Statistica Sinica **28** (2018), 743-759
doi:https://doi.org/10.5705/ss.202016.0209

PREDICTION BASED ON THE KENNEDY-O'HAGAN CALIBRATION MODEL: ASYMPTOTIC CONSISTENCY AND OTHER PROPERTIES

Rui Tuo and C. F. Jeff Wu

Chinese Academy of Sciences and Georgia Institute of Technology

Abstract: Kennedy and O'Hagan (2001) propose a model for calibrating some unknown parameters in a computer model and estimating the discrepancy between the computer output and physical response. This model is known to have certain identifiability issues. Tuo and Wu (2016) show that there are examples for which the Kennedy-O'Hagan method renders unreasonable results in calibration. In spite of its unstable performance in calibration, the Kennedy-O'Hagan approach has a more robust behavior in predicting the physical response. In this work, we present some theoretical analysis to show the consistency of predictor based on their calibration model in the context of radial basis functions.

Key words and phrases: Bayesian inference, computer experiments, kriging.

1. Introduction

With the development of mathematical models and computational technique, simulation programs or software have become increasingly powerful for the prediction, validation and control of many physical processes. A computer simulation run, based on a virtual platform, requires only computational resources that are rather inexpensive in today's computing environment. In contrast, a physical experiment usually requires more facilities, materials, and human labor. As a consequence, a typical computer simulation run is much cheaper than its corresponding physical experiment trial. The economic benefits of computer simulations make them particularly useful and attractive in scientific and engineering research. As a branch of statistics, *design of experiments* mainly studies the methodologies on the planning, analysis and optimization of physical experiments (Wu and Hamada (2011)). Given the rapid spread of computer simulations, it is beneficial to develop theory and methods for the design and analysis of computer simulation experiments. This emerging field is commonly referred to as *computer experiments*. We refer to Santner, Williams and Notz (2003) for

more details.

The input variables of a computer experiment normally consist of factors which can be controlled in the physical process, referred to as the *control variables*, as well as some model parameters. These model parameters represent certain intrinsic properties of the physical system. For example, to simulate a heat transfer process, we need to solve a heat equation. The formulation of the equation requires the environmental settings and the initial conditions of the system that can be controlled physically, as well as the thermal conductivity which is uncontrollable and cannot be measured directly in general. For most computer simulations, the prediction accuracy of the computer model is closely related to the choice of the model parameters. A standard method for determining the unknown model parameters is to estimate them by comparing the computer outputs and the physical responses. Such a procedure is known as *calibration* for computer models, and the model parameters to be identified are called the *calibration parameters*. Kennedy and O'Hagan (2001) first study the calibration problem using ideas and methods in computer experiments. They propose a Bayesian hierarchical model to estimate the calibration parameters by computing their posterior distributions. Tuo and Wu (2016) show that the Kennedy-O'Hagan method may render unreasonable estimates for the calibration parameters. Given the widespread use of the Kennedy-O'Hagan method, it is desirable to make a comprehensive assessment of this method. For brevity, we sometimes *refer to Kennedy-O'Hagan as KO*.

This paper endeavors to study the prediction performance of the Kennedy-O'Hagan approach. First, we adopt the framework of Tuo and Wu (2016) which assumes the physical observations to be non-random. Interpolation theory in the native spaces becomes the key mathematical tool in this part. Then, we study the more realistic situation where the physical data are noisy. We employ the asymptotic theory of the smoothing splines in Sobolev spaces to obtain the rate of convergence of the KO predictor in this case.

This article is organized as follows. In Section 2 we review the Bayesian method proposed by Kennedy and O'Hagan (2001) for calibrating the model parameters and predicting for new physical responses. In Section 3 we present our main results on the asymptotic theory on the prediction performance of the KO method. Concluding remarks and further discussions are made in Section 4. Some proofs are given in the supplementary material.

2. Review on the Kennedy-O'Hagan Method

In this section we review the Bayesian method proposed by Kennedy and O'Hagan (2001). The formulation of this approach can be generalized to some extent. See, for example, Higdon et al. (2004).

Denote the experimental region for the control variables as Ω. We suppose that Ω is a convex and compact subset of $\mathbb{R}^d$. Let $\{x_1, \ldots, x_n\} \subset \Omega$ be the set of design points for the physical experiment. Denote the responses of the n physical experimental runs by $y_1^p, \ldots, y_n^p$, respectively, with p standing for "physical". Let Θ be the domain of the calibration parameter. In this article, we suppose the computer model has computer output as a deterministic function of the control variables and the calibration parameters, denoted by $y^s(x, \theta)$ for $x \in \Omega, \theta \in \Theta$ with s standing for "simulation".

We consider two types of computer models. The first is called "cheap computer simulations". In these problems each run of the computer code takes only a short time so that we can call the computer simulation code inside our statistical analysis program which is usually based on an iterative algorithm like Markov Chain Monte Carlo (MCMC). The second is called "expensive computer simulations". In these problems each run of the computer code takes a long time so that it is unrealistic to embed the computer simulation code into an iterative algorithm. A standard approach in computer experiments is to run the computer code over a set of selected points, and build a *surrogate model* based on the obtained computer outputs to approximate the underlying true function. The surrogate model can be evaluated much faster. In the statistical analysis, the response values from the surrogate model are used instead of those from the original computer model.

2.1. The case of cheap computer simulations

We model the physical response y^p in the following nonparametric manner,

$$y_i^p = \zeta(x_i) + e_i, \tag{2.1}$$

where $\zeta(\cdot)$ is an underlying function, referred to as the *true process*, and the e_i's are the observation error. We assume the e_i's are independent and identically distributed normal random variables with mean zero and unknown variance σ^2. The computer output function and the physical true process are linked by

$$\zeta(\cdot) = y^s(\cdot, \theta) + \delta(\cdot), \tag{2.2}$$

where θ denotes the "true" calibration parameter in the sense of a "best fitting"

value of the calibration parameter (Kennedy and O'Hagan (2001)), and δ denotes an underlying discrepancy function between the physical process and the computer model under the true calibration parameters. It is reasonable to believe that in most computer experiment problems, the discrepancy function δ should be nonzero and possibly highly nonlinear because the computer codes are usually built under assumptions or simplifications that do not hold true.

To estimate θ and δ, we follow a standard Bayesian procedure by imposing certain prior distributions on the unknown parameters θ and σ^2 and the unknown function $\delta(\cdot)$. In the computer experiment literature, a prominent method is to use a Gaussian process as the prior for an unknown function (Santner, Williams and Notz (2003)). There are two major reasons for choosing Gaussian processes. First, the sample paths of a Gaussian process are smooth if a smooth covariance function is chosen, which can be beneficial when the target function is smooth as well. Second, the computational burden of the statistical inference and prediction for a Gaussian process model is relatively low. Specifically, we use a Gaussian process with mean zero and covariance function $\tau^2 C_\gamma(\cdot,\cdot)$ as the prior of $\delta(\cdot)$, where C_γ is a stationary kernel with hyper-parameter γ.

