Statistica Sinica Preprint No: SS-2016-0250R2

Title	Variable Selection for Gaussian Process Models using Experimental Design-Based Subagging
Manuscript ID	SS-2016.0250
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0250
Complete List of Authors	Ying Hung and Yibo Zhao
Corresponding Author	Ying Hung
E-mail	yhung@stat.rutgers.edu, hungying22@gmail.com
Notice: Accepted version subject to English editing.	

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

Efficient Gaussian Process Modeling using Experimental Design-Based Subagging

Yibo Zhao, †Yasuo Amemiya, and [1]Ying Hung

Department of Statistics and Biostatistics, Rutgers University

†Statistical Analysis & Forecasting, IBM T. J. Watson Research Center

Abstract

We address two important issues in Gaussian process (GP) modeling. One is how to reduce the computational complexity in GP modeling and the other is how to simultaneous perform variable selection and estimation for the mean function of GP models. Estimation is computationally intensive for GP models because it heavily involves manipulations of an n-by-n correlation matrix, where n is the sample size. Conventional penalized likelihood approaches are widely used for variable selection. However the computational cost of the penalized likelihood estimation (PMLE) or the corresponding one-step sparse estimation (OSE) can be prohibitively high as the sample size becomes large, especially for GP models. To address both issues, this article proposes an efficient subsample aggregating (subagging) approach with an experimental design-based subsampling scheme. The proposed method is computationally cheaper, yet it can be shown that the resulting subagging estimators achieve the same efficiency as the original PMLE and OSE asymptotically. The finite-sample performance is examined through simulation studies. Application of the proposed methodology to a data center thermal study reveals some interesting information, including identifying an efficient cooling mechanism.

Keywords: Bagging, Computer experiment, Experimental design, Gaussian process, Latin hypercube design, Model selection

1 Introduction

Gaussian process (GP) models, also known as kriging models, are widely used in many fields, including geostatistics (Cressie 1993, Stein 1999), machine learning (Smola and Bartlett 2001,

[1]Corresponding author. E-mail: *yhung@stat.rutgers.edu*

Snelson and Ghahramani 2006), and computer experiment modeling (Santner et al. 2003, Fang et al. 2006). In this article, we focus on two important issues in GP modeling. One is the study of simultaneous variable selection and estimation of GP models, for the mean function in particular, and the other is how to alleviate the computational complexity in GP modeling.

Various examples of variable selection in GP models can be found in the literature, such as in geostatistics (Hoeting et al. 2006, Huang and Chen 2007, Chu et al. 2011) and computer experiments (Welch et al. 1992, Linkletter et al. 2006, Joseph et al. 2008, Kaufman et al. 2011). In this article, we mainly focus on identifying active effects through the mean function. This is because it is reported in several empirical studies that by a proper selection of important variables in the mean function, the prediction accuracy of GP models can be significantly improved, especially when there are some strong trends (Joseph et al. 2008, Hung 2011, Kaufman et al. 2011). Furthermore, comparing with nonlinear effects identified from the covariance function (Linkletter et al. 2006), linear effects are relatively easier to interpret and of scientific interest in many applications. Conventional approaches based on penalized likelihood functions, such as the penalized likelihood estimators (PMLEs) and the corresponding one-step sparse estimators (OSEs), are conceptually attractive. However, they are computationally difficult in practice, especially with massive data observed on irregular grid. This is because estimation and variable selection heavily involve manipulations of an $n \times n$ correlation matrix that require $O(n^3)$ computations, where n is the sample size. The calculation is computationally intensive and often intractable for massive data.

The computational issue is well recognized in the literature and various methods are proposed. The proposed approaches may be characterized broadly as either changing the model to one that is computationally convenient or approximating the likelihood for the original data. Examples of the former includes Rue and Tjelmeland (2002), Rue and Held (2005), Cressie and Johannesson (2008), Banerjee et al. (2008), Gramacy and Lee (2008), Wikle (2010); while approximation approaches includes Nychka (2000), Smola and Bartlett (2001), Nychka et al. (2002), Stein et al. (2004), Furrer et al. (2006), Snelson and Ghahramani (2006), Fuentes (2007), Kaufman et al. (2008), Gramacy and Apley (2015). However, these methods focus mainly on estimation and prediction but not variable selection, and most of them are developed for datasets collected from a regular grid under a low-dimensional setting. Recent studies address the issues by imposing a sparsity constraint on the correlation matrix, including covariance tapering and compactly support correlation functions (Kaufman et al. 2008, 2011, Chu et al. 2011, Nychka et al. 2015). However, it has been shown that this method does not work well for purposes of parameter estimation (Stein 2013, Liang et al. 2013), which is crucial in selecting important variables. In addition, the connection between the degree of sparsity and computation time is nontrivial.

In this paper, we provide an alternative framework which alleviates the computational difficulties in estimation and variable selection by utilizing the idea of subsample aggregating, also known as subagging (Büchlmann and Yu 2002). This framework includes a subagging estimator and a new subsampling scheme based on a special class of experimental designs called Latin hypercube designs (LHDs), which is known to have a one-dimensional projection property. By borrowing the inherited one-dimensional projection property of LHDs and a block structure, the new subsampling scheme not only provides an efficient data reduction but also takes into account the spatial dependency in GP models. The computational complexity of the proposed subagging estimation is dramatically reduced, yet in the mean time, it is shown that the subagging estimators achieve the same efficiency as the original PMLE and OSE asymptotically.

The remainder of the paper is organized as follows. In Section 2, the conventional penalized likelihood approach is discussed. The new variable selection framework, including the new subsampling scheme and the subagging estimators are introduced in Section 3. Theoretical properties are derived in Section 4. In Section 5, finite-sample performance of the proposed framework is investigated in simulation studies. A data center example is illustrated in Section 6. Discussions are given in Section 7.

2 Variable selection in Gaussian process models

For a domain of interest Γ in R^d, we consider a Gaussian process $\{Y(\boldsymbol{x}) : \boldsymbol{x} \in R^d\}$ such that

$$Y(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta} + Z(\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{\beta}$ is a vector of unknown mean function coefficients and $Z(\boldsymbol{x})$ is a stationary Gaussian process with mean 0 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 $\psi(\boldsymbol{0};\boldsymbol{\theta}) = 1$ and $\psi(\boldsymbol{h};\boldsymbol{\theta}) = \psi(-\boldsymbol{h};\boldsymbol{\theta})$.

Suppose n observations are collected 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 $\boldsymbol{y}_n = (y_1, \ldots, y_n)^T$, $\boldsymbol{X}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$, $\boldsymbol{\phi} = (\boldsymbol{\theta}^T, \boldsymbol{\beta}^T, \sigma^2)^T$ be the vector of all the parameters, and Θ be the parameter space. Based on (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 $R_n(\boldsymbol{\theta})$ is an $n \times n$ correlation matrix with elements $\psi(\boldsymbol{x}_i - \boldsymbol{x}_j;\boldsymbol{\theta})]$. Thus, the log-likehood function, ignoring a constant, is

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

where $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and σ are the unknown parameters.

To achieve simultaneous variable selection and parameter estimation, we focus on penalized likelihood approaches, which are becoming increasingly popular in recent years. A penalized log-likelihood function for GP models can be written as

$$\ell_p(\boldsymbol{y}_n, \boldsymbol{X}_n, \boldsymbol{\phi}) = \ell(\boldsymbol{y}_n, \boldsymbol{X}_n, \boldsymbol{\phi}) - n\sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{3}$$

where $p_\lambda(\cdot)$ is a pre-specified penalty function with a tuning parameter λ. There are various choices of penalty functions such as LASSO (Donoho and Johnstone 1994, Tibshirani 1996), the adaptive LASSO (Zou 2006), and the minimax concave penalty (Zhang 2010). In this article, we focus on the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li 2001) defined by

$$p_\lambda(|\beta|) = \begin{cases} \lambda|\beta| & if|\beta| > \lambda, \\ \lambda^2 + (a-1)^{-1}(a\lambda|\beta| - \beta^2/2 - a\lambda^2 + \lambda^2/2) & if \lambda < |\beta| \leq a\lambda, \\ (a+1)\lambda^2/2 & if|\beta| > a\lambda, \end{cases}$$

for some $a > 2$. By maximizing (3), the penalized maximum likelihood estimators (PLMEs) of $\boldsymbol{\phi}$ can be obtained by $\hat{\boldsymbol{\phi}}_n = \arg\max_{\boldsymbol{\phi}} \ell_p(\boldsymbol{y}_n, \boldsymbol{X}_n, \boldsymbol{\phi})$.

To compute PMLEs under the SCAD penalty, Zou and Li (2008) develop a unified algorithm to improve computational efficiency by locally linear approximation (LLA) of the penalty function. They propose an one-step LLA estimation that approximates the solution after just one iteration in a Newton-Raphson-type algorithm starting at the maximum likelihood estimates (MLEs). Chu et al. (2011) extend the one-step LLA estimation to approximate the PMLEs for the spatial linear models and the resulting estimate is called the one-step sparse estimate (OSE).

Following the idea of Chu et al. (2011), the OSE of $\boldsymbol{\beta}$ in GP models, denoted by $\hat{\boldsymbol{\beta}}_{OSE}$, is obtained by maximizing

$$Q(\boldsymbol{\beta}) = -\frac{1}{2\hat{\sigma^2}^{(0)}}(\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta})^T R_n^{-1}(\hat{\boldsymbol{\theta}}^{(0)})(\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta}) - n\sum_{j=1}^{p} p'_\lambda(|\hat{\beta}_j^{(0)}|)|\beta_j|, \tag{4}$$

where $\hat{\boldsymbol{\beta}}^{(0)}$, $\hat{\boldsymbol{\theta}}^{(0)}$ and $\hat{\sigma^2}^{(0)}$ are the MLEs obtained from (2). We also update $\boldsymbol{\theta}$ and σ^2 by maximizing (4) evaluated at $\hat{\boldsymbol{\beta}}_{OSE}$ with respect to $\boldsymbol{\theta}$ and σ^2. The resulting OSE of $\boldsymbol{\theta}$ and σ^2 is denoted as $\hat{\boldsymbol{\theta}}_{OSE}$ and $\hat{\sigma}^2_{OSE}$. We fix the tuning parameter $a = 3.7$ as recommended by Fan and Li (2001). To determine λ, a Bayesian information criterion(BIC) proposed by Chu et al. (2011) in incorporated.

