Statistica Sinica **21** (2011), 1171-1190

PENALIZED BLIND KRIGING IN COMPUTER EXPERIMENTS

Ying Hung

Rutgers University

Abstract: Kriging models are popular in analyzing computer experiments. The most widely used kriging models apply a constant mean to capture the overall trend. This method can lead to a poor prediction when strong trends exist. To tackle this problem, a new modeling method is proposed, which incorporates a variable selection mechanism into kriging via a penalty function. An efficient algorithm is introduced and oracle properties in terms of selecting the correct mean function are derived according to fixed-domain asymptotics. The finite-sample performance is examined via a simulation study. Application of the proposed methodology to circuit-simulation experiments demonstrates a remarkable improvement in prediction, and the capability of identifying variables that most affect the system.

Key words and phrases: Computer model, fixed-domain asymptotics, Gaussian process model, geostatistics, oracle procedure, variable selection.

1. Introduction

Physical experimentations can be intensive in terms of material, time, and cost. The advent of modern computers and the advance of computer-aided design methodologies have given rise to another, more economical mode of experimentation. Computer experiments are done on computers using physical models and finite-element-based methods. The analysis of computer experiments has received a lot of attention in the past decades (Kennedy and O'Hagan (2000); Oakley and O'Hagan (2002); Higdon et al. (2008)).

Computer experiments have deterministic outputs, that is, replicates of the same inputs to the computer code produce identical outputs. To take this into account, Sacks et al. (1989a); Sacks, Schiller and Welch (1989b) proposed to model the deterministic response as a realization of a stochastic process. This approach is desirable because it can provide estimates of uncertainty of predictions. In this work, we focus specifically on modeling computer outputs as a Gaussian process (GP) with a spatial correlation structure. This is known as *kriging* and is particularly attractive because it can fit a large class of response surfaces and is widely used in the computer experiment literature (Santner, Williams, and Notz (2003); Fang, Li, and Sudjianto (2006); Linkletter et al. (2006)). Moreover,

several empirical studies have proved its superiority over the other interpolating techniques (Laslett (1994)).

A kriging model incorporates two parts: one is a stationary Gaussian process and the other is a mean function. Most of the kriging models used for analyzing computer experiments either assume some known important trends/variables in the mean function, known as *universal kriging* (UK), or use a grand mean to capture the overall trend, known as *ordinary kriging* (OK). Due to the fact that the important variables are rarely known before experiments, OK is commonly used in practice. It has been observed that the OK prediction can be poor if there are strong trends (Martin and Simpson (2005)) and prediction accuracy can be improved by selecting variables properly into the mean function (Joseph, Hung, and Sudjianto (2008)). This issue is important, but has not received much attention in the literature. Even though some studies have pointed out the benefits of using more complex mean functions (Martin and Simpson (2005); Qian et al. (2006)), there is no systematic methodology for obtaining them. A recent approach, *blind kriging* (Joseph, Hung, and Sudjianto (2008)), integrates a Bayesian forward selection procedure into the kriging model. This effectively reduces the prediction error and demonstrates the advantages of combining variable selection techniques with kriging; nevertheless, further improvements are needed for two reasons: the iterative Bayesian estimation is computationally intensive; identifying the correct mean function is an important element in ameliorating prediction accuracy. These are hard problems for the forward selection procedure. New methods with efficient estimation and rigorous theoretical studies of the selection property are called for.

The idea here is to incorporate a variable selection mechanism into kriging via penalized likelihood functions. Although penalized likelihood is a common technique, its utilization in the kriging models to achieve variable selection is new. It poses some challenges in that estimation and inferences with kriging are complicated. New estimation procedures and the corresponding algorithms are required because the standard maximum likelihood methods for kriging are not applicable. Moreover, the classical aymptotic results, such as variable selection consistency, cannot be applied because they are mainly based on independent observations. This assumption is violated in fixed-domain asymptotics for computer experiments (Ying (1993); Stein (1999); Zhang (2004)). Hence, in order to pursue large sample behavior, certain Markovian properties need to be exploited so that the correlations among observations can be taken into account.

There are numerous methods for variable selection in computer experiments. Welch et al. (1992) proposed an algorithm to screen important variables sequentially, and Linkletter et al. (2006) proposed a Bayesian variable selection procedure for Gaussian process models. These methods focus on identifying variables

with significant impact on the process being studied, the main goal in the early stages of experimentation (referred to as *screening* in Wu and Hamada (2000)). Our objective is to enhance the prediction accuracy with the help of variable selection. As well, existing variable selection criteria are based on estimated correlation parameters that appear to be numerically unstable (Li and Sudjianto (2005); Joseph (2006)); selected variables may not be reliable. In contrast, the proposed approach performs variable selection through the mean function, the Gaussian process part is used for interpolation. We note that though the Bayesian variable selection discussed in Linkletter et al. (2006) is easy to implement, it is computationally demanding especially with Gaussian process models.

The remainder of the paper is organized as follows. A modeling method is proposed in Section 2, and the properties of the estimators proposed are discussed in Section 3. An efficient algorithm is introduced in Section 4. In Section 5, simulation studies are carried out to examine the finite-sample performance. Application of the proposed method is illustrated in Section 6 for computer experiments based on a circuit-simulation code. Discussion is given in Section 7.

2. Penalized Blind Kriging

A new methodology for analyzing computer experiments is proposed, motivated by some existing methods. We first review some of the details.

2.1. Kriging preliminaries

The UK model assumes that the true function $y(x)$, $x \in \mathcal{R}^m$, is a realization from a stochastic process

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

where the mean function is $\mu(x) = \sum_{i=0}^{K} \mu_i \omega_i(x)$, the ω_i's are some known trends/variables, and the μ_i's are unknown parameters. Here $Z(x)$ is a weakly stationary Gaussian process with mean 0 and covariance function $\sigma^2\psi$, where $\sigma^2\psi(h) = cov\{Y(x+h), Y(x)\}$ is a positive semidefinite function with $\psi(0) = 1$ and $\psi(-h) = \psi(h)$. In this formulation, μ is used to capture the known trends, so that Z be a stationary process. In reality, however, trends are rarely known, and thus a special case, the OK model, is commonly used,

$$Y(x) = \mu_0 + Z(x). \tag{2.2}$$

Thus, instead of assuming known trends, this model uses a grand mean μ_0 to capture an overall trend. The OK model is one of the most popular methods in analyzing computer experiments (Santner, Williams, and Notz (2003)). It is easy to implement and works well in general, but not taking into account important

variables in the mean function can lead to a poor performance in prediction. Furthermore, including unimportant variables in the mean function can also lessen prediction performance (Joseph, Hung, and Sudjianto (2008)). Therefore, blind kriging (Joseph, Hung, and Sudjianto (2008)) is proposed: instead of assuming the ω_i to be known, they are identified by Bayesian forward selection. Even though blind kriging reduces the prediction error, the Bayesian forward selection is computationally intensive and theoretical properties are hard to obtain in the face of a forward selection procedure that brings stochastic errors from the stages of variable selection.

