Proceedings of the 2014 Winter Simulation Conference
A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, eds.

REGULARIZED RADIAL BASIS FUNCTION MODELS FOR STOCHASTIC SIMULATION

Yibo Ji
Sujin Kim

Department of Industrial and Systems Engineering
National University of Singapore
119260, Singapore

ABSTRACT

We propose a new radial basis function (RBF) model for stochastic simulation, called regularized RBF (R-RBF). We construct the R-RBF model by minimizing a regularized loss over a reproducing kernel Hilbert space (RKHS) associated with RBFs. The model can flexibly incorporate various types of RBFs including those with conditionally positive definite basis. To estimate the model prediction error, we first represent the RKHS as a stochastic process associated to the RBFs. We then show that the prediction model obtained from the stochastic process is equivalent to the R-RBF model and derive the associated mean squared error. We propose a new criterion for efficient parameter estimation based on the closed form of the leave-one-out cross validation error for R-RBF models. Numerical results show that R-RBF models are more robust, and yet fairly accurate compared to stochastic kriging models.

1 INTRODUCTION

In this paper, we aim to develop a radial basis function (RBF) model to approximate a real-valued smooth function using noisy observations obtained via black-box simulations. The observations are assumed to capture the information on the true function f through the commonly-used additive error model,

$$y = f(\mathbf{x}) + \varepsilon(\mathbf{x}), \quad \mathbf{x} \in \mathscr{X} \subset \mathbb{R}^d,$$

where the random noise is characterized by a normal random variable $\varepsilon(\mathbf{x}) \sim \mathsf{N}\left(0, \sigma_\varepsilon^2(\mathbf{x})\right)$, whose variance is determined by a continuously differentiable variance function $\sigma_\varepsilon : \mathscr{X} \to \mathbb{R}$. In practice, the variance function is unknown and needs to be estimated.

In general, interpolation models with outputs from stochastic simulations tend to overfit the data, typically exhibiting strong oscillations. In statistical and machine learning, the method of regularization has been considered as an effective way to cope with the overfitting issue arising in learning from noisy observations. The method searches for the optimal model by minimizing a regularized loss function over a hypothesis space of functions to which f presumably belongs. The regularized loss function consists of a conventional loss function and a regularizer on model complexity. Examples of such methods are Tikhonov regularization (Tikhonov and Arsenin 1977), ridge regression (Hoerl and Kennard 1970), and LASSO (Tibshirani 1996). Rather than specifying parametric models for the hypothesis space, contemporary nonparametric regression methods (Girosi et al. 1995, Evgeniou et al. 2000) consider a more flexible hypothesis space called Reproducing Kernel Hilbert Space (RKHS) generated by positive definite (PD) kernels. While the RKHS is, in fact, infinite-dimensional, the representation theorem (Kimeldorf and Wahba 1971) allows us to formulate the learning process by solving an optimization problem with a finite number of model parameters. RBFs can be considered as a class of kernel functions. In this paper, we adopt the regularization method over RKHS to construct an RBF-based metamodel for stochastic simulation, called the regularized RBF (R-RBF) model. Most kernel methods for noisy observations use PD kernels, whereas their generalizations, conditionally positive definite (CPD) kernels, are rarely considered (Steinke

and Schölkopf 2008) because of additional technical details. Some CPD kernels are globally supported and parameter-free. This property can improve the robustness of the resulting metamodels. With this regard, we consider both PD and CPD types of RBFs in this paper.

The literature lacks a practical measure to evaluate the prediction uncertainty of RBF models in stochastic simulations. The technique of bootstrap resampling (Efron and Tibshirani 1993) is a popular method to construct pointwise confidence intervals of metamodels. However, it may introduce a significant amount of overhead when incorporated into optimization algorithms. Also, most RBF-based metamodels studied in statistical and machine learning consider homoscedastic noise without explicitly characterizing the model prediction uncertainty (Haykin 1994, Poggio and Girosi 1990). The mean squared error (MSE) of kernel smoothing estimator (Takezawa 2005) can be applied to estimate the prediction error of an RBF model. But, it usually requires some information on the true function, which is not suitable for black-box problems. Due to the computational shortcuts for some regression models, leave-one-out cross-validation (LOOCV) (Hastie et al. 2008) has been widely applied as the model selection strategy in the literature. A variant of LOOCV known as generalized cross-validation (GCV) by Golub et al. (1979) is often used in choosing regularization parameters. GCV, however, is not applicable for R-RBF models because of different model representations.

In light of the aforementioned research gap, we make the following contributions: (a) We extend RBF interpolation models for stochastic simulations with heteroscedastic noise by applying regularization over a RKHS associated with both PD and CPD kernels. Under a set of mild conditions on the geometry of sample points and kernels, we compute model parameters as the unique solution to a system of linear equations. (b) We construct a stochastic process based on the kernels and derive the same model prediction obtained by means of regularization. Additionally, this provides us with the prediction uncertainty measured by mean squared error (MSE). This approach also allows us to unveil similarities and differences between R-RBF and stochastic kriging (SK) models. R-RBF models possess a lower number of degrees of freedom than SK in terms of the number of shape parameters in the basis functions. However, this *regularization* also has its benefits for stochastic simulations. R-RBF models are more robust and easier to build without sacrificing significant amount of model accuracy (see Section 6). (c) We perform mathematical analysis on R-RBF models. We show in Theorem 1 that models with PD and CPD kernels surprisingly share the same form of MSE-optimal prediction and the associated prediction uncertainty, even though the underlying assumptions and derivations differ a lot. (d) Lastly, we derive a closed-form LOOCV error for R-RBF models and use it in a new criterion to enhance computational efficiency of estimating the regularization and kernel parameters.

The rest of this paper is organized as follows. Section 2 presents the proposed R-RBF model based on the regularization theory. In Section 3, we analyze R-RBF models by constructing an appropriate stochastic process. The results also help unveil connections and differences between R-RBF and SK models in Section 4. Section 5 describes methods of estimating the regularization parameters using LOOCV. Section 6 compares R-RBF models with SK models via numerical examples. Section 7 concludes this paper.

2 MODEL FORMULATION

Augmented RBF interpolation models have been extensively studied by Powell (1992) and Wendland (2005). A key element in these models is incorporating the CPD type of RBF functions in order to add flexibility in modeling. As we want our results to be general, we present our results in terms of kernels that essentially are symmetric functions of its arguments.

Definition 1 Suppose that P_{m}^d is a $\widetilde{\mathsf{m}} = \binom{\mathsf{m}-1+d}{d}$-dimensional linear vector space with polynomials of degree less than m in $\mathbb{R}^d$. A real-valued continuous kernel function $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ is *conditionally positive definite* of order m if and only if for any p distinct sample points $\mathbf{X}_n = \{\mathbf{x}_i\}_{i=1}^p \subset \mathscr{X}$, $\sum_{i=1}^p \sum_{j=1}^p \lambda_i \lambda_j K(\mathbf{x}_i, \mathbf{x}_j) > 0$, for any nonzero vector $(\lambda_1, \ldots, \lambda_p) \in \mathbb{R}^p \setminus \{(0, \ldots, 0)\}$ satisfying $\sum_{i=1}^p \lambda_i \pi(\mathbf{x}_i) = 0$ for all $\pi \in P_{\mathsf{m}}^d$.

Table 1: Conditionally positive definite radial basis functions.

