Statistica Sinica Preprint No: SS-2016-0255R4	
Title	Single Nugget Kriging
Manuscript ID	SS-2016-0255R4
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0255R4
Complete List of Authors	Minyong Lee and Art Owen
Corresponding Author	Minyong Lee
E-mail	minyong@stanford.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

SINGLE NUGGET KRIGING

Minyong R. Lee and Art B. Owen

Stanford University

Abstract: We propose a method with better predictions at extreme values than the standard method of Kriging. We construct our predictor in two ways: by penalizing the mean squared error through conditional bias and by penalizing the conditional likelihood at the target function value. Our prediction exhibits robustness to the model mismatch in the covariance parameters, a desirable feature for computer simulations with a restricted number of data points. Applications on several functions show that our predictor is robust to the non-Gaussianity of the function.

Key words and phrases: Computer experiments, conditional bias, Gaussian process regression.

1 Introduction

In many fields of engineering and science, computer experiments have become an essential tool in studying physical processes. Kriging is a popular way to build metamodels in computer experiments, because Kriging exactly interpolates the experimental data and produces predictions at unobserved inputs. Sacks et al. (1989), Koehler and Owen (1996), Stein (1999), and

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

Switzer (2006) give summaries and in-depth discussions of computer experiments and Kriging.

There are several limitations of Kriging. Kriging prediction depends on the covariance hyperparameters that are usually unknown, and need to be estimated. The variability of the predicted process depends on the hyperparameters, and the likelihood of the hyperparameters, usually computationally expensive to compute, can be quite flat, so that the variance of the maximum likelihood estimator can be huge. Stein (1999) gives asymptotic results showing that there is asymptotically no loss using the misspecified covariance function if the true and assumed covariance functions yield equivalent Gaussian measures. In practice we might face functions that behave quite differently from the assumed covariance structure, and with only a small number of observations. Bachoc (2013) states that the fixed-domain asymptotics does not solve completely the issue of the estimation of the covariance function. There have been several approaches to stabilize the estimation of the hyperparameters, such as penalized Kriging by Li and Sudjianto (2005). We would like to find a predictor that is less affected by the hyperparameters.

Kriging prediction depends on the *mean function* whose specification needs form before looking at the data. In Kriging, there is a "regression

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

effect", in which the predictions are pulled towards the mean function. This can give bad predictions at extreme function values. Using a better design of input points, such as Maximum Projection Designs (Joseph, Gul, and Ba (2015)), can be effective. But, with a small number of observations, finding a good model can be challenging.

There are several approaches to mitigate the regression effect. Composite Gaussian process models (Ba and Joseph (2012)), Fractional Brownian Fields (Zhang and Apley (2014)), and Gaussian process model with Brownian integrated covariance functions (Zhang and Apley (2015)) use flexible covariance functions that capture the nonstationarity of functions. Blind Kriging (Joseph, Hung, and Sudjianto (2008)) incorporates variable selection procedure in the mean function. Limit Kriging (Joseph (2006)) and Kernel Interpolation (Kang and Joseph (2016)) modify the standard Gaussian process regression model or the predictor for more accurate prediction in certain situations. Conditional Bias-Penalized Kriging (CBPK) (Seo (2013)) suggests minimizing the mean squared error plus the squared conditional bias to improve the performance at the extreme values. All these methods, except Limit Kriging and CBPK, introduce more complexity in the model than the stationary Gaussian process model.

In this paper, we propose a new prediction method which we call Single

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

Nugget Kriging (SiNK). We show that SiNK has several desirable properties with the same model complexity and computational cost as standard Kriging. In Section 2, we briefly introduce Kriging. In Section 3, we discuss conditioning the likelihood at the target, a fundamental idea of the SiNK. In Section 4, we define SiNK, and show that it gives smaller mean squared prediction error than usual Kriging when the function value is far from the mean function. In other words, SiNK is robust to misspecifying the mean function. In Section 5, comparison between the performance of SiNK and the performance of usual Kriging and limit Kriging are given in several numerical experiments. Proofs are in the appendix, and additional proofs and details are provided in the supplementary material.

2 Kriging

Kriging treats the deterministic function $f(\mathbf{x})$ as a realization of a real-valued random field

$$Y(\mathbf{x}) = m(\mathbf{x}) + Z(\mathbf{x}) \tag{2.1}$$

where $\mathbf{x} \in \mathbb{R}^d$, $m(\mathbf{x})$ is a deterministic mean function, and $Z(\mathbf{x})$ is a centered (mean zero) square-integrable process with covariance function $K(\cdot, \cdot)$.

For the covariance function, stationary covariance functions that are tensor products of one-dimensional kernels are popular. Let $C_\theta : \mathbb{R} \to$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

$[-1, 1]$ be a covariance kernel with length-scale parameter θ. Let

$$K(\mathbf{x}, \mathbf{z}) = \sigma^2 C(\mathbf{h}) = \sigma^2 \prod_{j=1}^{d} C_{\theta_j}(|h_j|) = \sigma^2 \prod_{j=1}^{d} C_1\left(\frac{|h_j|}{\theta_j}\right)$$

where $\mathbf{h} = \mathbf{x} - \mathbf{z}$. Parameters σ^2 and $(\theta_1, \ldots, \theta_d)$ are usually estimated from the data. Matérn covariance kernels (Matérn (1986)) have the form

$$C_{\nu,\theta}(d) = \frac{(\sqrt{2\nu}\frac{d}{\theta})^\nu}{\Gamma(\nu)2^{\nu-1}} K_\nu\left(\sqrt{2\nu}\frac{d}{\theta}\right)$$

where $K_\nu(\cdot)$ is the modified Bessel function of the second kind. Matérn covariance kernels are commonly used in practice because of the smoothness of the associated Gaussian process, defined in terms of its mean square differentiability parametrized through ν. If there is a measurement error or noise in the function, then adding a nugget effect to the covariance function handles the discontinuity in the function.

Throughout the paper, we only consider deterministic computer experiments and we use the simple Kriging (or ordinary Kriging) model with a known (or estimated) constant mean β, for simplicity. The simplification of the mean function to a constant does not affect predictive performance in general; see Sacks et al. (1989). We assume that the hyperparameters of the covariance function are known (or estimated from the data), and we focus on the prediction at a new point $\mathbf{x}_0$.

Now suppose we observe $\mathbf{y} = (Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_n))$, and let $K = (K_{ij})$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

be the $n \times n$ covariance matrix of $\mathbf{y}$, $k(\mathbf{x}_0, \mathbf{x}_0)$ be the variance of $Y(\mathbf{x}_0)$, and $\mathbf{k}(\mathbf{x}_0)$ be the covariance vector between $\mathbf{y}$ and $Y(\mathbf{x}_0)$. Let $\mathbf{1}$ be the n-length vector of all ones. The Kriging predictor is the Best Linear Predictor (BLP) that minimizes the mean squared prediction error (MSPE) $\mathbb{E}[(Y(\mathbf{x}_0) - \hat{Y}(\mathbf{x}_0))^2]$.

The Kriging predictor can be also derived from the Gaussian process assumption on Y; this approach is called Gaussian process regression. Throughout this paper, we assume that Z in (2.1) is a centered Gaussian process with covariance function $K(\cdot, \cdot)$. Then,

$$Y(\mathbf{x}_0)\big|\ (Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_n)) = \mathbf{y}\ \sim N(m, s^2)$$

where $m = \beta + \mathbf{k}(\mathbf{x}_0)^T K^{-1}(\mathbf{y} - \beta\mathbf{1})$, and $s^2 = k(\mathbf{x}_0, \mathbf{x}_0) - \mathbf{k}(\mathbf{x}_0)^T K^{-1}\mathbf{k}(\mathbf{x}_0)$. The simple Kriging predictor is the conditional mean $\hat{Y}_{\mathrm{K}}(\mathbf{x}_0) = \mathbb{E}[Y(\mathbf{x}_0)\big|\mathbf{y}] = \beta + \mathbf{k}(\mathbf{x}_0)^T K^{-1}(\mathbf{y} - \beta\mathbf{1})$.