The implementation of the penalized likelihood approach, including the calculation of PMLEs and OSEs is computationally demanding. This is because it relies heavily on the calculation of $R_n^{-1}(\boldsymbol{\theta})$ and $|R_n(\boldsymbol{\theta})|$, which is computationally intensive and often intractable due to numerical issues. It is particularly difficult for massive data collected on irregular grids, because no Kronecker product techniques can be utilized for computational simplification (Bayarri et al. 2007, 2009, Rougier 2008). A similar issue has also been recognized in calculating the MLEs in GP models.

3 Variable selection for GP via subagging

3.1 A new block bootstrap subsampling scheme

Subagging, modified based upon bagging (bootstrap aggregating), is one of the most effective and computationally efficient procedure to improve on unstable estimators (Efron and Tibshirani 1993, Breiman 1996, Bühlmann and Yu 2002). Although it is originally proposed to reduce variance in estimations and predictions, the idea of subsampling is attractive in many applications to achieve computational reduction. It is particularly appealing to GP modeling because of its high computational demand in estimating PMLEs and OSEs. However, direct application of subagging with random bootstrap subsamples is not efficient in estimation and variable selection of GP because the data are assumed to be dependent. This is not surprising because similar issues occur in the conventional bootstrap when the data are dependent such as in time series and spatial data, and various block bootstrap techniques are introduced (Künsch 1989, Liu and Singh 1992, Lahiri 1995, 1999, 2003, Politis and Romano 1994). Therefore, as an analogous result to the conventional block bootstrap, a new subsample scheme for dependent data based on blocks is called for.

We introduce a new block bootstrap subsampling method based on Latin hypercube designs (LHDs). It is called LHD-based block bootstrap. LHD is a class of experimental designs which is known to have a one-dimensional projection property, i.e, the projection of an LHD onto any dimension has exactly one observation for each level and therefore the resulting design can spread out more uniformly over the space. An m-run LHD in a d-dimensional space, denoted by LHD(m, d) can be easily constructed by permuting $(0, 1, ..., m-1)$ for each dimension. Given the sample size, there are $(m!)^{d-1}$ LHDs. Two randomly generated LHD(6,2) are illustrated in Figure 1. It is clear that the projection onto either dimension has exactly one observation for each level. After decomposing the complete data into disjoint

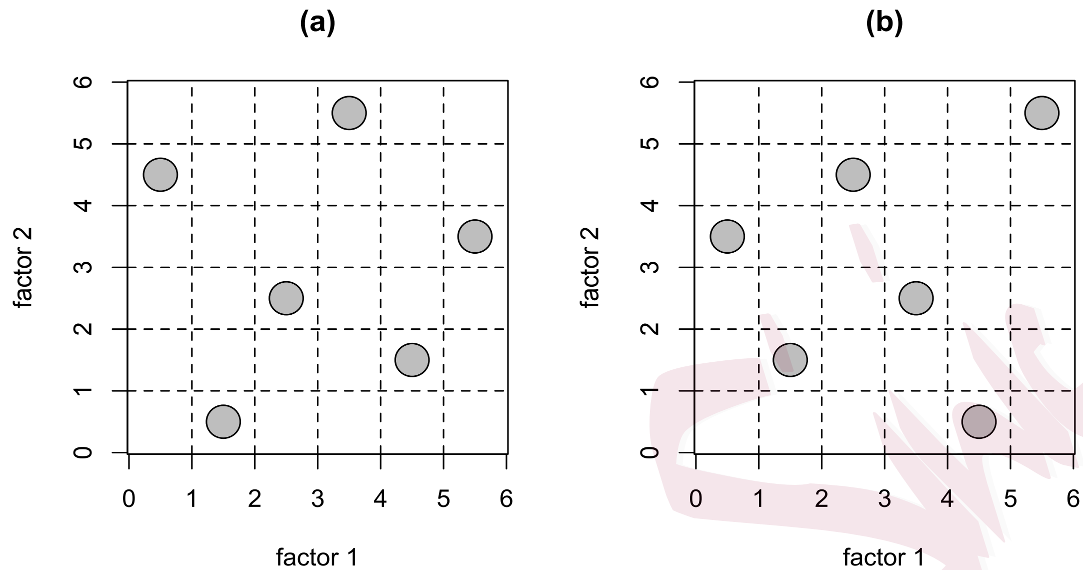

Figure 1: Two examples of LHDs

equally spaced hypercubes/blocks, a LHD-based block bootstrap subsample can be obtained by collecting blocks according to the structure of a randomly generated LHD. One example of a LHD-based block bootstrap subsample using the LHD in Figure 1(a) is given in Figure 2, where the circles are the observations, gray areas are the LHD-based blocks, and the red dots are the resulting subsamples. Formal definitions are given in the next section.

The LHD-based block bootstrap has the following advantages. First, the block structure takes into account the spatial dependency and therefore improves the estimation accuracy for correlation parameters in GP models. Second, because of the one-dimensional balance properties inherited from LHDs, the block bootstrap subsamples can be spread out more uniformly over the complete data and therefore the resulting subsamples can represent the complete data effectively. Third, it is shown that LHD can result in variance reduction in estimation compared with simple random samples (Mckay et al. 1979, Stein 1987). Therefore, the subagging estimates calculated by the proposed LHD-based subsamples are expected to outperform those calculated by the naive simple random subsamples in terms of estimation variance. This result is verified empirically by simulations in Section 5.2.

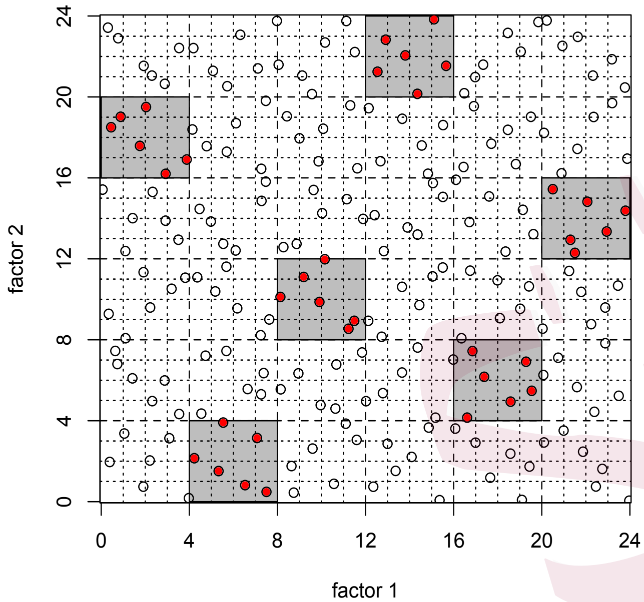

Figure 2: An example of LHD-based block bootstrap constructed from Figure 1(a)

3.2 Variable selection using LHD-based block subagging

The procedure can be described in three steps:

Step 1: *Divide each dimension of the interested region $\Gamma \in [0, l]^d$ 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 $\boldsymbol{i} = (i_1, ...i_d)$, $i_j \in (0, ..., m-1)$, represents the index of each 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 hypercube/block. For simplicity, assume the data points are equally distributed over the blocks, i.e. $|\mathcal{B}_n(\boldsymbol{i})| = n/m^d$.

Step 2: *Select m blocks according to a randomly generated LHD(m, d). Each column of the LHD is a random permutation of $\{0, \ldots, m-1\}$, denoted by $\boldsymbol{\pi}_i = (\pi_i(1), \ldots, \pi_i(m))^T$ for*

$1 \le i \le d$. So an m-run LHD is denoted by $\boldsymbol{i}_j^* = (\pi_1(j), \ldots, \pi_d(j))$, $j = 1, \ldots, m$, and the corresponding selected blocks are denoted 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, $\hat{\boldsymbol{\phi}}_N^*$ and its OSE $\hat{\boldsymbol{\phi}}_{N,OSE}^*$ is obtained by maximizing (3) and (4) respectively.

Step 3: Repeat Step 2 K times to obtain PMLEs $\hat{\boldsymbol{\phi}}_{N(j)}^*$ and the corresponding OSEs $\hat{\boldsymbol{\phi}}_{N,OSE(j)}^*$, where $j = 1, ..., K$. The subagging estimators are defined by $\hat{\boldsymbol{\phi}}_N = \frac{1}{K}\sum_{i=1}^K \hat{\boldsymbol{\phi}}_{N(i)}^*$ and $\hat{\boldsymbol{\phi}}_{N,OSE} = \frac{1}{K}\sum_{i=1}^K \hat{\boldsymbol{\phi}}_{N,OSE(i)}^*$.

Figure 2 is an example with experimental region $\Gamma \in [0, 24]^2$, i.e., $d = 2$, $l = 24$. A common practice is that the data are collected by normalizing the experimental region to a unit cube. In such a case, we have $l = 1$. The circles represent the settings in which the experiments are performed and the total sample size is $n = 216$. The design, LHD$(6, 2)$, implemented here is denoted by $\boldsymbol{i}_1^* = (0, 4)$, $\boldsymbol{i}_2^* = (1, 0)$, $\boldsymbol{i}_3^* = (2, 2)$, $\boldsymbol{i}_4^* = (3, 5)$, $\boldsymbol{i}_5^* = (4, 1)$, $\boldsymbol{i}_6^* = (5, 3)$ and $m = 6$. According to this design, the LHD-based blocks are presented by the gray areas with $b = 4$ and $|\mathcal{B}_n(\boldsymbol{i})| = 6$. The red dots are the resulting LHD-based block subsamples with size $N = 36$.