In view of the finite-dimensional distribution of a Gaussian process, given τ^2 and γ, $\delta(\mathbf{x}) = (\delta(x_1), \ldots, \delta(x_n))^{\mathrm{T}}$ follows the multivariate normal distribution $N(0, \tau^2\Sigma_\gamma)$, where $\Sigma_\gamma = (C_\gamma(x_i, x_j))_{ij}$. In order to discuss the prediction problem later, we apply the data augmentation algorithm of Tanner and Wong (1987) and consider the posterior distribution of $(\theta, \delta(\mathbf{x}), \sigma^2, \gamma)$ given by

$$\begin{aligned}
&\pi(\theta, \delta(\mathbf{x}), \tau^2, \sigma^2, \gamma|\mathbf{y}^p) \\
&\propto \pi(\mathbf{y}^p|\theta, \delta(\mathbf{x}), \tau^2, \sigma^2, \gamma)\pi(\delta(\mathbf{x})|\theta, \tau^2, \sigma^2, \gamma)\pi(\theta, \tau^2, \sigma^2, \gamma) \\
&\propto \sigma^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{y}^p - y^s(\mathbf{x}, \theta) - \delta(\mathbf{x})\|^2\right\} \\
&\times \tau^{-n/2} (\det \Sigma_\gamma)^{-1/2} \exp\left\{-\frac{\delta(\mathbf{x})^{\mathrm{T}}\Sigma_\gamma^{-1}\delta(\mathbf{x})}{2\tau^2}\right\} \pi(\theta, \tau^2, \sigma^2, \gamma), \qquad (2.3)
\end{aligned}$$

where $\mathbf{y}^p = (y_1^p, \ldots, y_n^p)^{\mathrm{T}}$, $y^s(\mathbf{x}, \theta) = (y^s(x_1, \theta), \ldots, y^s(x_n, \theta))^{\mathrm{T}}$. It is not time-consuming to evaluate the posterior density function $\pi(\cdot,\cdot,\cdot,\cdot,\cdot|\mathbf{y}^p)$ because the computer code is cheap to run. A standard MCMC procedure can then be employed to draw samples from the posterior distribution. We refer to Higdon et al. (2004) for further details.

In this work, we pay special attention to the prediction for a new physical reponse at an untried point x_{new}, denoted as $y^p(x_{new})$. Samples from the posterior predictive distribution of $y^p(x_{new})$ can be drawn along with the MCMC

sampling. To see this, we note that in view of the Gaussian process assumption, given $\delta(\mathbf{x})$ and γ, $\delta(x_{new})$ follows the normal distribution

$$N(\Sigma_1^{\mathrm{T}}\Sigma_\gamma^{-1}\delta(\mathbf{x}), \tau^2(C_\gamma(x_{new}, x_{new}) - \Sigma_1^{\mathrm{T}}\Sigma_\gamma^{-1}\Sigma_1)),$$

where $\Sigma_1 = (C_\gamma(x_1, x_{new}), \ldots, C_\gamma(x_n, x_{new}))^{\mathrm{T}}$. Because in each iteration of the MCMC procedure a sample of $(\delta(\mathbf{x}), \theta, \gamma, \sigma^2)$ is drawn, we can draw a sample of $y^p(x_{new})$ from its posterior distribution $\pi(y^p(x_{new})|\mathbf{y}^p, \delta(\mathbf{x}), \theta, \tau^2, \gamma, \sigma^2)$, which is the multivariate normal distribution

$$N(y^s(x_{new}, \theta) + \Sigma_1^{\mathrm{T}}\Sigma_\gamma^{-1}\delta(\mathbf{x}), \tau^2(C_\gamma(x_{new}, x_{new}) - \Sigma_1^{\mathrm{T}}\Sigma_\gamma^{-1}\Sigma_1) + \sigma^2). \quad (2.4)$$

2.2. The case of expensive computer simulations

When the computer code is expensive to run, it is intractable to run MCMC based on (2.3) directly. Instead, we need a *surrogate* model to approximate the computer output function $y^s(\cdot, \cdot)$. In this setting Kennedy and O'Hagan (2001) use the Gaussian process model again. Suppose we first run the computer simulation over a set of design points $\{(x_1^s, \theta_1^s), \ldots, (x_l^s, \theta_l^s)\} \subset \Omega \times \Theta$. We choose a Gaussian process with mean $m_\beta(\cdot)$ and covariance function $\tau'^2 C_{\gamma'}^s(\cdot, \cdot)$ as the prior for y^s, where β, τ' and γ' are hyper-parameters. Besides, the prior processes of y^s and δ are assumed to be independent.

The Bayesian analysis for the present model is similar to that in Section 2.1, but with more cumbersome derivations. We write $\mathbf{y}^s := (y^s(x_1^s, \theta_1^s), \ldots, y^s(x_l^s, \theta_l^s))^{\mathrm{T}}$ and define $(n+l)$-dimensional vectors

$$\mathbf{x}^E = (x_1^E, \ldots, x_{n+l}^E)^{\mathrm{T}} := (x_1, \ldots, x_n, x_1^s, \ldots, x_l^s)^{\mathrm{T}},$$
$$\theta^E = (\theta_1^E, \ldots, \theta_{n+l}^E)^{\mathrm{T}} := (\theta, \ldots, \theta, \theta_1^s, \ldots, \theta_l^s)^{\mathrm{T}}.$$

By (2.1) and (2.2), the joint distribution of $\mathbf{y}^p$ and $\mathbf{y}^s$ conditional on $\theta, \sigma^2, \gamma, \beta$, and τ is

$$(\mathbf{y}^p, \mathbf{y}^s)|\sigma^2, \gamma, \beta, \tau^2, \tau'^2, \gamma' \sim N\left(m_\beta(\mathbf{x}^E), \Sigma_E + \begin{pmatrix} \Sigma_{11} + \sigma^2 I_n & 0 \\ 0 & 0 \end{pmatrix}\right),$$

where $m_\beta(\mathbf{x}^E) = (m_\beta(x_1^E), \ldots, m_\beta(x_{n+l}^E))^{\mathrm{T}}$ and

$$\Sigma_E = \left(\tau'^2 C_{\gamma'}^s\left(\left(x_i^E, \theta_i^E\right), \left(x_j^E, \theta_j^E\right)\right)\right)_{ij},$$
$$\Sigma_{11} = \left(\tau^2 C_\gamma\left(x_i, x_j\right)\right)_{ij}.$$