With

$$\rho(\mathbf{x}_0) = \sqrt{\frac{\mathbf{k}(\mathbf{x}_0)^T K^{-1}\mathbf{k}(\mathbf{x}_0)}{k(\mathbf{x}_0, \mathbf{x}_0)}},$$

$\rho(\mathbf{x}_0)^2$ is the variance explained by conditioning divided by the marginal variance of y_0. The quantity $\rho(\mathbf{x}_0)$ always lies in $[0, 1]$, and can be understood as the correlation between the target function value and the data.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

3 Conditional likelihood at the target and conditional bias

In this section, we investigate the idea of maximizing the conditional likelihood given the target function value, which is the supporting idea of the SiNK. We also define a class of predictors by generalizing the Conditional Bias-Penalized Kriging.

3.1 Conditional likelihood at the target

We formulate the prediction problem as an estimation problem. From the Gaussian process assumption, the density (which can be also viewed as the *augmented likelihood* in Jones (2001)) of $(Y(\mathbf{x}_0), Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_n))$ is known. Instead of conditioning on the observed function values as in the Gaussian process regression, we condition on the unknown function value at the target point and compute the likelihood. We easily find that

$$(Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_n)) \big| \; Y(\mathbf{x}_0) = y_0 \; \sim N(\tilde{m}, \tilde{K}), \text{ where} \tag{3.1a}$$

$$\tilde{m} = \beta \mathbf{1} + k(\mathbf{x}_0, \mathbf{x}_0)^{-1}(y_0 - \beta)\mathbf{k}(\mathbf{x}_0) \text{ and} \tag{3.1b}$$

$$\tilde{K} = K - k(\mathbf{x}_0, \mathbf{x}_0)^{-1}\mathbf{k}(\mathbf{x}_0)\mathbf{k}(\mathbf{x}_0)^T. \tag{3.1c}$$

Thus, we reverse the perspective by seeing y_0 as a parameter. The conditional log likelihood is

$$l(y_0|\mathbf{y}) = l(y_0) = -\frac{1}{2}(\mathbf{y} - \tilde{m})^T \tilde{K}^{-1}(\mathbf{y} - \tilde{m}) + \text{constant}. \tag{3.2}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

The maximizer of the conditional likelihood with respect to y_0 with penalty $-(y_0-\beta)^2/(2k(\mathbf{x}_0,\mathbf{x}_0))$, which is the *maximum a posteriori* estimate of y_0 with the prior distribution $y_0 \sim N(\beta, k(\mathbf{x}_0,\mathbf{x}_0))$, is the simple Kriging predictor. However, when $\mathbf{k}(\mathbf{x}_0) \neq 0$, the maximizer of the conditional likelihood without penalty (CMLE) exists and it is

$$\hat{Y}_{\mathrm{CMLE}}(\mathbf{x}_0) = \beta + \frac{k(\mathbf{x}_0,\mathbf{x}_0)}{\mathbf{k}(\mathbf{x}_0)^T K^{-1}\mathbf{k}(\mathbf{x}_0)}\mathbf{k}(\mathbf{x}_0)^T K^{-1}(\mathbf{y}-\beta\mathbf{1}). \tag{3.3}$$

The derivation is in the supplementary material, section S1. The CMLE is obtained by inflating the residual term of the simple Kriging predictor by $1/\rho(\mathbf{x}_0)^2$.

3.2 Conditional Bias

In the geostatistical literature, there are two types of conditional bias (Katz and Murphy (1997), Seo (2013)). *Type 1 conditional bias* is defined as $\mathbb{E}[Y(\mathbf{x}_0)|\hat{Y}(\mathbf{x}_0)=\hat{y}]-\hat{y}$, which measures the degree of correspondence between the mean of the unknown function value given a particular prediction. This quantity has been used to measure the reliability of the forecast in geostatistics. For example, the simple Kriging predictor is type 1 conditionally unbiased. Type 1 conditionally unbiased predictors have been discussed with an interest in the issue of predicting tails better (Isaaks (2005), David, Marcotte, and Soulie (1984)).

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

Type 2 conditional bias is defined as $\mathbb{E}[\hat{Y}(\mathbf{x}_0)|Y(\mathbf{x}_0) = y_0] - y_0$, which is computed by conditioning the true function value. If this bias is large for some y_0, then it means that the prediction could be bad for these y_0. There has not been much discussion on type 2 conditional bias, until Seo (2013) explicitly defined the Conditional Bias-Penalized Kriging, which will be discussed in the following subsection. In this paper, we focus on type 2 conditional bias. Our intuition behind the new suggested predictor is that by reducing type 2 bias by the appropriate amount, it has better performance at the extreme values and it has analogous behavior as the nearest neighborhood regression.

For simple Kriging, we have $\mathbb{E}[\hat{Y}_{\mathrm{K}}(\mathbf{x}_0)\big|Y(\mathbf{x}_0) = y_0] = \beta + \rho(\mathbf{x}_0)^2(y_0 - \beta)$, which is not y_0 in general. Thus, $\hat{Y}_{\mathrm{K}}(\mathbf{x}_0)$ is type 2 conditionally biased in general. We can expect that for a given y_0 that is far from the prior mean, the performance of standard Kriging can be poor.

3.3 Conditional Bias-Penalized Kriging

Conditional Bias-Penalized Kriging (CBPK) is defined as the linear predictor $\hat{Y}(\mathbf{x}_0) = \beta + \lambda^T(\mathbf{y} - \beta\mathbf{1})$ that minimizes the MSPE plus a multiple of squared type 2 conditional bias (CB^2)

$$\mathbb{E}[(y_0 - \hat{Y}(\mathbf{x}_0))^2] + \delta\mathbb{E}[(y_0 - \mathbb{E}[\hat{Y}(\mathbf{x}_0)\big|y_0])^2] \ \text{(for some } \delta \geq 0) \qquad (3.4)$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

with respect to λ. Seo (2013) suggests that we use $\delta = 1$. We show in proposition 1 that using $\delta = 1$ in (3.4) leads to the predictor

$$\hat{Y}_{\mathrm{CBPK}}(\mathbf{x}_0) = \beta + \frac{2}{1+\rho(\mathbf{x}_0)^2}\mathbf{k}(\mathbf{x}_0)^T K^{-1}(\mathbf{y} - \beta\mathbf{1}). \tag{3.5}$$

We observe that it is again a predictor with an inflated residual term. Different choices of δ in (3.4) lead to different predictors. If $\delta = 0$, (3.4) is the objective for simple Kriging, and thus the minimizer $\hat{Y}_{\mathrm{CBPK}}(\mathbf{x}_0)$ is the simple Kriging predictor. If $\delta \to \infty$, the minimizing predictor approaches the CMLE. This accords with the fact that the CMLE is type 2 conditionally unbiased, $\mathbb{E}[\hat{Y}_{\mathrm{CMLE}}(\mathbf{x}_0)\big|Y(\mathbf{x}_0) = y_0] = y_0$.

The main question when using a CBPK is over the ratio to use between MSPE and CB^2. We seek an automatic way to choose δ instead of simply using $\delta = 1$ or applying a cross-validation-style approach. We suggest varying the ratio spatially, using an appropriate function of $\mathbf{x}_0$ as δ. To distinguish from Seo's CBPK ($\delta = 1$), we call the linear predictor $\hat{Y}$ that minimizes (3.4) (for general δ) the *generalized CBPK* predictor. The proof of the following is in the Appendix, Section 1.

Proposition 1. *For any nonnegative δ in* (3.4)*, the generalized CBPK predictor for a constant mean model is of the form*

$$\hat{Y}(\mathbf{x}_0) = \beta + w(\mathbf{x}_0)\mathbf{k}(\mathbf{x}_0)^T K^{-1}(\mathbf{y} - \beta\mathbf{1})$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

where $w(\mathbf{x}_0) \in [1, 1/\rho(\mathbf{x}_0)^2)$ *(if* $\rho(\mathbf{x}_0) = 0$*, then the range is* $[1, \infty)$*). For every nonnegative* δ*, there is a corresponding* $w(\mathbf{x}_0) \in [1, 1/\rho(\mathbf{x}_0)^2)$*.*

In the next section, we focus on a generalized CBPK with specific δ that has desirable properties.

4 Single Nugget Kriging

In this section, we define the Single Nugget Kriging and discuss its properties. Verification of Definition 1 is in the Appendix, Section 2.