Name	$K(\mathbf{x},\mathbf{x}') = \psi(r)$	Order m_K
Multiquadrics (MQ)	$-\sqrt{r^2+\theta}$	1
Thin plate splines (TPS)	$r^2 \ln(r)$	2
Cubic splines (CU)	r^3	2
Gaussian (GS)	$\exp(-r^2/\theta)$	0
Inverse multiquadrics (IMQ)	$(\theta+r^2)^{-1/2}$	0

According to its definition, a CPD kernel is only positive definite on the subspace $\{\lambda \in \mathbb{R}^p : \sum_{i=1}^p \lambda_i \pi(\mathbf{x}_i) = 0\}$. Table 1 summarizes the order m_K of CPD RBFs commonly-used in RBF-based models, where $\theta > 0$ is a tunable shape parameter and $K(\mathbf{x},\mathbf{x}')$ becomes a radial kernel $\psi(r)$ with $r = \|\mathbf{x}-\mathbf{x}'\|$. As a special case, PD kernels are basically CPD kernels of order $\mathsf{m} = 0$.

In this paper, the set of sample points is represented by $S_n = \{(\mathbf{x}_i, \bar{y}_i, n_i)\}_{i=1}^p$ with $\sum_{i=1}^p n_i = n$, where n_i is the number of replications of simulations performed at $\mathbf{x}_i$, and the sample mean is $\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$, where y_{ij} denotes the noisy observation from the jth replication of simulation at $\mathbf{x}_i$. We refer to $\mathbf{X}_p = \{\mathbf{x}_i\}_{i=1}^p$ as p design points.

RKHS H is a collection of all functions in the form of

$$g(\mathbf{x}) = \sum_{i=1}^{p} \alpha_i K(\mathbf{x},\mathbf{x}_i) + \sum_{j=1}^{\widetilde{\mathsf{m}}} \beta_j \pi_j(\mathbf{x}) = K_p(\mathbf{x})^\mathsf{T}\alpha + \pi_{\widetilde{\mathsf{m}}}(\mathbf{x})^\mathsf{T}\beta, \text{ for all } p \in \mathbb{N}\cup\{\infty\}, \mathbf{x}_i \in \mathscr{X}, \alpha_i, \beta_j \in \mathbb{R}, \quad (1)$$

where the weight vectors are $\alpha = \{\alpha_i\}_{i=1}^p$ and $\beta = \{\beta_j\}_{j=1}^{\widetilde{\mathsf{m}}}$, and $\{\alpha_i\}_{i=1}^p$ must satisfy $\sum_{i=1}^p \alpha_i \pi(\mathbf{x}_i) = 0$, for all $\pi \in P_\mathsf{m}^d$. The column vectors are defined as $K_p(\mathbf{x}) = \{K(\mathbf{x},\mathbf{x}_i)\}_{i=1}^p$, $\pi_{\widetilde{\mathsf{m}}}(\mathbf{x}) = \{\pi_j(\mathbf{x})\}_{j=1}^{\widetilde{\mathsf{m}}}$, and a set of functions $\{\pi_j(\cdot)\}_{j=1}^{\widetilde{\mathsf{m}}}$ forms a basis of P_m^d. In this paper, we use a linear tail, that is, $\mathsf{m} = 2$, $\widetilde{\mathsf{m}} = d+1$, and $\pi_{\widetilde{\mathsf{m}}}(\mathbf{x}_i) = (1, \mathbf{x}_i^\mathsf{T})^\mathsf{T}$. H is equipped with a seminorm, $|g|_H^2 = \sum_{i=1}^p \sum_{j=1}^p \alpha_i \alpha_j K(\mathbf{x}_i,\mathbf{x}_j) = \alpha^\mathsf{T}\mathbf{K}\alpha$, for all $g \in H$, where the constraint on α ensures the nonnegativeness of the seminorm. Notice that only when K is CPD of order 0, H carries with a norm $\|g\|_H^2 = \alpha^\mathsf{T}\mathbf{K}\alpha$, and has K as its reproducing kernel.

Throughout this article, we impose the following set of mild geometric conditions on $\mathbf{X}_p$:

Assumption 1 For a CPD kernel K of order m_K and a polynomial space P_m^d, the matrix Π, whose (i,j) element is $[\Pi]_{ij} = \pi_j(\mathbf{x}_i)$, has full column rank, equivalently, $\mathbf{X}_p$ is unisolvent for P_m^d with $\mathsf{m} \geq \mathsf{m}_K$.

Assumption 2 The unknown function f belongs to a RKHS H associated with a CPD kernel K and the polynomial space P_m^d.

Under the assumptions 1-2, and given S_n and P_m^d, an augmented RBF interpolation model takes form of (1) with $p = n$, and Powell (1992) proves that its weight vector is the unique solution to the linear system,

$$\begin{pmatrix} \mathbf{K} & \Pi \\ \Pi^\mathsf{T} & 0_{\widetilde{\mathsf{m}}\times\widetilde{\mathsf{m}}} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{y}}_p \\ 0_{\widetilde{\mathsf{m}}} \end{pmatrix}, \quad (2)$$

where $[\mathbf{K}]_{ij} = K(\mathbf{x}_i,\mathbf{x}_j)$, $i,j = 1,\ldots,p$. In the interpolation context, $\bar{\mathbf{y}}_p = \{f(\mathbf{x}_i)\}_{i=1}^p$ and $n_i = 1$ for all $i = 1,\ldots,p$. We show that the weight vector of a regularized RBF model satisfies a similar condition as (2), except that a diagonal matrix is added to $\mathbf{K}$.

We present our analysis by assuming that all model parameters are known. Issues regarding parameter estimation are resolved in Section 5. Given a set S_n of evaluated samples, a kernel K, and Assumption 2, the statistical learning theory suggests that one possible estimate of f can be obtained by minimizing a regularized loss over a RKHS H,

$$\min_{g\in H} \overline{\mathsf{R}}_n(g) := \frac{1}{p}\sum_{i=1}^{p}\left(\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{(y_{ij}-g(\mathbf{x}_i))^2}{\sigma_\varepsilon^2(\mathbf{x}_i)/n_i}\right) + \lambda_{\text{reg}}|g|_H^2 \quad (3)$$

The first term in (3) measures the squared Mahalanobis length of the residual vector that is also used in generalized least squares. This term represents the fidelity of the approximation model. In general, we have no access to the noise variance function $\sigma_\varepsilon^2(\cdot)$. Instead, we estimate $\sigma_\varepsilon^2(\mathbf{x}_i)$ with the sample variance of noisy observations,$\sigma_s^2(\mathbf{x}_i) = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$. We interpret the second term as a penalty imposed on the model complexity measured by RKHS seminorm. Finally, the trade-off between model fidelity and model complexity is controlled by a single regularization parameter λ_{reg}, the estimation of which is efficiently addressed via LOOCV in Section 5.