Then the posterior distribution of the parameters is given by

$$\pi(\theta, \sigma^2, \gamma, \beta, \tau^2, \tau'^2, \gamma'|\mathbf{y}^p, \mathbf{y}^s)$$
$$\propto \pi(\mathbf{y}^p, \mathbf{y}^s|\theta, \sigma^2, \gamma, \beta, \tau^2, \tau'^2, \gamma')\pi(\theta, \sigma^2, \gamma, \beta, \tau^2, \tau'^2, \gamma').$$

The parameter estimation proceeds in a similar manner to the MCMC scheme discussed in Section 2.1. As before, the prediction for the true process can be done along with the MCMC iterations. Noting the fact that $(y^p(x_{new}), \mathbf{y}^p, \mathbf{y}^s)$ follows a multivariate normal distribution given the model parameters, the posterior predictive distribution of $y^p(x_{new})$ can be obtained using the Bayes' theorem.

It can be seen that the modeling and analysis for the KO method with expensive computer code is much more complicated than that with cheap computer code. For the ease of mathematical analysis, our theoretical studies in the next section considers only the cases with cheap code. Hence, we omit the detailed formulae of the posterior density of the model parameters and the posterior predictive distribution of $y^p(x_{new})$ in this section.

3. Theoretical Studies

In this section we conduct some theoretical study of the power of prediction of the KO method. We consider the case of cheap computer code, but we believe that this simplification does not affect our general conclusion.

The asymptotic theory for the KO method depends on the choice of the correlation family C_γ. In the present work, we restrict ourselves to the Matérn family of kernel functions (Stein (1999)), defined as

$$C_{\upsilon,\gamma}(s,t) = \frac{1}{\Gamma(\upsilon)2^{\upsilon-1}} \left(2\sqrt{\upsilon}\gamma\|s-t\|\right)^{\upsilon} K_\upsilon \left(2\sqrt{\upsilon}\gamma\|s-t\|\right), \tag{3.1}$$

where K_υ is the modified Bessel function of the second kind. In the Matérn family, the model parameter υ dominates the smoothness of the process and γ is a scale parameter. Because the smoothness parameter υ has an effect on the rate of convergence of the prediction, for simplicity we suppose it is *fixed* in the entire data analysis.

All proofs in this section are given in the supplementary material.

3.1. A function approximation perspective

We follow the theoretical framework of Tuo and Wu (2016) to study the prediction performance of the KO method. Under this framework, the physical responses are assumed to have no random error, the e_i's in (2.1) are zero. This is an unrealistic assumption in practice, but it simplifies the model structure, so that we are able to find mathematical tools that help us to understand certain intrinsic properties of the KO method.

From (2.1), we have $y_i^p = \zeta(x_i)$, where ζ is a deterministic function (as the expectation of the physical response). Therefore, we regard the Gaussian

process modeling technique used in the KO method as a way of reconstructing the function ζ based on samples $\zeta(x_i)$.

An immediate consequence of the deterministic assumption is that $\delta(\mathbf{x}) = \mathbf{y}^p - y^s(\mathbf{x}, \theta)$ is determined by θ given the observations. Thus (2.3) is not applicable. Instead, we have

$$\begin{aligned}\pi(\theta, \tau^2, \gamma|\mathbf{y}^p) &\propto \pi(\mathbf{y}^p|\theta, \tau^2, \gamma)\pi(\theta, \tau^2, \gamma)\\ &\propto (\det \Sigma_\gamma)^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y}^p - y^s(\mathbf{x}, \theta))^{\mathrm{T}}\Sigma_\gamma^{-1}(\mathbf{y}^p - y^s(\mathbf{x}, \theta)\right\}\pi(\theta, \tau^2, \gamma).\end{aligned}$$

To differentiate between the true process ζ and its estimate based on the observations, we denote a draw from the predictive distribution $\pi(\zeta(x_{new}))$ by $\zeta^{\mathrm{rep}}(x_{new})$. Then the posterior predictive distribution $\pi(\zeta^{\mathrm{rep}}(x_{new})|\theta, \gamma, \mathbf{y}^p)$ is

$$N\left(y^s(x_{new}, \theta) + \Sigma_1^{\mathrm{T}}\Sigma_\gamma^{-1}(\mathbf{y}^p - y^s(\mathbf{x}, \theta)), \tau^2(C_{v,\gamma}(x_{new}, x_{new}) - \Sigma_1^{\mathrm{T}}\Sigma_\gamma^{-1}\Sigma_1)\right). \tag{3.2}$$

We now suppose the prior distribution $\pi(\theta, \tau^2, \gamma)$ is separable, $\pi(\theta, \tau^2, \gamma) = \pi(\theta)\pi(\tau^2)\pi(\gamma)$. Let S_θ, S_{τ^2}, and S_γ denote the supports of the distributions $\pi(\theta), \pi(\tau^2)$, and $\pi(\gamma)$, respectively. For the ease of mathematical treatment, we further suppose that S_θ is a compact subset of $\mathbf{R}$, and $S_{\tau^2} \subset [0, \tau_0^2], S_\gamma \subset [\gamma_1, \gamma_2]$ for some $0 < \tau_0^2 < +\infty$, $0 < \gamma_1 < \gamma_2 < +\infty$. The independence assumption of the prior distributions can be replaced with a more general assumption, which would not affect the validity of our theoretical analysis. However, the compact support assumption is technically unavoidable in the current treatment. Because we focus on the posterior mode, the use of the compact support assumption does not affect the practical applicability of the results.

The aim of this section is to study the asymptotic behavior of

$$\begin{aligned}\hat{\mu}_{\theta,\gamma} &= y^s(x_{new}, \theta) + \Sigma_1^{\mathrm{T}}\Sigma_\gamma^{-1}(\mathbf{y}^p - y^s(\mathbf{x}, \theta)),\\ \hat{\zeta}^2_{\tau^2,\gamma} &= \tau^2(C_{v,\gamma}(x_{new}, x_{new}) - \Sigma_1^{\mathrm{T}}\Sigma_\gamma^{-1}\Sigma_1),\end{aligned}$$

as the design points become dense in Ω, for $(\theta, \tau^2, \gamma) \in S_\theta, S_{\tau^2}, S_\gamma$. Clearly, the true posterior mean of $\zeta^{\mathrm{rep}}(x_{new})$ given by (3.2) is

$$E[\zeta^{\mathrm{rep}}(x_{new})|\mathbf{y}^p] = E[\hat{\mu}_{\hat{\theta},\hat{\gamma}}|\mathbf{y}^p],$$