4.1 Definition of SiNK

Definition 1. The Single Nugget Kriging (SiNK) predictor is defined as

$$\hat{Y}_{\text{SiNK}}(\mathbf{x}_0) = \begin{cases} \beta + \frac{1}{\rho(\mathbf{x}_0)}\mathbf{k}(\mathbf{x}_0)^T K^{-1}(\mathbf{y} - \beta\mathbf{1}) & \text{if } \rho(\mathbf{x}_0) \neq 0 \\ \beta & \text{otherwise,} \end{cases}$$

which is the maximizer of the conditional likelihood given $Y(\mathbf{x}_0) = y_0$ with penalty

$$pen(y_0) = -\frac{(y_0 - \beta)^2}{2k(\mathbf{x}_0, \mathbf{x}_0)} \frac{\rho(\mathbf{x}_0)}{1 + \rho(\mathbf{x}_0)}.$$

Thus, the implicit prior distribution on y_0 is $y_0 \sim N(\beta, k(\mathbf{x}_0, \mathbf{x}_0)(1 + 1/\rho(\mathbf{x}_0))$.

SiNK is defined as the *maximum a posteriori* estimator with a prior distribution on $Y(\mathbf{x}_0)$. We inflate the prior variance only at $\mathbf{x}_0$ by the

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

amount of uncertainty measured by $\rho(\mathbf{x}_0)$, to reduce the dependency on the prior. This is equivalent to assuming that there exist a nugget effect only on $Y(\mathbf{x}_0)$, so we call the method Single Nugget Kriging. We choose the specific penalty, or prior variance, because it yields two desirable properties which will be discussed in section 4.2.

In the geostatistical literature, the nugget effect is designed to model functions that are discontinuous (Stein (1999)). In our definition, we inflate the prior variance, like a nugget of size $k(\mathbf{x}_0, \mathbf{x}_0)/\rho(\mathbf{x}_0)$ would, but only on $Y(\mathbf{x}_0)$ to introduce the additional uncertainty. This nugget is not from any additional noise assumption, and thus SiNK is also an interpolator like simple Kriging. If one wants to introduce a nugget effect or Gaussian noise to the Gaussian process, the SiNK predictor can be adjusted accordingly.

Remark 1. The SiNK predictor is the CBPK predictor with $\delta = 1/\rho(\mathbf{x}_0)$ in (3.4); it is the linear predictor $\hat{Y}(\mathbf{x}_0) = \beta + \lambda^T(\mathbf{y} - \beta\mathbf{1})$ where λ is the solution of the optimization problem

$$\underset{\lambda}{\text{minimize}} \ \mathbb{E}[(y_0 - \hat{Y}(\mathbf{x}_0))^2] + \frac{1}{\rho(\mathbf{x}_0)}\mathbb{E}[(y_0 - \mathbb{E}[\hat{Y}(\mathbf{x}_0)\big|y_0])^2].$$

This can be verified by plugging in $\delta = 1/\rho(\mathbf{x}_0)$ to (1).

Remark 2. The SiNK prediction can be discontinuous at points where $\rho(\mathbf{x}_0) = 0$, which happens if and only if $\mathbf{k}(\mathbf{x}_0) = 0$. However, $\mathbf{k}(\mathbf{x}_0) = 0$ means that we do not have any information at the point $\mathbf{x}_0$, so predicting

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

with the prior mean is the best we can do. Note that $\mathbf{k}(\mathbf{x}_0) = 0$ could only happen if we use a kernel that has 0 in its range. If we use a strictly positive kernel, such as the Matérn kernel, then $\rho(\mathbf{x}_0) > 0$ for every $\mathbf{x}_0$. In practice, even though the prediction is theoretically well defined, dividing by $\rho(\mathbf{x}_0)$ can be numerically unstable when ρ is close to zero. A practical fix is to use

$$\hat{Y}_{\mathrm{SiNK},\epsilon}(\mathbf{x}_0) = \beta + \frac{1}{\max(\rho(\mathbf{x}_0), \epsilon)} \mathbf{k}(\mathbf{x}_0)^T K^{-1}(\mathbf{y} - \beta \mathbf{1}) \tag{4.1}$$

for a small ϵ. We use $\epsilon = 10^{-3}$ in our numerical work. A larger ϵ would protect from bad estimators of length-scale parameters, something that we did not encounter in our numerical experiments.

Remark 3. One can construct the credible interval around the SiNK predictor based on the posterior from the implicit new prior $y_0 \sim N(\beta, k(\mathbf{x}_0, \mathbf{x}_0)(1+1/\rho(\mathbf{x}_0))$ at $\mathbf{x}_0$, which is wider than the credible interval from standard Kriging. Further theoretical and empirical study on constructing prediction intervals of SiNK is of interest.

As mentioned in Section 3, the ratio δ is now a function of $\mathbf{x}_0$. The conditional bias penalty is larger when we have less information on the target function value. Penalizing by the conditional bias by an appropriate multiple of the conditional bias squared improves performance at extreme

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

values. The rationale of using $\delta = 1/\rho(\mathbf{x}_0)$ will be discussed in Section 4.2.

4.2 Properties

The main feature of SiNK is its stability which will be represented as *boundedness* and *localness* in this section. We find that the SiNK predictor is the unique predictor with these properties in the class of generalized CBPK predictors with MSPE-CB ratio δ as a function of $\rho(\mathbf{x}_0)$.

If the covariance function is stationary, then the SiNK predictor is bounded. On the contrary, CMLE is unbounded when $\rho(\mathbf{x}_0)$ approaches 0. The proof is given in the Appendix, Section 3.

Proposition 2 (Boundedness)**.**

$$|\hat{Y}_{\mathrm{SiNK}}(\mathbf{x}_0) - \beta| \le \sqrt{k(\mathbf{x}_0, \mathbf{x}_0)}\sqrt{(\mathbf{y} - \beta\mathbf{1})^T K^{-1}(\mathbf{y} - \beta\mathbf{1})} \tag{4.2}$$

Thus, if the covariance function is stationary, with probability 1,

$$\sup_{\mathbf{x}_0 \in \mathbb{R}^d} |\hat{Y}_{\mathrm{SiNK}}(\mathbf{x}_0)| < \infty. \tag{4.3}$$

For a predictor with inflated residual of simple Kriging predictor to be bounded, the maximum amount of inflation is of order $1/\rho(\mathbf{x}_0)$. Roughly speaking, SiNK is the predictor with maximum inflation of the residual term that satisfies boundedness.

To discuss the localness property, we consider a more specific covariance class that contains many widely used covariance functions. A measurable

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

function $f : \mathbb{R}^+ \to \mathbb{R}^+$ is called *rapidly varying* of index $-\infty$, in the sense of Haan (1970), if for any $t > 1$,

$$\lim_{x\to\infty} \frac{f(tx)}{f(x)} = 0$$

holds. The class of rapidly varying functions is important in asymptotic analysis (see Bingham, Goldie, and Teugels (1989)). The Matérn kernel with $\nu > 0$ is rapidly varying of index $-\infty$, because the modified Bessel function of the second kind satisfies $K_\nu(z) \propto \exp(-z)/\sqrt{z}\,(1 + O(1/z))$ as $z \to \infty$. The Gaussian kernel is also rapidly varying of index $-\infty$. This technical assumption is necessary only for proving the localness property stated in Proposition 3. There are kernels that are not rapidly varying of index $-\infty$ such as rational quadratic covariance kernel (Rasmussen (2006)).

Now let J_k be a set of points that have different distances from observations in the k'th coordinate,

$$J_k := \{\mathbf{x}_0 \mid |(\mathbf{x}_0 - \mathbf{x}_j)_k| \neq |(\mathbf{x}_0 - \mathbf{x}_l)_k| \text{ for all } j \neq l, j, l \in \{1, 2, \ldots, n\}\} \tag{4.4}$$

where $k \in \{1, 2, \cdots, d\}$. In Proposition 3 and Theorem 1, we assume that the new point $\mathbf{x}_0$ is in J_k to break the ties; we remove a measure zero set to simplify the argument. Also, define the neighborhood of an observation

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

$\mathbf{x}_j$ for $j \in \{1, 2, \ldots, n\}$ as

$$B(\mathbf{x}_j) := \{\mathbf{x}_0 \mid K(\mathbf{x}_0, \mathbf{x}_j) > K(\mathbf{x}_0, \mathbf{x}_l)\ \forall l \neq j, l \in \{1, 2, \ldots, n\}\}. \quad (4.5)$$

Thus, if $\mathbf{x}_0 \in B(\mathbf{x}_j)$, then $\mathbf{x}_j$ is the closest observation to $\mathbf{x}_0$ in terms of covariance.