By the semiparametric representer theorem (Schölkopf and Smola 2001), the optimal solution to (3) takes the form of (1), where the weight vectors α and β are computed by replacing g in (3) with $M_{K,S_n}(\mathbf{x})$ of the form (1) and minimizing the following optimization problem,

$$\min_{\alpha\in\mathbb{R}^p,\beta\in\mathbb{R}^{\tilde{\mathsf{m}}}} \quad \overline{\mathsf{R}}_n(M_{K,S_n}) \quad \text{Subject to} \quad \Pi^{\mathsf{T}}\alpha = 0_{\tilde{\mathsf{m}}}, \tag{4}$$

$$\begin{aligned}\text{where} \quad \overline{\mathsf{R}}_n(M_{K,S_n}) &= \frac{1}{p}\sum_{i=1}^{p}\left(\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{(y_{ij}-K_p(\mathbf{x}_i)^{\mathsf{T}}\alpha-\pi_{\tilde{\mathsf{m}}}(\mathbf{x}_i)^{\mathsf{T}}\beta)^2}{\sigma_\varepsilon^2(\mathbf{x}_i)/n_i}\right)+\lambda_{\text{reg}}\left|M_{K,S_n}\right|_H^2 \\ &= \frac{1}{p}(\mathbf{K}\alpha+\Pi\beta-\bar{\mathbf{y}}_p)^{\mathsf{T}}\Sigma_\varepsilon^{-1}(\mathbf{K}\alpha+\Pi\beta-\bar{\mathbf{y}}_p)+\lambda_{\text{reg}}\alpha^{\mathsf{T}}\mathbf{K}\alpha+\frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{n_i}\frac{(y_{ij}-\bar{y}_i)^2}{\sigma_\varepsilon^2(\mathbf{x}_i)},\end{aligned}$$

where the column vector $\bar{\mathbf{y}}_p = (\bar{y}_1,\bar{y}_2,\ldots,\bar{y}_p)^{\mathsf{T}}$ and the matrix Σ_ε is a diagonal matrix with its diagonal elements being $\{\sigma_\varepsilon^2(\mathbf{x}_i)/n_i\}_{i=1}^p$. A remarkable feature of the semiparametric representer theorem is that the optimal solution to (3) lies in a finite-dimensional space represented by (1) with a finite p, although the dimension of H is essentially infinite.

By assuming that λ_{reg} and Σ_ε is known, the KKT conditions for the solutions α and β give rise to the following linear system,

$$\left.\begin{aligned}\tfrac{2}{p}\mathbf{K}^{\mathsf{T}}\Sigma_\varepsilon^{-1}(\mathbf{K}\alpha+\Pi\beta-\bar{\mathbf{y}}_p)+2\lambda_{\text{reg}}\mathbf{K}\alpha+\Pi\eta_{\text{CPD}} &= 0_p,\\ \tfrac{2}{p}\Pi^{\mathsf{T}}\Sigma_\varepsilon^{-1}(\mathbf{K}\alpha+\Pi\beta-\bar{\mathbf{y}}_p) &= 0_{\tilde{\mathsf{m}}},\\ \Pi^{\mathsf{T}}\alpha &= 0_{\tilde{\mathsf{m}}},\end{aligned}\right\} \tag{5}$$

where $\eta_{\text{CPD}}\in\mathbb{R}^{\tilde{\mathsf{m}}}$ is a vector of Lagrange multipliers associated with the equality constraint $\Pi^{\mathsf{T}}\alpha = 0_{\tilde{\mathsf{m}}}$.

Proposition 1 The unique solution to (5) is given by $\eta_{\text{CPD}} = 0_{\tilde{\mathsf{m}}}$ and α and β satisfy

$$\begin{pmatrix}\mathbf{K}+p\lambda_{\text{reg}}\Sigma_\varepsilon & \Pi\\ \Pi^{\mathsf{T}} & 0_{\tilde{\mathsf{m}}\times\tilde{\mathsf{m}}}\end{pmatrix}\begin{pmatrix}\alpha\\ \beta\end{pmatrix} = \begin{pmatrix}\bar{\mathbf{y}}_p\\ 0_{\tilde{\mathsf{m}}}\end{pmatrix}, \tag{6}$$

and hence the R-RBF model prediction is $M_{K,S_n}(\mathbf{x}) = K_p(\mathbf{x})^{\mathsf{T}}\alpha+\pi_{\tilde{\mathsf{m}}}(\mathbf{x})^{\mathsf{T}}\beta$.

Proposition 2 below proves that the existence and the uniqueness of a solution to the linear system (6). Since there exists a unique solution to (5), this solution is also the global optimum of (4). Indeed, (6) shares a similar structure with (2) whose solution is the desired weight vector of an augmented RBF interpolation model. This similarity allows us to extend the LOOCV results obtained for RBF interpolation models to regularized RBF models (see Section 5)

Proposition 2 Under Assumption 1, $\widetilde{\mathbf{C}} = \begin{pmatrix}\mathbf{K}+p\lambda_{\text{reg}}\Sigma_\varepsilon & \Pi\\ \Pi^{\mathsf{T}} & 0_{\tilde{\mathsf{m}}\times\tilde{\mathsf{m}}}\end{pmatrix}$ is invertible and the linear system (6) is thus uniquely solvable.

3 PREDICTION ERROR ANALYSIS

In Section 2, we derived the R-RBF model prediction by applying regularization theory. A measure of the model prediction uncertainty is of great use in metamodel-based optimization to allocate computational

budgets for promising regions. By constructing an appropriate stochastic process, this section focuses on deriving the same model prediction as in Section 2 and additionally the prediction uncertainty measured by MSE. This section is divided into two parts. We first consider PD kernels that are qualified to be covariance functions of a stochastic process, and derive the MSE-optimal predictor and its associated prediction uncertainty. It is followed by a generalization of PD kernels to CPD kernels of arbitrary order, which cannot be directly used as covariance functions. Despite these crucial differences, we prove in Theorem 1 that models with PD and CPD kernels of arbitrary order share the same form of MSE-optimal prediction and the associated prediction uncertainty.

3.1 Positive Definite Kernels

We present the noisy observation on replication r at design point $\mathbf{x}$ using the following model,

$$Y_r(\mathbf{x}) = \sum_{j=1}^{\widetilde{\mathsf{m}}} \beta_j \pi_j(\mathbf{x}) + \sqrt{\kappa_Z} Z(\mathbf{x}) + \varepsilon_r(\mathbf{x}) \tag{7}$$

where $\kappa_Z > 0$ is a positive constant and κ_Z reflects our prior knowledge on the smoothness of the underlying function, $\{Z(\mathbf{x}), \mathbf{x} \in \mathscr{X}\}$ is a stochastic process with zero mean and a covariance function $\mathsf{Cov}[Z(\mathbf{x}), Z(\mathbf{x}')] = K(\mathbf{x}, \mathbf{x}')$. The noise $\varepsilon_r(\mathbf{x}), \varepsilon_r(\mathbf{x}), \ldots$ at $\mathbf{x}$ is assumed to be identically and independently distributed.

In the following, we try to derive the Best Linear Unbiased Prediction (BLUP) by following the procedures in Stein (1999) and Ankenman, Nelson, and Staum (2010). We consider the predictor of the form $\lambda_0 + \lambda^\mathsf{T} \overline{\mathbf{Y}}_p$, where $\overline{\mathbf{Y}}_p = (\frac{1}{n_1}\sum_{r=1}^{n_1} Y_r(\mathbf{x}_1), \ldots, \frac{1}{n_p}\sum_{r=1}^{n_p} Y_r(\mathbf{x}_p))$. The unbiasedness constraint implies that $\mathbb{E}[\lambda_0 + \lambda^\mathsf{T} \overline{\mathbf{Y}}_p] = \mathbb{E}[Y(\mathbf{x})]$, equivalently, $\lambda_0 + \lambda^\mathsf{T} \Pi \beta = \pi_{\widetilde{\mathsf{m}}}(\mathbf{x})^\mathsf{T} \beta$, for all $\beta \in \mathbb{R}^{\widetilde{\mathsf{m}}}$. This simply means that $\lambda_0 = 0$ and $\Pi^\mathsf{T} \lambda = \pi_{\widetilde{\mathsf{m}}}(\mathbf{x})$. The λ that yields the BLUP is obtained by minimizing the MSE of the prediction subject to the unbiasedness constraint,