2.2 Penalized blind kriging

We propose a model in which a variable selection procedure is incorporated in kriging. This is along the line of blind kriging, but the variable selection is achieved by a penalized likelihood. We call it *penalized blind kriging* (PBK). The PBK model is

$$Y(x) = f(x)'\beta + Z(x),$$

where $f(x)' = (1, f_1(x), \ldots, f_p(x))$ are candidate variables and $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$ are the corresponding coefficients. To select the important variables from the candidates, the coefficients β are estimated by maximizing the penalized log-likelihood

$$Q(\beta, \theta, \sigma^2) = L(\beta, \theta, \sigma^2) - \sum_{j=1}^{p} P_\lambda(|\beta_j|), \tag{2.3}$$

where the log-likelihood function can be written as

$$L(\beta, \theta, \sigma^2) = -\frac{1}{2}[N \log \sigma^2 + \log(\det(\Psi(\theta))) + (y - F\beta)^T \Psi(\theta)^{-1} \frac{(y - F\beta)}{\sigma^2}],$$

$y = (y_1, \ldots, y_N)'$ are the computer experiment outputs collected at N design points $\{x_1, \ldots, x_N\}$ with $x_i = (x_i^{(1)}, \ldots, x_i^{(m)})$ for $i = 1, \ldots, N$, $F = (f(x_1), \ldots, f(x_N))'$, $\Psi(\theta)$ is an $N \times N$ matrix with elements $\psi(x_i - x_j)$, θ represents the correlation coefficients involved in the correlation function, and P_λ is a penalty function. More discussion of various penalty functions can be found in Fan and Li (2001) and Zou and Hastie (2005). Here we focus on two popular penalty functions: the Lasso (Donoho and Johnstone (1994); Tibshirani (1996, 1997)), which is a technique used for simultaneous estimation and variable selection, and the adaptive Lasso (Zou (2006)).

Because the penalty term is not a function of θ and σ^2, the maximum likelihood estimate of θ can be obtained as in the universal kriging. That is,

$$\hat{\sigma}^2 = \frac{1}{N}(y - F\hat{\beta})'\Psi(\theta)(y - F\hat{\beta}). \tag{2.4}$$

Substituting the values of $\hat{\beta}$ and $\hat{\sigma}^2$ into (2.3), the maximum likelihood estimate of θ is

$$\hat{\theta} = \arg\max_{\theta}\{Q(\hat{\beta}, \theta, \hat{\sigma}^2)\}. \tag{2.5}$$

It follows that

$$\hat{\theta} = \arg\min_{\theta}\{N \log \hat{\sigma}^2 + \log(\det(\Psi(\theta))) + N\}. \tag{2.6}$$

The PBK predictor, which has the same form as that of the UK predictor (Santner, Williams, and Notz (2003)), is

$$\hat{y}(x) = f(x)'\hat{\beta} + \psi(x)'\Psi(\theta)^{-1}(y - F\hat{\beta}), \tag{2.7}$$

where $\psi(x)' = (\psi(x - x_1), \ldots, \psi(x - x_N))$.

The objective of PBK is to identify important mean functions from a set of candidate functions (or variables). If some simple functions are used in the candidate set, then the predictor can be easily interpreted using the first term $f(x)'\hat{\beta}$. The second term on the right hand side of (2.7) is used to achieve interpolation.

More than a simple extension of existing methods, the PBK models face some challenging tasks. To detect important variables automatically, a penalty function is included in the likelihood that makes the estimation complicated and different from the classical maximum likelihood methods in the GP modeling. Another important issue involves asymptotic properties. Capturing important variables in the mean function is important for prediction accuracy, so rigorous study of selection consistency is required. Existing results mainly focus on independent observations (Fan and Li (2001); Zou (2006)) and cannot be directly applied to the PBK models. New results are needed to account for correlated observations in the PBK models.

3. Asymptotic Properties

We are interested in variable selection properties, and in the asymptotic behavior of parameters such as the correlation parameter θ and variance σ^2. There are two types of asymptotics in spatial statistics: increase domain and fixed-domain asymptotics (Stein (1999)). Increasing domain asymptotics has more data collected by increasing the domain (Mardia and Marshall (1984)) while fixed-domain asymptotics has more data collected by sampling more densely in a fixed domain. We work with fixed-domain asymptotics because they are appropriate for the study of computer experiments, and we hope simulations can reveal how appropriate asymptotic results are in a specific finite-sample setting (Zhang (2004)). More discussion can be found in Ying (1993).

Numerous studies regarding variable selection properties are available in the literature, but they mainly focus on linear or generalized linear models and rely on the independence of observations. Such is not available in fixed-domain asymptotics. Instead, asymptotic behavior is pursued based on carefully constructed martingale difference sequences. The correlation function ψ is assumed to be the power exponential product correlation

$$\psi(h) = \exp(-\sum_{j=1}^{m} \theta_j |h_j|^a), \tag{3.1}$$

often invoked for computer experiments. The parameter a is usually taken as known, it indicates the smoothness of the underlying process. We take $a = 1$ help with Markovian properties (Ying (1993)). Under this assumption, let $\theta = (\theta_1, \ldots, \theta_m)$. The criteria used to examine the optimality of the variable selection procedure associated with the PBK models are known as *oracle properties*, popular criteria introduced by Fan and Li (2001).

Let $\beta = (\beta'_{(1)}, \beta'_{(2)})'$, where $\beta_{(1)} = (\beta_1, \ldots, \beta_q)'$ and $\beta_{(2)} = (\beta_{q+1}, \ldots, \beta_p)'$. Without loss of generality, assume that $\beta_{(2)} = 0$. Furthermore, assume $F'\Psi(\theta)^{-1}F/N \to C$, where C is positive definite and can be written as

$$C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}.$$

Assumption 1. The design points are taken from an experimental region that forms a complete lattice.

Assumption 2. There exists a positive vector η, such that $|C_{21}C_{11}^{-1}\text{sign}(\beta_{(1)})| \leq 1-\eta$, , where 1 is a $p-q$ by 1 vector of 1's, q is the length of $\beta_{(1)}$, and the inequality holds element-wise.

Assumption 3. Let $(v_1, \ldots, v_{p-q})' = (C_{21}C_{11}^{-1}F'(1) - F'(2))/\sqrt{N}$, where $F(1)$ and $F(2)$ denote the first q and the last $p-q$ columns of F. There exist $M_1, M_2 > 0$ so that $||v_i||_2^2 \leq M_1$ and $1'\Psi(\theta)^{-1}1 \leq M_2$.