Based on this procedure, the complexity is $O(n^3/m^{3(d-1)})$ for each subsample, which is computationally cheaper than $O(n^3)$ using the complete data especially for large d. Note that, we assume data points are equally distributed over blocks in order to simplify the notation in the proof and the results still hold as long as the number of observations in each block is in the same order, i.e. $|\mathcal{B}_n(\boldsymbol{i}_i^*)| = O(n/m^d)$. For example, if the original data is collected by an orthogonal array-based Latin hypercube design (Tang 1993), which is common in computer experiments, the proposed procedure can be successfully implemented. In practice, based on our empirical experience, as long as each bootstrap subsample contains a small amount of empty blocks, we can still have an efficient representation of the original data. Empty blocks often occur when the original design has only few levels for some particular variables, such as qualitative variables. This issue can be addressed by modifying the LHDs by space-filling designs for quantitative and qualitative factors (Qian and Wu 2009, Deng et al. 2015) and as a result, empty blocks can be avoided. Given the total sample size n, we have $1 \leq m \leq n^{\frac{1}{d-1}}$. This is because each bootstrap subsample has their size N in the order of $O(n/m^{d-1})$. If $N = n/m^{d-1}$, then we have $m \leq n^{\frac{1}{d-1}}$ to ensure $N \geq 1$. Clearly, $m = 1$ provides no computational reduction because the full data is utilized. As m increases, the subsample size N decreases and therefore a larger K is affordable given the same computational constraints.

It is also worth noting that, instead of selecting subsamples based on all the variables, this procedure can be modified to be based on a subset of variables. To do this, we can first select a subset of variables with dimension $\tilde{d}$, where $\tilde{d} < d$. This subset can be chosen randomly or according to some prior knowledge, such as some variables are of particular interest or the design of the variables spreads out more uniformly. Next, replace $LHD(m, d)$ in step 2 by $LHD(m, \tilde{d})$ and select the subsamples only according to the $\tilde{d}$ variables. This is practically useful when d is large (e.g., the data center example, Section 6, we have $d = 9$) because the size of each subsample, n/m^{d-1}, can be relatively small and it creases to $n/m^{\tilde{d}-1}$ by applying to a subset variables. Moreover, the proposed framework is constructed based on rectangular or hypercubic regions which is common in many applications. It can be easily extended to regions with irregular shape by replacing the LHD in step 2 by other space filling designs constructed for nonrectangular regions, such as Draguljić et al. (2012) and Hung et al. (2012).

4 Theoretical properties

To understand the asymptotic properties of the subagging estimators, there are two distinct frameworks: increasing domain (Cressie 1993, Mardia and Marshall 1984) asymptotics, where more and more data are collected in increasing domains while the sampling density stays constant, and fixed-domain asymptotics (Stein 1999, Liang et al. 2013), where data are collected by sampling more and more densely in a fixed domain. The results in this research focus on the increasing domain asymptotics. For some applications, such as some examples in computer experiments, studies under fixed-domain asymptotics are more appropriate. However, not surprisingly, the results under fixed-domain asymptotics are more difficult to derive in general and rely on stronger assumptions as discussed in the literature

(Ying 1993, Zhang 2004). It is shown by Zhang and Zimmerman (2005) that, given 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, although results given here are based on increasing domain asymptotics, they provide some insights about the subagging estimators in both frameworks. In ongoing work, we are exploring the theoretical properties under fixed domain asymptotics. More discussions are given in Section 7. Assumptions and the proofs are given in the Appendix and Supplemental material.

In the following theorem, we show that the subagging estimator $\hat{\boldsymbol{\phi}}_N$ converges to the original PMLE $\hat{\boldsymbol{\phi}}_n$ in probability. For any LHD-based block bootstrapped statistic $\hat{T}_N^*$, we write $\hat{T}_N^* \to 0$ if for any $\epsilon > 0$ and any $\delta > 0$, $\lim_{n\to\infty} P\{P_{N,\omega}^*(|\hat{T}_N^* > \epsilon| > \delta)\} = 0$.

Theorem 1. *Under the assumptions (A.1)- (A.6), if $m = o(n^{-1/d})$ and $m \to \infty$, then*

$$\hat{\boldsymbol{\phi}}_N - \hat{\boldsymbol{\phi}}_n \to 0.$$

Next we study the distributional consistency of the subagging estimators. Assume $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T$ to be the true regression coefficients, where, without loss of generality, $\boldsymbol{\beta}_{10}$ is an $s \times 1$ vector of nonzero regression coefficients and $\boldsymbol{\beta}_{20} = 0$ is a $(p-s)\times 1$ zero vector. Let $\gamma_0 = (\boldsymbol{\theta}_0, \sigma_0)$ denote the vector of true covariance parameters, $\hat{\boldsymbol{\phi}}_N^* = (\hat{\boldsymbol{\beta}}_{N,1}^*, \hat{\boldsymbol{\beta}}_{N,2}^*, \hat{\gamma}_N^*)$, $\hat{\boldsymbol{\phi}}_N = (\hat{\boldsymbol{\beta}}_{N,1}, \hat{\boldsymbol{\beta}}_{N,2}, \hat{\gamma}_N)$, and $\hat{\boldsymbol{\phi}}_n = (\hat{\boldsymbol{\beta}}_{n,1}, \hat{\boldsymbol{\beta}}_{n,2}, \hat{\gamma}_n)$. When the OSE approach is applied, we denote $\hat{\boldsymbol{\phi}}_{N,OSE}^* = (\hat{\boldsymbol{\beta}}_{N,1,OSE}^*, \hat{\boldsymbol{\beta}}_{N,2,OSE}^*, \hat{\gamma}_{N,OSE}^*)$, $\hat{\boldsymbol{\phi}}_N = (\hat{\boldsymbol{\beta}}_{N,1,OSE}, \hat{\boldsymbol{\beta}}_{N,2,OSE}, \hat{\gamma}_{N,OSE})$, and $\hat{\boldsymbol{\phi}}_{n,OSE} = (\hat{\boldsymbol{\beta}}_{n,1,OSE}, \hat{\boldsymbol{\beta}}_{n,2,OSE}, \hat{\gamma}_{n,OSE})$. Furthermore, we define $a_n = \max_j\{p'_{\lambda_n}(|\beta_j|) : \beta_j \neq 0\}$ and $b_n = \max_j\{p''_{\lambda_n}(|\beta_j|) : \beta_j \neq 0\}$. Also, let $\mathbf{g}(\boldsymbol{\phi}) = (p'_\lambda(\boldsymbol{\phi}))$ with $\mathbf{g}(\boldsymbol{\beta}) = (p'_\lambda(|\beta_1|sgn(\beta_1)), ..., p'_\lambda(|\beta_p|sgn(\beta_p)))$ and $\mathbf{g}(\gamma) = \mathbf{0}$. Let $\mathbf{G}(\boldsymbol{\phi}) = diag(p''_\lambda(\boldsymbol{\phi})$. Particularly $\mathbf{G}(\boldsymbol{\beta}) = diag(p''_\lambda(|\beta_1|), ..., p''_\lambda(|\beta_p|))$ and $\mathbf{G}(\boldsymbol{\beta}) = \mathbf{0}$.

The next theorem shows that, given an efficient computational reduction, this framework guarantees the asymptotic consistency of the subbagging estimators to the PMLE using the complete data.

Theorem 2. *Under assumptions (A.1)-(A.15), if $m = o(n^{-1/d})$ and $m \to \infty$, then*

(i) Sparsity: $\hat{\boldsymbol{\beta}}_{N,2} = 0$ with probability tending to 1.

(ii) Asymptotic normality:

For the mean function coefficients, we have

$$\sqrt{Kn/m^{d-1}}(\boldsymbol{J}(\boldsymbol{\beta}_{10}) + \boldsymbol{G}(\boldsymbol{\beta}_{10}))(\hat{\boldsymbol{\beta}}_{N,1} - \hat{\boldsymbol{\beta}}_{n,1}) \to N(0, \boldsymbol{J}(\boldsymbol{\beta}_{10})),$$

For the correlation parameters, we have

$$\sqrt{Kn/m^{d-1}}(\hat{\gamma}_N - \hat{\gamma}_n) \to N(0, \boldsymbol{J}(\gamma_0)^{-1}).$$

In Theorem 3, it shows that when the OSE algorithm is applied, the resulting subagging estimators are asymptotic consistency to the original OSEs using the complete data.

Theorem 3. *Under assumptions (A.1)-(A.15), if $m = o(n^{-1/d})$ and $m \to \infty$, then*

(i) Sparsity: $\hat{\boldsymbol{\beta}}_{N,2,OSE} = 0$ with probability tending to 1.

(ii) Asymptotic normality:

For the mean function coefficients, we have

$$\sqrt{Kn/m^{d-1}}(\hat{\boldsymbol{\beta}}_{N,1,OSE} - \hat{\boldsymbol{\beta}}_{n,1,OSE}) \to N(0, \boldsymbol{J}(\boldsymbol{\beta}_{10})^{-1}),$$

For the correlation parameters, we have

$$\sqrt{Kn/m^{d-1}}(\hat{\gamma}_{N,OSE} - \hat{\gamma}_{n,OSE}) \to N(0, \boldsymbol{J}(\gamma_0^{-1})).$$

5 Numerical studies

In this section, two sets of simulations are conducted to study the finite-sample performance of the proposed method. One is to demonstrate the performance of the subagging approach compared with the original approach using all the data. The other is to illustrate the advantages of the proposed experimental design-based subsampling scheme by comparing with a naive simple random sampling. The performance is evaluated in two aspects: the accuracy of variable selection and the parameter estimation, including the mean function coefficients and the correlation parameters using one-step sparse estimation as described in (4). The accuracy of variable selection is measured by two scores. One is the average number of the nonzero regression coefficients correctly identified in the repeated simulations, denoted by AC, and the other is the average number of the zero regression coefficients misspecified, denoted by AM. All the simulations are conducted by a 2.7GHz, 16G RAM workstation. Hereafter, we omit the subscript OSE for notation convenience.

5.1 Subagging vs. the estimation using all data

Three sample sizes, $n = 1000$, $n = 2000$ and $n = 3000$, are considered and the data are generated from a regular grid in a four-dimensional space, $[0, 1]^4$. Note that the proposed method is particularly useful for data collected from irregular grids. The reason to generate the simulations from a regular grid in this simulation is that the original PMLE calculation using full data can be further speed up by Kronecker product techniques and some matrix singularity can be avoided (Rougier 2008). These techniques are only applicable to data sets collected from a regular grid, therefore, a favorable comparison of the proposed method would make an even stronger case for the proposed procedure.