where $(\hat{\theta}, \hat{\gamma})$ follows the posterior distribution $\pi(\theta, \gamma|\mathbf{y}^p)$. Here

$$\begin{aligned}&|E[\zeta^{\mathrm{rep}}(x_{new})|\mathbf{y}^p] - \zeta(x_{new})|\\ &= \left|E\left\{E[\zeta^{\mathrm{rep}}(x_{new}) - \zeta(x_{new})|\mathbf{y}^p, \hat{\theta}, \hat{\gamma}]\middle|\mathbf{y}^p\right\}\right|\\ &\le \sup_{\theta\in S_\theta, \gamma\in S_\gamma} |E[\zeta^{\mathrm{rep}}(x_{new}) - \zeta(x_{new})|\mathbf{y}^p, \theta, \gamma]|\end{aligned}$$

$$= \sup_{\theta \in S_\theta, \gamma \in S_\gamma} |\hat{\mu}_{\theta,\gamma} - \zeta(x_{new})|,$$

thus the bias of the posterior predictive mean can be bounded by the supremum of $|\hat{\mu}_{\theta,\gamma} - \zeta(x_{new})|$. Similarly, we find

$$\mathrm{Var}(\zeta^{\mathrm{rep}}(x_{new})|\mathbf{y}^p) \le \sup_{\tau^2 \in S_{\tau^2}, \gamma \in S_\gamma} \hat{\varsigma}^2_{\tau^2,\gamma}.$$

In this section we bound $\sup_{\theta \in S_\theta, \gamma \in S_\gamma} |\hat{\mu}_{\theta,\gamma} - \zeta(x_{new})|$ and $\sup_{\tau^2 \in S_{\tau^2} \gamma \in S_\gamma} \hat{\varsigma}^2_{\tau^2,\gamma}$. To this end, we resort to the theory of native spaces. We refer to Wendland (2005) for detailed discussions. For a symmetric and positive definite function Φ over $\Omega \times \Omega$, consider the linear space

$$F_\Phi(\Omega) := \left\{ \sum_{i=1}^m \alpha_i \Phi(s_i, \cdot) : m \in \mathbb{N}^+, \alpha_i \in \mathbf{R} \right\},$$

equipped with the inner product

$$\left\langle \sum_{i=1}^m \alpha_i \Phi(s_i, \cdot), \sum_{j=1}^l \beta_j \Phi(t_j, \cdot) \right\rangle = \sum_{i=1}^m \sum_{j=1}^l \alpha_i \beta_j \Phi(s_i, t_j). \tag{3.3}$$

The completion of $F_\Phi(\Omega)$ with respect to its inner product is called the native space generated by Φ, denoted by $\mathcal{N}_\Phi(\Omega)$. Denote the inner product and the norm of $\mathcal{N}_\Phi(\Omega)$ by $\langle \cdot, \cdot \rangle_{\mathcal{N}_\Phi(\Omega)}$ and $\| \cdot \|_{\mathcal{N}_\Phi(\Omega)}$, respectively.

Now we state the interpolation scheme in the native space. Let $f \in \mathcal{N}_\Phi(\Omega)$ and $\mathbf{x} = \{x_1, \ldots, x_n\}$ be a set of distinct points in Ω. Let $\mathbf{y} = (f(x_1), \ldots, f(x_n))^{\mathrm{T}}$ be the observerd data. Define

$$s_{f,\mathbf{x}}(x) = \sum_{i=1}^n u_i \Phi(x_i, x), \tag{3.4}$$

where $\mathbf{u} = (u_1, \ldots, u_n)^{\mathrm{T}}$ is given by the linear equation

$$\mathbf{y} = \Phi(\mathbf{x}, \mathbf{x})\mathbf{u}$$

for $(\Phi(\mathbf{x}, \mathbf{x}))_{ij} = \Phi(x_i, x_j)$.

Clearly, $s_{f,\mathbf{x}} \in F_\Phi$ and thus $s_{f,\mathbf{x}} \in \mathcal{N}_\Phi(\Omega)$. The next lemma can be found in Wendland (2005). For the completeness, we provide its proof in the supplementary material.

Lemma 1. *For $f \in \mathcal{N}_\Phi(\Omega)$ and a set of design points $\mathbf{x} \subset \Omega$,*

$$\langle s_{f,\mathbf{x}}, f - s_{f,\mathbf{x}} \rangle_{\mathcal{N}_\Phi(\Omega)} = 0.$$

From Lemma 1 we can deduce the Pythagorean identity

$$\|s_{f,\mathbf{x}}\|^2_{\mathcal{N}_\Phi(\Omega)} + \|f - s_{f,\mathbf{x}}\|^2_{\mathcal{N}_\Phi(\Omega)} = \|f\|^2_{\mathcal{N}_\Phi(\Omega)}. \tag{3.5}$$

Now we consider an arbitrary function $h \in \mathcal{N}_\Phi(\Omega)$ that interpolates f over $\mathbf{x}$, denoted as $f|_{\mathbf{x}} = h|_{\mathbf{x}}$. Then we have $s_{f,\mathbf{x}} = s_{h,\mathbf{x}}$ and thus (3.5) also holds true if we replace f with h. This suggests $\|s_{f,\mathbf{x}}\|_{\mathcal{N}_\Phi(\Omega)} \le \|h\|_{\mathcal{N}_\Phi(\Omega)}$, which yields the optimality condition

$$s_{f,\mathbf{x}} = \operatorname*{argmin}_{\substack{h \in \mathcal{N}_\Phi(\Omega) \\ h|_{\mathbf{x}} = f|_{\mathbf{x}}}} \|h\|_{\mathcal{N}_\Phi(\Omega)}. \tag{3.6}$$

It can be shown that the native space generated by the Matérn kernel $C_{\upsilon,\gamma}$ for $\upsilon \ge 1$ coincides with the (fractional) Sobolev space $H^{\upsilon+d/2}(\Omega)$ (Adams and Fournier (2003)), and the norms are equivalent. See Tuo and Wu (2016) for details. We can also prove that the norms of the native spaces generated by $C_{\upsilon,\gamma}$ for a set of γ values bounded away from 0 and $+\infty$ are equivalent.

Lemma 2. *Suppose $\upsilon \ge 1$. There exist constants $c_1, c_2 > 1$, so that*

$$c_1\|f\|_{H^{\upsilon+d/2}(\Omega)} \le \|f\|_{\mathcal{N}_{C_{\upsilon,\gamma}}(\Omega)} \le c_2\|f\|_{H^{\upsilon+d/2}(\Omega)} \tag{3.7}$$

holds for all $f \in H^{\upsilon+d/2}(\Omega)$ and all $\gamma \in [\gamma_1, \gamma_2]$.