Assumption 1 adds notational convenience and avoids some technical difficulties, see Ying (1993). For lattice-type designs, correlation functions are easily written with the use of the Kronecker product. Assumptions 2 and 3 are similar to the necessary conditions for Lasso consistency with independent observations (Zou (2006); Zhao and Yu (2006)); they are required for Theorem 1. Proofs are given in the Appendix.

In Theorem 1, we show that the PBK estimates with the Lasso penalty ($P_\lambda(|\beta_j|) = \lambda|\beta_j|$) enjoy the oracle properties, which indicate the consistency in variable selection and the asymptotic normality.

Theorem 1. *Let* $\hat{\beta} = (\hat{\beta}'_{(1)}, \hat{\beta}'_{(2)})'$ *be the PBK estimates based on the Lasso penalty. Suppose* $\lambda \propto N^{(1+\varsigma)/2}$, *where* $1 > \varsigma > 0$. *Under Assumptions* 1 *to* 3, *as* $N \to \infty$;

(i) $\hat{\beta}_{(2)} = 0$ *with probability* 1,

(ii) $\sqrt{n}(\hat{\beta}_{(1)} - \beta_{(1)}) \to_D \mathcal{N}(0, C_{11}^{-1})$.

The adaptive Lasso (Zou (2006)) can be written as $P_\lambda(|\beta_j|) = \lambda\nu_j|\beta_j|$, where $\nu = (\nu_1, \ldots, \nu_p)$ is a known vector of weights. The specification of ν is flexible, more discussion can be found in Zou (2006). Here we consider a weight vector suggested in Zou (2006), $\hat{\nu} = |\hat{\beta}|^{-\gamma}$, where $\gamma > 0$ and $\hat{\beta}$ is a root-n-consistent estimator of β. For the adaptive Lasso, with a proper choice of λ and some regularity condition, the oracle properties hold as follows.

Theorem 2. *Denote the PBK estimates based on the adaptive Lasso penalty as* $\hat{\beta}^* = (\hat{\beta}^{*\prime}_{(1)}, \hat{\beta}^{*\prime}_{(2)})'$. *Suppose that* $\lambda/\sqrt{N} \to 0$ *and* $\lambda N^{(\gamma-1)/2} \to \infty$. *Under Assumption* 1, *as* $N \to \infty$;

(i) $\hat{\beta}^*_{(2)} = 0$ *with probability* 1,

(ii) $\sqrt{n}(\hat{\beta}^*_{(1)} - \beta^*_{(1)}) \to_D \mathcal{N}(0, C_{11}^{-1})$.

Theorem 3 addresses the consistency and asymptotic normality of θ and σ^2 in the PBK models. It is an extension of the results in Ying (1993), where the main focus is the zero mean OK models. The derivation is analogous to that in Ying (1993) and is omitted.

Theorem 3. *Under Assumption* 1, *the PBK estimates for* θ *and* σ^2 *satisfy, as* $N \to \infty$;

$$N^{(m-1)/2m} \begin{pmatrix} \hat{\theta} - \theta \\ \hat{\sigma}^2 - \sigma^2 \end{pmatrix} \to_D \mathcal{N}\left(0, \begin{bmatrix} \tilde{\Sigma}_\theta & b \\ b' & 2\sigma^4 \sum_{i=1}^m (1+\theta_i)^{-1} \end{bmatrix}\right), \tag{3.2}$$

where $\tilde{\Sigma}_\theta = \operatorname{diag}(2\theta_1^2/(1+\theta_1), \ldots, 2\theta_m^2/(1+\theta_m))$, $b = (b_1, \ldots, b_p)'$, *and* $b_i = -2\sigma^2\theta_i/(1+\theta_i)$.

4. New Algorithm

To estimate the parameters in the PBK models a new algorithm, iteratively reweighted least angle regression (IRLARS), is introduced. The idea is simple. For fixed θ and σ^2, the estimation of β can be formulated as a standard linear regression problem with unequal weights $\Psi(\theta)^{-1}/\sigma^2$, which can be easily solved by existing methods, such as the least angle regression algorithm (Efron et al. (2004)). The correlation parameters θ and variance σ^2 can then be estimated

by standard maximum likelihood methods. By iteratively reweighting F and y with respect to the updated correlation matrix $\Psi(\theta)$ and σ^2, the final estimation can be obtained. The computational details are illustrated below; the proof is straightforward and so is omitted. This algorithm is easy to implement and can be modified to take care of other penalty functions.

ALGORITHM: (*IRLARS algorithm for the Lasso penalty*)

Step 0: Set up initial values for $\hat{\beta}^{(0)}$, $\hat{\theta}^{(0)}$, and $\hat{\sigma}^{2(0)}$.

Step 1: With $\Psi(\hat{\theta}^{(l)})^{-1}/\hat{\sigma}^{2(l)} = R'R$, $y^* = Ry$, and $F^* = RF$, solve the Lasso problem for all λ,

$$\hat{\beta}^{(l+1)} = \arg\min||y^* - F^*\beta||^2 + \lambda\sum_{j=1}^{p}|\beta_j|.$$

Step 2: Apply the estimated $\hat{\beta}^{(l+1)}$ in Step 1 and solve θ and σ^2 by maximizing the likelihood as described in (2.4) and (2.6). Thus

$$\hat{\theta}^{(l+1)} = \arg\min_{\theta}\{N\log\hat{\sigma}^{2(l+1)} + \log(\det(\Psi(\theta)))\},$$

where

$$\hat{\sigma}^{2(l+1)} = \frac{1}{N}(y - F\hat{\beta}^{(l+1)})'\Psi(\theta)(y - F\hat{\beta}^{(l+1)}).$$

Step 3: If the convergence is achieved, declare $\hat{\beta}$, $\hat{\theta}$, and $\hat{\sigma}^2$ to be the estimates. Otherwise, return to Step 1.

To estimate the parameters with the adaptive Lasso penalty, Zou (2006) proposed a modification of the original LARS algorithm. The IRLARS algorithm can be similarly extended to the adaptive Lasso penalty by replacing Step 1 with Step 1*, as follows.

Step 1*: With $\Psi(\hat{\theta}^{(l)})^{-1}/\hat{\sigma}^{2(l)} = R'R$, $y^* = Ry$, $F^* = RF$, and $F^{**} = F^*/\nu = RF/\nu$, solve the Lasso problem for all λ,

$$\hat{\beta}^* = \text{argmin}||y^* - F^{**}\beta||^2 + \lambda\sum_{j=1}^{p}|\beta_j|.$$

Update $\hat{\beta}_j^{(l+1)} = \hat{\beta}_j^*/\nu_j$.

