity from $O(n^3)$ to $O(n^3/m^{3(d-1)})$, which is particularly useful for high-dimensional problems when d is large. This method also allows parallel computing for large data sets. Note too that this subsampling plan is flexible and can be modified to select subsamples based on a subset of variables, instead of all variables. To do so, we randomly select a subset of variables with dimension $\bar{d}$, where $\bar{d} < d$, and select subsamples only according to the $\bar{d}$ variables. This is typically useful when d is large, because the size

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

for each subsample is n/m^{d-1}, which can be too small to be representative; however, this increases to $n/m^{\bar{d}-1}$ if only a subset of the variables are implemented. The proposed procedure can also be extended to regions with irregular shapes by replacing the LHD in Step 2 with other space-filling designs constructed for nonrectangular regions, such as Draguljić, Dean and Santner (2012) and Hung, Qian and Wu (2012).

2.3 Two construction methods for predictive distribution

To construct a predictive distribution based on the LHD-based subsamples, we developed two bootstrap procedures. One is called the *direct density prediction* method, and the other is called the *normal approximation* method. Both procedures use the LHD-based subsamples to construct predictive distributions; therefore, compared with using the full sample, the computational complexity is reduced. The difference between these two methods is how the normal assumption is imposed. The direct density method imposes the normal assumption on each bootstrap iteration, which leads to the final predictive distribution following normal mixture. On the other hand, the normal approximation method assumes that the final predictive distribution is normal and the mean and variance are estimated using the LHD-based subsamples. The mathematical definitions of the two methods

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

are given as follows.

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 with empirical distribution P^*, and let $\hat{\boldsymbol{\phi}}_N^*$ be the maximizer of the log-likelihood $\ell(\boldsymbol{X}_N^*, \boldsymbol{y}_N^*, \boldsymbol{\phi})$. Then, a bootstrap predictive distribution is defined by

$$g^*(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{y}_n) = \int g(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_N^*, \boldsymbol{y}_N^*, \hat{\boldsymbol{\phi}}_N^*) dP^*(\boldsymbol{X}_N^*, \boldsymbol{y}_N^* \mid \boldsymbol{X}_n, \boldsymbol{y}_n), \quad (2.7)$$

where $g(\cdot)$ is the probability density function of the normal distribution with mean $\mu(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{y}_n, \hat{\phi}_n)$ and variance $\sigma^2(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{y}_n, \hat{\phi}_n)$.

Based on the LHD-based subsamples, a Monte Carlo estimate of (2.7) can be obtained by

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

where $\hat{\phi}_{N(u)}^*$, for $u = 1, ..., U$, are the MLEs obtained from each subsample. The resulting $\tilde{g}^*(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \mathbf{y}_n)$ follows a mixture distribution. When $U \to \infty$, $\tilde{g}^*(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \mathbf{y}_n)$ converges to $g^*(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{y}_n)$.

The conventional predictive distribution discussed in Section 2.1 is normal. Therefore, a reasonable alternative is to assume a normally distributed predictive distribution with mean and variance estimated as follows.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

Definition 2 (Normal approximation)**.** The predictive distribution is normal with mean

$$\mu^*(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{y}_n) = \int \mu(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_N^*, \boldsymbol{y}_N^*, \hat{\boldsymbol{\phi}}_N^*) dP^*(\boldsymbol{X}_N^*, \boldsymbol{y}_N^* \mid \boldsymbol{X}_n, \boldsymbol{y}_n) \quad (2.8)$$

and variance

$$\sigma^{2*}(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{y}_n) = \int \sigma^2(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_N^*, \boldsymbol{y}_N^*, \hat{\boldsymbol{\phi}}_N^*) dP^*(\boldsymbol{X}_N^*, \boldsymbol{y}_N^* \mid \boldsymbol{X}_n, \boldsymbol{y}_n) \quad (2.9)$$

Based on the LHD-based subsamples, the Monte Carlo estimates of (2.8) and (2.9) can be obtained by:

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

and

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

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

3. Theoretical properties and comparisons

In this section, we derive the theoretical properties, including the unbiasedness and the variance of the proposed predictors. The results discussed here focus only on GP prediction, assuming that the estimator $\hat{\boldsymbol{\phi}}_N^*$ converges to the original MLE $\hat{\boldsymbol{\phi}}_n$ in probability, as shown by Zhao, Amemiya

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

and Hung (2018). Note that there are two distinct asymptotics, namely, the fixed-domain (Stein (1999)) and the increasing domain (Cressie (1993), Mardia and Marshall (1984)) asymptotics. However, theoretical results under fixed-domain asymptotics are limited in the literature, owing to its generally complex correlation structure (Ying (1993)). It is shown by Zhang and Zimmerman (2005) that, given their quite different behavior under the two frameworks in a general setting, their approximation quality performs about equally well for the exponential correlation function under certain assumptions. Therefore, we focus here on the increasing domain asymptotics as a fundamental step in providing insights about the bootstrap estimators.