Next, we turn to the error estimate of the interpolant $s_{f,\mathbf{x}}$. Wendland (2005) shows that for $u \in H^\mu(\Omega)$ with $u|_{\mathbf{x}} = 0$ and $\lfloor \mu \rfloor > d/2$,

$$\|u\|_{L_\infty(\Omega)} \le C h_{\mathbf{x},\Omega}^{\mu - d/2} \|u\|_{H^\mu(\Omega)},$$

provided that $\mathbf{x}$ is "sufficiently dense", where C is independent of $\mathbf{x}$ and u; $h_{\mathbf{x},\Omega}$ is the fill distance of the design $\mathbf{x}$ defined as

$$h_{\mathbf{x},\Omega} = \sup_{x \in \Omega} \min_{x_j \in \mathbf{x}} \|x - x_j\|.$$

Here "$\mathbf{x}$ is sufficiently dense" means that its fill distance $h_{\mathbf{x},\Omega}$ is less than a constant h_0 depending only on Ω and μ. Noting that $(f - s_{f,\mathbf{x}})|_{\mathbf{x}} = 0$ and $f - s_{f,\mathbf{x}} \in H^{\upsilon+d/2}(\Omega)$, we obtain that, for $\upsilon \ge 1$,

$$\|f - s_{f,\mathbf{x}}\|_{L_\infty(\Omega)} \le C h_{\mathbf{x},\Omega}^{\upsilon} \|f - s_{f,\mathbf{x}}\|_{H^{\upsilon+d/2}(\Omega)},$$

which, together with (3.5), yields

$$\|f - s_{f,\mathbf{x}}\|_{L_\infty(\Omega)} \le C h_{\mathbf{x},\Omega}^{\upsilon} \|f\|_{H^{\upsilon+d/2}(\Omega)}. \tag{3.8}$$

Then we apply Lemma 2 to prove Lemma 3.

Lemma 3. *Suppose $\upsilon \ge 1$. For $f \in H^{\upsilon+d/2}(\Omega)$, let $s_{f,\mathbf{x}}$ be the interpolant of f over $\mathbf{x}$ with the kernel $C_{\gamma,\upsilon}$, $\gamma \in [\gamma_1, \gamma_2]$. Then for sufficiently dense $\mathbf{x}$*

$$\|f - s_{f,\mathbf{x}}\|_{L_\infty(\Omega)} \le C h_{\mathbf{x},\Omega}^{\upsilon} \|f\|_{\mathcal{N}_{C_{\upsilon,\gamma}}(\Omega)},$$

where C is independent of the choices of f, $\mathbf{x}$ and γ.

Following the notation of Tuo and Wu (2016), we define $\epsilon(x,\theta) = \zeta(x) - y^s(x,\theta)$. It is commented by Tuo and Wu (2016) that in general θ is not estimable due to the identifiability problem, and thus neither is $\delta(\cdot) = \epsilon(\cdot,\theta)$. However, as is shown later, the function $\epsilon(\cdot,\cdot)$ can be consistently estimated using KO calibration. Suppose $\epsilon(\cdot,\theta) \in H^{\upsilon+d/2}(\Omega)$ for each $\theta \in S_\theta$. Let $\epsilon(\mathbf{x},\theta) = (\epsilon(x_1,\theta),\ldots,\epsilon(x_n,\theta))^{\mathrm{T}}$. Clearly, $\mathbf{y}^p - y^s(\mathbf{x},\theta) = \epsilon(\mathbf{x},\theta)$ and thus

$$s_{\epsilon(\cdot,\theta),\mathbf{x}}(x_{new}) = \Sigma_1^{\mathrm{T}}\Sigma_\gamma^{-1}(\mathbf{y}^p - y^s(\mathbf{x},\theta)).$$

By (3.8) we obtain

$$\begin{aligned}|\hat{\mu}_{\theta,\gamma} - \zeta(x_{new})| &= |\epsilon(x_{new},\theta) - s_{\epsilon(\cdot,\theta),\mathbf{x}}(x_{new})| \\ &\le Ch_{\mathbf{x},\Omega}^{\upsilon}\|\epsilon(\cdot,\theta)\|_{H^{\upsilon+d/2}(\Omega)} \\ &\le Ch_{\mathbf{x},\Omega}^{\upsilon}\sup_{\theta\in S_\theta}\|\epsilon(\cdot,\theta)\|_{H^{\upsilon+d/2}(\Omega)}. \end{aligned} \tag{3.9}$$

The error bound for the variance term can be obtained similarly. Elementary calculations show that

$$\Sigma_1^{\mathrm{T}}\Sigma_\gamma^{-1}\Sigma_1 = s_{C_{\upsilon,\gamma}(\cdot,x_{new}),\mathbf{x}}(x_{new}).$$

Hence we apply Lemma 3 to find

$$\begin{aligned}|\tau^2(C_{\upsilon,\gamma}(x_{new},x_{new}) - \Sigma_1^{\mathrm{T}}\Sigma_\gamma^{-1}\Sigma_1)| &= \tau^2|C_{\upsilon,\gamma}(x_{new},x_{new}) - s_{C_{\upsilon,\gamma}(\cdot,x_{new}),\mathbf{x}}(x_{new})| \\ &\le \tau_0^2 Ch_{\mathbf{x},\Omega}^{\upsilon}\|C_{\upsilon,\gamma}(\cdot,x_{new})\|_{\mathcal{N}_{C_{\upsilon,\gamma}}(\Omega)} \\ &= \tau_0^2 Ch_{\mathbf{x},\Omega}^{\upsilon}, \end{aligned} \tag{3.10}$$

where the last equality follows from the fact that $\|C_{\upsilon,\gamma}(\cdot,x_{new})\|_{\mathcal{N}_{C_{\upsilon,\gamma}}(\Omega)} = 1$. We summarize our findings in (3.9) and (3.10) as Theorem 1.

Theorem 1. *If $\upsilon \ge 1, \gamma \in [\gamma_1,\gamma_2]$, and $\tau \le \tau_0$, then for a sufficiently dense design $\mathbf{x}$, we have the upper bound for the predictive mean as*

$$\sup_{\theta\in S_\theta,\gamma\in S_\gamma}|\hat{\mu}_{\theta,\gamma} - \zeta(x_{new})| \le Ch_{\mathbf{x},\Omega}^{\upsilon}\sup_{\theta\in S_\theta}\|\epsilon(\cdot,\theta)\|_{H^{\upsilon+d/2}(\Omega)},$$

and the upper bound for the predictive variance as

$$\sup_{\tau^2\in S_\tau^2,\gamma\in S_\gamma}\hat{\varsigma}_{\tau^2,\gamma}^2 \le \tau_0^2 Ch_{\mathbf{x},\Omega}^{\upsilon},$$

with a constant C depending only on $\Omega,\upsilon,\gamma_1,\gamma_2$.