Simulations are generated from a Gaussian process with the mean function coefficients $\boldsymbol{\beta} = (1, 0.5, 0, 0)$. Choose the correlation function to be

$$\psi(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp(-\sum_{i=1}^{4} \theta_i |x_{1i} - x_{2i}|)$$

where $\theta_1 = \theta_2 = \theta_3 = \theta_4 = 1$ and $\sigma = 0.1$. For each choice of sample size, a total of 50 data sets are simulated. For each simulated data set, K=10, that is 10 LHD-based block bootstrap samples are collected with $m = 4$. Due to the long computation time with the complete data, the tuning parameter $\lambda = 0.1$ is fixed for all simulations.

In Table 1, the parameter estimation and the computing time are reported. Standard deviations are given in parenthesis. The rows $AC/2$ and $AM/2$ represent the correct identification rate and the variable misspecification rate respectively. The results in Table 1 demonstrate that the estimated parameters using LHD-based subagging are consistent with those obtained using complete data. The variable selection performance is also compatible with the one using complete data. In terms of computing time, the proposed subagging is much faster to compute compared with the conventional approach especially when the

Table 1: Comparisons with all data

	$n = 1000$		$n = 2000$		$n = 3000$	
	LHD	*AllData*	*LHD*	*AllData*	*LHD*	*AllData*
θ_1	1.91 (0.55)	1.14 (0.05)	1.38 (0.35)	1.02 (0.02)	1.10(0.10)	0.97(0.02)
θ_2	1.94 (1.20)	1.08 (0.07)	1.16 (0.14)	1.00 (0.03)	1.17(0.08)	1.03(0.03)
θ_3	1.70 (0.68)	1.03 (0.04)	1.14 (0.20)	0.92 (0.03)	1.15(0.07)	1.06(0.02)
θ_4	1.77 (0.83)	1.04 (0.04)	1.37 (0.45)	1.02 (0.04)	1.10(0.03)	1.00(0.03)
β_1	$1.00(3.2 \times 10^{-3})$	$1.02(3.6 \times 10^{-3})$	$0.99(4.2 \times 10^{-3})$	$0.99(7.9 \times 10^{-3})$	$1.01(3.4 \times 10^{-3})$	$1.00(3.7 \times 10^{-3})$
β_2	$0.46(1.7 \times 10^{-2})$	$0.43(3.6 \times 10^{-2})$	$0.51(3.3 \times 10^{-3})$	$0.50(6.1 \times 10^{-3})$	$0.49(5.5 \times 10^{-3})$	$0.50(3.7 \times 10^{-3})$
$AC/2$	1	0.93	1	1	1	1
$AM/2$	0	0	0	0	0	0
time	243	464	990	2402	2524	8623

sample size of the complete data is large. For example when $n = 3000$, subagging provides more than 70% saving in computational time.

5.2 LHD-based block subsampling vs. random subsampling

One important feature of the proposed subsampling scheme is to borrow the idea of space-filling design to achieve an efficient data reduction. To demonstrate the advantage of this subsampling scheme, we compare its performance, denoted by LHD, with two naive alternatives, simple random sampling denoted by SRS and random blocks sampling denoted by RBS, with the same sample size. We first compare the performance of LHD with SRS in two different settings of subsampling scheme: $m = 4$ and $m = 6$.

The data are generated from a six-dimensional space, $[0, 1]^6$ with sample size $n = 3600$. We consider the same type of correlation function as before with the mean function coefficients set to be $\boldsymbol{\beta} = (1, 0.5, 0.3, 0, 0, 0)$, which indicates three non-zero coefficients with different signal strength and three zero coefficients. Results are summarized based on 100 simulations and K=20, which means 20 LHD-based block bootstrap samples are collected for each simulation. To focus on the capability of selecting active factors, the proposed subsampling is performed on the first three variables and the resulting sample sizes for $m = 4$

and $m = 6$ are approximate 225 and 100 respectively.

In Table 2, the estimated parameters, the correct identification rates and the variable misspecification rates are reported. In terms of parameter estimation, LHD performs similar to SRS in estimating the mean function coefficients. For estimating the correlation parameters, LHD outperforms SRS with a much smaller estimation variance, especially when the subsample size becomes smaller ($m = 6$). In general, it appears that the proposed subsampling based on LHDs provides an effective variance reduction in parameter estimation, which is consistent with the theoretical justifications in experimental design literature (Mckay et al. 1979, Stein 1987). In terms of variable selection, the correct identification rate for the LHD-based subsampling is 21% higher than SRS when $m = 4$ and 13% higher when $m = 6$. Both methods perform equally well with zero misspecification rate. To further assess the variable selection accuracy, the frequencies of individual variables identified from 100 simulations are reported in the last three rows of the table, denoted by $Fre(\beta_1)$, $Fre(\beta_2)$ and $Fre(\beta_3)$. The identification frequencies for β_3 decrease as expected due to its weak signal. But the proposed subsampling can still identify such a weak signal with at least 66% higher frequency compared with simple random subsamples.

In the next simulation, the proposed sampling scheme is compared with RBS in which blocks are selected randomly without the one-dimensional projection property. The data are generated from a 4-dimensional space with $n = 2000$. We consider the same type of correlation function as before with the mean function coefficients set to be $\boldsymbol{\beta} = (1, 0.5, 0.1, 0)$, which indicates three non-zero coefficients with different signal strength and one zero coefficients. Results are summarized in Table 3 based on 100 simulations and $K = 20$. The results of SRS with the same subsample size are also listed for comparison. In general, LHD outperforms the other two sampling and RBS performs slightly better than SRS. Comparing with RBS, the proposed method has a lower misspecification rate, i.e., a higher frequency of identifying the nonactive variable: 0.95 vs. 0.85. Moreover, LHD has less bias and a smaller variance in parameter estimation. These two observations empirically demonstrates the advantage of the one-dimensional balance property of LHD and consistent with the results derived by McKay et al. (1979) and Stein (1987).

Table 2: Comparisons with simple random subsampling

	$m=4$		$m=6$	
	LHD	SRS	LHD	SRS
θ_1	1.91 (0.60)	1.89 (4.11)	2.63 (1.63)	2.61 (9.93)
θ_2	2.24 (1.71)	1.90 (3.73)	2.64 (2.01)	2.95 (10.56)
θ_3	1.96 (0.79)	1.99 (2.66)	2.49 (1.14)	3.18 (10.97)
θ_4	1.93 (0.58)	1.92 (4.11)	2.69 (1.74)	2.90 (12.78)
θ_5	1.78 (0.35)	1.72 (1.91)	2.58 (0.84)	2.50 (12.55)
θ_6	1.89 (0.48)	1.94 (3.84)	2.74 (1.78)	1.80 (8.65)
β_1	$1.01(1.5 \times 10^{-3})$	$0.99(3.3 \times 10^{-3})$	$1.03(1.6 \times 10^{-3})$	$0.99(1.5 \times 10^{-3})$
β_2	$0.52(3.2 \times 10^{-3})$	$0.52(2.9 \times 10^{-3})$	$0.53(4.4 \times 10^{-3})$	$0.55(6.7 \times 10^{-3})$
β_3	$0.14(1.2 \times 10^{-2})$	$0.10(2.1 \times 10^{-2})$	$0.15(1.1 \times 10^{-2})$	$0.15(2.5 \times 10^{-2})$
AC/3	0.98	0.81	1	0.87
$AM/3$	0	0	0	0
$Fre(\beta_1)$	1	1	1	1
$Fre(\beta_2)$	1	1	1	1
$Fre(\beta_3)$	0.93	0.40	1	0.60

Table 3: Comparisons with simple random sampling of blocks

m=4	θ_1	θ_2	θ_3	θ_4	
LHD	1.21(0.26)	1.29(0.38)	1.27(0.32)	1.34(0.17)	
RBS	1.44(0.30)	1.50(0.34)	1.43(0.37)	1.50(0.33)	
SRS	1.77(0.88)	1.59(0.38)	1.55(0.72)	1.53(1.34)	
	β_1	β_2	β_3	$AC/3$	Freq($\beta_4 = 0$)
LHD	$1.00(1.9 \times 10^{-6})$	$0.50(2.3 \times 10^{-6})$	$0.09(1.8 \times 10^{-6})$	1.0	0.95
RBS	$1.00(7.5 \times 10^{-6})$	$0.51(3.0 \times 10^{-6})$	$0.08(3.1 \times 10^{-6})$	1.0	0.85
SRS	$1.00(3.7 \times 10^{-6})$	$0.51(1.1 \times 10^{-6})$	$0.09(1.2 \times 10^{-6})$	1.0	0.63

6 Data center thermal management

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 to 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. Since the thermal process in a data center is complex and depending on many factors, a crucial step is to model the thermal distribution at different experimental settings and in the mean time identify important factors that have significant impacts on the thermal distribution (Hung et al. 2012).

For a data center thermal study, physical experiments are not always feasible because some settings are highly dangerous and expensive to perform. Therefore, simulations based on computational fluid dynamics (CFD) are widely used. This type of simulations using complex mathematical models is often called computer experiments (Santner et al. 2003, Fang et al. 2006). In this example, CFD simulations are conducted at 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). There are 27,000 temperature outputs generated from the CFD simulator based on an irregular grid over an 9-dimensional space. The nine variables are listed in Table 4, 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 percentage of open areas (x_6), and spatial location in the data center (x_7 to x_9).

Gaussian process models are widely used for the analysis of computer experiments because it provides a flexible interpolator for the deterministic simulation outputs (Santner et al. 2003). However, in this example, it is computationally prohibitive to build a GP model based on the complete CFD data. So we implement the proposed LHD-based subagging approach with $m = 3$ for the first seven variables.

The fitted GP model is reported in the last two columns of Table 4, where $\hat{\boldsymbol{\beta}}$ represents the estimated mean function coefficients and $\hat{\boldsymbol{\theta}}$ represents the correlation parameters estimated using exponential covariance function. From the fitted model, it appears that seven out of the nine variables have significant effects to the mean function. The main effects plot based on the fitted GP model is given in Figure 3. It also appears that the two variables, x_5 and x_6, which are identified as nonactive have relatively small impacts on cooling. This result provides an important information regarding the efficiency of different cooling methods, because the variables are associated with two cooling mechanisms, a conventional cooling approach and a chilled water based cooling system. Among the active variables, the height (x_9) has a relatively large positive effect, which agrees with the general understanding of thermal dynamics that temperature increases significantly with height in a data center. The results also indicate that, among the four CRAC units in different locations of a data center, the first two CRAC units have significant effects on reducing the room temperature. This reveals important information that can help engineer locating the CRAC units more effectively and improve the efficiency of the cooling mechanism.