$$\min_{\lambda} \mathsf{MSE}_{K,S_n}(\mathbf{x}) := \mathbb{E}\left[(Y(\mathbf{x}) - \lambda^\mathsf{T} \overline{\mathbf{Y}}_p)^2\right] = \kappa_Z K(\mathbf{x}, \mathbf{x}) - 2\kappa_Z \lambda^\mathsf{T} K_p(\mathbf{x}) + \lambda^\mathsf{T} (\kappa_Z \mathbf{K} + \Sigma_\varepsilon) \lambda$$
$$\text{subject to } \Pi^\mathsf{T} \lambda = \pi_{\widetilde{\mathsf{m}}}(\mathbf{x}),$$

where the same notations from Section 2 are used. With the KKT conditions, we show that the optimal λ and the Lagrange multiplier η associated with the constraint $\Pi^\mathsf{T} \lambda = \pi_{\widetilde{\mathsf{m}}}(\mathbf{x})$ satisfy,

$$\begin{pmatrix} 2(\mathbf{K} + \kappa_Z^{-1}\Sigma_\varepsilon) & \Pi \\ \Pi^\mathsf{T} & 0_{\widetilde{\mathsf{m}} \times \widetilde{\mathsf{m}}} \end{pmatrix} \begin{pmatrix} \lambda \\ -\eta \end{pmatrix} = \begin{pmatrix} 2\kappa_Z K_p(\mathbf{x}) \\ \pi_{\widetilde{\mathsf{m}}}(\mathbf{x}) \end{pmatrix}.$$

Proposition 3 The R-RBF model prediction is given by

$$M_{K,S_n}(\mathbf{x}) = \begin{pmatrix} \lambda^\mathsf{T} & -\eta^\mathsf{T} \end{pmatrix} \begin{pmatrix} \mathbf{y}_n \\ 0_{\widetilde{\mathsf{m}}} \end{pmatrix} = \begin{pmatrix} K_p(\mathbf{x}) \\ \pi_{\widetilde{\mathsf{m}}}(\mathbf{x}) \end{pmatrix}^\mathsf{T} \begin{pmatrix} \mathbf{K} + \kappa_Z^{-1}\Sigma_\varepsilon & \Pi \\ \Pi^\mathsf{T} & 0_{\widetilde{\mathsf{m}} \times \widetilde{\mathsf{m}}} \end{pmatrix}^{-1} \begin{pmatrix} \overline{\mathbf{y}}_p \\ 0_{\widetilde{\mathsf{m}}} \end{pmatrix},$$

and the MSE associated with the prediction $M_{K,S_n}(\mathbf{x})$ is given by,

$$\mathsf{MSE}_{K,S_n}(\mathbf{x}) = \kappa_Z \left(K(\mathbf{x}, \mathbf{x}) - \begin{pmatrix} K_p(\mathbf{x}) \\ \pi_{\widetilde{\mathsf{m}}}(\mathbf{x}) \end{pmatrix}^\mathsf{T} \begin{pmatrix} \mathbf{K} + \kappa_Z^{-1}\Sigma_\varepsilon & \Pi \\ \Pi^\mathsf{T} & 0_{\widetilde{\mathsf{m}} \times \widetilde{\mathsf{m}}} \end{pmatrix}^{-1} \begin{pmatrix} K_p(\mathbf{x}) \\ \pi_{\widetilde{\mathsf{m}}}(\mathbf{x}) \end{pmatrix} \right).$$

Notice that if we define $\kappa_Z^{-1} = p\lambda_{\text{reg}}$, then the prediction $M_{K,S_n}(\mathbf{x})$ possesses the same form as the one computed by minimizing the regularized loss in Section 2. The advantage of using a stochastic process is that it provides the MSE associated with the prediction. Furthermore, if $Z(\mathbf{x})$ is Gaussian, then we obtain a conditional distribution of the unknown function value given S_n is $\mathsf{N}\left(M_{K,S_n}(\mathbf{x}), \mathsf{MSE}_{K,S_n}(\mathbf{x})\right)$.

3.2 Conditionally Positive Definite Kernels of Arbitrary Order

Except PD kernels, a CPD kernel K of arbitrary order is not the reproducing kernel for the RKHS H, since K is only positive definite on a subspace, thus is not a valid covariance function. Direct substitution of a CPD kernel into (7) is not suitable. One way of constructing the stochastic process is to transform a CPD kernel into a PD kernel. In other words, we need the reproducing kernel for H. The way of finding such a reproducing kernel has been discussed in the literature Wahba (1990) and Wendland (2005), and it is given by

$$\widetilde{K}(\mathbf{x},\mathbf{x}') := K(\mathbf{x},\mathbf{x}') - \sum_{i=1}^{\widetilde{\mathsf{m}}} l_i(\mathbf{x})K(\mathbf{x}',\xi_i) - \sum_{i=1}^{\widetilde{\mathsf{m}}} l_i(\mathbf{x}')K(\mathbf{x},\xi_i) + \sum_{i,j=1}^{\widetilde{\mathsf{m}}} l_i(\mathbf{x})l_j(\mathbf{x}')K(\xi_i,\xi_j) + \sum_{i}^{\widetilde{\mathsf{m}}} l_i(\mathbf{x})l_i(\mathbf{x}'), \quad (8)$$

where $\Xi = \{\xi_1,\ldots,\xi_{\widetilde{\mathsf{m}}}\}$ forms a P^d_{m}-unisolvent subset, and the set of $\{l_i(\cdot)\}_{i=1}^{\widetilde{\mathsf{m}}}$ is a Lagrange-type basis of P^d_{m}. One example of Ξ is $\Xi = \{\mathbf{x}_1,\ldots,\mathbf{x}_{\widetilde{\mathsf{m}}}\}$. Therefore, any $\pi \in P^d_{\mathsf{m}}$ can be uniquely reconstructed as $\pi(\mathbf{x}) = \sum_{i=1}^{\widetilde{\mathsf{m}}} l_i(\mathbf{x})\pi(\xi_i)$. Since this equality holds for arbitrary $\pi \in P^d_{\mathsf{m}}$, we plug in the set of function $\{\pi_j(\mathbf{x})\}_{j=1}^{\widetilde{\mathsf{m}}}$ used in defining the RKHS H and establish,

$$\begin{pmatrix} \pi_1(\xi_1) & \cdots & \pi_1(\xi_{\widetilde{\mathsf{m}}}) \\ \vdots & \vdots & \vdots \\ \pi_{\widetilde{\mathsf{m}}}(\xi_1) & \cdots & \pi_{\widetilde{\mathsf{m}}}(\xi_{\widetilde{\mathsf{m}}}) \end{pmatrix} \begin{pmatrix} l_1(\mathbf{x}) \\ \vdots \\ l_{\widetilde{\mathsf{m}}}(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \pi_1(\mathbf{x}) \\ \vdots \\ \pi_{\widetilde{\mathsf{m}}}(\mathbf{x}) \end{pmatrix}. \quad (9)$$

Assumption 1 guarantees the unique solution to (9). Now, the reproducing kernel $\widetilde{K}$ for H is available, thus the R-RBF model prediction is in the form of $M_{\widetilde{K},S_n}(\mathbf{x}) = \sum_{j=1}^{\widetilde{\mathsf{m}}} \widetilde{\beta}_j \pi_j(\mathbf{x}) + \sum_{i=1}^{p} \widetilde{\alpha}_i \widetilde{K}(\mathbf{x},\mathbf{x}_i)$, where $\widetilde{\alpha} = \{\widetilde{\alpha}_i\}_{i=1}^{p}$ and $\widetilde{\beta} = \{\widetilde{\beta}_j\}_{j=1}^{\widetilde{\mathsf{m}}}$ are the weight vectors of $M_{\widetilde{K},S_n}(\mathbf{x})$. Correspondingly, we can construct a stochastic process similar to (7) except that the covariance function of $\{Z(\mathbf{x}), \mathbf{x} \in \mathscr{X}\}$ is now $\widetilde{K}$. Following similar derivations in Section 3.1, we obtain the following proposition.