We first construct an asymptotic expansion of the predictive distributions, which is a fundamental tool for the theoretical development of the proposed method. Define the information matrix of the bootstrapped likelihood function evaluated at $\hat{\boldsymbol{\phi}}_n$ by

$$I = E^*\{-\nabla^2_{\boldsymbol{\phi}}\ell(\boldsymbol{X}^*_N, \boldsymbol{y}^*_N, \hat{\boldsymbol{\phi}}_n)\},$$

where I^{si} is the entry in the sth row and ith column of I^{-1}. The third-order derivative of the likelihood function evaluated at $\hat{\boldsymbol{\phi}}_n$ is then defined by

$$K_{ijk} = \frac{1}{2}E^*\{\frac{\partial^3\ell(\boldsymbol{X}^*_N, \boldsymbol{y}^*_N, \hat{\boldsymbol{\phi}}_n)}{\partial\phi_i\partial\phi_j\partial\phi_k}\}.$$

The cross products between the first- and second-order derivatives of the

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

predictive function and the second- and third-order derivatives of the likelihood function evaluated at $\hat{\boldsymbol{\phi}}_n$ are

$$L^j_{s,i}(h) = E^*\{\frac{\partial h(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}^*_N, \boldsymbol{y}^*_N, \hat{\boldsymbol{\phi}}_n)}{\partial \phi_s} \frac{\partial \ell(\boldsymbol{X}^*_N, \boldsymbol{y}^*_N, \hat{\boldsymbol{\phi}}_n)}{\partial \phi_i} \frac{\partial \ell(\boldsymbol{X}^*_N, \boldsymbol{y}^*_N, \hat{\boldsymbol{\phi}}_n)}{\partial \phi_j}\},$$

where $h_1(\boldsymbol{x}_{n+1} \mid ...) = I^{-1}h(\boldsymbol{x}_{n+1} \mid ...)$ and

$$J_{rs,ij}(h) = E^*\{\frac{\partial^2 h(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}^*_N, \boldsymbol{y}^*_N, \hat{\boldsymbol{\phi}}_n)}{\partial \phi_r \phi_s} \frac{\partial \ell(\boldsymbol{X}^*_N, \boldsymbol{y}^*_N, \hat{\boldsymbol{\phi}}_n)}{\partial \phi_i} \frac{\partial \ell(\boldsymbol{X}^*_N, \boldsymbol{y}^*_N, \hat{\boldsymbol{\phi}}_n)}{\partial \phi_j}\},$$

and

$$M_{s,j,ik}(h) = \frac{1}{2}E^*\{\frac{\partial h(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}^*_N, \boldsymbol{y}^*_N, \hat{\boldsymbol{\phi}}_n)}{\partial \phi_s} \frac{\partial \ell(\boldsymbol{X}^*_N, \boldsymbol{y}^*_N, \hat{\boldsymbol{\phi}}_n)}{\partial \phi_j} \frac{\partial^2 \ell(\boldsymbol{X}^*_N, \boldsymbol{y}^*_N, \hat{\boldsymbol{\phi}}_n)}{\partial \phi_i \partial \phi_k}\}.$$

The following theorem provides a third-order asymptotic expansion of the proposed predictive function. To facilitate the presentation, we use Einstein's summation convention hereafter: if an index appears twice in any one term, once as an upper and once as a lower index, summation over the index is applied.

Theorem 1. *Assume I is asymptotically nonsingular and the limit of $I^{-1/2}\nabla^2_{\boldsymbol{\phi}}\ell(\boldsymbol{X}^*_N, \boldsymbol{y}^*_N, \hat{\boldsymbol{\phi}}_n)I^{-1/2}$ is a unit matrix when $N \to \infty$. Then, the LHD-based bootstrap prediction function $h^*(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{Y}_n)$ has the following third-order asymptotic expansion:*

$$\begin{aligned} h^*(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{Y}_n) &= E^*h(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}^*_N, \boldsymbol{y}^*_N, \hat{\boldsymbol{\phi}}_n) + I^{si}I^{jk}M_{s,j,ik} \\ &\quad + \frac{1}{2}I^{ij}K_{irs}L^j_{r,s}(h) + I^{rj}I^{si}J_{rs,ij}(h) + O_{p^*}(N^{-2}). \end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

Owing to the correlation between $\boldsymbol{x}_{n+1}$ and $\boldsymbol{X}_n$, the first term $E^*h(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_N^*, \boldsymbol{y}_N^*, \hat{\boldsymbol{\phi}}_n)$ is not always equal to $h(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{y}_n, \hat{\boldsymbol{\phi}}_n)$. Assuming data independence, an important special case of Theorem 1, which agrees with the result in Fushiki, Komaki, and Aihara (2005) (Theorem 1), is the following.

