Gaussian processes with Input Location Error and Applications to the Composite Parts Assembly Process*

Wenjia Wang†, Xiaowei Yue‡, Benjamin Haaland§, and C. F. Jeff Wu¶

Abstract. This paper investigates Gaussian process modeling with input location error, where the inputs are corrupted by noise. Here, the best linear unbiased predictor for two cases is considered, according to whether there is noise at the target location or not. We show that the mean squared prediction error converges to a non-zero constant if there is noise at the target location, and provide an upper bound of the mean squared prediction error if there is no noise at the target location. We investigate the use of stochastic Kriging in the prediction of Gaussian processes with input location error, and show that stochastic Kriging is a good approximation when the sample size is large. Several numerical examples are given to illustrate the results, and a case study on the assembly of composite parts is presented. Technical proofs are provided in the Appendix.

Key words. Gaussian process; Input location error; Stochastic Kriging; Composite parts assembly.

AMS subject classifications. 62P30, 62F12, 62M40

1. Introduction.

Gaussian process (GP) modeling is widely used to recover underlying functions from scattered evaluations, possibly corrupted by noise. This method has been utilized in spatial statistics for several decades [7, 22]. Later, GP modeling has been applied in computer experiments to build emulators of their outputs [27]. In order to capture the randomness of real systems, it is natural to use stochastic simulation in computer experiments. For GP modeling, the output associated with each input can be decomposed as the sum of a mean GP output and a random error that is independent of the GP output. In stochastic simulation of computer experiments, the random error is typically i.i.d. on each input location [1]. We call the error added to the mean GP output as *output* noise. The output noise is usually from uncertainties associated with responses, such as measurement errors, computational errors, and other unquantified errors. The corresponding GP modeling with output noise is called *stochastic Kriging* (SK) [1].

*Submitted to the editors on (DATE).

Funding: Wang and Wu's work is supported by NSF grant DMS 1564438. Wang's work is also supported by NSFC grant 12101149. Wu's work is also supported by NSF grant DMS 1914632. Haaland's work is supported by NSF DMS 1621722 and DMS 1739097. Yue's work is supported by NSF CMMI 2035038 and ICTAS award.

†Data Science and Analytics Thrust, the Hong Kong University of Science and Technology (Guangzhou), and Department of Mathematics, the Hong Kong University of Science and Technology (wenjiawang@ust.hk).

‡Grado Department of Industrial & Systems Engineering, Virginia Polytechnic Institute and State University (xwy@vt.edu).

§The Department of Population Health Sciences, University of Utah (Benjamin.Haaland@hci.utah.edu).

¶The H. Milton School of Industrial and Systems Engineering, Georgia Institute of Technology (jeff.wu@isye.gatech.edu).

Besides output noise, in some cases, the input variables are also corrupted by noise. Noisy or uncertain inputs are quite common in spatial statistics, because geostatistical data are often indexed by imprecise locations. Detailed examples can be found in [2, 32]. We call the random error of input variables as *input location* noise. The input location noise comes from the natural uncertainties inherent to the complex systems, such as actuating uncertainty, controller fluctuation, and internal measurement error. In contrast to the output noise that is related to the response, input location noise is associated with input variables. If the input variables are corrupted by noise in a GP, it is known as a GP with input location error, and the corresponding best linear unbiased predictor is called Kriging adjusting for location error (KALE) [8]. Also see [4, 9, 15, 23] for more discussions. KALE has been applied in many areas, including robotics [10], wireless networks [24], and Wi-Fi fingerprinting [18].

KALE predicts the mean GP output at point $x \in \Omega$ *without* input location noise. In many applications, however, the prediction of the mean GP output at point $x \in \Omega$ *with* input location noise is desired. A motivating example is the composite aircraft fuselage assembly process. In this process, a model is needed to predict the dimensional deviations under noisy actuators' forces. Further, when new actuator forces are implemented in practice, there is an inevitable input location noise, i.e., uncertainty in the actually delivered actuator forces. Therefore, the output at point $x \in \Omega$ has input location noise. Under this scenario, we consider Kriging adjusting for location error and noise (KALEN), which is the best linear unbiased predictor of the mean GP output at point $x \in \Omega$ with input location noise. For another example, in the electric stability control system of vehicles, a model is developed to link the inputs (i.e., braking pressure and engine torque) and the outputs (i.e., stability control loss). Input location noise inevitably exists in this system due to the uncertainties in wheel pressure modulators, pressure reservoir, and electric pump. Other than the two examples mentioned above, KALEN fits many applications better than KALE due to the ubiquity of actuating errors in engineering systems.

In this paper, we discuss three predictors, KALE, KALEN, and SK, applied in prediction and uncertainty quantification of GP modeling with input location error. We show that unlike GP modeling without location error, the mean squared prediction error (MSPE) does not converge to zero as the sample size goes to infinity. Furthermore, we show that the limiting MSPE of KALEN and SK are equal if point $x \in \Omega$ has input location noise. We obtain an asymptotic upper bound on the MSPE of KALE and SK if there is no noise at point $x \in \Omega$. This upper bound is small if the input location noise at observed points is slight. Numerical results indicate that if the sample size is relatively small and noise is rather large, KALE or KALEN have a much smaller MSPE, and thus are desirable, compared with SK. If the sample size is large or the noise is quite small, then the performance of all three approaches is similar. We also compare the performance of KALEN and SK in the modeling of a composite parts assembly process problem. We find that the KALEN and SK are comparable across a range of small input location noise levels, corresponding to a range of actuator tolerances, which is consistent with the theoretical analysis.

The remainder of this article is structured as follows. In Section 2, we formally state the problem, introduce KALE and KALEN, and show some asymptotic properties of the MSPE

of KALE and KALEN. Section 3 presents some theoretical results when using SK in the prediction of GPs with input location error. Parameter estimation methods are discussed in Section 4. Numerical results are presented in Section 5. A case study of the composite parts assembly process is considered in Section 6. Technical details are given in the Appendix.

2. GPs with Input Location Error.

In this section, we introduce two predictors of GPs with input location error, KALE and KALEN. We also give several asymptotic properties of KALE and KALEN.

2.1. Two Predictors of GPs with Input Location Error.

Suppose f is an underlying function defined on $\mathbb{R}^d$, and the values of f on a convex and compact set Ω are of interest. Suppose we observe the responses $f(x_1), \ldots, f(x_n)$ on $X = \{x_1, \ldots, x_n\} \subset \Omega$. Following the terminology in design of experiments [38], we call $X = \{x_1, \ldots, x_n\}$ design points. A standard tool to build emulators based on observed data is GP modeling (see [13] and [28], for example). In GP modeling, the underlying function f is assumed to be a GP. We suppose f is *stationary*, which means that the covariance of $f(x)$ and $f(x')$ depends only on the difference $x - x'$ between the two input variables x and x'. We further assume $\mathrm{Cov}(f(x), f(x')) = \sigma^2\Psi(x - x')$, where σ^2 is the variance, and Ψ is the correlation function. Then Ψ must be positive definite and satisfy $\Psi(0) = 1$. Since f is defined on $\mathbb{R}^d$, Ψ should also be defined on $\mathbb{R}^d$. In GP modeling, one can assume that the mean of f is zero, a constant, or a linear combination of known functions. The corresponding methods are referred to as simple Kriging, ordinary Kriging, and universal Kriging, respectively. Ordinary Kriging and universal Kriging are more flexible and may improve the prediction performance, but the estimation of the mean function introduces more uncertainties. Moreover, Theorem 3 of [34] suggests that the estimation of the mean function can be inconsistent. These uncertainties and inconsistency make the theoretical analysis more cumbersome, and dilute the focus of the overall analysis. Therefore, for the ease of mathematical treatment, we assume the mean of f is zero in theoretical developments in Sections 2-4, which is equivalent to removing the mean surface. Nevertheless, we use a non-zero mean function in numerical and case studies to improve the prediction performance by introducing more degrees of freedom.

For a GP with input location error, the inputs are corrupted by noise. In this paper, we mainly focus on the input location error and assume the responses are not influenced by the output noise. It is worth noting that this assumption can be relaxed, and the GP with both input location error and output noise can be analyzed in a similar manner, as stated in Remark 2.1. Specifically, suppose the responses are perturbed by the input location error, that is, we observe $y_j = f(x_j + \epsilon_j)$ for $x_j \in X$, where the ϵ_j's are i.i.d. random vectors with mean 0, and have a probability density function $p(\cdot)$. Therefore, although x_j is known, the actual location $x_j + \epsilon_j$ is unknown and we observe the response $f(x_j + \epsilon_j)$ on this unknown location. It is possible to have replicates on some design points, i.e., for some $j \neq k$, $x_j = x_k$ for $x_j, x_k \in X$ but $\epsilon_j \neq \epsilon_k$. We assume $p(\cdot)$ is continuous and each element of ϵ_j has finite variance (note that ϵ_j is a vector).

Following the approach in [8], the best linear unbiased predictor of $f(x)$ on a point x is given

by

$$Q(Y;x) = \alpha_1^T Y + \alpha_2, \tag{1}$$

where $\alpha_1 \in \mathbb{R}^n, \alpha_2 \in \mathbb{R}$ are the solution to the optimization problem

$$\min_{(\alpha_1,\alpha_2)} \mathbb{E}(f(x) - Q(Y;x))^2 = \min_{(\alpha_1,\alpha_2)} \mathbb{E}(f(x) - \alpha_1^T Y - \alpha_2)^2, \tag{2}$$

and the responses on the design points are $Y = (y_1, \ldots, y_n)^T$. By minimizing (2) with respect to (α_1, α_2), we obtain the solution to (2) is $\alpha_1 = R^{-1}r(x)$ and $\alpha_2 = 0$, where $r(x) = (r(x, x_1), \ldots, r(x, x_n))^T$ denotes the covariance vector between $f(x)$ and Y with

$$r(x, x_j) = \mathbb{E}(f(x)y_j) = \sigma^2 \int_{\mathbb{R}^d} \Psi(x - (x_j + \epsilon_j))p(\epsilon_j)d\epsilon_j, \tag{3}$$

and $R = (R_{jk})_{jk}$ denotes the covariance matrix with

$$R_{jk} = \mathbb{E}(y_j y_k) = \begin{cases} \sigma^2\Psi(x_j - x_j) = \sigma^2, & j = k, \\ \sigma^2 \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \Psi(x_j + \epsilon_j - (x_k + \epsilon_k))p(\epsilon_j)p(\epsilon_k)d\epsilon_j d\epsilon_k, & j \neq k. \end{cases} \tag{4}$$

Plugging $\alpha_1 = R^{-1}r(x)$ and $\alpha_2 = 0$ into (1), we find the best linear unbiased predictor of $f(x)$ is

$$\hat{f}(x) = r(x)^T R^{-1} Y. \tag{5}$$

Remark 2.1. If the observations also have i.i.d. distributed output noise with mean zero and finite variance σ_δ^2, we only need to replace $\mathbb{E}(y_j y_j) = \sigma^2$ by $\mathbb{E}(y_j y_j) = \sigma^2 + \sigma_\delta^2$, and the rest of the theoretical analysis remains similar. Our theoretical analysis can also be generalized to the case that ϵ_i's are independent but not identically distributed. Although these generalizations do not influence the theoretical development a lot, they could dilute the main focus of this paper. Therefore, we focus on the GPs with only i.i.d. input location noise.