Proposition 3 (Localness)**.** *Suppose that the covariance function is a tensor product of stationary, rapidly varying of index $-\infty$, positive kernels with length scale parameter $\theta = (\theta_1, \ldots, \theta_d)$. Then*

$$\lim_{\theta_k \to 0} \sup_{\mathbf{x}_0 \in B(\mathbf{x}_j) \cap J_k} |\hat{Y}(\mathbf{x}_0) - Y(\mathbf{x}_j)| = 0$$

with probability 1, where J_k and $B(\mathbf{x}_j)$ are sets of points defined in (4.4) *and* (4.5)*, respectively.*

Thus, as $\theta_k \to 0$, if $\mathbf{x}_j$ is the closest observation (in k'th coordinate) to $\mathbf{x}_0$, then the SiNK predictor $\hat{Y}(\mathbf{x}_0)$ converges to $Y(\mathbf{x}_j)$. We can then show that the SiNK predictor is the only predictor that satisfies localness in the class of generalized CBPK predictors. As $\theta_k \to 0$, the simple Kriging predictor converges to the prior mean β.

Theorem 1 (Uniqueness)**.** *Consider a conditional biased penalized Kriging predictor*

$$\hat{Y}(\mathbf{x}_0) = \beta + w(\mathbf{x}_0)\mathbf{k}(\mathbf{x}_0)^T K^{-1}(\mathbf{y} - \beta\mathbf{1}),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

such that the covariance function is a tensor product of stationary, rapidly varying of index $-\infty$, positive kernels with length scale parameter $\theta = (\theta_1, \ldots, \theta_d)$, and $w(\mathbf{x}_0) \in [1, 1/\rho(\mathbf{x}_0)^2]$ is a continuous function of $\rho(\mathbf{x}_0)$. Suppose that $\mathbf{x}_1, \ldots, \mathbf{x}_n \in J_k$ (4.4). If there exists a $k \in \{1, 2, \ldots, d\}$ such that

$$\lim_{\theta_k \to 0} \sup_{\mathbf{x}_0 \in B(\mathbf{x}_j) \cap J_k} |\hat{Y}(\mathbf{x}_0) - Y(\mathbf{x}_j)| = 0 \tag{4.6}$$

holds with probability 1, where J_k and $B(\mathbf{x}_j)$ are sets of points defined in (4.4) and (4.5), respectively, then $w(\rho(\mathbf{x}_0)) = 1/\rho(\mathbf{x}_0)$, and $\hat{Y}(\mathbf{x}_0)$ is the SiNK predictor.

The proofs of Proposition 3 and Theorem 1 are given in the Appendix, Section 4. Restricting $w(\mathbf{x}_0)$ to be a function of $\rho(\mathbf{x}_0)$ enables us to guarantee that $w(\mathbf{x}_0) \in [1, 1/\rho(\mathbf{x}_0)^2]$. For example, $w(\mathbf{x}_0) = 1/\rho(\mathbf{x}_0)$ is always in $[1, 1/\rho(\mathbf{x}_0)^2]$. Another example for necessity of this condition is limit Kriging (Joseph (2006)) where the predictor has $w(\mathbf{x}_0) = 1/(\mathbf{k}(\mathbf{x}_0)^T K^{-1}\mathbf{1})$. The limit Kriging predictor has the localness property, but is not guaranteed to be a CBPK with nonnegative ratio δ, which means we cannot guarantee better performance at extreme values.

Figures 1 and 2 illustrate the property of SiNK and the difference from ordinary Kriging. The function used in Figure 1 is the 2-dimensional Za-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

kharov function in $[0,1]^2$ is

$$f(\mathbf{x}) = \sum_{i=1}^{d} x_i^2 + \left(\sum_{i=1}^{d} 0.5 i x_i\right)^2 + \left(\sum_{i=1}^{d} 0.5 i x_i\right)^4, \tag{4.7}$$

where $d = 2$, and the input points are 4 midpoints of the edges of a unit square.

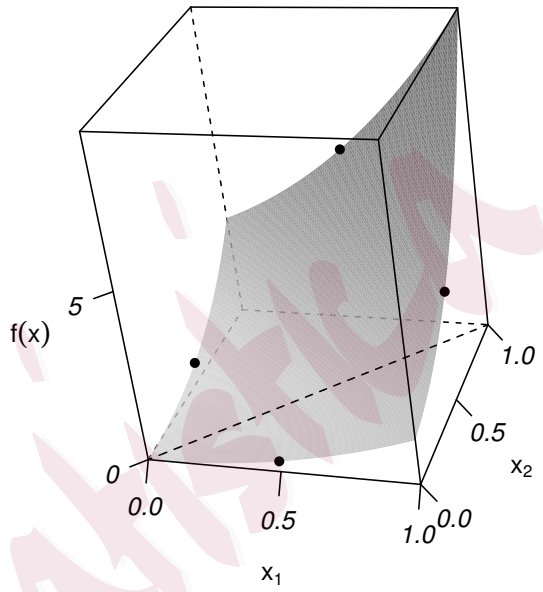
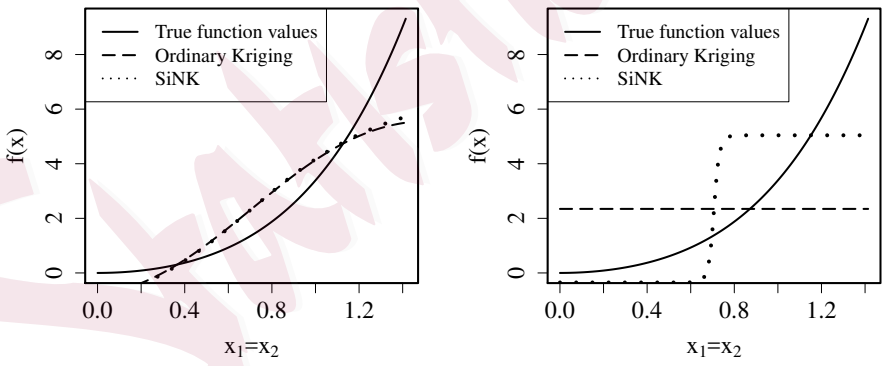


Figure 1: Two-dimensional Zakharov function. The gray surface is the true function, four points on the surface are the observations. We compare the SiNK predictions and ordinary Kriging predictions on the dotted line $x_1 = x_2$ in Figure 2.

We fitted ordinary Kriging and SiNK with an estimated constant mean and tensor product Matérn 5/2 covariance. To visualize the difference,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

we evaluated the predictors on $x_1 = x_2$, the dashed diagonal line on the square. Figure 2 shows the predictions with two different sets of parameters. For $\theta_1 = \theta_2 = 1$, the predictions are quite similar because $\rho(\mathbf{x}_0) \approx 1$ for all $\mathbf{x}_0 \in [0,1]^2$. However, when $\theta_1 = \theta_2$ are close to zero (0.05), we observe significant differences between the two predictions. We also observe the localness property of SiNK. The $\rho(\mathbf{x}_0)$ are close to zero for most of the plotted points, and thus the ordinary Kriging predictor is close to the estimated constant mean for points far from the observations. The SiNK predictor uses the function value of the observation that is the closest to the target point.



(a) $\theta = 1$. (b) $\theta = 0.05$.