Proposition 4 The regularized RBF model with a CPD kernel is given by

$$M_{\widetilde{K},S_n}(\mathbf{x}) = \begin{pmatrix} \widetilde{K}_p(\mathbf{x}) \\ \pi_{\widetilde{\mathsf{m}}}(\mathbf{x}) \end{pmatrix}^{\mathsf{T}} \begin{pmatrix} \widetilde{\mathbf{K}} + \kappa_Z^{-1}\Sigma_\varepsilon & \Pi \\ \Pi^{\mathsf{T}} & 0_{\widetilde{\mathsf{m}}\times\widetilde{\mathsf{m}}} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{y}_n \\ 0_{\widetilde{\mathsf{m}}} \end{pmatrix},$$

The MSE associated with the prediction $M_{\widetilde{K},S_n}(\mathbf{x})$ is given by,

$$\mathsf{MSE}_{\widetilde{K},S_n}(\mathbf{x}) = \kappa_Z \left(\widetilde{K}(\mathbf{x},\mathbf{x}) - \begin{pmatrix} \widetilde{K}_p(\mathbf{x}) \\ \pi_{\widetilde{\mathsf{m}}}(\mathbf{x}) \end{pmatrix}^{\mathsf{T}} \begin{pmatrix} \widetilde{\mathbf{K}} + \kappa_Z^{-1}\Sigma_\varepsilon & \Pi \\ \Pi^{\mathsf{T}} & 0_{\widetilde{\mathsf{m}}\times\widetilde{\mathsf{m}}} \end{pmatrix}^{-1} \begin{pmatrix} \widetilde{K}_p(\mathbf{x}) \\ \pi_{\widetilde{\mathsf{m}}}(\mathbf{x}) \end{pmatrix} \right).$$

In general, computation of $\widetilde{K}$ involves repeatedly solving (9) for $\{l_i(\mathbf{x})\}_{i=1}^{\widetilde{\mathsf{m}}}$, and the form of $\widetilde{K}$ is not suitable for further analysis. Theorem 1 proves that the computation of $M_{\widetilde{K},S_n}(\mathbf{x})$ and $\mathsf{MSE}_{\widetilde{K},S_n}(\mathbf{x})$ can be accomplished without resorting to $\widetilde{K}$, which is a new contribution of this paper.

Theorem 1 Given S_n, a CPD kernel K, and its corresponding PD kernel $\widetilde{K}$ obtained by Equation (8), we obtain the following equality,

$$M_{\widetilde{K},S_n}(\mathbf{x}) = M_{K,S_n}(\mathbf{x}), \quad \text{and} \quad \mathsf{MSE}_{\widetilde{K},S_n}(\mathbf{x}) = \mathsf{MSE}_{K,S_n}(\mathbf{x}).$$

Theorem 1 allows us to compute $M_{\widetilde{K},S_n}(\mathbf{x})$ and $\mathsf{MSE}_{\widetilde{K},S_n}(\mathbf{x})$ by simply replacing $\widetilde{K}$ with the CPD kernel K in Proposition 4.

4 CONNECTIONS BETWEEN R-RBF AND STOCHASTIC KRIGING

At first glance, R-RBF and SK models differ in many aspects. A R-RBF model approximates the true function with a weighted sum of radial kernels indexed by Euclidean distances between evaluated points, whereas a SK model makes predictions with a weighted sum of evaluated function values. The connections between R-RBF models and stochastic kriging are established in two parts depending on the properties of the kernel function. Because of Theorem 1, we suppress the explicit dependence on K in $M_{K,S_n}(\mathbf{x})$ and $\mathsf{MSE}_{K,S_n}(\mathbf{x})$ from now on. First, we reformulate the SK predictor given by Ankenman, Nelson, and Staum (2010) as follows,

$$\widehat{Y}(\mathbf{x}) = \beta_0 + \Sigma_{\mathsf{M}}(\mathbf{x},\cdot)^{\mathsf{T}}(\Sigma_{\mathsf{M}}+\Sigma_{\varepsilon})^{-1}(\overline{\mathbf{y}}_p - \beta_0 1_p) = \begin{pmatrix}\Sigma_{\mathsf{M}}(\mathbf{x},\cdot)\\ 1\end{pmatrix}^{\mathsf{T}}\begin{pmatrix}\Sigma_{\mathsf{M}}+\Sigma_{\varepsilon} & 1_p\\ 1_p^{\mathsf{T}} & 0\end{pmatrix}^{-1}\begin{pmatrix}\overline{\mathbf{y}}_p\\ 0\end{pmatrix},$$

with $\beta_0 = (1_p^{\mathsf{T}}(\Sigma_{\mathsf{M}}+\Sigma_{\varepsilon})^{-1}1_p)^{-1}1_p^{\mathsf{T}}(\Sigma_{\mathsf{M}}+\Sigma_{\varepsilon})^{-1}\overline{\mathbf{y}}_p$ and the optimal MSE associated with the SK predictor is

$$\mathsf{MSE}(\mathbf{x}) = \Sigma_{\mathsf{M}}(\mathbf{x},\mathbf{x}) - \Sigma_{\mathsf{M}}(\mathbf{x},\cdot)^{\mathsf{T}}(\Sigma_{\mathsf{M}}+\Sigma_{\varepsilon})^{-1}\Sigma_{\mathsf{M}}(\mathbf{x},\cdot) = \Sigma_{\mathsf{M}}(\mathbf{x},\mathbf{x}) - \begin{pmatrix}\Sigma_{\mathsf{M}}(\mathbf{x},\cdot)\\ 1\end{pmatrix}^{\mathsf{T}}\begin{pmatrix}\Sigma_{\mathsf{M}}+\Sigma_{\varepsilon} & 1_p\\ 1_p^{\mathsf{T}} & 0\end{pmatrix}^{-1}\begin{pmatrix}\Sigma_{\mathsf{M}}(\mathbf{x},\cdot)\\ 1\end{pmatrix}.$$

Analysis in Section 3 suggests that for R-RBF models using PD or CPD kernels, the MSE associated with $M_{S_n}(\mathbf{x})$ can be rewritten as

$$\mathsf{MSE}_{S_n}(\mathbf{x}) = \kappa_Z K(\mathbf{x},\mathbf{x}) - \begin{pmatrix}\kappa_Z K_p(\mathbf{x})\\ \pi_{\widetilde{\mathsf{m}}}(\mathbf{x})\end{pmatrix}^{\mathsf{T}}\begin{pmatrix}\kappa_Z\mathbf{K}+\Sigma_{\varepsilon} & \Pi\\ \Pi^{\mathsf{T}} & 0_{\widetilde{\mathsf{m}}\times\widetilde{\mathsf{m}}}\end{pmatrix}^{-1}\begin{pmatrix}\kappa_Z K_p(\mathbf{x})\\ \pi_{\widetilde{\mathsf{m}}}(\mathbf{x})\end{pmatrix}.$$

SK models consider a constant term β_0 representing the mean of the stochastic process, and $\Sigma_{\mathsf{M}}(\mathbf{x},\cdot)$ is the correlation vector between $\mathbf{x}$ and evaluated points, and Σ_{M} is the correlation matrix among evaluated points. From the above reformulations, we observe some differences and similarities between these two models.