Corollary 1. *If $\psi(\boldsymbol{x}_1, \boldsymbol{x}_2) = 0$ if $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$, the LHD-based bootstrap prediction function $h^*(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{Y}_n) = h^*(\boldsymbol{x}_{n+1} \mid \hat{\boldsymbol{\phi}}_n)$ has the following third-order asymptotic expansion:*

$$h^*(\boldsymbol{x}_{n+1}) = h(\boldsymbol{x}_{n+1} \mid \hat{\boldsymbol{\phi}}_n) + I^{si} I^{jk} M_{s,j,ik} + \frac{1}{2} I^{ij} K_{irs} L^j_{r,s} + I^{rj} I^{si} J_{rs,ij} + O_{p^*}(N^{-2}).$$

Based on the asymptotic expansion in Theorem 1, we show that the two new predictors are unbiased and their variances can be rewritten as in the next theorem. Denote the predictive mean and variance of the direct density method by $\mu_1^*(\cdot)$ and $\sigma_1^{2*}(\cdot)$, respectively. Similarly, denote these by $\mu_2^*(\cdot)$ and $\sigma_2^{2*}(\cdot)$, respectively, for the normal approximation method. Let $\sum_{\boldsymbol{i}}$ be the summation of all m^d blocks and $\sum_{\boldsymbol{\pi}}$ be the summation of independent permutations over $\{0, 1, \ldots, m-1\}$.

Theorem 2. *Under the regularity conditions given in the Supplementary Material, we have the following:*

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

i. The proposed predictors, μ_1^* and μ_2^*, are unbiased; that is,

$$\boldsymbol{E}\{\mu(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{y}_n, \hat{\phi}_n) - \mu_1^*\} = \boldsymbol{E}\{\mu(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{y}_n, \hat{\phi}_n) - \mu_2^*\} \quad \to 0$$

ii. The predictive variances have the following relationship:

$$P(\sigma_1^{2*} \geq \sigma_2^{2*}) \to 1,$$

$$\begin{aligned}
\sigma_1^{2*} &= \sigma_2^{2*} + \frac{1}{(m!)^{d-1}} \sum_{\boldsymbol{\pi}} [\mu(\boldsymbol{x}_{n+1} | \boldsymbol{X}_N^*, \boldsymbol{y}_N^*, \hat{\phi}_n) - \mu(\boldsymbol{x}_{n+1} | \boldsymbol{X}_n, \boldsymbol{y}_n, \hat{\phi}_n)]^2 \\
&\quad + o_p(1), \\
\sigma_2^{2*} &= \hat{\sigma}_n^2 \Big\{ 1 - \frac{1}{m^{d-1}} \sum_{\boldsymbol{i}} \gamma_{\boldsymbol{i}}(\hat{\boldsymbol{\theta}}_n)^T R_{\boldsymbol{i},\boldsymbol{i}}^{-1} \gamma_{\boldsymbol{i}}(\hat{\boldsymbol{\theta}}_n) \Big\} + o_p(1).
\end{aligned}$$

The next theorem compares of the predictive variance of the plug-in predictive distribution defined in (2.6) with those of the two new predictors.

Theorem 3. *Under the regularity assumptions given in the Supplementary Material, we have*

$$P(\sigma_1^{2*} \geq \sigma^2(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{y}_n, \hat{\phi}_n)) \to 1,$$

$$P(\sigma_2^{1*} \geq \sigma^2(\boldsymbol{x}_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{y}_n, \hat{\phi}_n)) \to 1.$$

It is known that the regular plug-in predictor interpolates the observed data. The next theorem shows that although this interpolation property cannot be guaranteed by the proposed predictors, the predictive variance on

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

an existing data point is smaller than the variance on an untried point. For the direct density approach, denote the variance within the sampled data by $\sigma_1^{2*(I)}$, and the variance for the out-of-sample data by $\sigma_1^{2*(O)}$. Similarly, we have $\sigma_2^{2*(I)}$ and $\sigma_2^{2*(O)}$, respectively, for the normal approximation method.

Theorem 4. *Under the regularity assumptions given in the Supplementary Material, we have:*

(i). The in-sample predictive variances are

$$\begin{aligned}\sigma_1^{2*(I)} &= (1-\frac{1}{m^{d-1}})\frac{1}{(m!)^{d-1}}\sum_{\boldsymbol{\pi}}[\mu(\boldsymbol{x}_{n+1}|\boldsymbol{X}_N^*,\boldsymbol{y}_N^*,\hat{\phi}_n)-\mu(\boldsymbol{x}_{n+1}|\boldsymbol{X}_n,\boldsymbol{y}_n,\hat{\phi}_n)]^2\\ &\quad +\sigma_2^{2*(I)}+o_p(1),\\ \sigma_2^{2*(I)} &= (1-\frac{1}{m^{d-1}})\hat{\sigma}_n^2[1-\frac{1}{m^{d-1}}\sum_{\boldsymbol{i}}\gamma_{n,\boldsymbol{i}}(\hat{\boldsymbol{\theta}}_n)^T R_{\boldsymbol{i},\boldsymbol{i}}^{-1}\gamma_{n,\boldsymbol{i}}(\hat{\boldsymbol{\theta}}_n)]+o_p(1).\end{aligned}$$