Figure 2: Illustration of the difference between ordinary Kriging and SiNK. SiNK performs better at extreme values than ordinary Kriging, more significantly when the correlations between function values are smaller.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

The localness property of SiNK is also related to the fact that the SiNK prediction at $\mathbf{x}_0$ only depends on the ratios of the correlations with observed function values. For instance, suppose that we predict at another point $\mathbf{x}_0'$ with covariance vector $\mathbf{k}(\mathbf{x}_0') = c\mathbf{k}(\mathbf{x}_0)$, where c is in $(0, 1)$. Then

$$\hat{Y}_{\mathrm{SiNK}}(\mathbf{x}_0') = \beta + \frac{\mathbf{k}(\mathbf{x}_0')^T K^{-1}(\mathbf{y} - \beta\mathbf{1})}{\sqrt{\mathbf{k}(\mathbf{x}_0')^T K^{-1}\mathbf{k}(\mathbf{x}_0')}} = \hat{Y}_{\mathrm{SiNK}}(\mathbf{x}_0).$$

Thus, the SiNK prediction at $\mathbf{x}_0'$ is the same as the prediction at $\mathbf{x}_0$. However, the simple Kriging prediction is shrunk to β by a factor of c. Thus, even if $\mathbf{x}_0'$ is far away from inputs, only the ratios of the correlation determine the SiNK prediction. Accordingly, SiNK does not automatically converge to the prior mean β as $\mathbf{k}(\mathbf{x}_0) \to 0$, for instance if one of the $\theta_j \to 0$.

4.3 Mean squared prediction error at extreme values

Since the simple Kriging predictor is the BLUP, the SiNK predictor has larger MSPE than the simple Kriging predictor. However, we can show that SiNK will be only slightly inferior; the ratio of MSPEs is bounded. The proof is given in the supplementary material, Section S2.

Proposition 4 (MSPE comparison)**.**

$$\mathbb{E}[(\hat{Y}_{\mathrm{SiNK}}(\mathbf{x}_0) - Y(\mathbf{x}_0))^2] = \frac{2}{1 + \rho(\mathbf{x}_0)}\mathbb{E}[(\hat{Y}_{\mathrm{K}}(\mathbf{x}_0) - Y(\mathbf{x}_0))^2],$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

the RMSPE of SiNK is at most $\sqrt{2}$ times larger than the RMSPE of Kriging.

Here we show that SiNK has improved performance at extreme values. This can be represented in two ways; conditioning on a single extreme value of $Y(x_0)$, and conditioning on a region of extreme $Y(x_0)$ values.

Proposition 5. *For a input $\mathbf{x}_0$ with $\rho(\mathbf{x}_0) > 0$, if*

$$\left| \frac{y_0 - \beta}{\sqrt{k(\mathbf{x}_0, \mathbf{x}_0)}} \right| \geq \sqrt{\frac{(1+\rho(\mathbf{x}_0))^2}{(1+\rho(\mathbf{x}_0))^2 - 1}}$$

holds, then

$$\mathbb{E}[\,(\hat{Y}_{\mathrm{SiNK}}(\mathbf{x}_0) - Y(\mathbf{x}_0))^2 \,\big|\, Y(\mathbf{x}_0) = y_0\,] \leq \mathbb{E}[\,(\hat{Y}_{\mathrm{K}}(\mathbf{x}_0) - Y(\mathbf{x}_0))^2 \,\big|\, Y(\mathbf{x}_0) = y_0\,].$$

Proof. Directly follows from (S2.1) in the proof of Proposition 4. □

Proposition 6. *Let $\phi(\cdot)$ and $\Phi(\cdot)$ be the density function and distribution function of the standard normal, respectively. Let $S(\mathbf{x}_0) = |(Y(\mathbf{x}_0) - \beta)/(\sqrt{k(\mathbf{x}_0, \mathbf{x}_0)})|$. For $M > 0$, if*

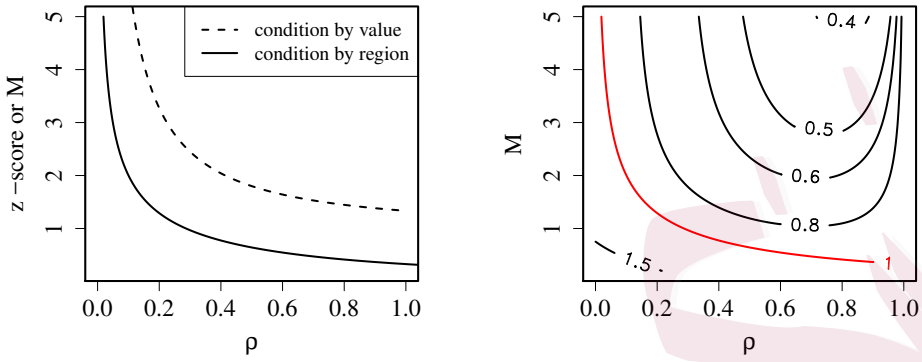
$$\rho(\mathbf{x}_0) \geq -1 + \sqrt{1 + (1 - \Phi(M))/(M\phi(M))}$$

holds, then

$$\mathbb{E}\left[(\hat{Y}_{\mathrm{SiNK}}(\mathbf{x}_0) - Y(\mathbf{x}_0))^2 \middle| S(\mathbf{x}_0) \geq M\right] \leq \mathbb{E}\left[(\hat{Y}_{\mathrm{K}}(\mathbf{x}_0) - Y(\mathbf{x}_0))^2 \middle| S(\mathbf{x}_0) \geq M\right].$$

The proof of Proposition 6 is given in the supplementary material, Section S3. $S(\mathbf{x}_0)$ represents the z-score of the function value.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255



(a) The level curve for the mean squared error. SiNK outperforms simple Kriging in the region above and to the right of the given curves.

(b) Contour plot of $\mathrm{CMSPE}_{\mathrm{SiNK}}/\mathrm{CMSPE}_{\mathrm{K}}$ (4.8). The ratio decreases as the threshold M increases.

Figure 3: Relation between $\rho(\mathbf{x}_0)$ and z-score.

Figure 3 shows the relation between $\rho(\mathbf{x}_0)$ and the critical z-score or the threshold M for the z-score. The ratio of the region-conditional mean squared prediction error

$$\frac{\mathrm{CMSPE}_{\mathrm{SiNK}}}{\mathrm{CMSPE}_{\mathrm{K}}} = \frac{\mathbb{E}\left[(\hat{Y}_{\mathrm{SiNK}}(\mathbf{x}_0) - Y(\mathbf{x}_0))^2 \middle| S(\mathbf{x}_0) \geq M\right]}{\mathbb{E}\left[(\hat{Y}_{\mathrm{K}}(\mathbf{x}_0) - Y(\mathbf{x}_0))^2 \middle| S(\mathbf{x}_0) \geq M\right]} \tag{4.8}$$

decreases as the threshold M increases.

5 Numerical experiments

For numerical simulations, we used the DiceKriging package in R by Roustant, Ginsbourger, and Deville (2012). We fit the constant mean model

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

for ordinary Kriging and SiNK, with the maximum likelihood estimator of the constant mean $\hat{\beta} = (\mathbf{1}^T K^{-1}\mathbf{1})^{-1}\mathbf{1}^T K^{-1}\mathbf{y}$. For the covariance function, we used tensor products of Matérn $\nu = 5/2$ kernels with maximum likelihood estimators of the length-scale parameters $\theta_1, \ldots, \theta_d$, unless specified otherwise. We used $\epsilon = 10^{-3}$ in (4.1).

To measure the performance of a predictor, we computed the empirical integrated squared error (EISE)

$$\frac{1}{n_T}\sum_{j=1}^{n_T}(\hat{Y}(\mathbf{x}_{test,j}) - Y(\mathbf{x}_{test,j}))^2$$

with an independent set of n_T test points. Note that EISE is different from the MSPE; MSPE is the expected squared prediction error at a fixed point $\mathbf{x}_0$. The EISE ratio of SiNK is computed by dividing the EISE of SiNK by the EISE of ordinary Kriging. We also report the test R^2 of the predictors, $1 -$ EISE/(sample variance of $Y(\mathbf{x}_{test})$), to understand the relative errors of predictors. To measure the performance of a predictor at extreme values, we also computed the extreme empirical integrated squared error (EEISE)

$$\frac{\sum_{j=1}^{n_T}(\hat{Y}(\mathbf{x}_{test,j}) - Y(\mathbf{x}_{test,j}))^2 1_{\{S(\mathbf{x}_{test,j})\geq 2\}}}{\sum_{j=1}^{n_T} 1_{\{S(\mathbf{x}_{test,j})\geq 2\}}}.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

5.1 Gaussian process

We generated a realization of a 7-dimensional Gaussian process with zero mean and Matérn covariance with length-scale hyperparameters $\theta = (1, 1, 1, 1, 1, 1, 1)$ and stationary variance $k(\mathbf{x}, \mathbf{x}) = \sigma^2 = 1$. The chosen observations were 100 points i.i.d. uniform in $[0, 1]^7$ and the test points were 2000 points i.i.d. uniform in $[0, 1]^7$.