	Variable	$\hat{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\theta}}$
x_1	CRAC unit 1 flow rate	-7.5	5.3
x_2	CRAC unit 2 flow rate	-13.1	1.3
x_3	CRAC unit 3 flow rate	-2.7	0.3
x_4	CRAC unit 4 flow rate	-7.1	13.2
x_5	Room temperature setting	0	0.9
x_6	Tile open area percentage	0	0.6
x_7	Location in x-axis	-11.3	21.44
x_8	Location in y-axis	2.1	9.5
x_9	Height	17.8	0.8

Table 4: *Analysis for the data center example*

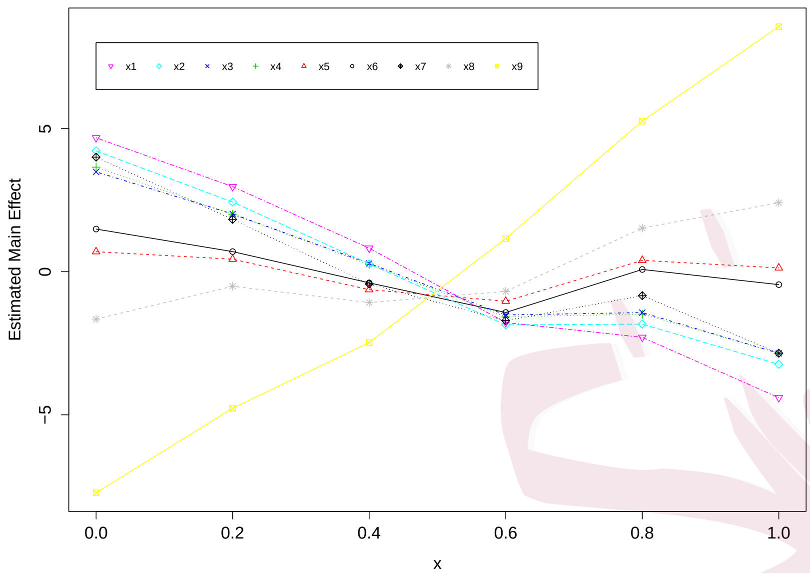

Figure 3: Main effect plot

7 Discussion

We propose a new framework to tackle computational difficulties in estimation and variable selection in GP models. This framework includes a subagging estimator and an efficient subsampling scheme that borrows the strength of experimental designs, particularly Latin hypercube designs in which the one-dimensional projection property is guaranted. The subagging estimation is computationally cheaper, yet it can be shown that the subagging estimators achieved the same efficiency as the original estimators using full data. Although one major focus of this framework is on simultaneous variable selection and estimation, the computational reduction introduced by the proposed framework remains effective for GP modeling in general. For example, for the conventional GP modeling, the proposed approach can be easily applied to alleviate the computational complexity and the theoretical results hold with a straightforward modification. Application of the proposed method to a data center thermal management study reveals important information for determining the most efficient cooling mechanism.

Future work will be explored in the following directions. First, extensions of the proposed procedure to optimal designs with better space-filling properties are appealing. For example, it is known that randomly generated LHDs can contain some structure. To further enhance desirable space-filling properties, various modifications are proposed. Numerical comparisons and theoretical developments of the generalization to different types of optimal space-filling designs will be carefully studied. Second, an interesting and important issue of the LHD-based block bootstrap is to determine the optimal block size. This topic has been discussed for conventional block bootstrap methods (Nordman et al. 2007), however the solutions therein are not directly applicable to GP models. We plan to study the optimal block size for the propose procedure based on a new criterion defined for GP. Third, theoretical development under fixed-domain asymptotics will be explored by extending the results of Ying (1993) and Hung (2011), and subagging predictors will also be developed. As pointed out by the referees, another interesting extension of the proposed work is to perform variable selection not only in the mean function but also in the correlation function. We are currently developing an extension to address this issue so that identification of linear effects in the mean function and nonlinear effects in the covariance function can be both achievable.

Acknowledgements

This research is supported by NSF DMS-1349415 grant. The authors are grateful to the Editor, an AE and two referees for their helpful comments and suggestions.

Appendix A: Assumptions

(A.1) $\frac{n}{m^d}\boldsymbol{Cov}\{(\bar{y}_{\boldsymbol{i}}-\mu)^2,(\bar{y}_{\boldsymbol{j}}-\mu)^2\}=O(1)$, $\boldsymbol{i}=(i_1,\ldots,i_d)\neq\boldsymbol{j}=(j_1,\ldots,j_d)$.

(A.2) $|\tau_n^2|=O(1)$.

(A.3) $\lim_{n\to\infty}\sup_{\boldsymbol{\theta}}\lambda_{\max}(E_n(\boldsymbol{\theta}))=0$ when the block space $b=l/m\to\infty$.

(A.4) $\forall\ \boldsymbol{\phi}_1,\ \boldsymbol{\phi}_2\in\Theta$, $|q_s(\cdot,\boldsymbol{\phi}_1)-q_s(\cdot,\boldsymbol{\phi}_2)|\leq L_s|\boldsymbol{\phi}_1-\boldsymbol{\phi}_2|a.s.P$, where L_s is Lipschitz constant and $\sup_n\{n^{-1}\sum_{s=1}^n\boldsymbol{E}L_s\}=O(1)$.

(A.5) Θ is compact.

(A.6) The functions $q_s(\omega, \boldsymbol{\phi})$ and $r_n(\omega, \boldsymbol{\phi})$ are such that $q_s(\cdot, \boldsymbol{\phi})$ and $r_n(\cdot, \boldsymbol{\phi})$ are measurable for all $\boldsymbol{\phi} \in \Theta$, a compact subset of R^p. In addition, $q_s(\omega, \cdot) : \Theta \longrightarrow R$ and $r_n(\omega, \cdot) : \Theta \longrightarrow R$ are continuous on Θ a.s.-P, $s = 1, \cdots, n$.

(A.7) $Q_n(\omega, \cdot) : \Theta \to R$ is continuously differentiable of order 2 on Θ a.s. P.

(A.8) There exists a sequence $J_n(\boldsymbol{\phi}) : \Theta \to R^{p\times p}$ such that $\nabla^2 Q_n(\cdot, \boldsymbol{\phi}) - J_n(\boldsymbol{\phi}) \xrightarrow{\text{P}} 0$ as $n \to \infty$ uniformly on Θ.

(A.9) $J_n(\boldsymbol{\phi}^0)$ is $O(1)$ and uniformly non-singular, i.e. $\lim_{n\to\infty} J_n^{-1}(\boldsymbol{\phi}^0) = 0$.

(A.10) $Q_N^*(\lambda, \omega, \cdot) : \Theta \to R$ are continuously differentiable of order 2 on Θ a.s. P. Also, function $\nabla^2 Q_n(\omega, \boldsymbol{\phi})$ is such that $\nabla^2 Q_n(\cdot, \boldsymbol{\phi})$ is measurable for all $\boldsymbol{\phi} \in \Theta$ and $\nabla^2 Q_n(\omega, \cdot) : \Theta \to R$ is continuous on Θ a.s.-P.

(A.11) $\forall\ \boldsymbol{\phi}_1,\ \boldsymbol{\phi}_2 \in \Theta, |\nabla^2 Q_n(\cdot, \boldsymbol{\phi}_1) - \nabla^2 Q_n(\cdot, \boldsymbol{\phi}_2)| \leq M_s|\boldsymbol{\phi}_1 - \boldsymbol{\phi}_2| a.s.P$, where M_s is Lipschitz constant and $\sup_n\{n^{-1}\sum_{s=1}^{n} \boldsymbol{E}M_s\} = O(1)$.

(A.12) $a_n = O(n^{-\frac{1}{2}})$ and $b_n \to 0$ as $n \to \infty$

(A.13) There exit positive constants c_1 and c_2 such that when $\beta_1, \beta_2 > c_1\lambda_n$, $|p''_{\lambda_n}(\beta_1) - p''_{\lambda_n}(\beta_2)| \leq c_2|\beta_1 - \beta_2|$.

(A.14) $\lambda_n \to 0$, $n^{\frac{1}{2}}\lambda_n \to \infty$ as $n \to \infty$.

(A.15) $\liminf_{n\to\infty} \liminf_{\boldsymbol{\beta}\to 0^+} \lambda_n^{-1} p'_{\lambda_n}(\boldsymbol{\beta}) > 0$.

Assumption (A.3) controls the correlation between bootstrapped blocks. (A.4) and (A.5) are required in order to achieve uniform convergency of the bootstrapped likelihood function. (A.6) ensures the existence of the estimators. (A.7)-(A.9) are regularity conditions for standard MLE consistency in GP models, which is analogue to the conditions in Mardia and Marshall (1984). (A.10) ensures the existence of covariance matrix. (A.11) is the global Lipschitz condition for $\nabla^2 Q_n(\omega, \cdot)$ which guarantees the convergence of the covariance matrix calculated based on the LHD-based block bootstrap. (A.12)-(A.15) are mild regularity conditions regarding the penalty function.

Appendix B: Consistency of the LHD-based block bootstrap mean

Before studying the asymptotic performance of MLEs, we first focus on understanding properties of the LHD-based block bootstrap mean, which is an important foundation to the theoretical development of $\hat{\boldsymbol{\phi}}_N^*$ later.

The LHD-based block bootstrap can be formulated mathematically as follows. Given the underlying probability space $(\Omega, \mathcal{F}, P)$ of a Gaussian process, a sample of size n with settings $\boldsymbol{x}_1(\omega), ..., \boldsymbol{x}_n(\omega)$ and responses $y(\boldsymbol{x})$'s are observed from a given realization $\omega \in \Omega$. Let $(\Lambda, \mathcal{G})$ be a measurable space on the realization. For each $\omega \in \Omega$, denote $P^*_{N,\omega}$ as the probability measure induced by the m-run LHD-based block bootstrap on $(\Lambda, \mathcal{G})$. The proposed bootstrap is a method to generate new dataset on $(\Lambda, \mathcal{G}, P^*_{N,\omega})$ conditional on the n original observations. Let $\tau_t : \Lambda \to \{1, ..., n\}$ denote a random index generated by the LHD-based block bootstrap. So, τ_t is the tth index in the intersect index of observations and $\{\mathcal{B}_n(\boldsymbol{i}_1^*), ..., \mathcal{B}_n(\boldsymbol{i}_m^*)\}$, where $(\boldsymbol{i}_1^*, ..., \boldsymbol{i}_m^*)$ is a randomly generated m-run LHD. Therefore, for $(\lambda, \omega) \in \Lambda \times \Omega$, we have the tth bootstrap sample: $\boldsymbol{x}_t^*(\lambda, \omega) \equiv \boldsymbol{x}_{\tau_t(\lambda)}(\omega)$.