For PD kernels, both $\Sigma_{\mathsf{M}}(\mathbf{x},\cdot)$ and $\kappa_Z K_p(\mathbf{x})$ can be interpreted as the covariance function of a stochastic process, although R-RBF models have less degrees of freedom since they assume the same degree of covariance for all dimensions. Our analysis, however, accommodate CPD kernels that are not analyzed by Ankenman, Nelson, and Staum (2010). Moreover, a CPD $\kappa_Z K_p(\mathbf{x})$ is no longer a valid covariance function since the corresponding covariance matrix is not positive definite. Indeed, in the case of CPD kernels, $\kappa_Z \widetilde{K}_p(\mathbf{x})$ forms the underlying covariance function. Surprisingly, Theorem 1 states that the choice of kernels does not alter the form of model prediction and MSE.

For CPD kernels, the polynomial space in R-RBF models needs to satisfy $\mathsf{m} \geq \mathsf{m}_K$, where the value of m_K for each radial kernel is given in Table 1. Meanwhile, for PD kernels, no restrictions are imposed on the choice of P_{m}^d for R-RBF models and the regression model in SK models. It is because PD kernels are essentially CPD of order $\mathsf{m}_K = 0$. Besides, CPD kernels are not allowed in SK models.

5 REGULARIZATION PARAMETER ESTIMATION

As discussed in Section 2, the optimal weight vector $(\alpha^{\mathsf{T}}, \beta^{\mathsf{T}})^{\mathsf{T}}$ is available through (6) but conditioned upon a known λ_{reg}. Selection of λ_{reg} is essentially the model selection problems. Maximum likelihood estimation is the most well-known method for estimating unknown parameters based on evaluated points; see Ankenman et al. (2010), Santner et al. (2003). In statistics, cross validation is another famous method for model selection and parameter estimation, especially in the regularization theory (Fang 2006). Wahba (1990) argued that cross validation methods should be more robust as they do not require the model assumption to be fulfilled. Therefore, we apply the frequentist cross validation method (Hastie, Tibshirani, and Friedman 2008) to estimate λ_{reg} and kernel parameters θ.

Consider the way of constructing R-RBF models in (6), we define $\overline{S}_p = \{\mathbf{x}_i, \overline{y}_i\}_{i=1}^p$. LOOCV partitions $\overline{S}_p$ into p parts. For each part, it constructs a model $M_{\overline{S}_{-i}}$ with the ith part of $\overline{S}_p$ removed, that is, $\overline{S}_{-i} = \overline{S}_p \setminus \{(\mathbf{x}_i, \overline{y}_i)\}$. Then the LOOCV cross validated residual of M_{S_n} is $\mathsf{V}(\overline{y}_i, M_{\overline{S}_{-i}}(\mathbf{x}_i)) = \left|\overline{y}_i - M_{\overline{S}_{-i}}(\mathbf{x}_i)\right|$, where the absolute loss function is applied. Theorem 2 defines a new criterion for parameter estimation based on LOOCV. Its main idea is that suppose M_{S_n} is reasonably accurate, then the $\frac{1}{p}\sum_{i=1}^p \left|\overline{y}_i - M_{\overline{S}_{-i}}(\mathbf{x}_i)\right| / \sqrt{\mathsf{MSE}_{K,\overline{S}_{-i}}(\mathbf{x}_i)}$ should be close to 1. A similar statement for interpolation models is given in (Cressie 1993, p. 102).

Theorem 2 The kernel parameters θ and regularization parameter λ_{reg} are determined by minimizing,

$$\mathsf{LOO}(M_{S_n}) = \frac{1}{p}\sum_{i=1}^p \left| 1 - \frac{\left|\overline{y}_i - M_{\overline{S}_{-i}}(\mathbf{x}_i)\right|}{\sqrt{\mathsf{MSE}_{\overline{S}_{-i}}(\mathbf{x}_i)}} \right| = \frac{1}{p}\sum_{i=1}^p \left| 1 - \frac{|\alpha_i|}{\left|[\widetilde{\mathbf{C}}^{-1}]_{i,i}\right|} \sqrt{\frac{1}{\mathsf{MSE}_{S_n}(\mathbf{x}_i)} - \frac{1}{\sigma_\varepsilon^2(\mathbf{x}_i)}} \right| \tag{10}$$

where $\mathsf{MSE}_{S_n}(\mathbf{x}_i)$ is the prediction MSE at sample point $\mathbf{x}_i$ and can be readily computed given M_{S_n}. The weight α_i and coefficient matrix $\widetilde{\mathbf{C}}^{-1}$ are known given M_{S_n}.

In the implementation, we replace $\sigma_\varepsilon^2(\mathbf{x}_i)$ with the sample variance $\sigma_s^2(\mathbf{x}_i)$. For kernels with parameters, not restricted to be radial kernels, we use conventional numerical algorithms (sequential quadratic programming (SQP) (Nocedal and Wright 2006) with multiple starting points) to search for $(\theta, \lambda_{\text{reg}})$. While for other radial kernels as CU and TPS, (10) reduces to a one-dimensional minimization problem of λ_{reg}, optimization of $\mathsf{LOO}(M_{S_n})$ requires repetitive evaluations of the inverse matrix $\widetilde{\mathbf{C}}^{-1}$. We apply Proposition 5, which is an immediate result from the Woodbury matrix identity (Woodbury 1950), to avoid such procedure and thus improve the computational efficiency.

Proposition 5 With eigendecomposition, we show that

$$\begin{pmatrix} \mathbf{K} + p\lambda_{\text{reg}}\Sigma_\varepsilon & \Pi \\ \Pi^\mathsf{T} & 0_{\widetilde{\mathsf{m}}\times\widetilde{\mathsf{m}}} \end{pmatrix}^{-1} = \mathbf{G}\mathsf{Diag}\left(\frac{1}{p\lambda_{\text{reg}}\delta_1 + 1}, \ldots, \frac{1}{p\lambda_{\text{reg}}\delta_{p+\widetilde{\mathsf{m}}} + 1} \right) \mathbf{G}^{-1} \begin{pmatrix} \mathbf{K} & \Pi \\ \Pi^\mathsf{T} & 0_{\widetilde{\mathsf{m}}\times\widetilde{\mathsf{m}}} \end{pmatrix}^{-1},$$

where $\mathbf{G}$ and $\mathsf{Diag}\left(\delta_1, \ldots, \delta_{p+\widetilde{\mathsf{m}}}\right)$ are from the eigendecomposition,

$$\begin{pmatrix} \mathbf{K} & \Pi \\ \Pi^\mathsf{T} & 0_{\widetilde{\mathsf{m}}\times\widetilde{\mathsf{m}}} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_\varepsilon & 0_{p\times\widetilde{\mathsf{m}}} \\ 0_{\widetilde{\mathsf{m}}\times p} & 0_{\widetilde{\mathsf{m}}\times\widetilde{\mathsf{m}}} \end{pmatrix} = \mathbf{G}\mathsf{Diag}\left(\delta_1, \ldots, \delta_{p+\widetilde{\mathsf{m}}}\right) \mathbf{G}^{-1}.$$

Proposition 5 extracts λ_{reg} from an inverse matrix thus multiple times of inverting matrices can be avoided. Matrices $\mathbf{G}, \{\delta_i\}_{i=1}^{p+\widetilde{\mathsf{m}}}$, and $\begin{pmatrix} \mathbf{K} & \Pi \\ \Pi^\mathsf{T} & 0_{\widetilde{\mathsf{m}}\times\widetilde{\mathsf{m}}} \end{pmatrix}^{-1}$ can be precomputed, significantly reducing computational efforts of estimating λ_{reg}. Instead of solving (10) using nonlinear optimization algorithms for CU and TPS radial kernels, we can approximate the optimal solution by evaluating $\mathsf{LOO}(M_{S_n})$ with a list of λ_{reg}, and choosing the one introducing the lowest LOOCV error.