In [8] equation (5) is referred to as Kriging adjusting for location error (KALE). If the prediction of $y(x) = f(x + \epsilon)$ on a point x with input location noise is of interest, it can be shown that we only need to replace $r(x)$ in (5) by $r_N(x) = (r_N(x, x_1), \ldots, r_N(x, x_n))^T$, where

$$r_N(x, x_j) = \sigma^2 \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \Psi(x + \epsilon - (x_j + \epsilon_j))p(\epsilon_j)p(\epsilon)d\epsilon_j d\epsilon. \tag{6}$$

We refer to the corresponding best linear unbiased predictor $\hat{y}(x) = r_N(x)^T R^{-1} Y$ as Kriging adjusting for location error and noise (KALEN). One direct relation between KALE and KALEN is $\hat{y}(x) = \int_{\mathbb{R}^d} \hat{f}(x + \epsilon)p(\epsilon)d\epsilon$.

In some cases, there exist closed forms of the integrals in (3)–(6). For example, if the correlation function $\Psi(s - t) = \exp(-\theta\|s - t\|_2^2)$, and the noise $\epsilon \sim N(0, \sigma_\epsilon^2 I_d)$, where $\theta > 0$ is

the correlation parameter, and $N(0, \sigma_\epsilon^2 I_d)$ is a mean zero normal distribution with covariance matrix $\sigma_\epsilon^2 I_d$, then (3)–(6) can be calculated respectively as [6]

$$
\begin{aligned}
R_{jk} &= \begin{cases} \sigma^2 & j = k, \\ \frac{\sigma^2}{(1+4\sigma_\epsilon^2\theta)^{d/2}} e^{\frac{-\theta\|x_j - x_k\|_2^2}{1+4\sigma_\epsilon^2\theta}} & j \neq k, \end{cases} \\
r(x, x_j) &= \frac{\sigma^2}{(1+2\sigma_\epsilon^2\theta)^{d/2}} e^{\frac{-\theta\|x - x_j\|_2^2}{1+2\sigma_\epsilon^2\theta}}, \\
r_N(x, x_j) &= \frac{\sigma^2}{(1+4\sigma_\epsilon^2\theta)^{d/2}} e^{\frac{-\theta\|x - x_j\|_2^2}{1+4\sigma_\epsilon^2\theta}}.
\end{aligned} \tag{7}
$$

We also include the calculation of (7) in Appendix C for readers' reference.

Unfortunately, in general, equations (3)–(6) are intractable and are typically estimated via Monte Carlo integration by sampling ϵ_j's from $p(\cdot)$, which can be computationally expensive. For example, if we choose the Matérn correlation function, then (5) does not have a closed form. In this case, the calculation of (5) will require much time, as we will see in Section 5.

With equations (3)–(6), the MSPE of KALE can be calculated by

$$
\begin{aligned}
\mathbb{E}(f(x) - \hat{f}(x))^2 &= \mathbb{E}(f(x) - r(x)^T R^{-1} Y)^2 \\
&= \mathbb{E}(f(x)^2) - 2r(x)^T R^{-1}\mathbb{E}(f(x)Y) + r(x)^T R^{-1}\mathbb{E}(YY^T)R^{-1}r(x) \\
&= \sigma^2 - r(x)^T R^{-1} r(x),
\end{aligned} \tag{8}
$$

where $\hat{f}$ is as in (5), and r and R are as defined in (3) and (4), respectively. The last equality is true because of (3) and (4), and $\mathbb{E}(f(x)^2) = \Psi(0) = 1$. Similarly, one can check the MSPE of KALEN is

$$
\mathbb{E}(y(x) - \hat{y}(x))^2 = \sigma^2 - r_N(x)^T R^{-1} r_N(x), \tag{9}
$$

where r_N is as defined in (6).

2.2. Asymptotic behaviors of KALEN. In this subsection, we consider asymptotic behaviors of KALEN. Define

$$
\Psi_S(s-t) = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \Psi(s + \epsilon_1 - (t + \epsilon_2))p(\epsilon_1)p(\epsilon_2)d\epsilon_1 d\epsilon_2. \tag{10}
$$

Notice that the MSPE of KALEN can be expressed as

$$
\begin{aligned}
\mathbb{E}(y(x) - \hat{y}(x))^2 &= \sigma^2 - r_N(x)R^{-1}r_N(x) \\
&= \sigma^2(1 - \Psi_S(0)) + \sigma^2\Psi_S(0) - r_N(x)R^{-1}r_N(x) \\
&= \underbrace{\sigma^2(1 - \Psi_S(0))}_{\text{a constant}} + \underbrace{\sigma^2\Psi_S(0) - r_N(x)(R_S + \sigma^2(1 - \Psi_S(0))I_n)^{-1}r_N(x)}_{\text{“MSPE of SK”}},
\end{aligned} \tag{11}
$$

where $R_S = \sigma^2(\Psi_S(x_j - x_k))_{jk}$ and I_n is an identity matrix. Intuitively, if the second term is indeed an MSPE of SK, then it converges to zero, and the MSPE of KALEN converges to a constant. However, the second term is an MSPE of SK unless Ψ_S is a valid correlation function (thus R_S is positive definite) and that is why we add quote marks in (11). In Proposition 3.1 of [6], it is shown that if a function $c(s,t) = \Psi_S(s-t)$ for $s \neq t$ and $c(s,s) = 1$, then $c(\cdot,\cdot)$ is a valid correlation function. Therefore, the covariance matrix R is positive definite. In order to show R_S in (11) is also positive definite, we assume the correlation function Ψ satisfies the following assumption, which is also assumed to be true in the rest of Section 2 and Section 3.

Assumption 2.2. *The correlation function Ψ is a radial basis function, i.e., $\Psi(s-t) = \phi(\|s-t\|_2)$ for $s,t \in \mathbb{R}^d$. Furthermore, $\phi(r) > 0$ is a strictly decreasing function of $r \in \mathbb{R}^+$, with $\phi(0) = 1$. The reproducing kernel Hilbert space generated by Ψ can be embedded into a Sobolev space $H^\eta(\Omega)$ with $\eta > d/2$.*

Remark 2.3. For a brief introduction to the reproducing kernel Hilbert space, see Appendix A.

Many widely used correlation functions, including isotropic Gaussian correlation functions and isotropic Matérn correlation functions, satisfy this assumption. See Appendix A for details. For an anisotropic correlation function that has form $\Psi(s-t) = \phi(\|A(s-t)\|_2)$ with A a diagonal positive definite matrix and $s,t \in \mathbb{R}^d$, we can stretch the space Ω to Ω' such that $\Psi_1(s'-t') := \Psi(s-t) = \phi(\|s'-t'\|_2)$ for $s',t' \in \Omega'$. Assumption 2.2 implies $\Psi_S(0) < 1$. With Assumption 2.2, we can show that Ψ_S is a positive definite function, which is stated in the following lemma whose proof is given in Appendix D.

Lemma 2.4. *Suppose Assumption 2.2 holds. Then Ψ_S is a positive definite function.*

Next, we consider the asymptotic properties of the MSPE of KALEN defined in (9) as the fill distance goes to zero, where the fill distance h_X of the design points X is defined by

$$h_X := \sup_{x\in\Omega} \min_{x_j \in X} \|x - x_j\|_2. \tag{12}$$

Specifically, we consider a sequence of designs X_m, $m = 1, 2, \ldots$ and we assume the following.

Assumption 2.5. *The sequence of design points $X_m = \{x_1, \ldots, x_{n_m}\}$ satisfies that there exists a constant $C > 0$ such that $h_{X_m} \leq C q_{X_m}$ for all m, where*

$$q_{X_m} = \min_{x_j \neq x_k, x_j, x_k \in X_m} \|x_j - x_k\|_2/2,$$

and h_{X_m} is the fill distance of X_m defined by (12).

Remark 2.6. Assumption 2.5 implies that the *distinct* design points are *quasi-uniform* [37].

It is not hard to find designs that satisfy this assumption. For example, grid designs satisfy Assumption 2.5. In the rest of the paper, we suppress the dependence of X on m for notational simplicity. It can be shown that if a GP has no input location noise, then the MSPE of the corresponding best linear unbiased predictor converges to zero as the fill distance goes to zero (see Lemma B.1 in Appendix B). Unlike a GP without input location error, we show that the

limit of the MSPE of KALEN is usually not zero. In fact, (11) and Lemma 2.4 imply that the MSPE of KALEN is the MSPE of SK plus a non-zero constant. These results are stated in Theorem 2.7, whose proof is provided in Appendix E.

Theorem 2.7. *Suppose Assumptions 2.2 and 2.5 hold. The MSPE of KALEN (9) converges to $\sigma^2(1-\Psi_S(0))$ as the fill distance of the design points h_X converges to zero, where Ψ_S is defined in (10).*

In Theorem 2.7, we present a limit of the MSPE of KALEN. The limit $\sigma^2(1-\Psi_S(0))$ is usually not zero. This is expected for KALEN since there is a random error at point x. The MSPE limit depends on two parts. One is the variance σ^2 and the other is the difference $1-\Psi_S(0)$. The variance σ^2 depends on the underlying process, while the difference depends on the probability density function of the noise $p(\cdot)$. Roughly speaking, the difference $1-\Psi_S(0)$ will be larger if the density $p(\cdot)$ is more spread out.

3. Comparison Between KALE/KALEN and SK. It is argued in [8] and [29] that using a nugget term is one way to counteract the influence of noise within the inputs. Therefore, it is natural to ask whether SK (or Kriging with a nugget term, see Remark 3.2 for discussion of the use of terminologies) is a good approximation method to predict the value at a point $x \in \Omega$, since it is not the best linear unbiased predictor under the settings of GP with input location error. In this paper, we show that the MSPE of SK has the same limit as the MSPE of KALEN, and provide an upper bound on the MSPE of SK if the target point x has no noise, as stated in Theorem 3.1. The proof can be found in Appendix F.

Theorem 3.1. *Suppose Assumptions 2.2 and 2.5 hold. Let $\mu > 0$ be any fixed constant. A SK predictor of a GP with input location error is defined as*

$$\hat{f}_S(x) = r_\Psi(x)(R_\Psi + \mu I_n)^{-1} Y, \tag{13}$$

where $r_\Psi(x) = (\Psi(x-x_1), \ldots, \Psi(x-x_n))^T$ and $R_\Psi = (\Psi(x_j - x_k))_{jk}$.

(i) Suppose there is noise at a point $x \in \Omega$ and $y(x)$ is to be predicted. The MSPE of the predictor (13), $\mathbb{E}(y(x)-\hat{f}_S(x))^2$, has the same limit as KALEN, which is $\sigma^2(1-\Psi_S(0))$, where Ψ_S is as defined in (10), when the fill distance of X goes to zero.

(ii) Suppose there is no noise at a point $x \in \Omega$ and $f(x)$ is to be predicted. An asymptotic upper bound on the MSPE of the predictor (13), $\mathbb{E}(f(x)-\hat{f}_S(x))^2$, is

$$\frac{1.04\sigma^2}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left|1-|b(t)|\right|^2 \mathcal{F}(\Psi)(t)dt, \tag{14}$$

where $\mathcal{F}(\Psi)$ is the Fourier transform of Ψ and $b(t) = \mathbb{E}(e^{i\epsilon^T t})$ is the characteristic function of $p(\cdot)$.

Remark 3.2. The form of the SK predictor is quite similar to that of the simple Kriging with an additional term μI_n. Following the terminology in computer experiments [16, 25], we call μI_n a "nugget" term. Despite a similar form, there are some distinct rationales for including a nugget term. In spatial statistics, the nugget term can accommodate discontinuities in the covariance

function (such variation is called the *nugget effect*) [26, 29], and the corresponding predictor is still an interpolator if there is no noise [26]. In deterministic computer experiments, the nugget term can be used to stabilize computation of the matrix inverse [16, 25]. The nugget term can also be used to counteract the influence of output noise in stochastic computer simulations and spatial statistics [1, 29]. In the latter two scenarios, the corresponding predictor is no longer an interpolator.