(ii). Comparison of the in-sample and out-of-sample predictive variance:

$$\begin{aligned}\sigma_1^{2*(O)}-\sigma_1^{2*(I)} &= \sigma_2^{2*(O)}-\sigma_2^{2*(I)}+o_p(1)\\ &\quad +(mm!)^{1-d}\sum_{\boldsymbol{\pi}}[\mu(\boldsymbol{x}_{n+1}|\boldsymbol{X}_N^*,\boldsymbol{y}_N^*,\hat{\phi}_n)-\mu(\boldsymbol{x}_{n+1}|\boldsymbol{X}_n,\boldsymbol{y}_n,\hat{\phi}_n)]^2,\\ \sigma_2^{2*(O)}-\sigma_2^{2*(I)} &= \frac{\hat{\sigma}_n^2}{m^{d-1}}[1-\frac{1}{m^{d-1}}\sum_{\boldsymbol{i}}\gamma_{\boldsymbol{i}}(\hat{\boldsymbol{\theta}}_n)^T R_{\boldsymbol{i},\boldsymbol{i}}^{-1}\gamma_{\boldsymbol{i}}(\hat{\boldsymbol{\theta}}_n)]+o_p(1)\end{aligned}$$

i.e.

$$P(\sigma_1^{2*(O)}\geq\sigma_1^{2*(I)})\to 1,\quad P(\sigma_2^{2*(O)}\geq\sigma_2^{2*(I)})\to 1.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

For Theorem 4, although the proposed predictors do not have the interpolation property, their in-sample predictive variances are, in general, smaller than their out-of-sample variances.

4. Simulation studies

The objective of this section is to demonstrate the finite-sample performance of the proposed method. This performance is compared with that of some existing methods, including the regular GP model, plug-in approach, and conventional bootstrap prediction. All simulations are conducted on a 2.4 GHz Intel Core i5, 8GB 1600 MHz DDR3 workstation under Python 3.5.2 running on MAC OS X.

4.1 Comparisons with regular MLE

The finite-sample performance of the proposed method is compared with that of the regular MLE using full data, denoted by "ALLData." Three settings of LHD-based block bootstrap are employed: $m = 3, 4$, and 5. The outputs are simulated from a GP with mean function coefficients $\boldsymbol{\beta} = (0, 2, -2, 1)$ and correlation function

$$\psi(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp(-\sum_{i=1}^{3} |x_{1i} - x_{2i}|/\theta_i),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

where $\theta_1 = \theta_2 = \theta_3 = 0.4$ and $\sigma = 1$. Two sample sizes, $n = 2000$ and 4000, are considered, and the design points are generated from a regular grid over the region $[0, 1]^3$. For each sample size, 50 training samples and 100 testing samples are generated. The performance of the parameter estimation is summarized in Table 1 in Appendix E of the Supplementary Material based on 100 replicates with 10 LHD-based block bootstrap samples implemented for the proposed method. In addition, the mean squared prediction errors (MSPEs) for the testing data sets and the average computing time are both reported.

The results demonstrate that the estimated parameters using the LHD-based block bootstrap are, in general, consistent with those obtained using the complete data. When $n = 2000$, the standard deviations increase with the number of blocks m, especially for the correlation parameters. This is not surprising, because the sample sizes are smaller for larger m and "ALLData" implies the special case of "m=1." The impact of m on the estimation variance appears to be smaller when the sample size increases to $n = 4000$. In terms of computing time, the LHD-based block bootstrap is much faster than the conventional GP modeling, especially for large n.

Note that the proposed method is particularly useful for data collected from irregular grids. The reason for generating the simulations from a reg-

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

ular grid is that the MLE calculation using full data, under this setting, can be simplified further using Kronecker product techniques, and some matrix singularity can be avoided (Rougier (2008)). However, these techniques are not applicable to data sets collected from an irregular grid. Therefore, the computational advantage of the proposed method is expected to be even more significant for data collected from irregular grids.

4.2 Comparisons of prediction variance

We compare the proposed predictive distributions with existing methods by looking at their predictive variance. Two existing methods, the regular bootstrap and the plug-in predictive distribution, are considered. The regular bootstrap, although computationally expensive, can serve as a benchmark for capturing the true prediction uncertainty. Simulations are generated from the same model given in Section 4.1. Owing to computational constraints in the regular bootstrap, we use relatively smaller sample sizes, $n = 1000$ and $n = 2000$, for the comparison. Both the LHD-based subsampling and the regular bootstrap are performed using the two construction approaches, direct density and normal approximation. The predictive variance is evaluated based on 100 untried settings with 50 replications.