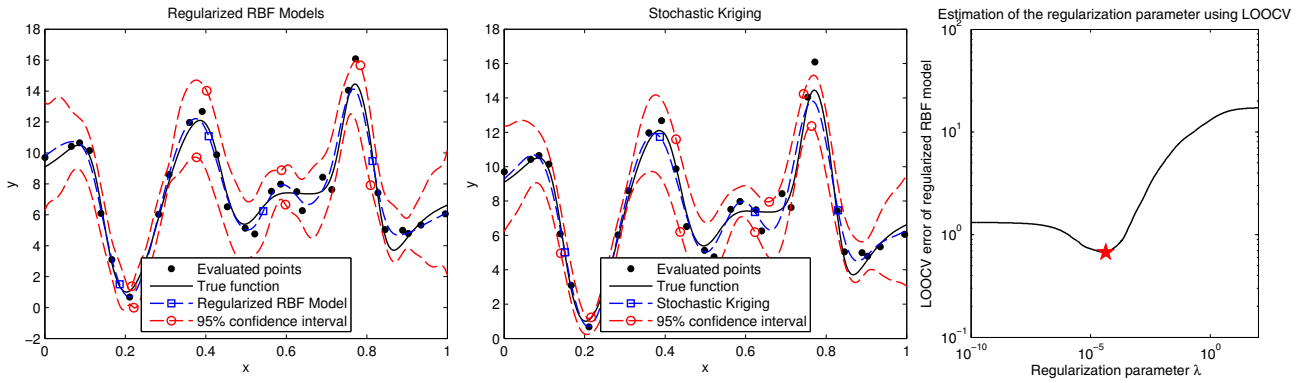
6 NUMERICAL EXPERIMENTS

6.1 One-dimensional Randomly-generated Function

We first illustrate R-RBF and SK models with a simple example. In this example, we assume the observations are from a randomly generated function and are distorted by white noise with a nonconstant variance function $\sigma_\varepsilon^2(\mathbf{x}) = 0.2f(\mathbf{x})$. The noise variance is estimated from replications of simulations. Both regularized RBF model and stochastic kriging are constructed with $p = 30$ design points, and $n_i = 5$ replications of simulation at each point $\{\mathbf{x}_i\}_{i=1}^p$.

Figure 1a portraits a regularized RBF model constructed using CU basis functions, and its 95% confidence interval. Figure 1b illustrates the stochastic kriging model based on a Gaussian correlation function. The

codes of constructing a stochastic kriging model are downloaded from http://stochastickriging.net/. Figure 1c demonstrates the estimation of the regularization parameter λ_{reg}, where the optimal λ_{reg} is shown in red star. As can be seen in Figure 1, the regularized RBF model fits well with the noisy observations and the true



(a) A regularized RBF Models. (b) A stochastic kriging model. (c) $\mathsf{LOO}(M_{S_n})$.

Figure 1: Model comparisons in the heteroscedastic case with an unknown noise variance function.

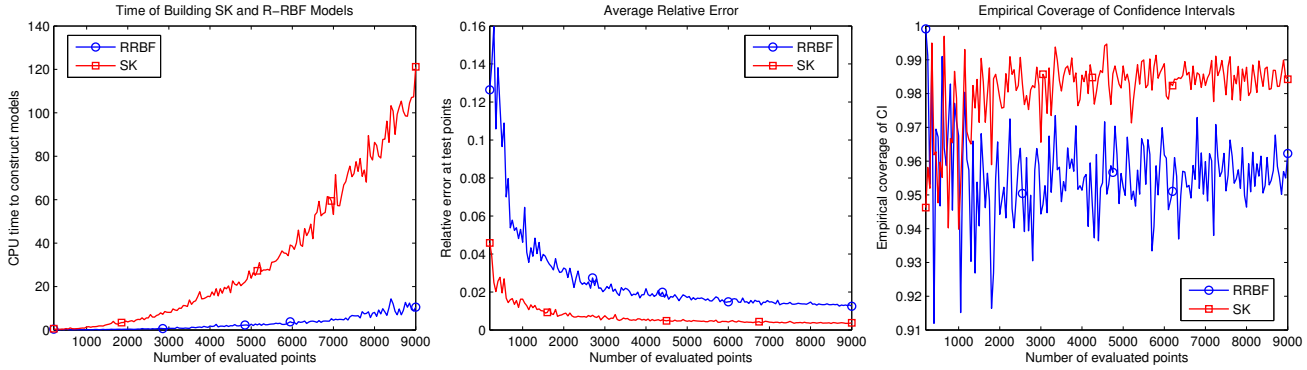
function is within the 95% confidence interval of the prediction. Notice that the confidence interval inflates at points close to the boundary, since these points are distant from evaluated points and its radial kernel function value increases quickly with the distance when using a CU kernel.

6.2 A Borehole Example

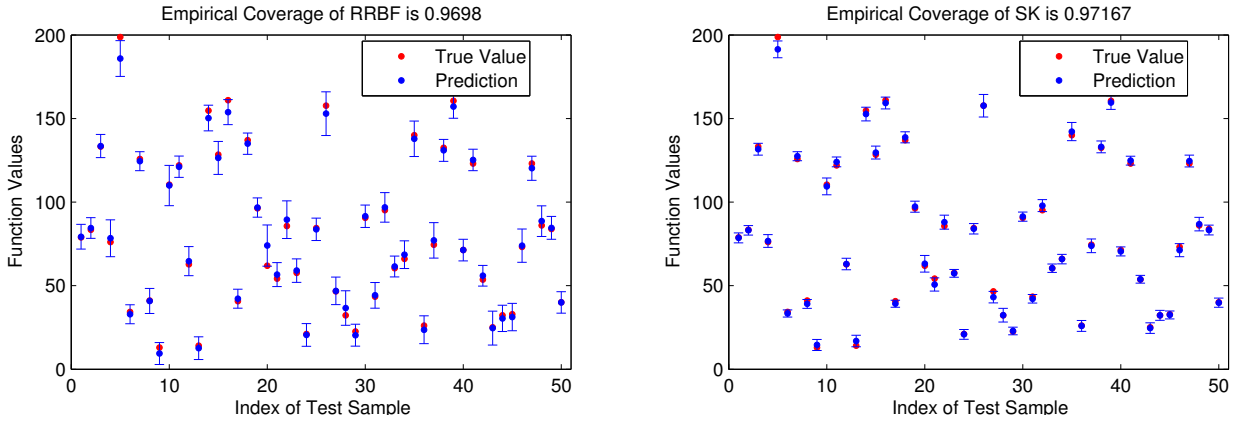
We also investigate a case study commonly used in the literature, such as Morris et al. (1993) and Fang (2006). This engineering example describes the flow rate of water through a borehole from an upper aquifer to a lower aquifer separated by an impermeable rock layer. To emulate noisy observations, we add heteroscedastic noise while evaluating sample points, that is, $y = f(\mathbf{x}) + \varepsilon(\mathbf{x})$, where $\varepsilon(\mathbf{x})$ is normally distributed with mean 0 and standard deviation $\sigma_\varepsilon^2(\mathbf{x}) = 0.1 f(\mathbf{x})$.