From Theorem 1, the rate of convergence is $O(h_{\mathbf{x},\Omega}^{\upsilon})$, known to be optimal in the current setting (Wendland (2005)). The predictive behavior of the KO calibration is more robust than in the case of estimation as shown by Tuo and Wu (2016), Theorem 4.2. Specifically, they show the KO calibration estimator tends to the minimizer of a norm involving the prior assumption. By comparison, the

predictive performance does not depend on the choice of the prior asymptotically.

3.2. A nonparametric regression perspective

Now we turn to the more realistic case, where the physical observations have random measurement errors. As before, we treat the true process $\zeta(\cdot)$ as a deterministic function. For the ease of mathematical treatment, in this section we fix the value of γ. Our analysis later will show that the resulting rate of convergence is not influenced by the choice of γ. Other parameters are either estimated or chosen to vary along with the sample size n.

To study the predictive behavior of the KO method asymptotically, the key is to understand the posterior mode of $\delta(\mathbf{x})$ in (2.3). We first introduce the representer theorem (Schölkopf, Herbrich and Smola (2001); Wahba (1990)), and give its proof, using Lemma 1, in the supplementary material.

Lemma 4 (Representer Theorem)**.** *Let $x_1, \ldots, x_n$ be a set of distinct points in Ω and $L : \mathbf{R}^n \to \mathbf{R}$ be an arbitrary function. If $\hat{f}$ is the minimizer of the problem*

$$\min_{f \in \mathcal{N}_\Phi(\Omega)} L(f(x_1), f(x_2), \ldots, f(x_n)) + \|f\|^2_{\mathcal{N}_\Phi(\Omega)},$$

then $\hat{f}$ possesses the representation

$$\hat{f} = \sum_{i=1}^{n} \alpha_i \Phi(x_i, \cdot),$$

with coefficients $\alpha_i \in \mathbf{R}, i = 1, \ldots, n$.

We first fix the values of τ^2, σ^2, and γ in their domain. Then we consider the profile posterior density function of $\delta(\mathbf{x})$ that, according to (2.3), is proportional to

$$\pi_{\tau^2,\sigma^2,\gamma}(\theta, \delta(\mathbf{x})) = \exp\left\{ -\frac{1}{2\sigma^2} \|\mathbf{y}^p - y^s(\mathbf{x}, \theta) - \delta(\mathbf{x})\|^2 - \frac{\delta(\mathbf{x})^{\mathrm{T}} \Sigma_\gamma^{-1} \delta(\mathbf{x})}{2\tau^2} \right\}. \tag{3.11}$$

The profile posterior mode $(\hat{\theta}_{KO}, \hat{\delta}(\mathbf{x}))$ maximizes $\pi_{\tau^2,\sigma^2,\gamma}(\cdot, \cdot)$. Using the representer theorem, we show an equality between $\hat{\delta}(\mathbf{x})$ and the solution to a penalized least squares problem.

Theorem 2. *Let $(\hat{\theta}, \hat{\Delta})$ be the solution to*

$$\operatorname*{argmin}_{\substack{\theta \in \Theta \\ f \in \mathcal{N}_{C_v,\gamma}(\Omega)}} \sum_{i=1}^{n} (y_i^p - y^s(x_i, \theta) - f(x_i))^2 + \frac{\sigma^2}{\tau^2} \|f\|^2_{\mathcal{N}_{C_v,\gamma}(\Omega)}. \tag{3.12}$$

Then $\hat{\theta} = \hat{\theta}_{KO}$ and $(\hat{\Delta}(x_1), \ldots, \hat{\Delta}(x_n))^T =: \hat{\Delta}(\mathbf{x}) = \hat{\delta}(\mathbf{x})$.

Now we are ready to state the main asymptotic theory. We first investigate the asymptotic properties of the predictive mean, then consider the consistency of the predictive variance.

From (2.4), the predictive mean of the KO model is

$$\zeta^{\mathrm{rep}}(x_{new}) = y^s(x_{new}, \hat{\theta}_{KO}) + \Sigma_1^{\mathrm{T}}(x_{new})\Sigma_\gamma^{-1}\hat{\delta}(\mathbf{x}), \tag{3.13}$$

where $\Sigma_1(x_{new}) = (C_{\upsilon,\gamma}(x_{new}, x_1), \ldots, C_{\upsilon,\gamma}(x_{new}, x_n))^{\mathrm{T}}$. Invoking Theorem 2, we have $\hat{\delta}(\mathbf{x}) = \hat{\Delta}(\mathbf{x})$ with $\hat{\Delta}$ defined in (3.12). Using (3.4), it can be seen that $\hat{\zeta}(\cdot) - y^s(\cdot, \hat{\theta}_{KO})$ is the kernel interpolant of the data $(\mathbf{x}, \hat{\Delta}(\mathbf{x}))$. Hence, from Lemma 4 and Theorem 2 we have

$$\hat{\zeta}(\cdot) - y^s(\cdot, \hat{\theta}_{KO}) = \hat{\Delta}(\cdot).$$

The ratio of the variances σ^2/τ^2 plays an important role in (3.12). In the nonparametric regression literature, such a quantity is commonly referred to as the *smoothing parameter*, a tuning parameter to balance the bias and variance of the estimator. It can be seen that as $\sigma^2/\tau^2 \to \infty$, $\hat{\epsilon}$ tends to 0, while as $\sigma^2/\tau^2 \downarrow 0$, $\hat{\epsilon}$ eventually interpolates $(x_i, y_i^p - y^s(x_i, \hat{\theta}_{KO}))$, typically an over-fit. We take $\sigma^2/\tau^2 = r_n$ when the sample size is n. According to Theorem 3, the optimal rate for r_n is $r_n \sim n^{d/(2\upsilon+2d)}$. van der Vaart and van Zanten (2008) have it that the optimal tuning rate can be automatically achieved by following a standard Bayesian analysis procedure. We do not pursue this approach here.

Some asymptotic theory for the penalized least squares problem (3.12) is available in van der Geer (2000). To use this, we need to choose the smoothing parameter r_n to diverge at an appropriate rate as n goes to infinity. For convenience, we suppose that the design points are randomly chosen. We consider the rate of convergence of the penalized least squares estimator under the L_2 metric. We assume that y^s is Lipschitz continuous. Then the metric entropy of $\{y^s(\cdot, \theta) : \theta \in \Theta\}$ is dominated by that of the unit ball of the nonparametric class $\mathcal{N}_{C_{\upsilon,\gamma}}(\Omega)$, see van der Vaart and Wellner (1996). Theorem 3 then is a direct consequence of Theorem 10.2 of van der Geer (2000), where the required upper bound for the metric entropy is obtained from (3.6) of Tuo and Wu (2015).