The algorithm can be generally used to estimate parameters in PBK models. Although only the Lasso and adaptive Lasso are illustrated in this study, the algorithm can be further extended and implemented to other penalty functions by modifying Step 1. In optimization, such an iterative algorithm is called a block

coordinate descent or nonlinear Gaussian-Seidel method (Bertsekas (1999)), and it is known that this type of algorithm converges under mild conditions. In particular, the optimization in Step 2 can be solved by standard nonlinear programming methods, such as quasi-Newton algorithms. When the dimensionality of θ is large, the result may be trapped in some local optimal solutions. To avoid this, multiple initial settings should be considered and further improvement can be achieved by combining the ideas in heuristic search (Aarts and Lenstra (2003)), such as simulated annealing (Kirkpatrick, Gelatt, and Vecchi (1983)) and genetic algorithm. Another important issue in practice is tuning, including the specification of λ, and sometimes the parameters involved in the definition of the weights vector ν. We suggest using the cross validation method along with the LARS algorithm to search for the optimal setting.

5. Finite-sample Performance and Empirical Application

In this section, the finite-sample performance of the PBK model is studied based on a known function, and is used in a circuit simulation experiment.

5.1. A known function

To evaluate the finite-sample performance of the PBK model, we first consider a known function. Even though it is not in a standard form for computer experiment models, it provides a better way to assess the performance of variable selection. Performance is evaluated in two aspects: the accuracy of variable selection and the size of prediction errors. The accuracy of variable selection is measured on two scores: the average number of the relevant variables correctly identified in the repeated simulations, the average number of the irrelevant variables misspecified. Prediction accuracy is measured by root mean square prediction errors calculated based on the randomly generated testing data. The PBK models with the Lasso and adaptive Lasso are illustrated and the results are compared with those according to the OK and UK models. To see the effect of sample size, simulations were conducted for different N given a fixed number of variables.

The known function is defined on a twelve-dimensional ($m = 12$) input space $[0, 1]^{12}$ where the first six variables, $(x^{(1)}, \ldots, x^{(6)})$, have decreasing effects on the computer experiment outputs, and the remaining variables, $(x^{(7)}, \ldots, x^{(12)})$, are irrelevant (i.e. zero coefficients) to the output. The function in question is

$$y(x) = 0.4x^{(1)} + 0.3x^{(2)} + 0.2x^{(3)} + 0.1x^{(4)} + 0.05x^{(5)} + 0.01x^{(6)} + e, \qquad (5.1)$$

where $e \sim \mathcal{N}(0, \sigma_e^2)$ and $\sigma_e = 0.05$. Response were generated independently using (5.1) and the experimental designs were Latin hypercube designs

Table 1. Comparison based on a known function.

	ACI (ACI/6)			AMC (AMC/6)		
Methods	$N = 50$	$N = 80$	$N = 100$	$N = 50$	$N = 80$	$N = 100$
PBK.L	4.49	4.50	4.51	0.60	0.22	0.05
	(0.7479)	(0.7500)	(0.7511)	(0.1002)	(0.0372)	(0.0089)
PBK.ada	4.49	4.52	4.52	0.57	0.27	0.05
	(0.7481)	(0.7528)	(0.7539)	(0.0954)	(0.0445)	(0.0078)
UK	6	6	6	6	6	5.98
	(1)	(1)	(1)	(1)	(1)	(0.9961)

(McKay, Beckman, and Conover (1979)) with 12 variables and sample sizes $N = 50, 80$, and 100. Latin hypercube designs are a popular choice for computer experiments because they can be generated with minimal computational effort and fill the design space relatively well. The four methods were used to analyze the simulation outputs and, for each fitted model, root mean square prediction errors (RMSPE) was calculated according to 100 randomly generated testing data.

Based on 500 iterations, the variable selection performances of the PBK models with the Lasso and adaptive Lasso are shown in Table 1. PBK.L and PBK.ada denote the results, respectively. Note that the vector ν used for the adaptive Lasso penalty was defined according to a suggestion in Zou (2006): $\nu = |\hat{\beta}_{ols}|^{-1}$. The column "ACI" provides the average number of variables correctly identified (i.e. the average number of estimated non-zero coefficients for $x^{(1)}, \ldots, x^{(6)}$) and the values in the parentheses, ACI/6, give the correct identification rate. The column "AMC" gives the average number of variables misspecified (i.e. average number of estimated non-zero coefficients for $x^{(7)}, \ldots, x^{(12)}$) and "AMC/6" states the variable misspecification rate. These performance measures are evaluated for all three sample sizes and the results are compared with UK, where all the 12 variables are considered in the mean function (i.e. $K = 12$, $\omega_1 = x^{(1)}, \ldots, \omega_{12} = x^{(12)}$). For the sake of comparison, estimated coefficients in UK were truncated to 0 if they were smaller than 0.001. Note that, the OK models are not included herein for the comparison of variable selection as they use an overall mean in the mean function without any selected variables.

In general, both PBK models performed well in terms of identifying the correct variables in the mean function. The average number of variables identified from $x^{(1)}, \ldots, x^{(6)}$ was 4.49 and the average number of misspecified variables was lower than 0.6. The misclassification rate was effectively reduced with an increase in sample size. Specifically, when the sample size was increased from 50 to 100, the misclassification rate was reduced by more than 91% ($= (0.57 - 0.05)/0.57$). The effect of the sample size was somewhat smaller on the correct identification rates. This is not surprising because the last two variables have relatively small

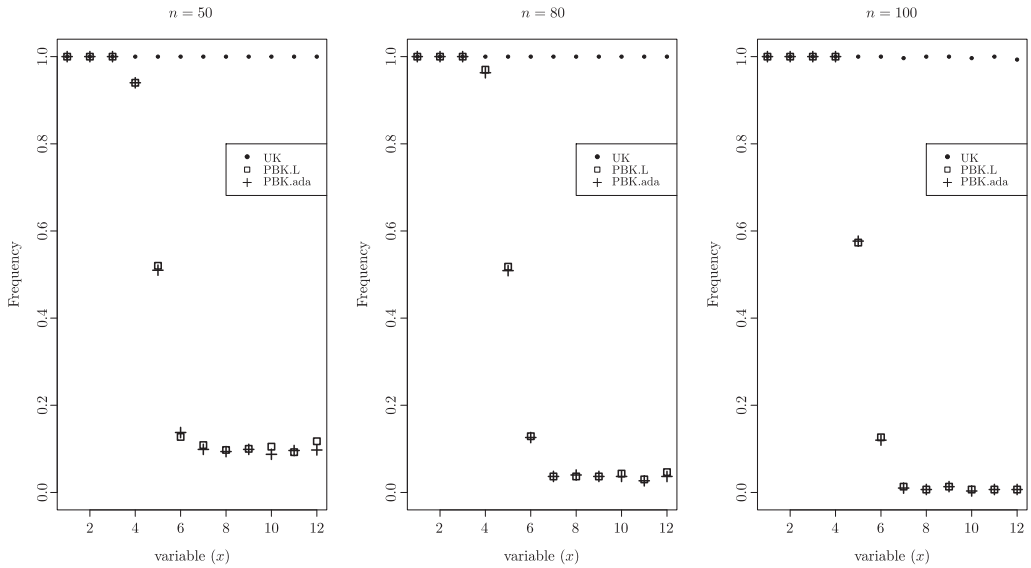

Figure 1. Comparison of individual variable frequency.

effects (0.05 and 0.01) and thus they are difficult to detect. On the other hand, both correct identification rates and misspecification rates of the UK models were close to 1; irrelevant variables could not be distinguished from important variables. Moreover, the selection property of UK was not sufficiently ameliorated with increased sample size. The PBK models appeal favorable with respect to UK in detecting important variables.