Suppose $\{Y(\boldsymbol{x}_t), t \in R\}$ follows a GP with mean μ. Given n observations, the sample estimation of mean μ is

$$\bar{y}_n = \frac{1}{n} \sum_{s=1}^{n} y_s,$$

and the LHD-based block bootstrap mean with N samples is given by

$$\bar{y}_N^* = \frac{1}{N} \sum_{s=1}^{N} y_s^*.$$

With a slight abuse of notation, we replace the notation of random variable Y by its realization y unless otherwise specified. The following theorem shows the asymptotic consistency of the LHD-based block bootstrap mean.

Theorem 4. *Under (A.1)-(A.2), if $m \to \infty$ and $m = o(n^{1/d})$, then*

$$\sup_x |P^*_{N,\omega}(\sqrt{n/m^{d-1}}(\bar{y}_N^* - \bar{y}_n)/\tau_n \le x) - P(\sqrt{n}(\bar{y}_n - \mu)/\tau_n \le x)| \xrightarrow{\text{P}} 0,$$

when $n \longrightarrow \infty$.

The proof consists of several Lemmas that culminate in the final proof. To save space, the proof for the Lemmas are given in the supplemental material. Note that $\boldsymbol{E}(\cdot)$ and $\boldsymbol{Cov}(\cdot,\cdot)$ denote the expectation and variance under P while $\boldsymbol{E}^*_{N,\omega}(\cdot)$ and $\boldsymbol{Cov}^*_{N,\omega}(\cdot,\cdot)$ denote the expectation and variance under $P^*_{N,\omega}$.

Lemma 1. *LHD-based block bootstrap mean is unbiased, i.e.,*

$$\boldsymbol{E}^*_{N,\omega}(\bar{y}^*_N) = \bar{y}_n.$$

Lemma 2. *Let* $\bar{y}_{\boldsymbol{i}} = \frac{1}{\mathcal{B}_n(\boldsymbol{i})}\sum_{\boldsymbol{x}_s \in \mathcal{B}_n(\boldsymbol{i})} y_s$, $\forall \boldsymbol{i} = (i_1, \ldots, i_d)$. *Assuming (A.1), (A.2) and* $m = o(n^{1/d})$, *we have*

$$\frac{n}{m^{2d}} \sum_{i_1,\ldots,i_d} (\bar{y}_{i_1,\ldots,i_d} - \mu)^2 - \tau_n^2 \xrightarrow{\text{P}} 0,$$

where $\tau_n^2 = \frac{1}{n}\sum_{s,t=1}^n \boldsymbol{Cov}(Y_s(\boldsymbol{x}_s), Y_t(\boldsymbol{x}_t)).$

Lemma 3. *Assume (A.1)- (A.2), then*

$$n{\tau^*_N}^2/m^{d-1} - \tau_n^2 \xrightarrow{\text{P}} 0,$$

where ${\tau^*_N}^2 = \boldsymbol{Cov}^*_{N,\omega}(\bar{y}^*_N, \bar{y}^*_N)$.

Proof of Theorem 4: It suffices to show that (1) $\boldsymbol{E}^*_{N,\omega}(\bar{y}^*_N) = \bar{y}_n$; (2) $n{\tau^*_N}^2/m^{d-1} - \tau_n^2 \xrightarrow{\text{P}} 0$; and (3) $\sup_x |P^*_{N,\omega}((\bar{y}^*_N - \boldsymbol{E}^*_{N,\omega}(\bar{y}^*_N))/\tau^*_N \le x) - \Phi(x)| \xrightarrow{\text{P}} 0$, where $\Phi(\cdot)$ denotes standard normal distribution function and ${\tau^*_N}^2 = \boldsymbol{Cov}^*_{N,\omega}(\bar{y}^*_N, \bar{y}^*_N)$.

Lemmas 1 and 3 imply the results in (1) and (2). Note that $\bar{y}^*_N = \frac{1}{m}\sum_{j=1}^m \bar{y}_{\boldsymbol{i}^*_j}$ and $(\bar{y}_{\boldsymbol{i}^*_1}, \ldots, \bar{y}_{\boldsymbol{i}^*_m})$ follows Latin Hypercube sampling distribution. According to Loh (1996), we have the Berry-Essen type of bound for Latin Hypercube sampling

$$\sup_x |P^*_{N,\omega}((\bar{y}^*_N - \bar{y}_n)/\tau^*_N \le x) - \Phi(x)| \le c^* m^{-1/2},$$

where c^* is a constant that depends only on d, given $\boldsymbol{E}^*_{N,\omega}\|\bar{y}_{\boldsymbol{i}^*_1}\|^3 < \infty$. So we only need to show that $\boldsymbol{E}^*_{N,\omega}\|\bar{y}_{\boldsymbol{i}^*_1}\|^3$ is bounded uniformly in probability under P. Since $\boldsymbol{E}^*_{N,\omega}\|\bar{y}_{\boldsymbol{i}_1}\|^3 = \frac{1}{m^d}\sum_{\boldsymbol{i}} \bar{y}^3_{\boldsymbol{i}}$ and according to Minkowski's inequality, it follows that

$$\frac{1}{m^d}\sum_{\boldsymbol{i}} \boldsymbol{E}\{\bar{y}^3_{\boldsymbol{i}}\} \le \frac{1}{m^d}\sum_{\boldsymbol{i}} \frac{1}{|\mathcal{B}_n(\boldsymbol{i})|^3}\{\sum_{\boldsymbol{x}_s \in \mathcal{B}_n(\boldsymbol{i})} \boldsymbol{E}(y_s)\}^3 < \infty.$$

□

Appendix C: Proof of Theorem 1

To investigate the asymptotic properties of the estimators from LHD-based block bootstrap, we decompose the likelihood function into blocks. For each block, denote $\boldsymbol{y_i} = (y_s(\boldsymbol{x}_s), \boldsymbol{x}_s \in \mathcal{B}_n(\boldsymbol{i}))$, $\boldsymbol{X_i} = (\boldsymbol{x}_s, \boldsymbol{x}_s \in \mathcal{B}_n(\boldsymbol{i}))^T$, $R_{\boldsymbol{i},\boldsymbol{j}}(\boldsymbol{\theta})=\big[\psi(y(\boldsymbol{x}_s), y(\boldsymbol{x}_t); \boldsymbol{\theta}),\ \boldsymbol{x}_s \in \mathcal{B}_n(\boldsymbol{i}), \boldsymbol{x}_t \in \mathcal{B}_n(\boldsymbol{j})\big]$ and $\boldsymbol{z_i} = R_{\boldsymbol{i},\boldsymbol{i}}^{-1/2}(\boldsymbol{\theta})(\boldsymbol{y_i} - \boldsymbol{X_i}\boldsymbol{\beta})$. Then, we can rewrite the penalized log-likelihood function $n^{-1}\ell(\boldsymbol{X}_n, \boldsymbol{y}_n, \boldsymbol{\phi})$ as

$$\begin{aligned} Q_n(\boldsymbol{X}_n, \boldsymbol{y}_n, \boldsymbol{\phi}) = \;& -(2n\sigma^2)^{-1}\textstyle\sum_{s=1}^n z_s^2 - (2n)^{-1}\sum_{s=1}^n \log(\lambda_s) \\ & -(2n)^{-1}\textstyle\sum_{s=1}^n \log(\sigma^2) + n^{-1} r_n(\boldsymbol{X}_n, \boldsymbol{y}_n, \boldsymbol{\phi}) \\ & -\textstyle\sum_{s=1}^p p_\lambda(|\beta_s|) \\ = \;& n^{-1}\textstyle\sum_{s=1}^n q_s(\omega, \boldsymbol{\phi}) + n^{-1} r_n(\omega, \boldsymbol{\phi}) - \sum_{s=1}^p p_\lambda(|\beta_s|) \end{aligned} \tag{5}$$

where $\{\lambda_s, s = 1, \ldots, n\} = \{\text{eigenvalues of } |R_{\boldsymbol{i},\boldsymbol{i}}(\boldsymbol{\theta})|, \boldsymbol{i} = (i_1, \ldots, i_d)\}$ with $(i_1, \ldots, i_d)$ in lexicographical order and eigenvalues from the largest to the smallest. Note that $r_n(\omega, \boldsymbol{\phi}) = \ell(\boldsymbol{X}_n, \boldsymbol{y}_n, \boldsymbol{\phi}) - \sum_{s=1}^n q_s(z_s, \boldsymbol{\phi})$ contains all terms involving the off block-diagonal terms. Define $D_n(\boldsymbol{\theta}) = \mathrm{diag}(R_{\boldsymbol{i},\boldsymbol{i}}(\boldsymbol{\theta}))$ and $E_n(\boldsymbol{\theta}) = R_n(\boldsymbol{\theta}) - D_n(\boldsymbol{\theta})$. Assuming that $E_n(\boldsymbol{\theta}) = U_n(\boldsymbol{\theta})U_n^T(\boldsymbol{\theta})$, we have

$$\begin{aligned} r_n(\omega, \boldsymbol{\phi}) \;=\;& \frac{1}{2\sigma^2(1+g)}(\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta})^T D_n^{-1}(\boldsymbol{\theta}) E_n(\boldsymbol{\theta}) D_n^{-1}(\boldsymbol{\theta})(\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta}) \\ & + \frac{1}{2}\log|I_n + U_n^T(\boldsymbol{\theta}) D_n^{-1}(\boldsymbol{\theta}) U_n(\boldsymbol{\theta})|, \end{aligned}$$

where $g = \mathrm{trace}(E_n(\boldsymbol{\theta}) D_n^{-1}(\boldsymbol{\theta}))$.