The performance on predictive variance is summarized in Table 2 in

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

Appendix E of the Supplementary Material. LHD-based subsampling is denoted by "LHD." In general, using LHD subsampling, the predictive variance constructed using direct density is larger than that using the normal approximation, which is consistent with the theoretical results in Theorem 2. It is also not surprising to see that the predictive variances obtained from LHD subsampling are larger than the benchmark results from the regular bootstrap, and the differences become smaller when the sample size increases. On the other hand, the plug-in approach offers the smallest prediction variances for the two sample sizes. The proposed methods provide a significant computational reduction compared with the regular bootstrap approach and the required computing time is even smaller than the plug-in approach, especially when sample sizes increase. This result suggests that, by using LHD-based subsampling, the proposed predictive distributions offer computationally efficient analogues of the conventional bootstrap methods.

4.3 Comparisons of predictive covering by the borehole function

This numerical study compares the predictive coverage probabilities obtained by the two proposed construction methods for predictive distributions with that of the plug-in approach. Data are generated from the bore-

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

hole function, which is a benchmark example commonly used in the computer experiment literature (Morris, Mitchell and Ylvisaker (1993)). The LHD block procedure is implemented with $m = 2$ and $U = 10$. The comparisons are performed based on 2000 training data and 2000 testing data with 100 replications. The empirical performance of the predictive coverage for the 95% and 90% confidence intervals is illustrated in Table 1. In general, the two proposed construction methods provide empirical coverage probabilities closer to the nominal levels than the plug-in approach does. The plug-in approach tends to have a much smaller empirical coverage, which may be due to the underestimation of the prediction uncertainty in the construction of the confidence intervals. The confidence intervals constructed using the direct density approach appear to produce empirical coverage probabilities larger than those of the normal approximation approach. This is expected, because the direct density approach does not rely on the normal assumption, and therefore has a larger confidence interval, which leads to a larger empirical coverage than that of the normal approximation.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

Table 1: Comparisons of predictive coverage probabilities.

	95%CI	90%CI
LHD (Direct Density)	99.88%	99.55%
LHD (Normal Approximation)	97.12%	94.59%
Plug-in	78.97%	70.97%

4.4 Comparisons of prediction accuracy

In this section, we compare the prediction performance of the proposed method with a computationally efficient approximation of a GP using local GP models (Gramacy and Apley (2015)). The data are generated from a similar GP model to that in Section 4.1, with $n = 1500$ and a slightly larger dimension $p = 4$. The mean function coefficients are set to (0, 4,-4,-1,-3), and the exponential covariance function is assumed with parameters $\boldsymbol{\theta} = (0.3, 1.6, 3, 2)$ and σ =1. The proposed method is implemented using $m = 3$ and $U = 20$. The local GP is implemented using the *laGP* package in R, with the initial number of nearest neighbors set to six, the total size of the local designs set to 100, and the default settings in prediction minimizing the predictive variance. The prediction performance is evaluated based on untried testing data with sample size 1500.

Based on 100 replicates, the prediction performance is demonstrated

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

by box plots in Figure 1, based on two measurements, namely, the RMSPE and the Mahalanobis distance to the underlying truth Bastos and O'Hagan (2009). Note that the two proposed approaches produce the same predictive mean, and thus have the same RMSPE. However, they have different prediction variance. Therefore, the corresponding Mahalanobis distances are not necessarily the same. According to Figure 1, the two proposed methods outperform *laGP* by producing smaller RMSPEs and smaller Mahalanobis distances to the true responses.

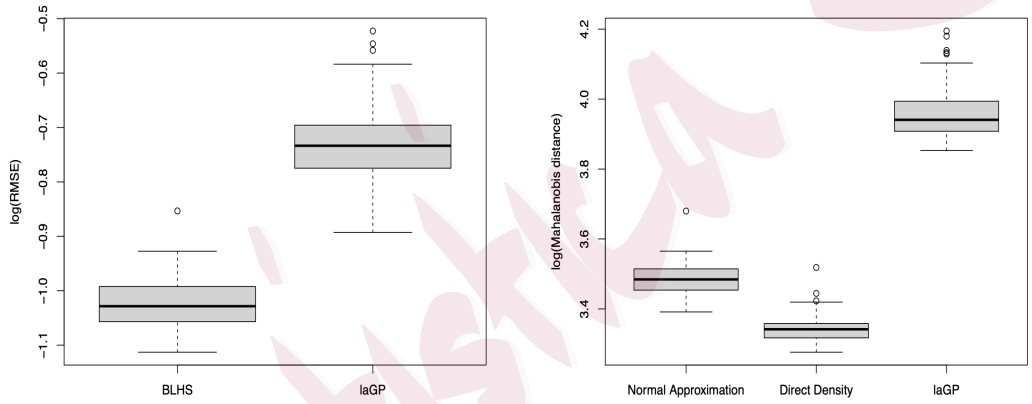

Figure 1: Prediction comparisons with laGP. The left panel is based on RMSPEs, and the right panel is based on the Mahalanobis distances.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

5. Real Examples

5.1 A data center thermal management example

A data center is a computing infrastructure facility that houses large amounts of information technology equipment used to process, store, and transmit digital information. Data center facilities constantly generate large amounts of heat in the room, which must be maintained at an acceptable temperature for reliable operation of the equipment. A significant fraction of the total power consumption in a data center is for heat removal; therefore, determining the most efficient cooling mechanism has become a major challenge. To solve the problem, a crucial step is to model the thermal distribution at different experimental settings (Hung, Qian and Wu (2012)).