To further assess the fitted PBK models in terms of variable selection accuracy, the frequencies of individual variables identified (with non-zero estimated coefficients) from 500 simulations are plotted in Figure 1. The results from three sample sizes, $n = 50$, $n = 80$, and $n = 100$, are considered. The PBK models with both penalty functions performed similarly. For the first six variables, their identification frequencies were decreasing as expected. Among the six variables, the top three were successfully detected (with frequency 1) even for the smaller sample sizes. The frequency of detecting the fourth variable increased from 0.94 to 1 with the sample size increased from 50 to 100, while relatively smaller increases were seen for the fifth and six variables. For the last sixth variables misspecification frequencies were low and become even lower when the sample size increased. For the UK models, Figure 1 shows that all candidate variables were specified to be important. These misleading results went uncorrected with increased sample size.

The RMSPEs based on the 500 simulations are summarized in Table 2 to assess the prediction accuracy of the fitted models. The values in the parentheses

Table 2. Compare RMSPE based on a known function.

Methods	$N = 50$	$N = 80$	$N = 100$
PBK.L	0.2133(0.0065)	0.2108(0.0062)	0.2084(0.0062)
PBK.ada	0.2115(0.0063)	0.2113(0.0063)	0.2092(0.0061)
OK	0.2237(0.0065)	0.2220(0.0062)	0.2200(0.0061)
UK	0.2282(0.0067)	0.2266(0.0061)	0.2238(0.0061)

are the corresponding standard errors. The results are compared with the OK and UK predictors. Table 2 indicates that the PBK models consistently outperformed the OK and UK predictors in terms of RMSPE. In particular, the RMSPE for the UK predictors were up to 8%(= $(0.2282 - 0.2115)/0.2115$) larger than the PBK predictors. Note that the RMSPEs for UK were in general higher than those for the OK and PBK models. Thus, including unimportant variables in the mean function can worsen the prediction performance.

5.2. Circuit simulation

The proposed method is illustrated on the circuit-simulation code (Sacks et al. (1989a)). There are six (transistor widths) inputs in this experiment and the response is a clock asynchronization or "skew". A total of 32 runs of the simulation experiments were conducted. This data set is also analyzed by other methods in the literature with the nature of relationship between skew and variables $x^{(1)}, \ldots, x^{(6)}$ well-understood (Welch et al. (1992)); therefore, it can be useful for illustrating the accuracy of variable selection and prediction. In order to compare the performance of different methods, we randomly split the data set into two parts: half of the data (16 runs) were used for training, the other half for testing.

To predict the clock skew as a function of the six input variables, the PBK models with two penalty functions were applied, assuming the candidates in the mean function to be $f(x)' = (1, x^{(1)}, \ldots, x^{(6)})$. With the Lasso penalty, the fitted PBK model can be written as

$$\hat{y}(x) = f(x)'\hat{\beta} + \psi(x)'\Psi^{-1}(y - F\hat{\beta}), \tag{5.2}$$

where $F = (f(x_1), \ldots, f(x_{16}))'$, $\hat{\theta} = (0.01, 0.01, 0.01, 3.21, 0.01, 1.75)$, $\hat{\sigma}^2 = 0.04$, the coefficients estimated by (2.3) were $\hat{\beta} = (-0.77, \beta^{PBK.L})'$, with $\beta^{PBK.L}$ listed in Table 3 for further comparison. When the penalty function is the adaptive Lasso, the fitted PBK model was similar to (5.2) with estimated parameters $\hat{\sigma}^2 = 0.14$, and $\hat{\beta} = (-0.79, \beta^{PBK.ada})'$, where $\beta^{PBK.ada}$ can also be found in Table 3. The correlation parameters were the same as the estimates in PBK with the Lasso penalty.

Table 3. Compare the estimated coefficients.

Methods	x_1	x_2	x_3	x_4	x_5	x_6
$\beta^{PBK.L}$	0	0.39	−0.58	0.19	0.63	−0.61
$\beta^{PBK.ada}$	0	0.65	−0.64	0	0.70	−0.67
β^{UK}	0.19	0.39	−0.64	0.10	0.66	−0.53

In addition, the OK and UK models were fitted. The OK predictor based on (2.2) can be written as $\hat{y}(x) = \hat{\mu}_0 + \psi(x)'\Psi^{-1}(y - \hat{\mu}_0)$, where $\hat{\mu}_0 = (1'\Psi^{-1}1)^{-1}1'\Psi^{-1}y = -0.77$, $\hat{\theta} = (0.02, 0.07, 0.12, 0.06, 0.10, 0.18)$ and $\hat{\sigma}^2 = 0.70$. If prior knowledge were available to suggest that all six variables have strong linear effects on the response, the fitted UK model should include these in the mean function. Therefore, we have the fitted UK predictor as $\hat{y}(x) = f(x)'\hat{\beta} + \psi(x)'\Psi^{-1}(y - F\hat{\beta})$, where $\hat{\sigma}^2 = 0.02$, $\hat{\beta} = (F'\Psi^{-1}F)^{-1}F'\Psi^{-1}y = (-0.77, \beta^{UK})'$, and $\hat{\theta}$ is the same as the estimate in the PBK models.

The fitted coefficients from the foregoing methods (excluding the OK model) are listed in Table 3, by which the performance of variable selection in the mean function can be compared. According to the fitted coefficients, the PBK predictor with the Lasso penalty identified five important variables (out of six candidates) in the mean function by removing $x^{(1)}$. This result is similar to that found in Welch et al. (1992), where a sequential variable selection algorithm was performed and five variables were identified in the order $x^{(5)}, x^{(3)}, x^{(6)}, x^{(2)}, x^{(4)}$. A slight difference is that the orders of $x^{(3)}$ and $x^{(6)}$ were exchanged if we rank the estimated coefficients. This is reasonable because their effects were not significantly different from each other based on the sensitivity plot in Welch et al. (1992). With the adaptive Lasso, the PBK model identified four variables in the mean function by removing $x^{(1)}$ and $x^{(4)}$, the least important one among the five selected variables in Welch et al. (1992). In contrast, the UK model has non-zero coefficients for all of the candidates which means no variable is removed from the mean function.