Remark 3.3. We say b is an asymptotic upper bound on a sequence a_n, if there exists a sequence b_n such that $a_n \leq b_n$ for all $n = 1, 2, \ldots$, and $\lim_{n\to\infty} b_n = b$.

Remark 3.4. The constant 1.04 in (14) is not essential. It can be changed to any constant greater than one, but a smaller constant leads to a "slower" convergence speed.

Remark 3.5. Note that KALE is the best linear unbiased predictor when a point $x \in \Omega$ has no noise. Therefore, the upper bound of MSPE for SK is also an upper bound of MSPE for KALE. For an illustration of the upper bound and lower bound of the MSPE of KALE, see Example 3.7.

Theorem 3.1 shows that the predictor (13) is as good as KALEN asymptotically. The following proposition states that if the noise is small, then (14) can be controlled. The proof of Proposition 3.6 can be found in Appendix G.

Proposition 3.6. *Suppose Assumption 2.2 holds, and* $\{\epsilon_n\}$ *is a sequence of independent random vectors that converges to* 0 *in distribution. Let*

$$a_n = \frac{\sigma^2}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left|1 - |b_n(t)|\right|^2 \mathcal{F}(\Psi)(t)dt, \tag{15}$$

where $b_n(t) = \mathbb{E}(e^{i\epsilon_n^T t})$. *Then* a_n *converges to zero.*

Example 3.7. Consider a GP f with mean zero and covariance function $\sigma^2\Psi$. Suppose the correlation function $\Psi(s-t) = \exp(-\theta\|s-t\|_2^2)$ with $\theta > 0$, and the input location noise $\epsilon_j \sim N(0, \sigma_\epsilon^2 I_d)$ are i.i.d., where $N(0, \sigma_\epsilon^2 I_d)$ is a mean zero normal distribution with covariance matrix $\sigma_\epsilon^2 I_d$. By Theorem 3.1, the limit of the MSPE of KALEN $\mathbb{E}(y(x) - \hat{y}(x))^2$ and SK $\mathbb{E}(y(x) - \hat{f}_S(x))^2$ is $\sigma^2(1 - \Psi_S(0))$, which can be computed by

$$\begin{aligned}\sigma^2(1-\Psi_S(0)) =& \sigma^2\left(1 - \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \Psi(x+\epsilon_1-(x+\epsilon_2))p(\epsilon_1)p(\epsilon_2)d\epsilon_1 d\epsilon_2\right)\\ =& \sigma^2 - r_N(x,x) = \sigma^2 - r_N(x_j,x_j) = \sigma^2\left(\frac{(1+4\sigma_\epsilon^2\theta)^{d/2}-1}{(1+4\sigma_\epsilon^2\theta)^{d/2}}\right),\end{aligned} \tag{16}$$

where $r_N(x_j, x_j)$ is as in (7) with $x = x_j$.

If there is no noise at point x, Theorem 3.1 states that an asymptotic upper bound of MSPE $\mathbb{E}(f(x) - \hat{f}_S(x))^2$ for SK is

$$\frac{1.04\sigma^2}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left|1 - |b(t)|\right|^2 \mathcal{F}(\Psi)(t)dt.$$

Note that the characteristic function of $N(0, \sigma_\epsilon^2 I_d)$ is $b(t) = \mathbb{E}(e^{i\epsilon^T t}) = e^{-\frac{1}{2}\sigma_\epsilon^2 t^T t}$, and $\mathcal{F}(\Psi)(t) = \theta^{-d/2} e^{-\frac{t^T t}{4\theta}}$. Thus, the upper bound can be computed by

$$\begin{aligned} \frac{1.04\sigma^2}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left|1 - |b(t)|\right|^2 \mathcal{F}(\Psi)(t) dt &= \frac{1.04\sigma^2}{(2\pi\theta)^{d/2}} \int_{\mathbb{R}^d} (1 - e^{-\sigma_\epsilon^2 t^T t/2})^2 e^{-\frac{t^T t}{4\theta}} dt \\ &= 1.04\sigma^2 \left(1 + \frac{1}{(1 + 4\sigma_\epsilon^2\theta)^{d/2}} - \frac{2}{(1 + 2\sigma_\epsilon^2\theta)^{d/2}}\right). \end{aligned} \tag{17}$$

Figure 1 shows the plot of limit (16) and the asymptotic upper bound (17) with $\theta = 1$ and $\sigma^2 = 1$. It can be seen that as the variance of noise increases, both (16) and (17) increase, and (17) is larger than (16). From Panel 1 and Panel 2 of Figure 1, the error is more prominent if the dimension of the space is larger. This indicates that GP with input location error is also influenced by the dimension as in many statistic problems.

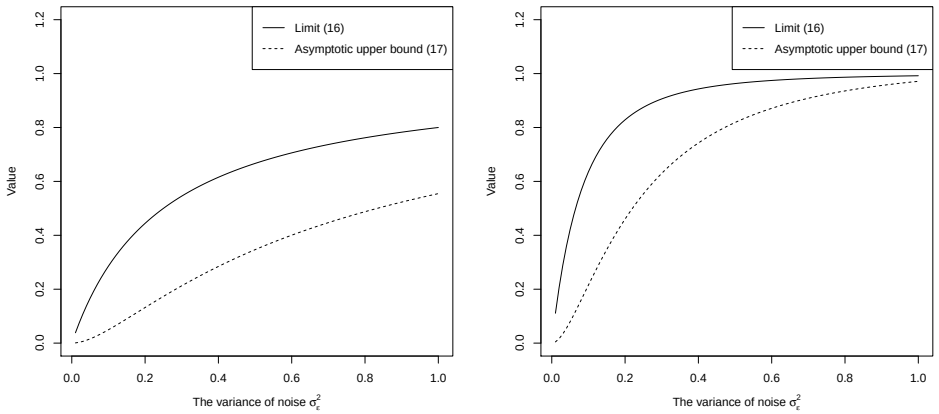

Figure 1. *The limit* (16) *and the asymptotic upper bound* (17) *with* $\theta = 1$ *and* $\sigma^2 = 1$. ***Panel 1:*** $d = 2$. ***Panel 2:*** $d = 6$.

One advantage of SK is that we can simplify the calculation since we do not need to calculate the integrals in (3), (4), and (6). If the noise is small and the fill distance is small, Theorem 3.1 and Proposition 3.6 state that the MSPE of the SK predictor (13) can be comparable with the best linear unbiased predictor.

It is argued in [6] that since the integrated covariance function in (4) is not the same as the covariance function in the original GP without location error, a nugget term alone cannot capture the effect of location error. While it is true that the MSPE of KALE or KALEN is the smallest among all the linear unbiased predictors, our results also show that with any fixed constant nugget term, the predictor (13) is as good as KALEN asymptotically (i.e., has the same limit as that of KALEN). The results indicate that there is little absolute difference between KALE and the predictor (13) if the variance of the input location noise

and the fill distance are small, because the same asymptotic upper bound for both MSPEs $\mathbb{E}(f(x)-\hat{f}(x))^2$ and $\mathbb{E}(f(x)-\hat{f}_S(x))^2$ is small. If the sample size n is large, the computational cost of KALE/KALEN and SK will be high, because the computation of a dense matrix inverse is $O(n^3)$. Note that the dense matrix inverse also appears in ordinary GP modeling. If the sample size is small and the variance of the input location noise is large, as suggested by numerical studies, the difference between the MSPE of KALE or KALEN and SK is large. Thus SK with a single nugget term may not lead to a good predictor in this case.

4. Parameter Estimation. Let $\Psi_{\theta^{(1)}}$ be a class of correlation functions and $p_{\theta^{(2)}}(\cdot)$ be a class of probability density functions indexed by $(\theta^{(1)},\theta^{(2)}) \in \Theta$, respectively, where $\theta^{(j)} \in \Theta_j \subset \mathbb{R}^{q_j}$ for $j=1,2$. Thus, $\Theta=\Theta_1\times\Theta_2$. Suppose Θ is a compact subregion of $\mathbb{R}^{q_1+q_2}$. An intuitive approach to estimate the parameters is maximum likelihood estimation. Up to a multiplicative constant, the likelihood function is

$$\ell(\sigma^2,\theta^{(1)},\theta^{(2)};X,Y) \propto \int_{\mathbb{R}^d}\dots\int_{\mathbb{R}^d} \det(\Sigma_1)^{-1/2} e^{-\frac{1}{2}Y^T\Sigma_1^{-1}Y} p_{\theta^{(2)}}(\epsilon_1)\dots p_{\theta^{(2)}}(\epsilon_n)d\epsilon_1\dots d\epsilon_n, \tag{18}$$

where $\Sigma_1 = (\sigma^2\Psi_{\theta^{(1)}}(x_j+\epsilon_j-(x_k+\epsilon_k)))_{jk}$, and $\det(A)$ is the determinant of a matrix A. Unfortunately, the integral in (18) is difficult to calculate, because the dimension of the integral increases as the sample size increases. In this work, we use a pseudo-likelihood approach proposed by [8]. Define

$$\ell_g(\sigma^2,\theta^{(1)},\theta^{(2)};X,Y) = (2\pi)^{-n/2}\det(R_{(\theta^{(1)},\theta^{(2)})})^{-1/2}\exp\left(-\frac{1}{2}Y^T R^{-1}_{(\theta^{(1)},\theta^{(2)})}Y\right), \tag{19}$$

where $\sigma^2,\theta^{(1)},\theta^{(2)}$ are parameters we want to estimate, and $R_{(\theta^{(1)},\theta^{(2)})}$ is defined in (4) by replacing Ψ and $p(\cdot)$ with $\Psi_{\theta^{(1)}}$ and $p_{\theta^{(2)}}(\cdot)$, respectively. The maximum pseudo-likelihood estimator can be defined as

$$(\hat{\sigma}_1^2,\hat{\theta}_1^{(1)},\hat{\theta}_1^{(2)}) = \underset{(\sigma^2,\theta^{(1)},\theta^{(2)})}{\operatorname{argsup}} \ell_g(\sigma^2,\theta^{(1)},\theta^{(2)};X,Y). \tag{20}$$

If (20) has multiple solutions, we choose any one from them. Because of non-identifiability, parameters inside the GP $(\sigma^2,\theta^{(1)})$ and parameters inside the probability density function of input variable noise $\theta^{(2)}$ cannot be estimated simultaneously [6].

The properties of the pseudo-likelihood approach are discussed in [6]. Here we list a few of them. First, the pseudo-score provides an unbiased estimation equation, i.e.,

$$\mathbb{E}(S(\sigma^2,\theta^{(1)},\theta^{(2)};X,Y)) = \mathbb{E}(\nabla\log(\ell_g(\sigma^2,\theta^{(1)},\theta^{(2)};X,Y))) = 0.$$

Second, the covariance matrix of the pseudo-score $\mathbb{E}(S(\sigma^2,\theta^{(1)},\theta^{(2)};X,Y)S(\sigma^2,\theta^{(1)},\theta^{(2)};X,Y)^T)$ and the expected negative Hessian of the log pseudo-likelihood $\mathbb{E}\big(\frac{\partial^2}{\partial\vartheta_j\partial\vartheta_k}\log(\ell_g(\sigma^2,\theta^{(1)},\theta^{(2)};X,Y))\big)$ can be calculated, where ϑ_j and ϑ_k are elements in $(\sigma^2,\theta^{(1)},\theta^{(2)})$, i.e., $(\sigma^2,\theta^{(1)},\theta^{(2)}) = (\vartheta_1,\vartheta_2,...,\vartheta_{1+q_1+q_2})$. However, the consistency of parameters estimated by pseudo-likelihood in the case of GP has not been theoretically justified to the best of our knowledge.