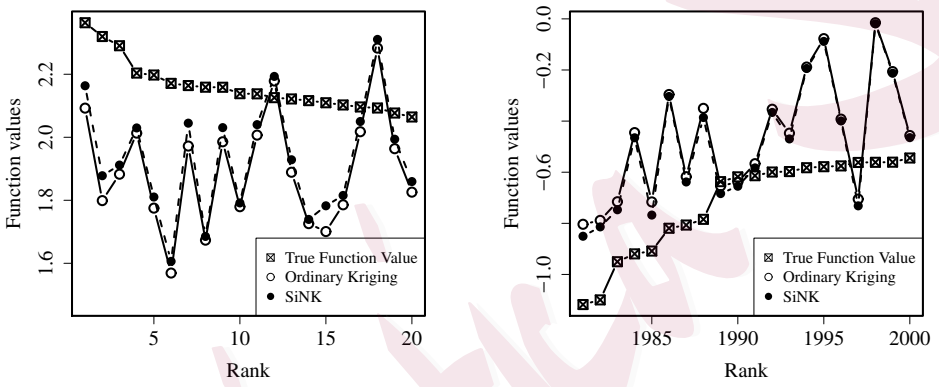
To emulate the situation where the hyperparameters are unknown, we estimated the hyperparameters by maximizing the likelihood. The estimated mean was $\hat{\beta} = 0.143$, the estimated length-scale hyperparameters were $\hat{\theta} = (1.29, 0.92, 1.18, 1.41, 0.95, 0.76, 1.32)$, and the estimated stationary variance was $\hat{\sigma}^2 = 0.94$. The performance comparison between SiNK and ordinary Kriging is in Table 1. We observe that SiNK had slightly inferior EISE, but showed better performance at extreme values.

Table 1: Performance comparison of ordinary Kriging and SiNK for a realization of Gaussian process and piston function. The EISE ratios are the EISE of SiNK divided by the EISE of ordinary Kriging.

Function	Gaussian Process	Piston Function
Number of observations	100	14
EEISE Ratio (SiNK/Ordinary Kriging)	0.820	0.814
EISE Ratio (SiNK/Ordinary Kriging)	1.020	0.887
R^2 Ordinary Kriging	0.818	0.674

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

Figure 4 shows the prediction at test points with extreme function values. We first sorted the test points by the true function values and see the 1% largest and smallest function values. We observe that SiNK reduces the conditional bias by inflating the residual term. Differences are small but consistently in the right direction.



(a) Prediction at test points with 1% largest function values.

(b) Prediction at test points with 1% smallest function values.

Figure 4: Ordinary Kriging and SiNK for a realization of 7-dimensional Gaussian process. Rank is the order of the true function values of the test points.

5.2 Piston function

We examined the performance of SiNK in a computer experiment. The piston simulation function in Zacks (1998) models the circular motion of a piston within a cylinder. The response C is the time it takes to complete

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

one cycle, in seconds. The formula for the function is

$$C(\mathbf{x}) = 2\pi\sqrt{\frac{M}{k + S^2\frac{P_0V_0}{T_0}\frac{T_a}{V^2}}},$$

where

$$V = \frac{S}{2k}\left(\sqrt{A^2 + 4k\frac{P_0V_0}{T_0}T_a} - A\right) \text{ and } A = P_0S + 19.62M - \frac{kV_0}{S}.$$

The description of the input variables is in Table 2. For this example, we adopted the Randomized QMC design (Faure sequence base 7) for observations and test points, because the number of observation is small. In Table 1, we see that in this case SiNK performs better not only at extreme values but also overall. This result possibly comes from the model mismatch of Gaussian process for the piston function; more specifically, the reduction of conditional bias may have had a large effect on the test error here.

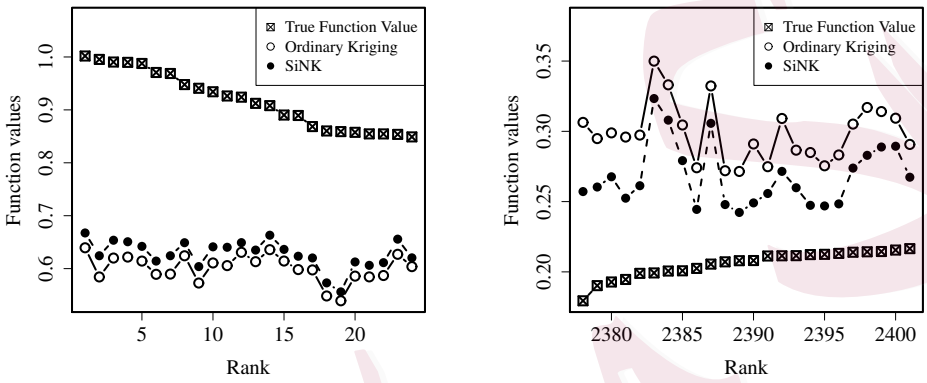
Table 2: Input variables $\mathbf{x}$ for the piston function.

$M \in [30, 60]$	piston weight (kg)
$S \in [0.005, 0.020]$	piston surface area (m^2)
$V_0 \in [0.002, 0.010]$	initial gas volume (m^3)
$k \in [1000, 5000]$	spring coefficient (N/m)
$P_0 \in [90000, 110000]$	atmospheric pressure (N/m^2)
$T_a \in [290, 296]$	ambient temperature (K)
$T_0 \in [340, 360]$	filling gas temperature (K)

Again, in Figure 5 the SiNK predictions are better at the test points with extreme function values than the ordinary Kriging predictions, and the

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

difference is significant at the test points with 1% smallest function values. The inflation of the residual is consistently in the right direction, and larger than that of the Gaussian process example.



(a) Prediction at the test points with 1% largest function values.

(b) Prediction at the test points with 1% smallest function values.

Figure 5: ordinary Kriging and SiNK for the piston function. Rank is the order of the true function values of the test points.

5.3 Other functions

We fitted ordinary Kriging, limit Kriging, and SiNK for several deterministic functions and compared the performances. The test function codes are from Bingham's website (Bingham (2013)). See the supplementary materials, Section S4 for the details of the test functions. For each test function, we trained with 100 independent sets of input points and computed the

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

averages and standard deviations of metrics. The training points and test points were independent and uniformly distributed in the domain of inputs. Table 3 shows the dimension of the function, the number of observed points and test points, covariance type, R^2, EISE ratio, and EISE ratio at extreme values for each function. The number of training points for fitting each function was chosen so that the R^2 of ordinary Kriging is roughly 0.95, for all but the Robot Arm function which is a comparably difficult function to fit with our prediction methods.

Table 3: Performance comparison among ordinary Kriging, limit Kriging and SiNK. The average and standard deviation of metrics from 100 independent set of inputs are reported. The EISE ratios are the EISE of SiNK or limit Kriging divided by the EISE of ordinary Kriging. The Friedman function did not have extreme values for most of the simulations.