The maximum likelihood estimator is obtained by $\hat{\boldsymbol{\phi}}_n = \arg\max_{\boldsymbol{\phi}} Q_n(\boldsymbol{X}_n, \boldsymbol{y}_n, \boldsymbol{\phi})$. Analogue to the decomposition for $Q_n(\boldsymbol{X}_n, \boldsymbol{y}_n, \boldsymbol{\phi})$, the log-likelihood function for LHD-based block bootstrap samples can be written as

$$Q_N^*(\boldsymbol{X}_N^*, \boldsymbol{y}_N^*, \boldsymbol{\phi}) \;=\; N^{-1}\sum_{s=1}^N q_s^*(\cdot, \omega, \boldsymbol{\phi}) + N^{-1} r_N^*(\cdot, \omega, \boldsymbol{\phi}) - \sum_{s=1}^p p_\lambda(|\beta_s|) \tag{6}$$

where $r_N^*(\cdot, \omega, \boldsymbol{\phi})$ contains all terms involving the off block-diagonal terms with bootstrapped samples. Specifically,

$$\begin{aligned} r_N^*(\cdot, \omega, \boldsymbol{\phi}) \;=\;& \frac{1}{2\sigma^2(1+g^*)}(\boldsymbol{y}_N^* - \boldsymbol{X}_N^*\boldsymbol{\beta})^T {D_N^*}^{-1}(\boldsymbol{\theta}) E_N^*(\boldsymbol{\theta}) {D_N^*}^{-1}(\boldsymbol{\theta})(\boldsymbol{y}_N^* - \boldsymbol{X}_N^*\boldsymbol{\beta}) \\ & + \frac{1}{2}\log|I_N + {U_N^*}^T(\boldsymbol{\theta}) {D_N^*}^{-1}(\boldsymbol{\theta}) U_N^*(\boldsymbol{\theta})|, \end{aligned}$$

where $D_N^*(\boldsymbol{\theta}) = \text{diag}(R_{\boldsymbol{i}_j^*,\boldsymbol{i}_j^*}(\boldsymbol{\theta}), j = 1,\dots,m)$ and $E_N^*(\boldsymbol{\theta}) = R_N^*(\boldsymbol{\theta}) - D_N^*(\boldsymbol{\theta})$ with $E_N^*(\boldsymbol{\theta}) = U_N^*(\boldsymbol{\theta}) U_N^{*T}(\boldsymbol{\theta})$; $g^* = \text{trace}(E_N^*(\boldsymbol{\theta}) D_N^{*-1}(\boldsymbol{\theta}))$. The bootstrapped version of $\hat{\boldsymbol{\phi}}_n$ is $\hat{\boldsymbol{\phi}}_N^* = \arg\max_{\boldsymbol{\phi}} Q_N^*(\boldsymbol{X}_N^*, \boldsymbol{y}_N^*, \boldsymbol{\phi})$. Theoretical properties of the LHD-based block bootstrap likelihood function (6) are established in the following two lemmas, which leads to a proof of convergence properties of the bootstrap estimator $\hat{\boldsymbol{\phi}}_N^*$. Lemma 4 below first established the pointwise weak law of large numbers for the LHD-based block bootstrap likelihood functions.

Lemma 4. *Under (A.1)-(A.3), for each* $\boldsymbol{\phi} \in \Theta$,

$$\lim_{n\to\infty} P\Big[P_{N,\omega}^*\big(|N^{-1}\sum_{s=1}^N q_s^*(\cdot,\omega,\boldsymbol{\phi}) + N^{-1} r_N^*(\cdot,\omega,\boldsymbol{\phi}) - n^{-1}\sum_{s=1}^n q_s(\omega,\boldsymbol{\phi}) - n^{-1} r_n(\omega,\boldsymbol{\phi})| > \delta\big) > \xi\Big] = 0.$$

The next lemma further extends Lemma 4 to the uniform weak law of large numbers for the LHD-based block bootstrap likelihood functions.

Lemma 5. *(Uniform Weak Law of Large Numbers) Under (A.1)-(A.5),* $\forall\, \delta, \xi > 0$,

$$\lim_{n\to\infty} P\Big[P_{N,\omega}^*(\sup_{\boldsymbol{\phi}\in\Theta} |N^{-1}\sum_{s=1}^N q_s^*(\cdot,\omega,\boldsymbol{\phi}) + N^{-1} r_N^*(\cdot,\omega,\boldsymbol{\phi}) \\ -n^{-1}\sum_{s=1}^n q_s(\omega,\boldsymbol{\phi}) - n^{-1} r_n(\omega,\boldsymbol{\phi})| > \delta) > \xi\Big] = 0.$$

Proof of Theorem 1: Based on Lemma 5, we have

$$\lim_{n\to\infty} P[P_{N,w}^*(\sup_{\boldsymbol{\phi}\in\Theta} |Q_n - Q_N^*| > \delta) > \xi] = 0,$$

where Q_n and Q_N^* are given in (5) and (6). With the full preparation of the likelihood convergence developed in Lemmas 4 and 5, the convergence of bootstrap parameter estimation follows immediately given the existence of $\hat{\boldsymbol{\phi}}_n$ and $\hat{\boldsymbol{\phi}}_N^*$.

Denote $\bar{q}_N^*(\cdot,\omega,\boldsymbol{\phi}) = N^{-1}\sum_{i=1}^N q_i^*(\cdot,\omega,\boldsymbol{\phi})$ and $\bar{q}_n(\omega,\boldsymbol{\phi}) = n^{-1}\sum_{i=1}^n q_i(\omega,\boldsymbol{\phi})$. By (A.6), $q_s^*(\cdot,\omega,\cdot) : \Lambda\times\Theta \to R$ and $r_N^*(\cdot,\omega,\cdot) : \Lambda\times\Theta \to R$ are measurable-$\mathcal{G}$ for each $\boldsymbol{\phi}\in\Theta$. In addition, $q_s^*(\lambda,\omega,\cdot)$ and $r_N^*(\lambda,\omega,\cdot)$ are continuous on Θ for all λ. Thus, we have $\hat{\boldsymbol{\phi}}_N^*(\cdot,\omega)$ exists as a measurable-$\mathcal{G}$ function by Jennrich (1969).

Following the procedure in Goncalves and White (2004), for any subsequence $\{n'\}$, given that $\hat{\boldsymbol{\phi}}_{n'}$ is identifiable and unique, there exists a further subsequence $\{n''\}$ such that $\hat{\boldsymbol{\phi}}_{n''}$ is identifiably unique with respect to $\{Q_{n''}\}$ for all $\omega \in F$ in some $F \in \mathcal{F}$ with $P(F) = 1$. By condition (A.6), there exists $G \in \mathcal{F}$ with $P(G) = 1$ such that for all $\omega \in G$, $\{Q^*_{N''}(\cdot, \omega, \boldsymbol{\phi})\}$ (N'' is corresponding bootstrapped sample size of n'') is a sequence of random function on $(\Lambda, \mathcal{G}, P^*_{N,\omega})$ continuous on Θ for all $\lambda \in \Lambda$. Hence, by White (1996), for fixed $\omega \in G$, there exists $\hat{\boldsymbol{\phi}}^*_{N''}(\cdot, \omega) : \Lambda \to \Theta$ measurable-$\mathcal{G}$ and $\hat{\boldsymbol{\phi}}^*_{N''}(\cdot, \omega) = \arg\max_{\boldsymbol{\phi}} Q^*_{N''}(\cdot, \omega, \boldsymbol{\phi})$. By the uniform weak law of large numbers for $Q^*_N(\boldsymbol{X}^*_N, \boldsymbol{y}^*_N, \boldsymbol{\phi})$ obtained from Lemma 5, we have $Q^*_{N''}(\cdot, \omega, \boldsymbol{\phi}) - Q_{n''}(\omega, \boldsymbol{\phi}) \to 0$ as $n'' \to \infty$ $prob{-}P^*_{N,\omega}$ $prob{-}P$ uniformly on Θ, where we write $\hat{Q}^*_N \to 0$ $prob{-}P^*_{N,\omega}, prob{-}P$ if, for any $\epsilon > 0$ and $\delta > 0$, $\lim_{n\to\infty} P\{P^*_{N,\omega}(|\hat{Q}^*_N > \epsilon| > \delta)\} = 0$ and omit $prob{-}P^*_{N,\omega}, prob{-}P$ in the text for notation simplicity. Hence, there exists a further subsequence $\{n'''\}$ such that $Q^*_{N'''}(\cdot, \omega, \boldsymbol{\phi}) - Q_{n'''}(\omega, \boldsymbol{\phi}) \to 0$ as $n'' \to \infty$ $prob - P^*_{N,\omega}$ $prob - P$ for all ω in some $H \in \mathcal{F}$ with $P(H) = 1$. Choose $\omega \in F \cap G \cap H$, by White (1996), we have $\hat{\boldsymbol{\phi}}^*_{N'''} - \hat{\boldsymbol{\phi}}_{n'''} \to 0$ as $n''' \to \infty$ $prob - P^*_{N,\omega}$ $prob - P$. Since this is true for any subsequence $\{n'\}$, we have $P(F \cap G \cap H) = 1$. Thus, $\hat{\boldsymbol{\phi}}^*_N - \hat{\boldsymbol{\phi}}_n \to 0$ $prob - P^*_{N,\omega}, prob - P$. Then $\hat{\boldsymbol{\phi}}_N = \frac{1}{K}\sum_{i=1}^K \hat{\boldsymbol{\phi}}^*_N(i) - \hat{\boldsymbol{\phi}}_n \to 0$ $prob - P^*_{N,\omega}, prob - P$. □