If we use SK, the corresponding (misspecified) log likelihood function is, up to an additive constant,

$$\ell_{nug}(\sigma^2, \theta^{(1)}, \mu; X, Y) = -\frac{1}{2}\log(\det(R_{\theta^{(1)}} + \mu I_n)) - \frac{1}{2}Y^T(R_{\theta^{(1)}} + \mu I_n)^{-1}Y, \tag{21}$$

where $R_{\theta^{(1)}} = (\Psi_{\theta^{(1)}}(x_j - x_k))_{jk}$. The maximum likelihood estimator of $(\sigma^2, \theta^{(1)}, \mu)$ is defined by

$$(\hat{\sigma}_2^2, \hat{\theta}_2^{(1)}, \hat{\mu}) = \operatorname*{argsup}_{(\sigma^2, \theta^{(1)}, \mu)} \ell_{nug}(\sigma^2, \theta^{(1)}, \mu; X, Y). \tag{22}$$

Note that (21) is the log likelihood function for a GP with only output noise. Thus it is misspecified, and the estimated parameters may also be misspecified. However, it has been shown by the well-known works [40] and [42] that the GP model parameters in the covariance functions may not have consistent estimators. Therefore, using GP models for prediction may be more meaningful than for parameter estimation. The following theorem indicates that the change of parameters do not significantly influence our theoretical results on the MSPE of KALE, KALEN and SK. The proof is presented in Appendix H.

Theorem 4.1. *Suppose for some constant $C > 0$, $1/C \le \tilde{\mu} \le C$ holds for all n, and parameters $\tilde{\sigma}^2, \tilde{\theta}_1^{(1)}, \tilde{\theta}_1^{(2)}, \tilde{\theta}_2^{(1)}$ are deterministic (but possibly depending on n). Let $\tilde{\Psi}_1$ and $\tilde{\Psi}_2$ be the correlation functions with parameters $\tilde{\theta}_1^{(1)}, \tilde{\theta}_2^{(1)} \in \Theta_1$, respectively. Let $\tilde{p}(\cdot)$ be the probability density function with parameters $\tilde{\theta}_1^{(2)} \in \Theta_2$. Let $\tilde{\Psi}_S$ be as in (10) with parameters $\tilde{\theta}_1^{(1)}$ and $\tilde{\theta}_1^{(2)}$. Potential dependency of $\tilde{\mu}$, $\tilde{\Psi}_1$, $\tilde{\Psi}_2$, $\tilde{p}(\cdot)$, and $\tilde{\Psi}_S$ on n is suppressed for notational simplicity. Assume the following.*

(1) There exists a constant A_1 such that for all n

$$\max\left\{\left\|\frac{\mathcal{F}(\Psi)}{\mathcal{F}(\tilde{\Psi}_S)}\right\|_{L_\infty}, \left\|\frac{\mathcal{F}(\Psi)}{\mathcal{F}(\tilde{\Psi}_1)}\right\|_{L_\infty}, \left\|\frac{\mathcal{F}(\Psi)}{\mathcal{F}(\tilde{\Psi}_2)}\right\|_{L_\infty}\right\} \le A_1. \tag{23}$$

(2) There exists a Sobolev space $H^m(\Omega)$ such that Assumption 2.2 holds for all $\tilde{\Psi}_1$ and $\tilde{\Psi}_2$, and the embedding constants have a uniform upper bound for all n, i.e., there exists a constant C such that $\|f\|_{H^m(\Omega)} \le C\|f\|_{\mathcal{N}_{\tilde{\Psi}_1}(\Omega)}$ and $\|f\|_{H^m(\Omega)} \le C\|f\|_{\mathcal{N}_{\tilde{\Psi}_2}(\Omega)}$ holds for all $\tilde{\Psi}_1$ and $\tilde{\Psi}_2$.

(3) Assumption 2.5 holds for the sequence of designs X.

(4) All probability density functions $\tilde{p}(\cdot)$ are continuous, have mean zero and second moment. The second moments of all $\tilde{p}(\cdot)$ have a uniform positive lower bound and upper bound for all n.

Then the following statements are true.

(i) Suppose there is noise at point x. Then the MSPE of KALEN $\mathbb{E}(y(x) - \hat{y}(x))^2$ and the MSPE of SK $\mathbb{E}(y(x) - \hat{f}_S(x))^2$ have the limit $\sigma^2(1 - \Psi_S(0))$ when the fill distance of X goes to zero, where Ψ_S is defined in (10).

(ii) Suppose there is no noise at point x. An asymptotic upper bound on the MSPE of SK $\mathbb{E}(f(x)-\hat{f}_S(x))^2$ is

$$\frac{1.04\sigma^2}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}|1-|b(t)|^2|\mathcal{F}(\Psi)(t)dt,$$

where $b(t)=\mathbb{E}(e^{i\epsilon^T t})$ is the characteristic function of $p(\cdot)$. Furthermore, if $\tilde{p}(\cdot)=p(\cdot)$ and $\left\|\frac{\mathcal{F}(\tilde{\Psi}_1)}{\mathcal{F}(\Psi)}\right\|_{L_\infty}\le A_2$, an asymptotic upper bound on the MSPE of KALE $\mathbb{E}(f(x)-\hat{f}(x))^2$ is

$$\frac{1.04A_1A_2\sigma^2}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}|1-|b(t)|^2|\mathcal{F}(\Psi)(t)dt.$$

Theorem 4.1 states if we have a reasonable sequence of parameters, then we have the following: (i) If point x has noise (i.e., predicting $f(x+\epsilon)$), the limit of the MSPE of KALEN and SK remains the same; and (ii) If point x has no noise (i.e., predicting $f(x)$), the upper bounds on the MSPE of KALE and SK can be obtained. The limit and upper bounds are small if the noise is small. The upper bound for the MSPE of SK is the same as the bound in Theorem 3.1. However, the upper bound for the MSPE of KALE is inflated by A_1A_2. We believe this inflation is not necessary and can be improved.

Remark 4.2. Note that the parameters in Theorem 4.1 are *deterministic*. Therefore, there is still a gap between Theorem 4.1 and the convergence results of KALE/KALEN/SK with estimated parameters. The authors cannot confirm if the results hold for estimated parameters which depend on the random observations Y. Nevertheless, given the parameters have sufficient flexibility, we believe that Theorem 4.1 can still provide some insights on the influence of the parameter estimation. We thank one reviewer for pointing out the incorrectness in the previous version of Theorem 4.1.

The computation complexity of (22) is about the same as that of (20), if (4) can be calculated analytically. Unfortunately, (4) usually does not have a closed form, which substantially increases the computation time of solving (20).

5. Numerical Results.

In this section, we report some simulation studies to investigate the numerical performance of KALE, KALEN and SK. In Example 1, we use Gaussian correlation functions to fit a 1-d function, where the predictor (5) has analytic form. In Example 2, we use Matérn correlation functions to fit a 2-d function, where the integrals in (3) and (4) are typically estimated by Monte Carlo sampling [8].

5.1. Example 1.

Suppose the underlying function is $f(x)=\sin(2\pi x/10)+0.2\sin(2\pi x/2.5)$, $x\in[0,8]$ [19]. The design points are selected to be 161 evenly spaced points on $[0,8]$. The input location noise is chosen to be mean zero normally distributed with the variances $0.05k$, for $k=1,2,3,4$. We use a Gaussian covariance function $\Psi(s-t)=\sigma^2\exp(-\theta\|s-t\|_2^2)$ to make predictions, and use the pseudo-likelihood approach presented in Section 4 to estimate the unknown parameters σ^2,θ and the variance of noise σ_ϵ^2. For each variance of input location

noise, we approximate the squared L_2 error $\|f-\hat{f}\|_2^2$ by $\frac{8}{n}\sum_{i=1}^{n}(f(x_i)-\hat{f}(x_i))^2$, where the x_i's are 8001 evenly spaced points on $[0,8]$. Then we run 100 simulations and take the average of $\frac{8}{n}\sum_{i=1}^{n}(f(x_i)-\hat{f}(x_i))^2$ to estimate $\mathbb{E}\|f-\hat{f}\|_2^2$. We estimate $\mathbb{E}\|y-\hat{y}\|_2^2$ by a similar approach, i.e., estimate $\mathbb{E}\|y-\hat{y}\|_2^2$ by the average of $\frac{8}{n}\sum_{i=1}^{n}(y(x_i)-\hat{y}(x_i))^2$ of 100 simulations, where $y(x_i)=f(x_i+\epsilon_i)$ and ϵ_i's are input location noise. Recall that $\mathbb{E}\|f-\hat{f}\|_2^2$ and $\mathbb{E}\|y-\hat{y}\|_2^2$ are related to KALE and KALEN, respectively. With abuse of terminology, we still call $\mathbb{E}\|f-\hat{f}\|_2^2$ and $\mathbb{E}\|y-\hat{y}\|_2^2$ MSPE.

The RMSPEs, which are the square roots of MSPEs, for KALE/KALEN and SK, are shown in Table 1/Table 2, respectively.

σ_ϵ^2	RMSPE (SD) of KALE	RMSPE (SD) of stochastic Kriging	Difference
0.05	0.1147(0.0287)	0.1209(0.0288)	0.0062
0.10	0.1528(0.0372)	0.1764(0.0387)	0.0236
0.15	0.1917(0.0475)	0.2364(0.0418)	0.0448
0.20	0.2380(0.0597)	0.3149(0.0773)	0.0769

Table 1

Comparison of the RMSPE for KALE and SK: 1-d function with Gaussian covariance function. SD stands for standard deviation of RMPSE. In the fourth column, difference = 3rd column − 2nd column, i.e., the RMSPE of SK − the RMSPE of KALE.

σ_ϵ^2	RMSPE (SD) of KALEN	RMSPE (SD) of stochastic Kriging	Difference
0.05	0.3627(0.0076)	0.3619(0.0073)	−0.0014
0.10	0.4940(0.0095)	0.4931(0.0092)	−0.0009
0.15	0.5884(0.0107)	0.5885(0.0108)	0.0001
0.20	0.6651(0.0127)	0.6704(0.0164)	0.0053

Table 2

Comparison of the RMSPE for KALEN and SK: 1-d function with Gaussian covariance function. SD stands for standard deviation of RMPSE. In fourth column, difference = 3rd column − 2nd column, i.e., the RMSPE of SK − the RMSPE of KALEN.

It can be seen from Tables 1 and 2 that the RMSPE (standard deviations) of KALE/KALEN and SK decreases as the variance of the input location noise drops. This corroborates the results in Theorem 3.1 and Proposition 3.6. The difference of RMSPE between KALE/KALEN and SK also decreases when the variance of the input location noise decreases. Comparing Table 2 with Table 1, it can be seen that the RMSPE of KALEN is larger than that of KALE. This is reasonable because KALEN predicts $y(x)$, which includes an error term while $f(x)$ does not. The computation of KALE/KALEN has the same complexity as the SK in this example, because a Gaussian covariance function is used, and the integrals in (4) and (6) can be calculated analytically.

In order to further understand the performance of KALE/KALEN and SK, two realizations among the 100 simulations for Table 1 and Table 2 are illustrated in Panel 1 and Panel 2 of

Figure 2, respectively, where the variance of the input location noise is chosen to be 0.05. In Panel 1 of Figure 2, the circles are the collected data points. The true function, the prediction curves of KALE and SK are denoted by the solid line, the dashed line and the dotted line, respectively. It can be seen from the figure that both KALE and SK approximate the true function well. In Panel 2 of Figure 2, the dots are the samples of $y(x)$ on 8001 testing points. It can be seen that the samples are around the predictions of KALEN and SK, but with much more fluctuations. This shows that the RMSPE in Table 2 is larger than those in Table 1.