With the help of selecting important variables in the mean function, the main objective of the PBK model is to increase prediction accuracy. Hence, RMSPEs for the testing data are summarized and compared in Table 4. It appears that the PBK predictor with the adaptive Lasso performed slightly better than with the Lasso penalty in this example and both OK and UK predictors gave larger RMSPEs compared with the PBK predictors. In particular, the OK predictor was 27%(= (0.123−0.097)/0.097) and 29%(= (0.123−0.095)/0.095) larger in RMSPE than the two PBK predictors. Due to the fact that most of the variables (except the first variable) have significant effects on the response , the UK predictor would be expected to work reasonably well in this example, but had larger RMSPE than the PBK predictor.

Table 4. Compare RMSPE.

Methods	Number of variables in mean function	RMSPE
PGP with Lasso	5	0.097
PGP with adaLasso	4	0.095
OK	0	0.123
UK	6	0.101

6. Discussion

The naive two-stage method of estimating the mean function by various variable selection methods in the first step and then using them as a known trend in the UK models, does not work well in general; the performance of GP model is quite sensitive to the choice of mean function. The proposed approach uses penalty functions to identify important variables and includes the constant mean as a special case. If the constant mean in the GP model is the optimal predictor, this cannot be detected in the naive two-stage method. The PBK method has advantages.

An interesting extension of the penalized blind kriging would select important variables based on both the mean function and the Gaussian process. This requires a careful study of the relationship between the mean functions and the Gaussian process. Research on this topic is currently ongoing and will be reported elsewhere.

Acknowledgement

The research was supported by U.S. NSF grants DMS 0905753 and CMMI 0927572. The author thanks the Editor, an associate editor, and two referees for their helpful comments and suggestions.

Appendix

Several lemmas culminate in the proof of Theorem 1.

Lemma 1. *Let g and h are $N \times 1$ vectors with elements denoted by $g_{d_1,\ldots,d_m}$ and $h_{d_1,\ldots,d_m}$, where $1 \le d_i \le n$, for all $i = 1, \ldots m$, and $N = n^m$. From Assumption 1, the N experimental points $(x_1, \ldots, x_N)$ in m-dimensional space can be rewritten as $\{(s^1_{d_1}, \ldots, s^m_{d_m}) : 1 \le d_i \le n, 1 \le i \le m\}$ and the elements in the $N \times N$ correlation function $\Psi(\theta)$ can be rewritten as $\exp(-\sum_{i=1}^m (\theta_i |s^i_k - s^i_l|))$, where $1 \le k, l \le n$ and $\theta_i > 0$. For simplicity, assume $s^i_k - s^i_{k-1} = \xi$ for all i. Then*

$$
\begin{aligned}
&g'\Psi(\theta)^{-1}h = g'(\otimes_{i=1}^{m}\Psi_i)^{-1}h \\
&= g_{1,\ldots,1}h_{1,\ldots,1} + \sum_{d_1=2}^{n}\left[\frac{(g_{d_1,1,\ldots,1}-\exp(-\theta_1\xi)g_{d_1-1,1,\ldots,1})(h_{d_1,1,\ldots,1}-\exp(-\theta_1\xi)h_{d_1-1,1,\ldots,1})}{1-\exp(-2\theta_1\xi)}\right. \\
&\quad \left. + \ldots + \frac{(g_{1,\ldots,1,d_1}-\exp(-\theta_m\xi)g_{1,\ldots,1,d_1-1})(h_{1,\ldots,1,d_1}-\exp(-\theta_p\xi)h_{1,\ldots,1,d_1-1})}{1-\exp(-2\theta_m\xi)}\right] \\
&\quad + \sum_{d_1=2}^{n}\sum_{d_2=2}^{n}\left[\frac{(g^{(1)}_{d_1,d_2,1,\ldots,1}-\exp(-\theta_2\xi)g^{(1)}_{d_1,d_2-1,1,\ldots,1})(h^{(1)}_{d_1,d_2,1,\ldots,1}-\exp(-\theta_2\xi)h^{(1)}_{d_1,d_2-1,1,\ldots,1})}{(1-\exp(-2\theta_1\xi))(1-\exp(-2\theta_2\xi))}\right. \\
&\quad \left. + \ldots + \frac{(g^{(1)}_{1,\ldots,1,d_1,d_2}-\exp(-\theta_m\xi)g^{(1)}_{1,\ldots,1,d_1,d_2-1})(h^{(1)}_{1,\ldots,1,d_1,d_2}-\exp(-\theta_m\xi)h_{1,\ldots,1,d_1,d_2-1})}{(1-\exp(-2\theta_{m-1}\xi))(1-\exp(-2\theta_m\xi))}\right] \\
&\quad + \sum_{d_1=2}^{n}\cdots\sum_{d_m=2}^{n}\frac{(g^{(m-1)}_{d_1,d_2,\ldots,d_m}-\exp(-\theta_m\xi)g^{(m-1)}_{d_1,d_2,\ldots,d_m-1})(h^{(m-1)}_{d_1,d_2,\ldots,d_m}-\exp(-\theta_m\xi)h^{(m-1)}_{d_1,d_2,\ldots,d_m-1})}{(1-\exp(-2\theta_1\xi))\ldots(1-\exp(-2\theta_m\xi))},
\end{aligned}
\tag{A.1}
$$

where $\otimes$ *denotes the Kronecker product,* Ψ_i *is a* $n \times n$ *matrix with elements* $\exp(-(\theta_i|s_k^i - s_l^i|))$, $1 \le k, l \le n$,

$$
\begin{aligned}
g^{(1)}_{d_1,\ldots,d_m} &= (g_{d_1,\ldots,d_m} - \exp(-\theta_1\xi)g_{d_1-1,\ldots,d_m}), \\
&\vdots \\
g^{(m-1)}_{d_1,\ldots,d_m} &= (g^{(m-2)}_{d_1,\ldots,d_{m-1},d_m} - \exp(-\theta_{m-1}\xi)g^{(m-2)}_{d_1,\ldots,d_{m-1}-1,d_m}),
\end{aligned}
$$

and similar definitions apply to $h^{(1)}_{d_1,\ldots,d_m}, \ldots, h^{(m-1)}_{d_1,\ldots,d_m}$.

Proof of Lemma 1. The result is an extension of the two-dimensional case in Ying (1993). It can be derived by mathematical induction and the detail of the proof is analogous to that in Ying (1993).