Appendix D: Proof of Theorem 2

Proof. Define $B = Var\{n^{-1/2}\sum_{s=1}^n \nabla q_s(\cdot, \omega, \boldsymbol{\phi}_0)\}$. We first show that $\sqrt{n/m^{d-1}} B^{-1/2} \nabla Q^*_N(\cdot, \omega, \hat{\boldsymbol{\phi}}_n) \to N(0, I)$. Denote $\bar{h}^*_N(\boldsymbol{\phi}) = N^{-1}\sum_{s=1}^N \nabla q^*_s(z^*_s, \boldsymbol{\phi})$ and $\bar{h}_n(\boldsymbol{\phi}) = n^{-1}\sum_{s=1}^n \nabla q_s(z_s, \boldsymbol{\phi})$. We have

$$\begin{aligned}
\sqrt{n/m^{d-1}}[\bar{h}^*_N(\hat{\boldsymbol{\phi}}_n) - \bar{h}_n(\hat{\boldsymbol{\phi}}_n)] &= +\sqrt{n/m^{d-1}}[\bar{h}^*_N(\hat{\boldsymbol{\phi}}_n) - \bar{h}^*_N(\boldsymbol{\phi}^0)] + \sqrt{n/m^{d-1}}[\bar{h}^*_N(\boldsymbol{\phi}^0) - \bar{h}_n(\boldsymbol{\phi}^0)] \\
&\quad + \sqrt{n/m^{d-1}}[\bar{h}_n(\boldsymbol{\phi}^0) - \bar{h}_n(\hat{\boldsymbol{\phi}}_n) \\
&= J_1 + J_2 + J_3.
\end{aligned}$$

Since $\bar{h}_n$ and $\bar{h}^*_N$ are functions whose secondary derivative are continuous, $J_1 + J_3 \to 0$ as $\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0 \to 0$ by Theorem 3.1 in Chu (2011). Moreover, the two terms in J_2 are both evaluated at $\boldsymbol{\phi}_0$ which is a fixed value, then by Theorem 4, we have $B^{-1/2} J_2 \to N(0, I)$.
By condition (A.10) and follow a similar proof as Lemma 5, we have

$$\nabla^2 Q^*_N(\cdot, \omega, \boldsymbol{\phi}) - \nabla^2 Q_n(\omega, \boldsymbol{\phi}) \to 0 \ \ prob - P^*_{N,\omega}, prob - P.$$

Let $\hat{H}_n(\omega) = \nabla^2 Q_n(\omega, \hat{\boldsymbol{\phi}}_n)$. According to White (1996), given the result $\hat{\boldsymbol{\phi}}^*_N - \hat{\boldsymbol{\phi}}_n \to 0$

$prob - P^*_{N,\omega}, prob - P$ and assumption (A.8), we have

$$\begin{aligned}\sqrt{N}(\hat{\boldsymbol{\phi}}^*_N - \hat{\boldsymbol{\phi}}_n) &= -\hat{H}_n^{-1}(\omega)\sqrt{N}\nabla Q^*_N(\cdot,\omega,\hat{\boldsymbol{\phi}}_n) + o_{P^*_{N,\omega}}(1)\\ &= -H_n(\boldsymbol{\phi}_0)^{-1}(\omega)\sqrt{N}\nabla Q^*_N(\cdot\omega,\hat{\boldsymbol{\phi}}_n) + o_{P^*_{N,\omega}}(1).\end{aligned}$$

Given the fact that

$$\sqrt{n/m^{d-1}}B^{-1/2}\nabla Q^*_N(\cdot,\omega,\hat{\boldsymbol{\phi}}_n) \to N(0,I) \;\; prob - P^*_{N,\omega}, prob - P.$$

we have

$$B^{-1/2}H_n(\boldsymbol{\phi}_0)\sqrt{N}(\hat{\boldsymbol{\phi}}^*_N - \hat{\boldsymbol{\phi}}_n) \to N(0,I).$$

For $\boldsymbol{\beta}_{10}$, B and H can be written as $\mathbf{J}(\boldsymbol{\beta}_{10})$ and $\mathbf{J}(\boldsymbol{\beta}_{10}) + \mathbf{G}(\boldsymbol{\beta}_{10})$. For $\hat{\boldsymbol{\beta}}^*_{N,1}$, we have

$$\sqrt{N}[\mathbf{J}(\boldsymbol{\beta}_{10}) + \mathbf{G}(\boldsymbol{\beta}_{10})]\{\hat{\boldsymbol{\beta}}^*_{N,1} - \hat{\boldsymbol{\beta}}_{n,1}\} \to N(0, J(\boldsymbol{\beta}_{10})).$$

For sub-bagging estimator $\hat{\boldsymbol{\beta}}_{N,1} = \sum_{i=1}^{K}\hat{\boldsymbol{\beta}}^*_{N,1}(i)$, we have

$$\sqrt{KN}[\mathbf{J}(\boldsymbol{\beta}_{10}) + \mathbf{G}(\boldsymbol{\beta}_{10})]\{\hat{\boldsymbol{\beta}}_{N,1} - \hat{\boldsymbol{\beta}}_{n,1}\} \to N(0, J(\boldsymbol{\beta}_{10})),$$

then the result follows.

Appendix E: Proof of Theorem 3

Using the same technique before, we decompose the log-likelihood by blocks and rewrite the likelihood of $\boldsymbol{\beta}$ based on the OSE approach as follows:

$$Q_n(\boldsymbol{\beta}) = n^{-1}\sum_{s=1}^{n} q_s(\omega,\boldsymbol{\beta},\hat{\boldsymbol{\theta}}_n^{(0)},\hat{\sigma_n^2}^{(0)}) + n^{-1}r_n(\omega,\boldsymbol{\beta},\hat{\boldsymbol{\theta}}_n^{(0)}\hat{\sigma_n^2}^{(0)}) - \sum_{j=1}^{p} p'_\lambda(|\hat{\beta}_j^{(0)}|)|\beta_j|.$$

The likelihood based on subsampled data can be written as:

$$Q^*_N(\boldsymbol{\beta}) = N^{-1}\sum_{s=1}^{N} q^*_s(\omega,\boldsymbol{\beta},\hat{\boldsymbol{\theta}}_N^{*(0)},\hat{\sigma_N^2}^{*(0)}) + N^{-1}r^*_N(\omega,\boldsymbol{\beta},\hat{\boldsymbol{\theta}}_N^{*(0)},\hat{\sigma_N^2}^{*(0)}) - \sum_{j=1}^{p} p'_\lambda(\hat{\beta}_j^{*(0)}|)|\beta_j|.$$

By the fact that $\hat{\boldsymbol{\phi}}^*_N - \hat{\boldsymbol{\phi}}_n \to 0$ and the results in Lemma 2, Lemma 3 and Theorem 4 still hold, we have $\hat{\boldsymbol{\phi}}^*_{N,OSE} - \hat{\boldsymbol{\phi}}_{n,OSE} \to 0$. Then follows the same technique in the proof of Theorem 2, the result follows. $\square$

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.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24, 123–140.

Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *Annals of Statistics*, 30, 927–961.

Chu, T., Zhu, J. and Wang, H. (2011). Penalized maximum likelihood estimation and variable selection in geostatistics. *Annals of Statistics*, 39, 2607–2625.

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.

Deng, X., Hung, Y., and Lin, C. D. (2015). Design for computer experiments with qualitative and quantitative factors. *Statistica Sinica*, 25, 1567-1581.

Donoho, D. and Johnstone, I. (1994). Ideal spatial adaption by wavelet shrinkage. *Biometrika*, 81, 425–455.

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

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, 96:13481360.

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.

Goncalves, S. and White, H. (2004). Maximum likelihood and the bootstrap for nonlinear dynamic models. *Journal of Econometrics*, 119, 199–219.

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

Hoeting, J., Davis, R., Merton, A. and Thompson, S. (2006). Model Selection For Geostatistical Models. *Ecological Applications*, 16:8798

Huang, H. and Chen, C. (2007). Optimal Geostatistical Model Selection. *Journal of the American Statistical Association*, 102, 1009-1024.

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.

Hung, Y. (2011). Penalized Blind Kriging in Computer Experiments. *Statistica Sinica*, 21, 1171-1190

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, 40, 633–643.

Joseph, V., Hung, Y. and Sudjianto, A. (2008), Blind Kriging: A New Method for Developing Metamodels. *Journal of Mechanical Design*, 130(3), 031102

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.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17, 1217–1241.

Lahiri, S. N. (1995). On the asymptotic behavior of the moving block bootstrap for normalized sums of heavy-tail random variables. *The Annals of Statistics*, 23, 1331–1349.

Lahiri, S. N. (1999). Theoretical comparisons of block bootstrap methods. *The Annals of Statistics*, 27, 386–404.

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*, Springer, New York.

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.

Linkletter, C., Bingham, D., Hengartner, N., Higdon, D. and Ye, K. Q. (2006). Variable selection for Gaussian process models in computer experiments. *Technometrics* **48** 478–490.

Liu, R. Y. and Singh, K. (1992). Moving Blocks Jackknife and Bootstrap Capture Weak Dependence. In *Exploring the Limits of Bootstrap* (R. LePage and L. Billard, eds.), 225–248, Wiley, New York.

Loh, W.-L. (1996). On Latin hypercube sampling. *The annals of statistics*, 24, 2058–2080.

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

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.

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. In *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 multi-resolution Gaussian process model for the analysis of large spatial data sets. *Journal of Computational and Graphical Statistics*, to appear.

Nychka, D. W., Wikle, C. and Royle, J. A. (2002). Multiresolution models for non-stationary spatial covariance functions. *Statistical Modeling*, 2, 315–331.

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89, 1303–1313.

Qian, P. Z. G. and Jeff, C. F. J. (2009). Sliced space-filling designs. *Biometrika*, 96, 945-956.

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.

Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics*, 29, 31–49.

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

Smola, A. J. and Bartlett, P. L. (2001). Sparse greedy Gaussian process regression. In *Advances in Neural Information Processing Systems*, 13, (T. K. Leen, T. G. Dietterich, and V. Tresp eds.) 619–625.

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. (1987). Large sample properties of simulations using Latin hypercube sampling. *Technometrics*, 29, 143–151.

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.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the Lasso. *Journal of Royal Statistical Society Series B*, 58, 267–288.

Tang, B. (1993). Orthogonal array-based Latin hypercubes. *Journal of the American Statistical Association*, 88, 1392–1397.

Welch, W. J., Buck, R. J., Sacks, J., Wynn, H. P., Mitchell, T. J., and Morris, M. D. (1992). Screening, predicting, and computer experiments. *Technometrics*, 34, 15–25.

White, H. (1996). *Estimation, inference and specification analysis*, Cambridge university press, New York.

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. *Annals of Statistics*, 21, 1567–1590.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38, 894–942.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99, 250–261.

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

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36, 15091533.