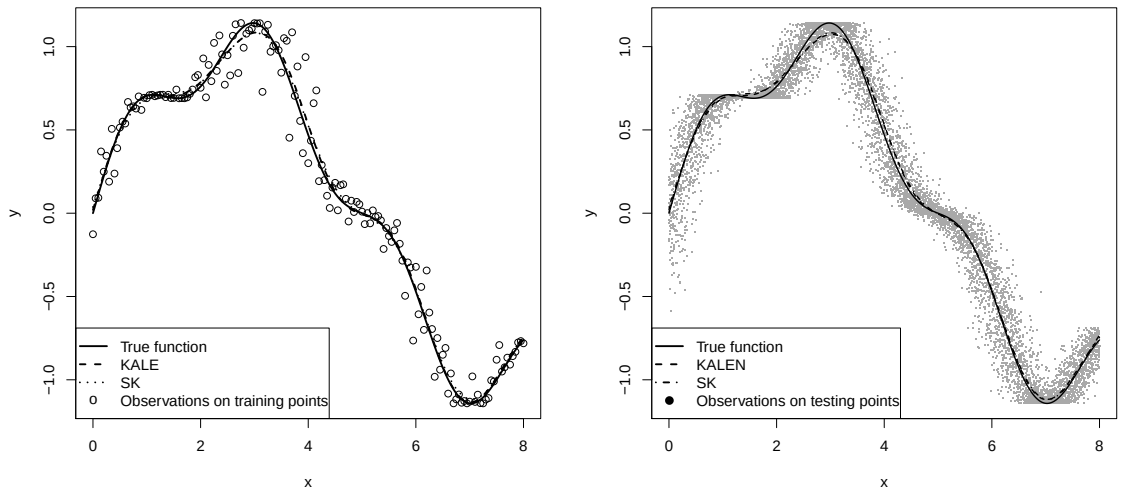

Figure 2. ***Panel 1:*** *An illustration of KALE and SK. The true function, the prediction curves of KALE and SK are denoted by solid line, dashed line and dotted line, respectively. The circles are the observed data points.* ***Panel 2:*** *An illustration of KALEN and SK. The dots are the samples of $y(x)$ on testing points. The true function, prediction curves of KALEN and SK are denoted by solid line, dashed line and dotted line, respectively.*

We also include the confidence interval results in this subsection. It is known [6] that there is no nontrivial structure for ϵ (that is, ϵ is not identical to zero) such that $f(x+\epsilon)$ is a GP on Ω. Since there is no closed form for the distribution of KALE $\hat{f}(x)$ (or KALEN $\hat{y}(x)$), we use Gaussian approximation. Specifically, we treat $f(x)$ (or $y(x)$) as normally distributed and compute the pointwise conditional variance $\hat{\sigma}_f(x)^2$ (or $\hat{\sigma}_y(x)^2$). Then we compute the pointwise confidence interval of GP, defined by $[\hat{f}(x) - q_\beta \hat{\sigma}_f(x), \hat{f}(x) + q_\beta \hat{\sigma}_f(x)]$ (or $[\hat{y}(x) - q_\beta \hat{\sigma}_y(x), \hat{y}(x) + q_\beta \hat{\sigma}_y(x)]$) with confidence level $(1-\beta)100\%$, where q_β denote the $(1-\beta/2)$th quantile of standard normal distribution. We select $\beta = 0.05$ and use coverage rate to quantify the quality of the confidence interval, where the coverage rate is the proportion of the time that the interval contains the true value. However, the length of the confidence interval of SK for GP with only output error converges to zero, which does not reflect the fact that the actual MSPE of SK does not converge to zero. Because of this, we adjust the estimated conditional variance of the SK by adding the limit value $\sigma^2(1-\Psi_S(0))$. The results

are reported in Table 3.

σ_ϵ^2	KALE	SK_1	Adjusted SK_1	KALEN	SK_2	Adjusted SK_2
0.05	0.9179	0.8547	0.9630	0.9292	0.4903	0.6328
0.10	0.9268	0.7906	0.9754	0.9296	0.4432	0.6490
0.15	0.9202	0.6987	0.9670	0.9345	0.4033	0.6677
0.20	0.9163	0.5834	0.9213	0.9358	0.3494	0.6545

Table 3

Coverage rate of pointwise confidence interval of KALE and SK (when there is no noise on target point), and KALEN and SK (when there is noise on target point). The following notation is used: (Adjusted) SK_1 = (Adjusted) SK without noise at the target point; (Adjusted) SK_2 = (Adjusted) SK with noise at the target point. The nominal level is selected to be 95%.

From Table 3, it can be seen that the (misspecified) pointwise confidence interval does not achieve the nominal level. It is expected that the SK has poor coverage because the model is misspecified. KALE and KALEN, on the other hand, can provide more reliable confidence intervals. In fact, even for GP without error, it is often observed that GP models have poor coverage of their confidence intervals [16, 20, 39]. Therefore, a better uncertainty quantification methodology for GP with input location error is needed.

5.2. Example 2. In this example, we compare the calculation time of SK and KALE, where the predictor (5) of KALE does not have an analytic form. Suppose the underlying function is $f(x) = [(30 + 5x_1 \sin(5x_1))(4 + \exp(-5x_2)) - 100]/6$ for $x_1, x_2 \in [0, 1]$ [21]. We use Matérn correlation functions [29]

$$\Psi_M(x; \nu, \phi) = \frac{1}{\Gamma(\nu)2^{\nu-1}}(2\sqrt{\nu}\phi\|x\|_2)^\nu K_\nu(2\sqrt{\nu}\phi\|x\|_2) \tag{24}$$

to make predictions, where K_ν is the modified Bessel function of the second kind, and $\nu, \phi > 0$ are model parameters. The Matérn correlation function can control the smoothness of the predictor by ν and thus is more robust than a Gaussian correlation function [34]. The covariance function is chosen to be $\Psi(x - y) = \Psi_M(x - y; \nu, \phi)$. The input location noise is chosen to be mean zero normally distributed with the variances $0.01k$, for $k = 2, 3, 4, 5$. We use maximin Latin hypercube design with 20 points to estimate parameters, and choose the first 100 points in the Halton sequence [17] as testing points. The smoothness parameter ν is chosen to be 3, which can provide a robust estimator of f. In order to improve the prediction performance, we use ordinary Kriging, where the mean in GP model is assumed to be an unknown constant instead of zero, i.e., f is a realization of GP with unknown mean β and covariance function $\sigma^2\Psi_M$.

If we use a Matérn correlation function, the integrals in (3) and (4) do not have analytic forms and are calculated by Monte-Carlo sampling. We randomly choose 30 points to approximate the integral in (3), and 900 points to approximate the integral in (4). Preliminary results show that, if we use Monte-Carlo sampling with different points every time in the evaluation of the integrals in (3) and (4), it is not possible to use maximum pseudo-likelihood estimation to estimate the unknown parameters, consisting of ϕ in (24), σ^2, the variance of noise σ_ϵ^2 and the

mean β. The reason is that at each step of the optimization in maximum pseudo-likelihood estimation, we need to calculate the integral, whose computational cost is high. Therefore, we generate 900 points and 30 points randomly at one time and use these 900 points and 30 points for evaluations of (4) and (3), respectively. Then we use maximum pseudo-likelihood estimation to estimate the unknown parameters. We run 20 simulations to obtain different realizations of input location noise. In each simulation, we compute the processing time and the approximated MSPE $\frac{1}{100}\sum_{i=1}^{100}(f(x_i)-\hat{f}(x_i))^2$, where $\hat{f}$ is the KALE predictor, and x_i's are testing points. Then we compute the average processing time and the average approximated MSPE.

For SK, we use (misspecified) maximum likelihood estimation to estimate the unknown parameters, which are ϕ in (24), σ^2, the nugget term μ and the mean β. We run 100 simulations and compute the average processing time and the average approximated MSPE $\frac{1}{100}\sum_{i=1}^{100}(f(x_i)-\hat{f}(x_i))^2$, where $\hat{f}$ is the SK predictor, and x_i's are the same testing points as in KALE. The RMSPE, which is the square root of MSPE, and the processing time of KALE and SK are shown in Table 4.

σ_ϵ^2	RMSPE of KALE	PT of KALE	RMSPE of SK	PT of SK	Difference
0.02	1.5292	648.86	1.9852	0.6261	0.4559
0.03	1.7899	633.55	2.2346	0.5947	0.4446
0.04	1.9734	695.27	2.5226	0.5848	0.5492
0.05	2.4501	748.33	3.3415	0.5803	0.8915

Table 4

The RMSPE of KALE and SK: 2-d function with Matérn correlation function. The processing time is in seconds. In sixth column, difference = 4th column − 2nd column, i.e., the RMSPE of SK − the RMSPE of KALE. We use PT = Processing time.

It can be seen that KALE has some improvement on prediction accuracy over SK. However, KALE takes too much computation time, even though the numbers of design points and testing points are relatively small. The comparison would get worse as the number of points became larger. Therefore, if the integrals in (3) and (4) do not have analytic forms, SK is preferred, especially when the sample size is large and the variance of input location noise is small.

6. Case Study: Application to Composite Parts Assembly Process. To illustrate the performance of KALEN and SK, we apply them to a real case study, the composite parts assembly process. As shown in Figure 3 (a) and Figure 3 (b), ten adjustable actuators are installed at the edge of a composite part [36, 41]. These actuators can provide push or pull forces to adjust the shape of the composite part to the target dimensions. The locations of these actuators can be optimized by sparse learning method [12]. The dimensional shape adjustment of composite parts is one of the most important steps in the aircraft assembly process. It reduces the gap between the composite parts and decreases the assembly time with improved dimensional quality. Detailed descriptions about the shape adjustment of composite parts can be found in [36]. Modeling of composite parts is the key for shape adjustment. The

objective is to build a model that has the capability to predict the dimensional deviations accurately under specific actuators' forces. In this model, the input variables are ten actuators' forces. The responses are the dimensional deviations of multiple critical points along the edge plane near the actuators, shown in Figure 3 (c). We consider responses at 91 critical points around the composite edge in the case study.

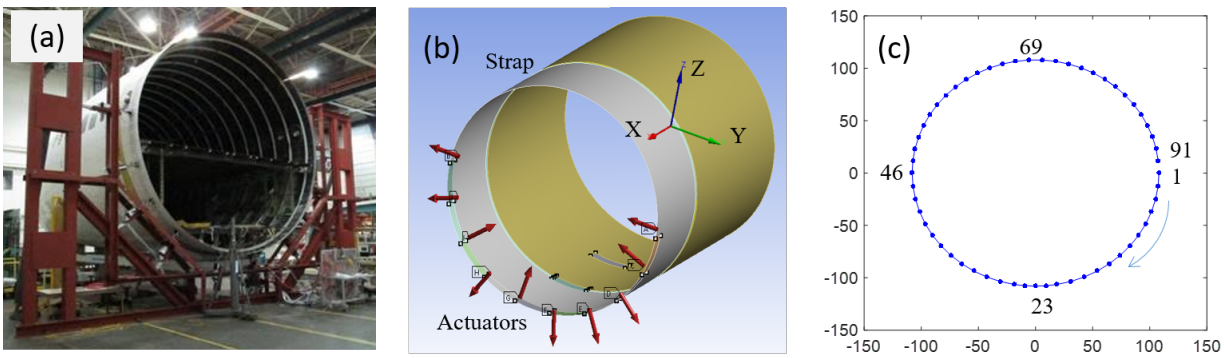

Figure 3. *Schematic diagram for composite part shape adjustment: (a) composite part shape adjustment [36], (b) layout of ten actuators, (c) multiple critical points.*

In the shape control of composite parts, input location noise commonly exists in the actuators' forces [41]. When a force is implemented by an actuator, the actual force may not be exactly the same as the target force. The magnitudes of forces may have uncertainties naturally due to the device tolerances of the hydraulic or electromechanical system of actuators. Uncertainties in the directions and application points of forces come from the deviations of contact geometry of actuators and their installations. For the modeling of composite parts, there are two steps: (i) training the parameters using experimental data; (ii) predicting dimensional deviations for new actuators' forces. In the training step, we need to consider input error in the experimental data. Additionally, when new actuator forces are implemented in practice, the uncertainty in the actual delivered forces inevitably exists. This suggests that KALEN is suitable for this application scenario. We will show the performance of KALEN and compare it with SK as follows.