All experiments are carried out using Matlab 7.11 installed on a Windows machine with Intel(R) Xeon(R) 2.80GHz processor. We construct R-RBF models with a CU radial kernel and SK models with a Gaussian correlation function. Throughout the experiments, we conduct a fixed $n_i = 10$ replications of simulations at each sample point $\mathbf{x}_i$, and the maximum number of function evaluations is $N_{\max} = 9000$. For a given number p of design points, we construct the SK and R-RBF model using the same data. We compare these two types of models based on performance measures computed using a set $\mathbf{X}_{\text{test}}$ of 3×10^4 randomly-generated test points. The performance measures include (a) CPU time for constructing metamodels; (b) The average relative error at all test points; (c) The average empirical coverage of the confidence interval measured by the percentage of test points in $\mathbf{X}_{\text{test}}$ whose true function values lie in the 95% predictive confidence interval.

Figure 2a presents the CPU time of constructing metamodels when the sample size increases. Overall, the construction time of SK models increases much faster than R-RBF models. The reason is straightforward: here SK models estimate $d+2$ correlation parameters using maximum likelihood estimation. These parameters include d parameters in the correlation function, variance of the random field and the overall surface mean. In contrast, R-RBF models estimate 2 parameters (the shape and regularization parameters) by minimizing (10). The worst-case complexity of nonconvex global optimization is exponential in the problem dimension (Vavasis 1991). Therefore, we expect that the time complexity of building a R-RBF model to be smaller than that of building a SK model. In case of CPD basis functions such as CU and TPS, the time complexity is further reduced since only the regularization parameter needs to be estimated.



(a) Time to build models. (b) Average relative error at test points. (c) Empirical coverage probabilities.



(d) 95% confidence intervals at test points.

Figure 2: Comparisons between R-RBF and SK models.

A reduced number of parameters increases the robustness of metamodels for stochastic simulations. This comes at the price of certain loss in model fidelity. However numerical results suggest that the resulted R-RBF models are still of acceptable quality compared with SK models. Figure 2b measures the average relative error at test points. Overall, R-RBF models makes predictions with relative error less than 5%.

Figure 2c illustrates the empirical coverage probabilities of 95% confidence intervals of each metamodel. Figure 2d presents the confidence intervals at a subset of test points, demonstrating the model prediction and the true function values of test points. Overall, confidence intervals of R-RBF models predict the true function values with high coverage probabilities. Although R-RBF models exhibit lower model fidelity than SK models, the confidence intervals of R-RBF models can compensate such disadvantages in metamodel-based optimization. Appropriate utilizations of both model predictions and associated prediction errors measured by MSE can guide the algorithm to explore promising regions more effectively.

7 CONCLUSIONS

In this paper, we propose RBF-based metamodels for stochastic simulations with heteroscedastic noise, creating regularized RBF models. We first develop the R-RBF model prediction by minimizing the regularized loss over an infinite-dimensional function space H associated with RBF functions. The R-RBF model is unique under some mild assumptions. Furthermore, by constructing an appropriate stochastic process, we derive the same model as well as an estimate of its prediction error measured by MSE. We contribute to

mathematical analysis of R-RBF models with CPD kernels of arbitrary order, showing that they share a unified form of model prediction and pointwise prediction errors. We also develop an efficient method to estimate the model parameters $(\theta, \lambda_{\text{reg}})$ by proposing a new criterion based on LOOCV error of R-RBF models. When kernels are parameter-free, we further exploit the structure of the estimator and enhance the estimation process by means of eigen-decomposition. Finally, two numerical results are used to compare R-RBF and SK models, showing that R-RBF models are faster and more robust to construct while being reasonably accurate.

ACKNOWLEDGMENTS

This research is supported by Singapore MOE Academic Research Fund R-266-000-075-112.

REFERENCES

Ankenman, B., B. Nelson, and J. Staum. 2010. "Stochastic kriging for simulation metamodeling". *Operations Research* 58 (2): 371–382.

Cressie, N. 1993. *Statistics for spatial data*. Wiley series in probability and mathematical statistics: Applied probability and statistics. J. Wiley.

Efron, B., and R. Tibshirani. 1993. *An Introduction to the Bootstrap*. Monographs on statistics and applied probability. Chapman & Hall.

Evgeniou, T., M. Pontil, and T. Poggio. 2000. "Regularization networks and support vector machines". *Advances in Computational Mathematics* 13 (1): 1–50.

Fang, K. 2006. *Design and Modeling for Computer Experiments*, Volume 6. CRC Press.

Girosi, F., M. Jones, and T. Poggio. 1995. "Regularization theory and neural networks architectures". *Neural computation* 7 (2): 219–269.

Golub, G. H., M. Heath, and G. Wahba. 1979. "Generalized cross-validation as a method for choosing a good ridge parameter". *Technometrics* 21 (2): 215–223.

Hastie, T., R. Tibshirani, and J. H. Friedman. 2008. *The Elements of Statistical Learning*. 2nd ed. Springer-Verlag.

Haykin, S. 1994. *Neural Networks: a Comprehensive Foundation*. Prentice Hall PTR.

Hoerl, A., and R. Kennard. 1970. "Ridge regression: Biased estimation for nonorthogonal problems". *Technometrics* 12 (1): 55–67.

Kimeldorf, G., and G. Wahba. 1971. "Some results on Tchebycheffian spline functions". *Journal of Mathematical Analysis and Applications* 33 (1): 82–95.

Morris, M. D., T. J. Mitchell, and D. Ylvisaker. 1993. "Bayesian design and analysis of computer experiments: use of derivatives in surface prediction". *Technometrics* 35 (3): 243–255.

Nocedal, J., and S. Wright. 2006. *Numerical Optimization*. Springer.

Poggio, T., and F. Girosi. 1990. "Regularization algorithms for learning that are equivalent to multilayer networks". *Science* 247 (4945): 978–982.

Powell, M. 1992. "The theory of Radial Basis Function Approximation". *Advances in Numerical Analysis* 2:105–210.

Santner, T. J., B. Williams, and W. Notz. 2003. *The Design and Analysis of Computer Experiments*. Springer-Verlag.

Schölkopf, B., and A. Smola. 2001. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press.

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

Steinke, F., and B. Schölkopf. 2008. "Kernels, regularization and differential equations". *Pattern Recognition* 41 (11): 3271–3286.

Takezawa, K. 2005. *Introduction to Nonparametric Regression*. Wiley Series in Probability and Statistics. Wiley.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso". *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1): pp. 267–288.
Tikhonov, A., and V. Arsenin. 1977. *Solutions of Ill-posed Problems*. Scripta series in mathematics. Winston.
Vavasis, S. A. 1991. *Nonlinear optimization: complexity issues*. Oxford University Press, Inc.
Wahba, G. 1990. *Spline Models for Observational Data*, Volume 59. Society for Industrial Mathematics.
Wendland, H. 2005. *Scattered Data Approximation*. Cambridge University Press.
Woodbury, M. A. 1950. "Inverting modified matrices". *Memorandum report* 42:106.

AUTHOR BIOGRAPHIES

YIBO JI is a Ph.D. candidate in the Department of Industrial and Systems Engineering, National University of Singapore. He received his B.Eng. degree in Industrial Engineering at the School of Management and Engineering of Nanjing University in 2009. His research interests include simulation optimization, and metamodel-based global optimization problems. His email address is jiyibo@nus.edu.sg.

SUJIN KIM is an assistant professor in the Department of Industrial and Systems Engineering at the National University of Singapore. She received her Ph.D. degree in Operations Research from Cornell University in 2006. Before she joined the National University of Singapore, she was a visiting assistant professor in the Department of Industrial Engineering at Purdue University. Her research concerns simulation methodology and stochastic simulation-based optimization, with applications in electric power and health service systems. Her e-mail address is iseks@nus.edu.sg.