For a data center thermal study, physical experiments are not always feasible, because some settings are highly dangerous and expensive. Therefore, simulations based on computational fluid dynamics (CFD) are widely used. In this example, CFD simulations are conducted at the IBM T. J. Watson Research Center based on a real data center layout. Detailed discussions about the CFD simulations can be found in (Lopez and Hamann (2011)). The first three columns in Table 3 in Appendix E of the Supplementary Material list nine variables and their levels in the CFD simulations,

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

including four computer room air conditioning (CRAC) units with different flow rates $(x_1, ..., x_4)$, the overall room temperature setting (x_5), the perforated floor tiles with different percentages of open areas (x_6), and the spatial location in the data center (x_7 to x_9). There are 27,000 temperatures simulated from the CFD simulator, obtained from an irregular grid over the nine-dimensional experimental space.

It is computationally intensive to build a GP model based on the complete CFD data. Therefore, we implement the proposed LHD-based block bootstrap approach with $m = 3$ for variables x_6, x_7, and x_9, which are the top three factors with the highest levels. The fitted GP model is summarized in the last two columns of Table 3 in Appendix E of the Supplementary Material, where $\hat{\beta}$ represents the estimated mean function coefficients, and $\hat{\theta}$ represents the correlation parameters estimated based on the exponential covariance function. From the fitted model, it appears that the height (x_9) in a data center has a relatively larger effect, particularly on the mean function. Furthermore, we find that the temperatures increase dramatically with height, based on the predicted heat map at three different heights (Figure 2 in the Supplementary Material) with an untried setting (i.e., CRAC unit 1 flow rate 6500, unit 2 flow rate 6500, unit 3 flow rate 2750, unit 4 flow rate 2750, room temperature 70 (F), and tile percentage 59). These

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

findings can be validated by a general understanding of thermodynamics.

5.2 Ice sheet thickness modeling

The second application examines ice sheet thickness using the community ice sheet model (CISM; Rutt et al. (2009)). The main objective of this model is to understand ice sheet behavior and its impact on climate. The CISM mimics the effects of past climate on the current ice sheet state by considering a model of an idealized ice sheet over a rectangular region that is flowing out to sea on one side, while accumulating ice from prescribed precipitation over a period of 1000 years. There are two control variables in the CISM, namely, a constant term in the Glen–Nye flow law (Greve and Blatter (2009)) controlling the deformation of the ice sheet, denoted by x_1, and the heat conductivity in the ice sheet, denoted by x_2. The simulated thickness is produced on a 27×32 rectangular lattice of the spatial locations, denoted by x_3 and x_4. We focus on the central part of the icebergs by taking the middle 13×16 rectangular lattice in this analysis. A set of simulations with 20 combinations of the two control variables is considered; therefore, the total sample size is $n = 4160$. The detailed variable settings can be found in Higdon, Mitra and Johnson (2013).

The study compares the performance of the proposed method with that

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

of a conventional GP using full data in real applications. A four-dimensional GP is considered for the analysis of the simulation results from the CISM. The estimation and prediction performance is evaluated based on a 10-fold cross-validation. The LHD-based subsamples are obtained using the setting ($m = 3$, $U = 10$), and each LHD-based subsample has size $N = 139$. The results are summarized in Table 2, with the conventional GP denoted by "Alldata" and the proposed method denoted by "LHD." RMSPE is the root mean squared prediction error calculated from the 10-fold cross-validation.

From the results in Table 2, it appears that even with only 3.7%($\approx$ 1/27) of the data in each subsample, the LHD-based approach provides a reasonable performance in terms of parameter estimation and prediction. The computational time is reduced by more than 99.3% using the proposed method. In general, the estimation for x_3 seems to be more challenging than for the other variables, owing to its relatively smaller effect. One example of the iceberg thickness prediction is demonstrated in Figure 3 of the Supplementary Material over the entire spatial location with the parameter setting $x_1 = 2.40$ and $x_2 = 6.53 \times 10^4$. The left panel is the original simulation outputs from the CISM. The middle panel is the plug-in prediction using the full data. The right panel is the prediction obtained from the LHD-based approach. It shows that, given some roughness owing

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

Table 2: LHD bootstrap analysis of CISM data

		x_1	x_2	x_3	x_4	RMSPE	Time (Sec)
Alldata	$\hat{\beta}$	-0.80(4.0×10^{-3})	0.31(6.0×10^{-3})	7.3×10^{-5}(6.0×10^{-4})	0.11(4.6×10^{-4})	0.01	10097.39
	$\hat{\theta}$	7.26(11.00)	3.41(8.15)	0.03(9.5×10^{-3})	0.65(0.21)		
LHD	$\hat{\beta}$	-0.84(0.08)	0.42(0.05)	-1.2 $\times 10^{-4}$(2.2×10^{-3})	0.08(0.01)	0.07	66.36
	$\hat{\theta}$	23.71(19.56)	3.82(4.00)	3.4×10^{-3}(1.1×10^{-3})	0.21(0.58)		

to the small subsample size, the prediction using the LHD-based approach efficiently captures the underlying structure.