The model we use in this case study is $Y^{(j)} = F^T\beta^{(j)} + Z^{(j)}(F)$ for $j = 1, \ldots, 91$, where $Y^{(j)}$ is the dimensional deviation vector of the composite part at the critical point j, $F = (F_{(1)}, \ldots, F_{(10)})^T \in \mathbb{R}^{10}$ is the vector of actuators' forces, and $Z^{(j)}(\cdot)$ is a mean zero GP, with input variables in $\mathbb{R}^{10}$. The covariance of $Z^{(j)}(F_1)$ and $Z^{(j)}(F_2)$ for any forces $F_1 = (F_{1,(1)}, \ldots, F_{1,(10)})^T$ and $F_2 = (F_{2,(1)}, \ldots, F_{2,(10)})^T$ is assumed to be $\sigma_j^2 \exp(-\sum_{k=1}^{10} \theta_{jk}(F_{1,(k)} - F_{2,(k)})^2)$, where $\sigma_j, \theta_{jk} > 0$ are parameters. We assume the input location noise $\epsilon \sim N(0, \sigma_\epsilon^2 I_{10})$, where $N(0, \sigma_\epsilon^2 I_{10})$ is a mean zero normal distribution with covariance matrix $\sigma_\epsilon^2 I_{10}$. The parameters $\beta^{(j)}$, θ_{jk}, σ_ϵ^2, and σ_j^2 are estimated by maximum (pseudo-)likelihood estimation as described in Section 4. The mean function $F^T\beta^{(j)}$ we use in this model is to represent the

linear component in dimensional shape control of composite fuselage, which follows the approach in [41]. Specifically, according to the mechanics of composite material and classical lamination theory, there is a linear relationship between dimensional deviations and actuators' forces within the elastic zone. The term $F^T\beta^{(j)}$ describes how the actuators' forces impact the part deviations linearly, and $Z^{(j)}(\cdot)$ represents the nonlinear components so as to obtain accurate predictions.

For the computer experiments, we generated 50 training samples and 30 testing samples based on a maximin Latin hypercube design. The designed experiments are conducted in the finite element simulation platform developed by [36]. This platform was developed based on the ANSYS Composite PrepPost workbench. It has been calibrated and validated via a sensible variable identification approach [35]. It is worth mentioning that the computer simulation here is not a deterministic simulation because we add the input location noise at the input points in simulation to simulate the randomness in the real process. Therefore, repeated runs with the same input points will have different outputs. The input location noise is added to the actuators' forces to mimic real actuators. The standard deviations (SD) of actuators' forces are chosen to be 0.005, 0.01, 0.02, 0.03, and 0.04 lbf (lbf is a unit of pound-force), which is determined by the tolerance of different kinds of actuators according to engineering domain knowledge. The maximum actuators' force is set to 600 lbf. After we have the computer experiment data, we can estimate the parameters of KALEN by solving the pseudo-likelihood equation (20), and the parameters of SK by solving the maximum likelihood equation (22). Then, we can use the model to predict dimensional deviations at the target points in the testing dataset.

The performance of KALEN and SK are compared in terms of mean absolute error (MAE). This is an index that has been commonly used in the composite parts assembly domain to evaluate the modeling performance. We also compare RMSPE of KALEN and SK, and the processing time of generating each output. The RMSPE is the square root of MSPE, which is approximated by the average of $\frac{1}{30}\sum_{i=1}^{30}(Y^{(j)}(F_i) - \hat{Y}^{(j)}(F_i))^2$ on the 91 points, where F_i's are the inputs of testing samples, $Y^{(j)}(F_i)$ is the observed testing data, and $\hat{Y}^{(j)}(F_i)$ is the KALEN predictor. The MAE is approximated by $\frac{1}{30}\sum_{i=1}^{30}|Y^{(j)}(F_i) - \hat{Y}^{(j)}(F_i)|$ on the 91 points.

SD of AF	MAE (RMSPE) of KALEN	MAE (RMSPE) of SK	Difference	PT of KALEN	PT of SK
0.005	0.0059 (0.0081)	0.0059 (0.0081)	7.1×10^{-7} (1.9×10^{-6})	0.1500	0.3415
0.01	0.0117 (0.0147)	0.0119 (0.0151)	1.7×10^{-4} (3.7×10^{-4})	0.4691	0.3938
0.02	0.0216 (0.0265)	0.0217 (0.0264)	9.5×10^{-5} (-8.7×10^{-5})	0.5048	0.3964
0.03	0.0286 (0.0335)	0.0304 (0.0376)	1.7×10^{-3} (4.1×10^{-3})	0.6746	0.4115
0.04	0.0389 (0.0478)	0.0486 (0.0610)	9.7×10^{-3} (1.3×10^{-2})	0.6529	0.4302

Table 5

The MAE (RMSPE) of KALEN and SK in the composite part modeling. In 4th column, difference = 3rd column − 2nd column. The processing time is in seconds. The following abbreviation is used: AF = actuators' forces, PT = Processing time for each output.

The MAE and RMSPE of KALEN and SK are summarized in Table 5. As the SD of actuators' forces changes from 0.04 lbf to 0.005 lbf, the MAE and RMSPE of KALEN and SK also decrease. This result is consistent with the conclusions in Theorem 3.1 and Proposition 3.6. The MAE and RMSPE of KALEN are slightly smaller than the MAE and RMSPE of SK. Generally speaking, their performances are comparable, especially when the SD of actuators' forces is small. The main reason is that, when the uncertainty in the input variables is small, SK can approximate the best linear unbiased predictor KALEN very well. Since a Gaussian correlation function is used, the computational complexity of KALEN and SK are the same. The computation time of KALEN is smaller than that of the SK in this example. We conjecture this is because of the different computation time of maximum (pseudo-) likelihood estimation. In summary, if high-quality actuators are used and the input location noise in the actuators is therefore small, then both KALEN and SK can realize very good prediction performance. When the input location noise in the actuators' forces becomes larger, KALEN outperforms SK.

7. Conclusions and Discussion. We first summarize our contributions in this work. We have investigated three predictors, KALE, KALEN and SK, as applied to GPs with input location error. When predicting the mean GP output at a point with input location noise, we prove that the limits of MSPE of KALEN and SK are the same as the fill distance of the design points goes to zero. If there is no noise at point $x \in \Omega$, we provide an upper bound on the MSPE of KALE and SK. The upper bound is close to zero if the noise is small, which implies the MSPE of KALE and SK are close. We also provide an asymptotic upper bound on the MSPE of KALE/KALEN and SK with estimated parameters. These results indicate that if the number of data points is large or the variance of the input location noise is small, then there is not much difference between KALE/KALEN and SK in terms of prediction accuracy. The numerical results corroborate our theory. A case study is presented to illustrate the performance of KALEN and SK for modeling in the composite parts assembly process.

The calculation of the predictor (5) is not computationally efficient if the integrals in (3) and (4) do not have an analytic form, where Monte-Carlo integration is typically to be used. If the sample size is large, then using pseudo maximum likelihood to estimate the unknown parameters is challenging, especially when the integrals in (3) and (4) do not have analytic forms. In this case, using SK as an alternative would be more desirable.

There are several problems that remain to be solved. In this paper, the MSPE of KALE, KALEN, and SK are primarily considered asymptotically, i.e., the number of design points goes to infinity. The theory does not cover the results under non-asymptotic cases, i.e., the number of design points is fixed. It can be expected that the difference between the MSPE of KALE/KALEN and SK will decrease as the fill distance decreases. If there is no noise on point $x \in \Omega$, only upper bounds are obtained for KALE and SK. The asymptotic performance of KALE and SK when target point has no noise will be pursued in the future work.

Acknowledgements. The authors are grateful to the AE and all the reviewers for their very helpful comments and suggestions.

Appendix A. Reproducing kernel Hilbert space, Sobolev space and kernel ridge regression.

Suppose $\Omega \subset \mathbb{R}^d$ is convex and compact. Assume that $K : \Omega \times \Omega \to \mathbb{R}$ is a symmetric positive definite kernel function. Define the linear space

$$F_K(\Omega) = \left\{ \sum_{k=1}^{n} \beta_k K(\cdot, x_k) : \beta_k \in \mathbb{R}, x_k \in \Omega, n \in \mathbb{N} \right\}, \tag{25}$$

and equip this space with the bilinear form

$$\left\langle \sum_{k=1}^{n} \beta_k K(\cdot, x_k), \sum_{j=1}^{m} \gamma_j K(\cdot, x'_j) \right\rangle_K := \sum_{k=1}^{n} \sum_{j=1}^{m} \beta_k \gamma_j K(x_k, x'_j).$$

Then the *reproducing kernel Hilbert space* $\mathcal{N}_K(\Omega)$ generated by the kernel function K is defined as the closure of $F_K(\Omega)$ under the inner product $\langle \cdot, \cdot \rangle_K$, and the norm of $\mathcal{N}_K(\Omega)$ is $\|f\|_{\mathcal{N}_K(\Omega)} = \sqrt{\langle f, f \rangle_{\mathcal{N}_K(\Omega)}}$, where $\langle \cdot, \cdot \rangle_{\mathcal{N}_K(\Omega)}$ is induced by $\langle \cdot, \cdot \rangle_K$. The following theorem gives another characterization of the reproducing kernel Hilbert space when K is defined by a stationary kernel function Ψ, via the Fourier transform. Note that a kernel function Ψ is said to be stationary if the value $\Psi(x, x')$ only depends on the difference $x - x'$. Thus, we can write $\Psi(x - x') := \Psi(x, x')$.

Theorem A.1 (Theorem 10.12 of [37]). *Let Ψ be a positive definite kernel function which is stationary, continuous and integrable in $\mathbb{R}^d$. Define*

$$\mathcal{G} := \{ f \in L_2(\mathbb{R}^d) \cap C(\mathbb{R}^d) : \mathcal{F}(f)/\sqrt{\mathcal{F}(\Psi)} \in L_2(\mathbb{R}^d) \},$$

with the inner product

$$\langle f, g \rangle_{\mathcal{N}_\Psi(\mathbb{R}^d)} = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \frac{\mathcal{F}(f)(\omega)\overline{\mathcal{F}(g)(\omega)}}{\mathcal{F}(\Psi)(\omega)} d\omega.$$

Then $\mathcal{G} = \mathcal{N}_\Psi(\mathbb{R}^d)$, and both inner products coincide.

By Bochner's theorem (Page 208 of [14]; Theorem 6.6 of [37]) and Theorem 6.11 of [37], if Ψ is a correlation function (thus positive definite), there exists a function f_Ψ such that

$$\Psi(x) = \int_{\mathbb{R}^d} e^{i\omega^T x} f_\Psi(\omega) d\omega$$

for any $x \in \mathbb{R}^d$. The function f_Ψ is known as the *spectral density* of Ψ.