	Borehole	Welch	Piston	Friedman	Robot Arm
Dimension	8	20	7	5	8
Training, test	32, 5000	320, 5000	49, 5000	50, 5000	512, 5000
Covariance type	Matérn 5/2	Matérn 5/2	Matérn 5/2	Matérn 5/2	Matérn 5/2
Fraction of extremes	0.091 (± 0.050)	0.249 (± 0.026)	0.072 (± 0.037)	0.000 (± 0.001)	0.051 (± 0.015)
EEISE Ratio (SiNK)	0.718 (± 0.084)	0.554 (± 0.068)	0.819 (± 0.063)	NA	0.696 (± 0.040)
EEISE Ratio (Limit)	0.694 (± 0.134)	0.675 (± 0.033)	0.802 (± 0.121)	NA	0.877 (± 0.035)
EISE Ratio (SiNK)	0.838 (± 0.095)	0.813 (± 0.070)	0.939 (± 0.064)	0.978 (± 0.021)	1.008 (± 0.027)
EISE Ratio (Limit)	0.754 (± 0.102)	0.781 (± 0.023)	0.903 (± 0.083)	0.966 (± 0.033)	0.985 (± 0.012)
R^2 (OK)	0.970 (± 0.015)	0.948 (± 0.005)	0.967 (± 0.011)	0.973 (± 0.013)	0.838 (± 0.011)

We observe that for the five functions that we considered, SiNK performed better than ordinary Kriging in terms of EISE, and the performance

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

gets even better for extreme values in terms of EISE. For instance, for the Welch function, the SiNK predictions at points with extreme function values (function values such that $|z\text{-score}| > 2$) have roughly half EISE of the EISE of ordinary Kriging predictions. In addition, we observe that the performance of limit Kriging and SiNK is very similar in terms of EISE. Limit Kriging also shows improved performance at extreme values compared to ordinary Kriging, but the improvement is smaller or no different than the improvement of SiNK. For the Friedman function, there was not a test point function value which had $|z\text{-score}|$ larger than 2 for most of the simulations, due to the large estimate of the stationary variance $\sigma^2 = k(\mathbf{x}, \mathbf{x})$. A suspicious estimate of the stationary variance can be found occasionally in practice, but it is not a problem for prediction because all three predictors that we are comparing do not depend on the estimate of σ^2.

6 Discussion

We have presented an alternative to Kriging with improved predictions at the extreme values. We first found a link between conditional likelihood at the target and CBPK, and used it to define SiNK. In addition, we showed that SiNK has a boundedness and a localness property. In numerical experiments, we observed that SiNK generally performs better not only at

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

extreme values but also in terms of overall integrated squared error. This result is possibly because the functions used in the examples may not behave like typical realizations of stationary Gaussian processes.

Supplementary Materials

The online supplementary materials contain proofs of the theoretical results and details on the test functions used in the numerical experiments.

Acknowledgements

The authors are very grateful to the editors and referees for helpful comments and suggestions. This work was supported by NSF grants DMS-1407397 and DMS-1521145.

Appendix

1 Proof of Proposition 1 (Generalized CBPK)

Proof. Without loss of generality, let $\beta = 0$. Expanding (3.4), we get

$$\begin{aligned}
&\mathbb{E}[(y_0 - \lambda^T \mathbf{y})^2] + \delta \mathbb{E}[(y_0 - \mathbb{E}[\lambda^T \mathbf{y}|y_0])^2] \\
&= k(\mathbf{x}_0, \mathbf{x}_0) - 2\lambda^T \mathbf{k}(\mathbf{x}_0) + \lambda^T K \lambda + \delta \mathbb{E}[(y_0 - \lambda^T \tilde{m})^2] \\
&= k(\mathbf{x}_0, \mathbf{x}_0) - 2\lambda^T \mathbf{k}(\mathbf{x}_0) + \lambda^T K \lambda + \delta \left(1 - \frac{\lambda^T \mathbf{k}(\mathbf{x}_0)}{k(\mathbf{x}_0, \mathbf{x}_0)}\right)^2 k(\mathbf{x}_0, \mathbf{x}_0).
\end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

This is a quadratic function of λ. Thus, the minimzing λ is

$$\hat{\lambda} = \left(K + \frac{\delta}{k(\mathbf{x}_0, \mathbf{x}_0)} \mathbf{k}(\mathbf{x}_0)\mathbf{k}(\mathbf{x}_0)^T \right)^{-1} (\mathbf{k}(\mathbf{x}_0) + \delta \mathbf{k}(\mathbf{x}_0)) = \frac{\delta + 1}{\delta \rho(\mathbf{x}_0)^2 + 1} K^{-1} \mathbf{k}(\mathbf{x}_0)$$

by the Woodbury formula. Then $w(\mathbf{x}_0) = (\delta + 1)/(\delta \rho(\mathbf{x}_0)^2 + 1)$, and

$$\hat{Y}(\mathbf{x}_0) = \beta + \frac{\delta + 1}{\delta \rho(\mathbf{x}_0)^2 + 1} \mathbf{k}(\mathbf{x}_0)^T K^{-1} (\mathbf{y} - \beta \mathbf{1}). \tag{1}$$

If $\rho(\mathbf{x}_0) \neq 0$, then for $\delta \geq 0$, $w(\mathbf{x}_0)) \in [1, 1/\rho(\mathbf{x}_0)^2)$, and $\lim_{\delta \to \infty} w(\mathbf{x}_0) = 1/\rho(\mathbf{x}_0)^2$.

If $\rho(\mathbf{x}_0) = 0$, then for $\delta \geq 0$, $w(\mathbf{x}_0) \in [1, \infty)$, and $\lim_{\delta \to \infty} w(\mathbf{x}_0) = \infty$. □

2 Definition of SiNK

The logarithm of the posterior probability (up to a constant) is

$$\log p(y_0|\mathbf{y}) = -\frac{1}{2}(\mathbf{y} - \tilde{m})^T \tilde{K}^{-1} (\mathbf{y} - \tilde{m}) - \frac{\rho(\mathbf{x}_0)}{2(1 + \rho(\mathbf{x}_0))} \frac{(y_0 - \beta)^2}{k(\mathbf{x}_0, \mathbf{x}_0)}.$$

Differentiating with respect to y_0, we get

$$\begin{aligned} \frac{\partial \log p(y_0|\mathbf{y})}{\partial y_0} &= \frac{1}{k(\mathbf{x}_0, \mathbf{x}_0)} (\mathbf{y} - \tilde{m})^T \tilde{K}^{-1} \mathbf{k}(\mathbf{x}_0) - \frac{\rho(\mathbf{x}_0)}{1 + \rho(\mathbf{x}_0)} \frac{(y_0 - \beta)}{k(\mathbf{x}_0, \mathbf{x}_0)} \\ &= \frac{1}{(1 - \rho(\mathbf{x}_0)^2) k(\mathbf{x}_0, \mathbf{x}_0)} (\mathbf{y} - \tilde{m})^T K^{-1} \mathbf{k}(\mathbf{x}_0) - \frac{\rho(\mathbf{x}_0)}{1 + \rho(\mathbf{x}_0)} \frac{(y_0 - \beta)}{k(\mathbf{x}_0, \mathbf{x}_0)} \end{aligned}$$

from (S1.1). If $\rho(\mathbf{x}_0) \neq 0$, then solving $\partial \log p(y_0|\mathbf{y})/\partial y_0 = 0$ leads to

$$\hat{y_0} = \beta + \frac{1}{\rho(\mathbf{x}_0)} \mathbf{k}(\mathbf{x}_0)^T K^{-1} (\mathbf{y} - \beta \mathbf{1}).$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

3 Proof of Proposition 2 (Boundedness)

Proof. By the Cauchy-Schwarz inequality,

$$\begin{aligned}|\hat{Y}_{\mathrm{SiNK}}(\mathbf{x}_0) - \beta| &= \frac{1}{\rho(\mathbf{x}_0)}|\mathbf{k}(\mathbf{x}_0)^T K^{-1}(\mathbf{y} - \beta\mathbf{1})| \\ &\leq \frac{1}{\rho(\mathbf{x}_0)}\sqrt{\mathbf{k}(\mathbf{x}_0)^T K^{-1}\mathbf{k}(\mathbf{x}_0)}\sqrt{(\mathbf{y} - \beta\mathbf{1})^T K^{-1}(\mathbf{y} - \beta\mathbf{1})} \\ &= \sqrt{k(\mathbf{x}_0, \mathbf{x}_0)}\sqrt{(\mathbf{y} - \beta\mathbf{1})^T K^{-1}(\mathbf{y} - \beta\mathbf{1})}\end{aligned}$$

and equality holds when $K^{-1/2}\mathbf{k}(\mathbf{x}_0)$ and $K^{-1/2}(\mathbf{y} - \beta\mathbf{1})$ are parallel. If the covariance function is stationary, then the right hand side of (4.2) does not depend on $\mathbf{x}_0$, thus (4.3) holds. □