6. Conclusion

We present an LHD-based block subsampling procedure with two prediction methods to tackle the computational difficulties and uncertainty quantification issues in GP prediction. The new procedure borrows the strength of space-filling designs to provide an efficient subsampling plan and a reduction in computational complexity. Theoretical properties of the proposed predictive distributions are discussed. The proposed procedure is applied to two complex computer experiments with high-dimensional inputs and massive outputs.

The following areas offer potential for future work. First, extensions of

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

the proposed procedure to optimal designs with better space-filling properties are intuitively appealing. For example, it is known that randomly generated LHDs can contain some structure. To further enhance the desirable space-filling properties, various modifications are proposed. Numerical comparisons and theoretical developments of the generalization to different types of optimal space-filling designs should be studied carefully. Second, an interesting and important issue with the LHD-based block bootstrap is to determine the optimal block size. This topic has been discussed for conventional block bootstrap methods (Hall, Horowitz and Jing (1995), Lahiri (1999), Nordman, Lahiri and Fridley (2007)). However, their solutions are not directly applicable to GP models. We plan to study the optimal block size for the propose procedure based on some new criteria defined for a GP.

Supplementary Material

The online Supplementary Material contains the proofs of Theorem 1 to Theorem 4, as well as the figures and tables.

Acknowledgments

The authors gratefully acknowledge the constructive advice from the associate editor and the referee. This research was supported by NSF grants.

References

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

References

Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B*, **70**, 825–848.

Bastos, L. and O'Hagan, A. (2009). Diagnostics for Gaussian Process Emulators. *Technometrics*, **51**, 4, 425-438.

Castrillon, J. E., Genton, M. G., and Yokota, R. (2015). Multi-level restricted maximum likelihood covariance estimation and kriging for large non-gridded spatial datasets. *Spatial Statistics*, **18**, 105–124.

Chang, W., Haran, M., Olson, R., and Keller, K. (2014). Fast dimension-reduced climate model calibration and the effect of data aggregation. *Annals of Applied Statistics*, **8**, 649–673.

Cressie, N. (1993). *Statistics for Spatial Data*, Wiley, New York.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B*, **70**, 209–226.

DiCiccio, T. J. and Efron, B. (1992). More accurate confidence intervals in exponential families. *Biometrika*, **79**, 231–245.

DiCiccio, T. J. and Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, **11**, 189–228.

Draguljić, D., Dean, A. M., and Santner, T. J. (2012). Noncollapsing space-filling designs for

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

bounded nonrectangular regions. *Technometrics*, **54**, 169–178.

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap*, Chapman and Hall/CRC press, New York.

Fang, K.-T., Li, R. and Sudjianto, A. (2006). *Design and modeling for computer experiments*, Chapman and Hall/CRC press, New York.

Fuentes, M. (2007). Approximate likelihood for large irregularly spaced spatial data. *Journal of the American Statistical Association*, **102**, 321–331.

Furrer, R., Genton,M. G. and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, **15**, 502–523.

Fushiki, T., Komaki, F., and Aihara, K. (2005). Nonparametric bootstrap prediction, *Bernoulli*, **11**, 293–307.

Gramacy R. B. and Apley D. W. (2015). Local Gaussian process approximation for large computer experiments. *Journal of Computational and Graphical Statistics*, **24:2**, 561–578.

Gramacy, R. B. and Lee, H. K. (2008). Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, **103**, 1119–1130.

Greve, R. and Blatter, H. (2009). *Dynamics of ice sheets and glaciers*. Springer, Berlin.

Hall, P., Horowitz, J. L. and Jing, B.-Y. (1995). On blocking rules for the bootstrap with dependent data. *Biometrika*, **82**, 561–574.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

Handcock, M. S. and Stein, M. L. (1993). A Bayesian Analysis of Kriging. *Technometrics*, **35(4)**: 403-410.

Higdon, C.W., Mitra R.D. and Johnson, S.L. (2013). Gene Expression Analysis of Zebrafish Melanocytes, Iridophores, and Retinal Pigmented Epithelium Reveals Indicators of Biological Function and Developmental Origin. *PLoS ONE*, **8(7)**: e67801.

Hung, Y., Qian, P. Z. G., and Wu, C. F. J. (2012). Statistical design and analysis methods for data center thermal management. In *Energy efficient thermal management of data centers* (J. Yogendra and K. Pramod eds.), Springer, New York.

Irvine, K. M., Gitelman, A. I. and Hoeting, J. A. (2007). Spatial designs and properties of spatial correlation: Effects on covariance estimation. *Journal of Agricultural, Biological, and Environmental Statistics*, **12**, 450–469.

Kaufman, C. G., Schervish, M. J. and Nychka, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, **103**, 1545–1555.