Lemma 2. *Under Assumption* 2,

$$
P(\hat{\beta}_{(2)} = 0) \ge 1 - P\left(\left|(C_{21}C_{11}^{-1}F'(1) - F'(2))\Psi(\theta)^{-1}\frac{y - F\beta}{\sqrt{N}}\right| \ge \frac{\lambda}{\sqrt{N}}\eta\right).
$$

Proof of Lemma 2. The penalized log likelihood estimators is

$$
\begin{aligned}
\hat{\beta} &= \arg\min_\beta \tfrac{1}{2}[\log(\det(\Psi(\theta))) + (y - F\beta)'\Psi(\theta)^{-1}(y - F\beta)] + \lambda\textstyle\sum_{i=1}^{p}|\beta_i| \\
&= \arg\min_\beta\Big(NL(\beta) + \lambda\textstyle\sum_{i=1}^{p}|\beta_i|\Big).
\end{aligned}
$$

If $u = \hat{\beta} - \beta$, we have

$$
\hat{u} = \arg\min_u NL(\beta + u) + \lambda||\beta + u||_1. \tag{A.2}
$$

Applying a Taylor expansion, the first term on the right hand side can be written as

$$
\begin{aligned}
NL(\beta + u) &= NL(\beta) - u'F'\Psi(\theta)^{-1}(y - F\beta) + \frac{1}{2}u'F'\Psi(\theta)^{-1}Fu \\
&= NL(\beta) - u'F'\Psi(\theta)^{-1}(y - F\beta) + \frac{(\sqrt{N}u)'C(\sqrt{N}u)}{2},
\end{aligned}
\tag{A.3}
$$

If there exists $\hat{u}$, the following result holds, based on the Karush-Kuhn-Tucker (KKT) optimality condition,

$$C_{11}\sqrt{N}\hat{u}_{(1)} - \frac{F(1)'\Psi(\theta)^{-1}(y-F\beta)}{\sqrt{N}} = -\frac{\lambda}{\sqrt{N}}\text{sign}(\beta_{(1)}), \tag{A.4}$$

$$\left|C_{21}\left(\sqrt{N}\hat{u}_{(1)} - \frac{F(2)'\Psi(\theta)^{-1}(y-F\beta)}{\sqrt{N}}\right)\right| \leq \frac{\lambda}{\sqrt{N}}1. \tag{A.5}$$

From (A.4) and (A.5), we have

$$\left|\left(C_{21}C_{11}^{-1}F(1)' - F(2)'\right)\frac{\Psi(\theta)^{-1}(y-F\beta)}{\sqrt{N}}\right| \leq \frac{\lambda}{\sqrt{N}}\left(1 - |C_{21}C_{11}^{-1}\text{sign}(\beta_{(1)})|\right). \tag{A.6}$$

Thus, Lemma 2 holds by Assumption 2.

Proof of Theorem 1. We first focus on consistency. Based on Lemma 2,

$$\begin{aligned} P(\hat{\beta}_{(2)} = 0) &\geq 1 - P\left(\left|(C_{21}C_{11}^{-1}F(1)' - F(2)')\frac{\Psi(\theta)^{-1}(y-F\beta)}{\sqrt{N}}\right| \geq \frac{\lambda}{\sqrt{N}}\eta\right) \\ &\geq 1 - \sum_{i=1}^{p-q} P\left(\left|v_i'\Psi(\theta)^{-1}(y-F\beta)\right| \geq \frac{\lambda}{\sqrt{N}}\eta_i\right), \end{aligned} \tag{A.7}$$

where $v_1, \ldots, v_{p-q}$ are defined in Assumption 3. Next, we need to show that

$$v_i'\Psi(\theta)^{-1}(y-F\beta) \to_D \mathcal{N}(0, v_i'\Psi(\theta)^{-1}v_i). \tag{A.8}$$

For notational convenience, the case when $m = 2$ is illustrated, namely $\Psi(\theta) = \Psi_1 \otimes \Psi_2$. The result is easily extended to higher dimensions. Define the $N \times 1$ vector $(y - F\beta)$ by $(z_{1,1}, \ldots, z_{n,n})'$, where $N = n^2$, and denote the $N \times 1$ vector v_i by $\{(w_{k,l}) : 1 \leq k, l, \leq n\}$. Based on Lemma 2, we have

$$\begin{aligned} &v_i'(\Psi_1 \otimes \Psi_2)^{-1}(y-F\beta) \\ &\quad = w_{1,1}z_{1,1} + \sum_{j=2}^{n}\Bigg[\frac{(w_{1,j} - \exp(-\theta_1\xi)w_{1,j-1})(z_{1,j} - \exp(-\theta_1\xi)z_{1,j-1})}{1 - \exp(-2\theta_1\xi)} \\ &\qquad + \frac{(w_{j,1} - \exp(-\theta_2\xi)w_{j-1,1})(z_{j,1} - \exp(-\theta_2\xi)z_{j-1,1})}{1 - \exp(-2\theta_2\xi)} \\ &\qquad + \sum_{l=2}^{n}\frac{(w_{j,l}^{(1)} - \exp(-\theta_2\xi)w_{j,l-1}^{(1)})(z_{j,l}^{(1)} - \exp(-\theta_2\xi)z_{j,l-1}^{(1)})}{(1 - \exp(-2\theta_1\xi))(1 - \exp(-2\theta_2\xi))}\Bigg] \\ &\quad = w_{1,1}z_{11} + \sum_{j=2}^{n}\tau_j, \end{aligned} \tag{A.9}$$

where $z_{j,l}^{(1)} = (z_{j,l} - \exp(-\theta_1\xi)z_{j-1,l})$ and $w_{j,l}^{(1)} = (w_{j,l} - \exp(-\theta_1\xi)w_{j-1,l})$. It is not difficulty to see that $\{\tau_1, \ldots, \tau_n\}$ is a martingale difference sequence with respect to the σ-filtration $\mathcal{F}_j = \sigma\{z_{i,l}, l \le j, i = 1, \ldots n\}$, and the following results hold.

$$\sum_{j=1}^{n} E(\tau_j^2|\mathcal{F}_{j-1}) \to_{\mathcal{P}} \sum_{j=2}^{n} \Bigg[\frac{(w_{1,j} - \exp(-\theta_1\xi)w_{1,j-1})^2}{1 - \exp(-2\theta_1\xi)} + \frac{(w_{j,1} - \exp(-\theta_2\xi)w_{j-1,1})^2}{1 - \exp(-2\theta_2\xi)} + \sum_{l=2}^{n} \frac{(w_{j,l}^{(1)} - \exp(-\theta_2\xi)w_{j,l-1}^{(1)})^2}{(1 - \exp(-2\theta_1\xi))(1 - \exp(-2\theta_2\xi))} \Bigg],$$

$$\sum_{j=1}^{n} E(\tau_j^4|\mathcal{F}_{j-1}) \to_{\mathcal{P}} 0. \tag{A.10}$$

Hence, based on (A.10), Lemma 2, and the Martingale Central Limit Theorem (Pollard (1984)), (A.8) follows. Then apply this to (A.7) to get

$$\sum_{i=1}^{p-q} P\Big(\Big| v_i' \Psi(\theta)^{-1}(y - F\beta) \Big| \ge \frac{\lambda}{\sqrt{N}} \eta_i \Big) = (p-q) O\Big(1 - \Phi(\frac{1}{M_1} \frac{\lambda}{\sqrt{N}} \frac{\min_i \eta_i}{M_2}) \Big).$$

The proof of consistency is complete.