Theorem 3. *Suppose the design points $\{x_i\}$ are independently uniform over Ω. Let $\upsilon \geq 1$, y^s be Lipschitz continuous, and $r_n \sim n^{d/(2\upsilon+2d)}$. Under* (2.1) *with $\sigma^2 > 0$, the KO predictor $\hat{\zeta}$ at* (3.13) *satisfies*

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\zeta}(x_i) - \zeta(x_i))^2 = O_p(n^{-(2\upsilon+d)/(2\upsilon+2d)}), \tag{3.14}$$

$$\|\hat{\zeta} - \zeta\|_{L_2(\Omega)} = O_p(n^{-(\upsilon+d/2)/(2\upsilon+2d)}), \tag{3.15}$$

$$\|\hat{\zeta} - \zeta\|_{H^{\upsilon+d/2}(\Omega)} = O_p(1). \tag{3.16}$$

Since the native space $\mathcal{N}_{C_{\upsilon,\gamma}(\Omega)}$ is equivalent to the Sobolev space $H^{\upsilon+d/2}(\Omega)$, the rate of convergence in (3.15) is optimal according to Stone (1982). According to Adams and Fournier (2003),

$$\|\hat{\zeta} - \zeta\|_{L_\infty(\Omega)} \le K\|\hat{\zeta} - \zeta\|^{d/(2(\upsilon+d/2))}_{H^{\upsilon+d/2}(\Omega)}\|\hat{\zeta} - \zeta\|^{1-d/(2(\upsilon+d/2))}_{L_2(\Omega)},$$

with constant K depending only on Ω and υ. In view of (3.15) and (3.16), we have

$$\|\hat{\zeta} - \zeta\|_{L_\infty(\Omega)} = O_p(n^{-\upsilon/(2\upsilon+2d)}), \tag{3.17}$$

which gives the rate of convergence of the predictive mean under the uniform metric.

For treating the consistency of the predictive variance, we denote the true value of σ^2 by σ_0^2. Consider the profile posterior mode of σ^2 in (2.3) with the non-informative prior for σ^2, $\pi(\sigma^2) \propto 1$. The limiting value of the posterior mode of σ^2 is not affected by the choice of $\pi(\sigma^2)$, provided σ_0^2 is contained in the support of $\pi(\sigma^2)$. From (2.3) the posterior mode of σ^2 is

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{\|\mathbf{y}^p - y^s(\mathbf{x}, \hat{\theta}_{KO}) - \hat{\delta}(\mathbf{x})\|^2}{n} \\
&= \frac{1}{n}\sum_{i=1}^n \{e_i + (\hat{\zeta}(x_i) - \zeta(x_i))\}^2 \\
&= \frac{1}{n}\sum_{i=1}^n e_i^2 + \frac{2}{n}\sum_{i=1}^n e_i(\hat{\zeta}(x_i) - \zeta(x_i)) + \frac{1}{n}\sum_{i=1}^n (\hat{\zeta}(x_i) - \zeta(x_i))^2,
\end{aligned}$$

which yields the inequality

$$\left|\hat{\sigma}^2 - \frac{1}{n}\sum_{i=1}^n e_i^2\right| \le \left|\frac{2}{n}\sum_{i=1}^n e_i(\hat{\zeta}(x_i) - \zeta(x_i))\right| + \left|\frac{1}{n}\sum_{i=1}^n (\hat{\zeta}(x_i) - \zeta(x_i))^2\right|. \tag{3.18}$$

If

$$I_{\max} = \sup_{\zeta' \in H^{\upsilon+d/2}(\Omega)} \frac{1/n\sum_{i=1}^n e_i(\hat{\zeta}(x_i) - \zeta(x_i))}{(1/n\sum_{i=1}^n (\zeta'(x_i) - \zeta(x_i))^2)^{(2\upsilon)/(2\upsilon+d)}\|\zeta' - \zeta\|^{d/(2\upsilon+d)}_{H^{\upsilon+d/2}(\Omega)}},$$

then

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^n e_i(\hat{\zeta}(x_i) - \zeta(x_i))\right| \\
&\le I_{\max}\left(\frac{1}{n}\sum_{i=1}^n (\hat{\zeta}(x_i) - \zeta(x_i))^2\right)^{(2\upsilon)/(2\upsilon+d)} \|\hat{\zeta} - \zeta\|^{d/(2\upsilon+d)}_{H^{\upsilon+d/2}(\Omega)}
\end{aligned}$$

$$\leq I_{\max} O_p(n^{-(2\upsilon+d)/(2\upsilon+2d)\cdot(2\upsilon)/(2\upsilon+d)}) = I_{\max} O_p(n^{-\upsilon/(\upsilon+d)}), \tag{3.19}$$

where the second inequality follows from Theorem 3. According to the standard theory of empirical processes (see (10.6) of van der Geer (2000)),

$$I_{\max} = O_p(n^{-1/2}). \tag{3.20}$$

Combining Theorem 3, (3.18), and (3.19), we obtain

$$\left| \hat{\sigma}^2 - \frac{1}{n} \sum_{i=1}^{n} e_i^2 \right| = O_p(n^{(2\upsilon+d)/(2\upsilon+2d)}) = o_p(n^{-1/2}),$$

which, together with the Central Limit Theorem, implies

$$|\hat{\sigma}^2 - \sigma^2| = O_p(n^{-1/2}). \tag{3.21}$$

From (2.4), the predictive variance of the KO model is

$$\hat{\varsigma}^2(x_{new}) = \tau^2(C_\gamma(x_{new}, x_{new}) - \Sigma_1^{\mathrm{T}} \Sigma_\gamma^{-1} \Sigma_1) + \hat{\sigma}^2. \tag{3.22}$$

As discussed in Section 3.2, $C_\gamma(x_{new}, x_{new}) - \Sigma_1^{\mathrm{T}} \Sigma_\gamma^{-1} \Sigma_1$ is the approximation error of the kernel interpolation for the function $C_\gamma(\cdot, x_{new})$. Clearly, the error from the interpolation problem discussed in Section 3.2 should be no more than that for the smoothing problem discussed in the current section, because of the presence of the random error in the latter situation. Thus we have

$$\begin{aligned} &\sup_{x_{new} \in \Omega} |\tau^2(C_\gamma(x_{new}, x_{new}) - \Sigma_1^{\mathrm{T}} \Sigma_\gamma^{-1} \Sigma_1)| \\ &= \sup_{x_{new} \in \Omega} |r_n^{-1} \hat{\sigma}^2(C_\gamma(x_{new}, x_{new}) - \Sigma_1^{\mathrm{T}} \Sigma_\gamma^{-1} \Sigma_1)| \\ &= O_p(n^{-d/(2\upsilon+2d)} n^{-\upsilon/(2\upsilon+2d)}) = O_p(n^{-1/2}), \end{aligned} \tag{3.23}$$

where the second equality follows from the assumption $r_n \sim n^{d/(4\upsilon+4d)}$ in Theorem 3, (3.22) and (3.17). Combining (3.22) and (3.23) we obtain Theorem 4.