Kaufman, C. G., Bingham, D., Habib, S., Heitmann, K. and Frieman, J. A (2011). Efficient emulators of computer experiments using compactly supported correlation functions, with an application to cosmology. *The Annals of Applied Statistics*, **5**, 2470–2492.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**: 425–464.

Lahiri, S. N. (1999). Theoretical comparisons of block bootstrap methods. *The Annals of Statis-*

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

tics, **27**, 386–404.

Li, R. and Sudjianto, A. (2005). Analysis of computer experiments using penalized likelihood in Gaussian kriging models *Technometrics*, **47**, 111–120.

Liang, F., Cheng, Y., Song, Q., Park, J. and Yang, P. (2013). A resampling-based stochastic approximation method for analysis of large geostatistical data. *Journal of the American Statistical Association*, **108**, 325–339.

Liu, Y. and Hung, Y. (2015). Latin Hypercube Design-based Block Bootstrap for Computer Experiment Modeling. *Tech. rep., Rutgers, The State University of New Jersey, New Brunswick, New Jersey.*

Lopez V. and Hamann, H. F. (2011). Heat transfer modeling in data centers. *International Journal of Heat and Mass Transfer*, **54**, 5306–5318.

Luna, S. S. and Young, A. (2003). The bootstrap and kriging prediction intervals. *the Scandinavian Journal of Statistics*, **30**, 175–192.

Mak, S. and Joseph, V. R. (2018). Support points. *The Annals of Statistics*, **46(6A)**:2562-2592.

Mardia, K.V. and Marshall, R. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, **71**, 135–146.

McKay, M. D., Beckman, R. J. and Conover, W. J. (1979). Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, **21**, 239–245.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

Morris, D., Mitchell, T., and Ylvisaker, D. (1993). Bayesian Design and Analysis of Computer Experiments: Use of Derivatives in Surface Prediction. *Technometrics*, **35**, 243–255.

Nordman, D. J., Lahiri, S. N. and Fridley, B. L. (2007). Optimal block size for variance estimation by a spatial block bootstrap method. *Sankhyā*, **69**, 468–493.

Nychka, D. W. (2000). Spatial-process estimates as smoothers. *Smoothing and regression: approaches, computation, and application*, (M. G. Schimek ed.). 393–424, Wiley, New York.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2015). A multiresolution Gaussian process model for the analysis of large spatial data sets. *Journal of Computational and Graphical Statistics*, **24(2)**, 579–599.

Rougier, J. (2008). Efficient emulators for multivariate deterministic functions. *Journal of Computational and Graphical Statistics*, **17**, 827–843.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*, Chapman and Hall/CRC Press, Boca Raton.

Rutt, I. C., M. Hagdorn, N. R. J. Hulton, and A. J. Payne (2009). The Glimmer community ice sheet model, *J. Geophys. Res.*, **114**, F02004.

Santner, T. J., Williams, B. J. and Notz, W. (2003). *The design and analysis of computer experiments*, Springer, New York.

Schmidt, A. M. and O'Hagan, A. (2003). Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society: Series B*

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

(Statistical Methodology), **65**: 743–758.

Snelson, E. and Ghahramani, Z. (2006). Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems*, **18**, 1257–1264.

Stein, M. L. (1999). *Interpolation of spatial data: some theory for kriging*, Springer, New York.

Stein, M. L. (2013). Statistical properties of covariance tapers. *Journal of Computational and Graphical Statistics*, **22**, 866–885.

Stein, M. L., Chi, Z. and Welty, L. J. (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B*, **66**, 275–296.

Sun, F., Gramacy, R. B., Haaland, B., Lawrence, E. and Walker, A. (2019). Emulating Satellite Drag from Large Simulation Experiments. *SIAM/ASA Journal on Uncertainty Quantification* **7:2**, 720-759.

Wang, H., Yang, M., and Stufken, J. (2019). Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association*, **114(525)**:393-405.

Wikle, C. K. (2010). Low-rank representations for spatial processes. In *Handbook of Spatial Statistics* (A. E. Gelfand, P. Diggle, M. Fuentes and P. Guttorp eds.), 107–118, Chapman and Hall/CRC Press, Boca Raton.

Ying, Z.-L. (1993). Maximum likelihood estimation of parameters under a spatial sampling scheme. *The Annals of Statistics*, **21**, 1567–1590.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0376

Zhang, H. and Zimmerman, D. L. (2005). Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika*, **92**, 921–936.

Zhao, Y., Amemiya, Y. and Hung, Y. (2018). Efficient Gaussian Process modeling using experimental design-based subagging. *Statistica Sinica*, **28**, 1459–1479.

Zhu, Z. and Stein, M.L. (2006). Spatial sampling design for prediction with estimated parameters. *Journal of Agricultural, Biological, and Environmental Statistics*, 11: 24.

Department of Statistics, Rutgers University

E-mail: linglin.he@rutgers.edu

Department of Statistics, Rutgers University

Corresponding author: E-mail: yhung@stat.rutgers.edu