Condition A.2. *There exist constants $c_2 \geq c_1 > 0$ and $\eta > d/2$ such that, for all $\omega \in \mathbb{R}^d$,*

$$c_1(1 + \|\omega\|_2^2)^{-\eta} \leq f_\Psi(\omega) \leq c_2(1 + \|\omega\|_2^2)^{-\eta}.$$

We say a Hilbert function space $\mathcal{G}_1$ can be (continuously) embedded into another Hilbert function space $\mathcal{G}_2$, if there exists a constant C such that

$$\|g_1\|_{\mathcal{G}_2} \leq C \|g_1\|_{\mathcal{G}_1}, \forall g_1 \in \mathcal{G}_1,$$

where $\|\cdot\|_{\mathcal{G}_1}$ and $\|\cdot\|_{\mathcal{G}_2}$ are the norms of the function spaces $\mathcal{G}_1$ and $\mathcal{G}_2$, respectively. Therefore, it can be seen from Theorem A.1 that if, for two positive definite functions Φ_1 and Φ_2, the spectral densities f_{Φ_1} and f_{Φ_2} satisfy $f_{\Phi_1} \leq C f_{\Phi_2}$, then the reproducing kernel Hilbert space $\mathcal{N}_{\Phi_1}(\mathbb{R}^d)$ can be embedded into $\mathcal{N}_{\Phi_2}(\mathbb{R}^d)$.

For a positive number $\eta > d/2$, the Sobolev space on $\mathbb{R}^d$ with smoothness η can be defined as

$$H^\eta(\mathbb{R}^d) = \{f \in L_2(\mathbb{R}^d) : |\mathcal{F}(f)(\cdot)|(1 + \|\cdot\|_2^2)^{m/2} \in L_2(\mathbb{R}^d)\},$$

equipped with an inner product

$$\langle f, g\rangle_{H^\eta(\mathbb{R}^d)} = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \mathcal{F}(f)(\omega)\overline{\mathcal{F}(g)(\omega)}(1 + \|\omega\|_2^2)^\eta d\omega.$$

It can be shown that $H^\eta(\mathbb{R}^d)$ coincides with the reproducing kernel Hilbert space $\mathcal{N}_\Psi(\mathbb{R}^d)$, if Ψ satisfies Condition A.2 ([37], Corollary 10.13).

Remark A.3. In this work, we are only interested in Sobolev spaces with $\eta > d/2$ because these spaces contain only continuous function according to the Sobolev embedding theorem.

The isotropic Matérn correlation function (24) has the spectral density [30]

$$f_{\Psi_M}(\omega; \nu, \phi) = \pi^{-d/2} \frac{\Gamma(\nu + d/2)}{\Gamma(\nu)} (4\nu\phi^2)^\nu (4\nu\phi^2 + \|\omega\|_2^2)^{-(\nu+d/2)}.$$

We can see Ψ_M satisfies Condition A.2. Thus, the reproducing kernel Hilbert space generated by Ψ_M coincides with the Sobolev space $H^{\nu+d/2}$, which implies Ψ_M fulfills Assumption 2.2.

The isotropic Gaussian correlation function $\Psi_G(x) = e^{-\theta\|x\|^2}$ has the spectral density (Theorem 5.20 of [37])

$$f_{\Psi_G}(\omega) = (4\pi\theta)^{-d/2} e^{-\|\omega\|_2^2/(4\theta)}.$$

Since for any fixed ν, $f_{\Psi_G}(\omega) \leq C(1 + \|\omega\|_2^2)^{-\nu-d/2}$ for some constant C not depending on ω, the reproducing kernel Hilbert space generated by Ψ_G can be embedded the Sobolev space $H^{\nu+d/2}(\mathbb{R}^d)$. This implies Ψ_G fulfills Assumption 2.2.

A reproducing kernel Hilbert space can also be defined on a suitable subset (for example, convex and compact) $\Omega \subset \mathbb{R}^d$, denoted by $\mathcal{N}_\Psi(\Omega)$, with norm

$$\|f\|_{\mathcal{N}_\Psi(\Omega)} = \inf\{\|f_E\|_{\mathcal{N}_\Psi(\mathbb{R}^d)} : f_E \in \mathcal{N}_\Psi(\mathbb{R}^d), f_E|_\Omega = f\},$$

where $f_E|_\Omega$ denotes the restriction of f_E to Ω. A Sobolev space on Ω can be defined in a similar way. By the extension theorem [11], the reproducing kernel Hilbert space defined on space Ω generated by Ψ_M and Ψ_G can be embedded into the Sobolev space $H^{\nu+d/2}(\Omega)$.

In the rest of the Appendix, we use $C, C_j, j \geq 0$ to denote generic positive constants, whose value can change from line to line.

Appendix B. A Lemma about MSPE of stochastic Kriging.

Lemma B.1. *Let Φ be a radial basis function, positive definite, and stationary. Suppose the reproducing kernel Hilbert space generated by Φ can be embedded into a Sobolev space $H^\eta(\Omega)$ with $\eta > d/2$. Assume Assumption 2.5 is true for a sequence of designs $X = \{x_1, \ldots, x_n\}$. Then for any fixed constant $\mu > 0$, $\Phi(0) - r_\Phi(x)^T(R_\Phi + \mu I_n)^{-1} r_\Phi(x)$ converges to zero pointwisely as the fill distance of X goes to zero, where $r_\Phi(x) = (\Phi(x - x_1), \ldots, \Phi(x - x_n))^T$ and $R_\Phi = (\Phi(x_j - x_k))_{jk}$.*

Proof. Let $\bar{X} = \{\bar{x}_1, ..., \bar{x}_{n'}\}$ be the distinct design points corresponding to X. At each design point $\bar{x}_j \in \bar{X}$, suppose there are a_j replicates, thus,

$$X = \{\underbrace{\bar{x}_1^{(1)}, \ldots, \bar{x}_1^{(a_1)}}_{a_1 \text{ replications}}, \underbrace{\bar{x}_2^{(1)}, \ldots, \bar{x}_2^{(a_2)}}_{a_2 \text{ replications}}, \ldots, \underbrace{\bar{x}_{n'}^{(1)}, \ldots, \bar{x}_{n'}^{(a_{n'})}}_{a_{n'} \text{ replications}}\}.$$

It can be shown that $\Phi(0) - r_\Phi(x)^T(R_\Phi + \mu I_n)^{-1} r_\Phi(x) = \Phi(0) - \bar{r}_\Phi(x)^T(\bar{R}_\Phi + \Lambda I_{n'})^{-1}\bar{r}_\Phi(x)$, where $\bar{r}_\Phi(x) = (\Phi(x - x_1), \ldots, \Phi(x - \bar{x}_{n'}))^T$, $\bar{R}_\Phi = (\Phi(\bar{x}_j - \bar{x}_k))_{jk}$, and $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_m)$ with $\lambda_j = \mu/a_j$ (See Lemma 3.1 of [3] and the proof of Proposition 3.1 of [33]). Let $a = \min_j a_j$ and fix a point x. We have

$$\begin{aligned}
&\Phi(0) - r_\Phi(x)^T(R_\Phi + \mu I_n)^{-1} r_\Phi(x) \\
=&\Phi(0) - \bar{r}_\Phi(x)^T(\bar{R}_\Phi + \Lambda I_{n'})^{-1}\bar{r}_\Phi(x) \\
\leq&\Phi(0) - \bar{r}_\Phi(x)^T(\bar{R}_\Phi + \mu/a I_{n'})^{-1}\bar{r}_\Phi(x) \\
\leq&\|g_x\|_{L_\infty(\Omega)},
\end{aligned}$$

where the first inequality is because $(\bar{R}_\Phi + \Lambda I_{n'})^{-1} \succeq (\bar{R}_\Phi + \mu/a I_{n'})^{-1}$, and $g_x(t) = \Phi(t - x) - \bar{r}_\Phi(t)^T(\bar{R}_\Phi + \mu/a I_{n'})^{-1}\bar{r}_\Phi(x)$. Here $A \succeq B$ denotes that for any vector b, $b^T(A - B)b \geq 0$.

Since $\mathcal{N}_\Phi(\Omega)$ can be embedded into a Sobolev space $H^\eta(\Omega)$, we have $g_x \in H^\eta(\Omega)$, where $H^\eta(\Omega)$ is the Sobolev space with smoothness η. By the interpolation inequality [5], $\|g_x\|_{L_\infty(\Omega)} \leq C_1\|g_x\|_{L_2(\Omega)}^{1-\frac{d}{2\eta}}\|g_x\|_{H^\eta(\Omega)}^{\frac{d}{2\eta}}$. By Corollary 10.25 in [37] and the fact that $\bar{R}_\Phi^{-1} \succeq (\bar{R}_\Phi + \mu/a I_{n'})^{-1}$, it can be shown that

$$\begin{aligned}
&\|g_x\|_{H^\eta(\Omega)}^2 \leq C_2\|g_x\|_{\mathcal{N}_\Phi(\Omega)}^2 \\
\leq&C_2(1 - 2\bar{r}_\Phi(x)^T(\bar{R}_\Phi + \mu/a I_{n'})^{-1}\bar{r}_\Phi(x) \\
&+ \bar{r}_\Phi(x)^T(\bar{R}_\Phi + \mu/a I_{n'})^{-1}\bar{R}_\Phi(\bar{R}_\Phi + \mu/a I_{n'})^{-1}\bar{r}_\Phi(x)) \\
\leq&C_2(1 - \bar{r}_\Phi(x)(\bar{R}_\Phi + \mu/a I_{n'})^{-1}\bar{r}_\Phi(x)^T) \leq C_2,
\end{aligned}$$

where $\|g_x\|_{\mathcal{N}_\Phi(\Omega)}$ is the norm of g in the reproducing kernel Hilbert space $\mathcal{N}_\Phi(\Omega)$. Thus, the result follows if we can show $\|g_x\|_{L_2(\Omega)}$ converges to zero. By the representer theorem, $\hat{g}_1(t) := \bar{r}_\Phi(t)^T(\bar{R}_\Phi + \mu/a I_{n'})^{-1}\bar{r}_\Phi(x)$ is the solution to the optimization problem

$$\min_{g_1 \in \mathcal{N}_\Phi(\Omega)} \frac{1}{n}\sum_{j=1}^{n}(g_1(\bar{x}_j) - \Phi(x - \bar{x}_j))^2 + \frac{\mu}{an}\|g_1\|_{\mathcal{N}_\Phi(\Omega)}^2. \tag{26}$$

Note $g_x(t) = \Phi(t-x) - \hat{g}_1(t)$. Under Assumption 2.5, by Lemma 3.4 of [31], the result follows from

$$\begin{aligned}\|g_x\|_{L_2}^2 \leq & C_3\left(\frac{1}{n}\sum_{j=1}^{n}(\hat{g}_1(\bar{x}_j) - \Phi(x-\bar{x}_j))^2 + h_{\bar{X}}^{2\eta}\|g_x\|_{H^\eta(\Omega)}^2\right)\\ \leq & C_3\left(\frac{1}{n}\sum_{j=1}^{n}(\hat{g}_1(\bar{x}_j) - \Phi(x-\bar{x}_j))^2 + \frac{\mu}{an}\|\hat{g}_1\|_{\mathcal{N}_\Phi(\Omega)}^2 + h_{\bar{X}}^{2\eta}\|g_x\|_{H^\eta(\Omega)}^2\right)\\ \leq & C_3\left(\frac{1}{n}\sum_{j=1}^{n}(\Phi(x-\bar{x}_j) - \Phi(x-\bar{x}_j))^2 + \frac{\mu}{an}\|\Phi(x-\cdot)\|_{\mathcal{N}_\Phi(\Omega)}^2 + h_{\bar{X}}^{2\eta}\|g_x\|_{H^\eta(\Omega)}^2\right) \to 0,\end{aligned}$$

where the last inequality is true because $\hat{g}_1$ is the solution to (26). ∎

Appendix C. Calculation of (7). In this section, we show that if the correlation function is $\Psi(s-t) = \exp(-\theta\|s-t\|_2^2)$, and the noise $\epsilon \sim N(0, \sigma_\epsilon^2 I_d)$, where $\theta > 0$ is the correlation parameter, and $N(0, \sigma_\epsilon^2 I_d)$ is the mean zero normal distribution with covariance matrix $\sigma_\epsilon^2 I_d$, then (3)–(6) can be calculated respectively as in (7). Let $p_N(t)$ be the probability density function of normal distribution $N(0, \sigma_\epsilon^2 I_d)$, i.e.,

$$p_N(t) = \frac{1}{\sqrt{(2\pi\sigma_\epsilon^2)^d}}\exp\left(-\frac{t^T t}{2\sigma_\epsilon^2}\right).$$