Theorem 4. *Under the conditions of Theorem* 3*, the error bound for the predictive variance under the uniform metric is*

$$\|\hat{\varsigma}^2(\cdot) - \sigma_0^2\|_{L_\infty(\Omega)} = O_p(n^{-1/2}).$$

Here σ_0^2 is the variance of the random noise, which is present in prediction for a new physical response. Theorems 3 and 4 reveal that the predictive distribution given by the KO method can capture the true uncertainty of the physical data in the asymptotic sense.

4. Discussion

In this work, we prove some error bounds for the predictive error given by

the Kennedy-O'Hagan method when the physical observations have no random error, and when they are noisy. We only consider the Matérn correlation family, but, were a different covariance structure used, we believe that the consistency for the predictive mean and the predictive variance still holds. Additional study is required to obtain the appropriate rate of convergence. We have ignored the estimation of some model parameters, like γ and τ^2. It is still unclear how the theory can be developed to account for this.

We have taken the smoothness parameter υ as given. From Theorems 1 and 3, a better rate of convergence comes from a larger υ, provided the target function still lies in $\mathcal{N}_{C_{\upsilon,\gamma}}(\Omega)$. Ideally, one should choose υ to be close to, but no more than the true degree of smoothness of the target function. There are different ways of choosing data-dependent υ, but the mathematical analysis is much more involved. We refer to Loh (2015) and the references therein for some related discussions.

We have assumed that the design points x_i's are random samples over Ω. In practice, one might choose design points using a systematic (deterministic) scheme. In general, if a sequence of fixed designs is used, the same (optimal) rate of convergence is retained if these designs satisfy certain space-filling conditions. We refer to Utreras (1988) for the results and necessary mathematical tools.

We have allowed the number of physical measurements to grow to infinity to obtain the rate of convergence. By comparing our results and the standard ones using radial basis functions or smoothing spline approximation, we find that the rate of convergence is not elevated by doing calibration. But there are heuristics to suggest that by doing KO calibration the predictive error can be improved by a constant factor. From the proof of Theorem 1, it can be seen that if we fix Φ and γ, the predictive error is bounded by

$$|\hat{\mu}_{\theta,\gamma} - \zeta(x_{new})| \leq C h_{\mathbf{x},\Omega}^{\upsilon} \|\epsilon(\cdot,\theta)\|_{H^{\upsilon+d/2}(\Omega)}, \tag{4.1}$$

for an arbitrarily chosen $\theta \in \Theta$. So the rate of convergence is given by $O(h_{\mathbf{x},\Omega}^{\upsilon})$, and $\|\epsilon(\cdot,\theta)\|_{H^{\upsilon+d/2}(\Omega)}$ acts as a constant factor. Tuo and Wu (2016) show that, under certain conditions, the KO estimator for the calibration parameter converges to

$$\theta' = \operatorname*{argmin}_{\theta\in\Theta} \|\epsilon(\cdot,\theta)\|_{\mathcal{N}_\Phi(\Omega)},$$

as the design points become dense over Ω. Since $\|\cdot\|_{\mathcal{N}_\Phi(\Omega)}$ is equivalent to $\|\cdot\|_{H^{\upsilon+d/2}(\Omega)}$, estimating the calibration parameter via the KO method is apparently beneficial for prediction in the sense that the upper error bound is reduced

because $\|\epsilon(\cdot,\theta')\|_{\mathcal{N}_\Phi(\Omega)} \leq \|\epsilon(\cdot,\theta)\|_{\mathcal{N}_\Phi(\Omega)}$ for all $\theta \in \Theta$. This also holds for the stochastic case, using the arguments in the proof of Theorem 10.2 of van der Geer (2000).

Supplementary Materials

Proofs of Lemma 1, Lemma 2, Lemma 4, and Theorem 2 are in the online supplement.

Acknowledgment

Tuo's work is supported by the National Center for Mathematics and Interdisciplinary Sciences in CAS and NSFC grants 11501551, 11271355, and 11671386. Wu's work is supported by NSF grant DMS 1564438. The authors are grateful to an associate editor and the referees for very helpful comments.

References

Adams, R. A. and Fournier J. J. F. (2003). *Sobolev Spaces* **140**, Access Online via Elsevier.

Higdon, D., Kennedy, M., Cavendish, J. C., Cafeo, J. A. and Ryne, R. D. (2004). Combining field data and computer simulations for calibration and prediction. *SIAM Journal of Scientific Computing* **26**, 448–466.

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

Loh, W. L. (2015). Estimating the smoothness of a Gaussian random field from irregularly spaced data via higher-order quadratic variations. *The Annals of Statistics* **43** 2766–2794.

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

Schölkopf, B., Herbrich, R. and Smola A. J. (2001). A generalized representer theorem. In *Computational Learning Theory*, 416–426. Springer.

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

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics* **10**, 1040–1053.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82**, 528–540.

Tuo, R. and Wu, C. F. J. (2016). A theoretical framework for calibration in computer models: parametrization, estimation and convergence properties. *SIAM/ASA Journal on Uncertainty Quantification* **4**, 767–795.

Tuo, R. and Wu C. F. J. (2015). Efficient calibration for imperfect computer models. *The Annals of Statistics* **43**, 2331–2352.

Utreras, F. I. (1988). Convergence rates for multivariate smoothing spline functions. *Journal of Approximation Theory* **52**, 1–27.

van der Geer, S. A. (2000). *Empirical Processes in M-estimation* **6**, Cambridge university press.

van der Vaart, A. W. and van Zanten, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics* **36**, 1435–1463.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.

Wahba, G. (1990). *Spline Models for Observational Data* **59**, Society for Industrial Mathematics.

Wendland, H. (2005). *Scattered Data Approximation*. Cambridge University Press.

Wu, C. F. J. and Hamada, M. S. (2011). *Experiments: Planning, Analysis, and Optimization* **552**, John Wiley & Sons.

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China 100190.

E-mail: tuorui@amss.ac.cn

School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA.

E-mail: jeffwu@isye.gatech.edu

(Received April 2016; accepted May 2017)