4 Proof of Theorem 1 and Proposition 3 (Localness and Uniqueness)

Proof. Let the stationary variance $K(\mathbf{x}, \mathbf{x}) = \sigma^2$. Now for a target point $\mathbf{x}_0 \in B(\mathbf{x}_j) \cap J_k$, for $l \neq j$, from the rapidly varying of index $-\infty$ condition,

$$\lim_{\theta_k \to 0} \frac{K(\mathbf{x}_0, \mathbf{x}_l)}{K(\mathbf{x}_0, \mathbf{x}_j)} = \lim_{\theta_k \to 0} \prod_{i=1}^{d} \frac{C_{\theta_i}(|(\mathbf{x}_l - \mathbf{x}_0)_i|)}{C_{\theta_i}(|(\mathbf{x}_j - \mathbf{x}_0)_i|)} = \lim_{\theta_k \to 0} \prod_{i=1}^{d} \frac{C_1\left(\frac{|(\mathbf{x}_l - \mathbf{x}_0)_i|}{\theta_i}\right)}{C_1\left(\frac{|(\mathbf{x}_j - \mathbf{x}_0)_i|}{\theta_i}\right)} = 0.$$

Thus we obtain

$$\lim_{\theta_k \to 0} \frac{1}{K(\mathbf{x}_0, \mathbf{x}_j)}\mathbf{k}(\mathbf{x}_0) = \mathbf{e}_j$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

where $\mathbf{e}_j$ is the j-th unit vector. Noting that $\mathbf{x}_j \in B(\mathbf{x}_j) \cap J_k$, we have

$$\lim_{\theta_k \to 0} \frac{1}{\sigma^2} K = I_n,$$

where I_n is the $n \times n$ identity matrix. Thus,

$$\lim_{\theta_k \to 0} \frac{\rho^2}{K(\mathbf{x}_0, \mathbf{x}_j)^2} = \lim_{\theta_k \to 0} \frac{\mathbf{k}(\mathbf{x}_0)^T K^{-1} \mathbf{k}(\mathbf{x}_0)}{\sigma^2 K(\mathbf{x}_0, \mathbf{x}_j)^2} = \frac{1}{\sigma^4} \text{ and}$$
$$\lim_{\theta_k \to 0} \frac{\sigma^2 \mathbf{k}(\mathbf{x}_0)^T K^{-1}(\mathbf{y} - \beta \mathbf{1})}{K(\mathbf{x}_0, \mathbf{x}_j)} = y_j - \beta.$$

Now note that

$$\begin{aligned} \hat{Y}(\mathbf{x}_0) &= \beta + w(\rho) \mathbf{k}(\mathbf{x}_0)^T K^{-1}(\mathbf{y} - \beta \mathbf{1}) \\ &= \beta + w(\rho) \rho \, \frac{K(\mathbf{x}_0, \mathbf{x}_j)}{\rho \sigma^2} \frac{\sigma^2 \mathbf{k}(\mathbf{x}_0)^T K^{-1}(\mathbf{y} - \beta \mathbf{1})}{K(\mathbf{x}_0, \mathbf{x}_j)}. \end{aligned}$$

Thus, to satisfy (4.6),

$$\lim_{\theta_k \to 0} w(\rho) \rho = 1 \tag{2}$$

is the condition that needs to hold. For the SiNK predictor, $w(\rho) = 1/\rho$, so the condition holds, and therefore SiNK has the localness property and Proposition 3 holds.

The limit range of ρ as $\theta_k \to 0$ needs to be determined. For fixed $\mathbf{x}_0 \in B(\mathbf{x}_j) \cap J_k$, $\rho \to 0$ as $\theta_k \to 0$. Now for any $\delta \in (0, 1]$, let $\epsilon = C_1^{-1}(\delta)$ and $\mathbf{x}_0 = \mathbf{x}_j + \epsilon \theta_k \mathbf{e}_k$. For all sufficiently small and positive θ_k, we have $\mathbf{x}_0 \in B(\mathbf{x}_j) \cap J_k$. Then

$$\lim_{\theta_k \to 0} \frac{K(\mathbf{x}_0, \mathbf{x}_j)}{\sigma^2} = \lim_{\theta_k \to 0} \prod_{i=1}^{d} C_{\theta_i}((\mathbf{x}_j - \mathbf{x}_0)_i) = \lim_{\theta_k \to 0} C_{\theta_k}(\epsilon \theta_k) = C_1(\epsilon) = \delta$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

Thus, $\lim_{\theta_k \to 0} \rho = \delta$ for our selection of $\mathbf{x}_0$. For (2) to hold, since w is a continuous function of ρ, $w(\delta)\delta = 1$ must hold for all $\delta \in (0, 1]$. To put it differently, if (4.6) holds, then it is the SiNK predictor. □

References

Ba, S. and Joseph, V. R. (2012). Composite Gaussian process models for emulating expensive functions. *The Annals of Applied Statistics* **6**, 1838–1860.

Bachoc, F. (2013). Cross Validation and Maximum Likelihood estimations of hyper-parameters of Gaussian processes with model misspecification. *Computational Statistics & Data Analysis* **66**, 55–69.

Bingham, D. (2013). *Virtual Library of Simulation Experiments: Test Functions and Datasets*. `http://www.sfu.ca/~ssurjano`. Accessed: 2015-06-01.

Bingham, N. H., Goldie, C. M., and Teugels, J. L. (1989). *Regular variation*. Cambridge University Press.

David, M., Marcotte, D., and Soulie, M. (1984). Conditional bias in kriging and a suggested correction. *Geostatistics for Natural Resources Characterization*. Springer Netherlands, Dordrecht.

Haan, L. de (1970). *On regular variation and its application to the weak convergence of sample extremes*. Mathematisch Centrum.

Isaaks, E. (2005). The kriging oxymoron: a conditionally unbiased and accurate predictor. *Geostatistics Banff 2004*. Springer Netherlands.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

Jones, D. R. (2001). A taxonomy of global optimization methods based on response surfaces. *Journal of global optimization* **21**, 345–383.

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

Joseph, V. R., Gul, E., and Ba, S. (2015). Maximum projection designs for computer experiments. *Biometrika* **102**, 371–380.

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

Kang, L. and Joseph, V. R. (2016). Kernel Approximation: From Regression to Interpolation. *SIAM/ASA Journal on Uncertainty Quantification* **4**, 112–129.

Katz, R. W. and Murphy, A. H. (1997). *Economic value of weather and climate forecasts*. Cambridge University Press.

Koehler, J. R. and Owen, A. B. (1996). Computer experiments. *Handbook of statistics* **13**, 261–308.

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

Matérn, B. (1986). *Spatial variation*. Lecture notes in statistics. Springer, New York. ISBN: 9783540963653. URL: `https://books.google.com/books?id=s-xczaXRptoC`.

Rasmussen, C. E. (2006). *Gaussian processes for machine learning*. MIT Press, Cambridge, MA.

Roustant, O., Ginsbourger, D., and Deville, Y. (2012). Dicekriging, Diceoptim: Two R packages for the analysis of computer experiments by Kriging-based metamodelling and optimization. *Journal of Statistical Software* **51**, 1–55.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0255

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

Seo, D. (2013). Conditional bias-penalized Kriging (CBPK). *Stochastic Environmental Research and Risk Assessment* **27**, 43–58.

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

Switzer, P. (2006). Kriging. *Encyclopedia of Environmetrics*.

Zacks, S. (1998). *Modern industrial statistics: Design and control of quality and reliability*. Cengage Learning.

Zhang, N. and Apley, D. W. (2014). Fractional brownian fields for response surface metamodeling. *Journal of Quality Technology* **46**, 285–301.

Zhang, N. and Apley, D. W. (2015). Brownian Integrated Covariance Functions for Gaussian Process Modeling: Sigmoidal Versus Localized Basis Functions. *Journal of the American Statistical Association* **111**, 1182–1195.

Department of Statistics, Stanford University, Stanford, California 94305, U. S. A.

E-mail: minyong@stanford.edu

Department of Statistics, Stanford University, Stanford, California 94305, U. S. A.

E-mail: owen@stanford.edu