Now we prove the asymptotic normality. Based on Lemma 1 and the argument leading to (A.8), it can be shown that

$$u' F' \Psi(\theta)^{-1}(y - F\beta) \to_D (\sqrt{N}u)' \mathcal{N}(0, C).$$

Therefore, according to (A.2), (A.3), and the argument at Theorem 2 of Zou (2006), we have

$$\sqrt{N}\hat{u}_{(1)} \to_D C_{11}^{-1} \mathcal{N}(0, C_{11}) \text{ and } \sqrt{N}\hat{u}_{(2)} \to_D 0. \tag{A.11}$$

Thus, asymptotic normality holds.

Proof of Theorem 2. The proof is along the lines of Theorem 2 in Zou (2006). Based on the adaptive Lasso penalty, $\hat{u} = \arg\min_u \Upsilon(u)$, where $\Upsilon(u) = NL(\beta + u) + \lambda \sum_{j=1}^{p} \nu_j |\beta_j + u_j|$. By Taylor expansion, we have

$$\Upsilon(u) = \Upsilon(0) - u' F' \Psi(\theta)^{-1}(y - F\beta) + \frac{(\sqrt{N}u)' C (\sqrt{N}u)}{2} + \lambda \sum_{j=1}^{p} \nu_j (|\beta_j + u_j| - |\beta_j|).$$

Borrowing the result in Theorem 2 of Zou (2006), we have

$$\lambda \sum_{j=1}^{p} \nu_j (|\beta_j + u_j| - |\beta_j|) \to_{\mathcal{P}} \begin{cases} 0 & \text{if } \beta_j \ne 0, \\ 0 & \text{if } \beta_j = 0 \text{ and } u_j = 0, \\ \infty & \text{if } \beta_j = 0 \text{ and } u_j \ne 0. \end{cases} \tag{A.12}$$

Then, by Slutsky's Theorem, for every u,

$$\Upsilon(u)-\Upsilon(0) \to_D \begin{cases} -u'_{(1)}F(1)'\Psi(\theta)^{-1}(y-F\beta)+\frac{(\sqrt{N}u_{(1)})'C_{11}(\sqrt{N}u_{(1)})}{2} & \text{if } u_{(2)}=0, \\ \infty & \text{otherwise.} \end{cases}$$

With the same construction of martingale differences in Theorem 1 and (A.8), asymptotic normality holds by the Martingale Central Limit Theorem.

For consistency, it suffices to show that $P(\hat{\beta}_{(2)} \neq 0) \to 0$. Again, using the KKT conditions, it follows that

$$2F(1)'\Psi(\theta)^{-1}(y - F\hat{\beta}) = \lambda\nu_{(1)}, \tag{A.13}$$

where $\nu_{(1)}$ are the weights corresponding to the first q variables. The left hand side of (A.13) can be written as

$$2\frac{F(1)'\Psi(\theta)^{-1}(y - F\hat{\beta})}{\sqrt{N}} = 2\frac{F(1)'\Psi(\theta)^{-1}F(\beta - \hat{\beta})}{\sqrt{N}} + 2\frac{F(1)'\Psi(\theta)^{-1}(y - F\theta)}{\sqrt{N}}.$$

For normal distribution follows as at (A.8), while the right hand side of (A.13) satisfies $\lambda(\nu_{(1)}/\sqrt{N}) \to_{\mathcal{P}} \infty$ (Theorem 2, Zou (2006)). Therefore,

$$P(\hat{\beta}_{(2)} \neq 0) \leq P\left(2F(1)'\Psi(\theta)^{-1}(y - F\hat{\beta}) = \lambda\nu_{(1)}\right) \to 0,$$

and Theorem 2 holds.

References

Aarts, E. and Lenstra, J. K. (2003). *Local Search in Combinatorial Optimization*, Princeton University Press, Princeton, NJ.

Bertsekas, D. P. (1999). *Nonlinear Programming*. Athena Scientific, Belmont, MA.

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

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression (with discussion). *Ann. Statist.* **32**, 407-499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fang, K. T., Li, R. and Sudjianto, A. (2006). *Design and Modeling for Computer Experiments*. CRC Press, New York.

Higdon, D., Gattiker, J., Williams, B. and Rightley, M. (2008). Computer model calibration using high dimensional output. *J. Amer. Statist. Assoc.* **103**, 570-583.

Joseph, V. R. (2006). Limit Kriging. *Technometrics* **48**, 458-466.

Joseph, V. R., Hung, Y. and Sudjianto, A. (2008). Blind Kriging: A new method for developing metamodels. *ASME Journal of Mechanical Design* **130**, 031102-1-8.

Kennedy, M. C. and O'Hagan, A. (2000). Predicting the output from a complex computer code when fast approximations are available. *Biometrika* **87**,1-13.

Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* **220**, 671-680.

Laslett, G. M. (1994). Kriging and splines: An empirical comparison of their predictive performance in some applications. *J. Amer. Statist. Assoc.* **89**, 391-400.

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

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

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

Martin, J. D. and Simpson, T. W. (2005). On the use of Kriging models to approximate deterministic computer models. *AIAA Journal* **43**, 853-863.

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

Oakley, J. E. and O'Hagan, A. (2002). Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika* **89**, 769-784.

Pollard, D. (1984), *Convergence of Stochastic Processes*, Springer, New York.

Qian, Z., Seepersad, C. C., Joseph, V. R., Allen, J. K. and Wu, C. F. J. (2006). Building surrogate models based on detailed and approximate simulations. *ASME J. Mechanical Design* **128**, 668-677.

Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P. (1989a). Design and analysis of computer experiments. *Statistical Science* **4**, 409-423.

Sacks, J., Schiller, S.B. and Welch, W. J. (1989b). Designs for computer experiments. *Technometrics* **31**, 41-47.

Santner, T. J., Williams, B. J. and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*. Springer, New York.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Tibshirani, R. J. (1997). The Lasso method for variable selection in the Cox model. *Statistics in Medicine* **16**, 385-395.

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.

Wu, C. F. J. and Hamada, M. (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization*. Wiley, New York.

Ying, Z.-L. (1993). Maximum likelihood estimation of parameters under a spatial sampling scheme. *Ann. Statist.* **21**, 1567-1590.

Zhang, H. (2004). Inconsistent estimation and asymptotically interpolations in model-based geostatistics. *J. Amer. Statist. Assoc.* **99**, 250-261.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Machine Learning Research* **7**, 2541-2563.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B* **67**, 301-320.

Department of Statistics, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ 08854, USA.

E-mail: yhung@stat.rutgers.edu

(Received September 2009; accepted April 2010)