The idea of calculating (3)–(6) is to utilize

$$\int_{\mathbb{R}^d}\frac{1}{(2\pi a^2)^{d/2}}\exp\left(-\frac{\|s-b\|_2^2}{2a^2}\right)ds = 1,$$

for $a > 0$ multiple times. By direct calculation, we have

$$\begin{aligned}r_N(x, x_j) =& \sigma^2\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\Psi(x+\epsilon-(x_j+\epsilon_j))p(\epsilon_j)p(\epsilon)d\epsilon_j d\epsilon\\ =& \sigma^2\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\exp(-\theta\|x+\epsilon-(x_j+\epsilon_j)\|_2^2)\frac{1}{\sqrt{(2\pi\sigma_\epsilon^2)^d}}\exp\left(-\frac{\epsilon_j^T\epsilon_j}{2\sigma_\epsilon^2}\right)\frac{1}{\sqrt{(2\pi\sigma_\epsilon^2)^d}}\exp\left(-\frac{\epsilon^T\epsilon}{2\sigma_\epsilon^2}\right)d\epsilon_j d\epsilon\\ =& \sigma^2\frac{\exp(-\theta\|x-x_j\|_2^2)}{(2\pi\sigma_\epsilon^2)^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\exp\left(-\left(\theta+\frac{1}{2\sigma_\epsilon^2}\right)\epsilon^T\epsilon - 2\theta(x-x_j-\epsilon_j)^T\epsilon\right)d\epsilon\\ &\times\exp\left(-\left(\theta+\frac{1}{2\sigma_\epsilon^2}\right)\epsilon_j^T\epsilon_j + 2\theta(x-x_j)^T\epsilon_j\right)d\epsilon_j.\end{aligned} \tag{27}$$

We first compute

$$\int_{\mathbb{R}^d}\exp\left(-\left(\theta+\frac{1}{2\sigma_\epsilon^2}\right)\epsilon^T\epsilon - 2\theta(x-x_j-\epsilon_j)^T\epsilon\right)d\epsilon$$

$$= \int_{\mathbb{R}^d} \exp\left(-\left(\theta + \frac{1}{2\sigma_\epsilon^2}\right) \left\| \epsilon + \frac{\theta(x - x_j - \epsilon_j)}{\theta + \frac{1}{2\sigma_\epsilon^2}} \right\|_2^2 + \frac{\theta^2}{\left(\theta + \frac{1}{2\sigma_\epsilon^2}\right)} \|x - x_j - \epsilon_j\|_2^2 \right) d\epsilon$$

$$= \exp\left(\frac{2\sigma_\epsilon^2 \theta^2}{1 + 2\sigma_\epsilon^2 \theta} \|x - x_j - \epsilon_j\|_2^2 \right) \sqrt{\left(2\pi \frac{\sigma_\epsilon^2}{1 + 2\theta\sigma_\epsilon^2} \right)^d}. \tag{28}$$

Plugging (28) into (27) yields

$$\begin{aligned} r_N(x, x_j) =& \sigma^2 \frac{\exp(-\theta \|x - x_j\|_2^2)}{(2\pi\sigma_\epsilon^2)^d} \sqrt{\left(2\pi \frac{\sigma_\epsilon^2}{1 + 2\theta\sigma_\epsilon^2} \right)^d} \int_{\mathbb{R}^d} \exp\left(\frac{2\sigma_\epsilon^2 \theta^2}{1 + 2\sigma_\epsilon^2 \theta} \|x - x_j - \epsilon_j\|_2^2 \right) \\ & \times \exp\left(-\left(\theta + \frac{1}{2\sigma_\epsilon^2}\right) \epsilon_j^T \epsilon_j + 2\theta (x - x_j)^T \epsilon_j \right) d\epsilon_j. \end{aligned} \tag{29}$$

We next compute

$$\begin{aligned} & \int_{\mathbb{R}^d} \exp\left(\frac{2\sigma_\epsilon^2 \theta^2}{1 + 2\sigma_\epsilon^2 \theta} \|x - x_j - \epsilon_j\|_2^2 \right) \exp\left(-\left(\theta + \frac{1}{2\sigma_\epsilon^2}\right) \epsilon_j^T \epsilon_j + 2\theta (x - x_j)^T \epsilon_j \right) d\epsilon_j \\ =& \exp\left(\frac{2\sigma_\epsilon^2 \theta^2}{1 + 2\sigma_\epsilon^2 \theta} \|x - x_j\|_2^2 \right) \int_{\mathbb{R}^d} \exp\left(-\left(\theta + \frac{1}{2\sigma_\epsilon^2} - \frac{2\sigma_\epsilon^2 \theta^2}{1 + 2\sigma_\epsilon^2 \theta}\right) \epsilon_j^T \epsilon_j + 2\left(\theta - \frac{2\sigma_\epsilon^2 \theta^2}{1 + 2\sigma_\epsilon^2 \theta}\right) (x - x_j)^T \epsilon_j \right) d\epsilon_j \\ =& \exp\left(\frac{2\sigma_\epsilon^2 \theta^2}{1 + 2\sigma_\epsilon^2 \theta} \|x - x_j\|_2^2 \right) \int_{\mathbb{R}^d} \exp\left(-\left(\frac{1 + 4\sigma_\epsilon^2 \theta}{(1 + 2\sigma_\epsilon^2 \theta)\sigma_\epsilon^2}\right) \epsilon_j^T \epsilon_j + \frac{2\theta}{1 + 2\sigma_\epsilon^2 \theta} (x - x_j)^T \epsilon_j \right) d\epsilon_j \\ =& \exp\left(\frac{2\sigma_\epsilon^2 \theta^2}{1 + 2\sigma_\epsilon^2 \theta} \|x - x_j\|_2^2 \right) \sqrt{\left(2\pi \frac{(1 + 2\sigma_\epsilon^2 \theta)\sigma_\epsilon^2}{1 + 4\sigma_\epsilon^2 \theta} \right)^d} \exp\left(\frac{(1 + 2\sigma_\epsilon^2 \theta)\sigma_\epsilon^2}{1 + 4\sigma_\epsilon^2 \theta} \frac{\theta^2}{(1 + 2\sigma_\epsilon^2 \theta)^2} \|x - x_j\|_2^2 \right). \end{aligned} \tag{30}$$

By plugging (30) into (29), we obtain

$$\begin{aligned} r_N(x, x_j) =& \sigma^2 \frac{\exp(-\theta \|x - x_j\|_2^2)}{(2\pi\sigma_\epsilon^2)^d} \sqrt{\left(2\pi \frac{\sigma_\epsilon^2}{1 + 2\theta\sigma_\epsilon^2} \right)^d} \\ & \times \exp\left(\frac{2\sigma_\epsilon^2 \theta^2}{1 + 2\sigma_\epsilon^2 \theta} \|x - x_j\|_2^2 \right) \sqrt{\left(2\pi \frac{(1 + 2\sigma_\epsilon^2 \theta)\sigma_\epsilon^2}{1 + 4\sigma_\epsilon^2 \theta} \right)^d} \exp\left(\frac{2(1 + 2\sigma_\epsilon^2 \theta)\sigma_\epsilon^2}{1 + 4\sigma_\epsilon^2 \theta} \frac{\theta^2}{(1 + 2\sigma_\epsilon^2 \theta)^2} \|x - x_j\|_2^2 \right) \\ =& \frac{\sigma^2}{(1 + 4\sigma_\epsilon^2 \theta)^{d/2}} \exp\left(\frac{-\theta \|x - x_j\|_2^2}{1 + 4\sigma_\epsilon^2 \theta} \right), \end{aligned} \tag{31}$$

which is desired. The term $r(x, x_j)$ can be computed by

$$\begin{aligned} r(x, x_j) =& \sigma^2 \int_{\mathbb{R}^d} \Psi(x - (x_j + \epsilon_j)) p(\epsilon_j) d\epsilon_j \\ =& \sigma^2 \int_{\mathbb{R}^d} \exp(-\theta \|x - (x_j + \epsilon_j)\|_2^2) \frac{1}{\sqrt{(2\pi\sigma_\epsilon^2)^d}} \exp\left(-\frac{\epsilon_j^T \epsilon_j}{2\sigma_\epsilon^2} \right) \end{aligned}$$

$$
\begin{aligned}
&= \sigma^2 \frac{\exp(-\theta\|x-x_j\|_2^2)}{\sqrt{(2\pi\sigma_\epsilon^2)^d}} \int_{\mathbb{R}^d} \exp\left(-\left(\theta+\frac{1}{2\sigma_\epsilon^2}\right)\epsilon_j^T\epsilon_j + 2\theta(x-x_j)^T\epsilon_j\right) d\epsilon_j \\
&= \sigma^2 \frac{\exp(-\theta\|x-x_j\|_2^2)}{\sqrt{(2\pi\sigma_\epsilon^2)^d}} \exp\left(\frac{2\sigma_\epsilon^2\theta^2}{1+2\sigma_\epsilon^2\theta}\|x-x_j\|_2^2\right)\sqrt{\left(2\pi\frac{\sigma_\epsilon^2}{1+2\theta\sigma_\epsilon^2}\right)^d} \\
&= \frac{\sigma^2}{(1+2\sigma_\epsilon^2\theta)^{d/2}} \exp\left(\frac{-\theta\|x-x_j\|_2^2}{1+2\sigma_\epsilon^2\theta}\right). \qquad (32)
\end{aligned}
$$

Note $K_{jk} = r_N(x_j, x_k)$ if $j \neq k$. Together with (31) and (32), we obtain (7).

Appendix D. Proof of Lemma 2.4. By Fourier transform [37], we have

$$
\Psi(x_j - x_k) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{i\langle x_j - x_k, t\rangle} \mathcal{F}(\Psi)(t) dt, \qquad (33)
$$

where $\langle s, t\rangle = s^T t$ is the inner product in $\mathbb{R}^d$. Therefore, by Fubini's theorem, direct calculation leads to

$$
\begin{aligned}
\Psi_S(x_j - x_k) &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{i\langle x_j+\epsilon_1-(x_k+\epsilon_2),t\rangle} \mathcal{F}(\Psi)(t) p(\epsilon_1)p(\epsilon_2) dt d\epsilon_1 d\epsilon_2 \\
&= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{i\langle x_j+\epsilon_1-(x_k+\epsilon_2),t\rangle} p(\epsilon_1)p(\epsilon_2) d\epsilon_1 d\epsilon_2\right) \mathcal{F}(\Psi)(t) dt \\
&= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{i\langle x_j - x_k, t\rangle} \left(\int_{\mathbb{R}^d} e^{i\langle \epsilon_1, t\rangle} \int_{\mathbb{R}^d} e^{i\langle -\epsilon_2, t\rangle} p(\epsilon_1)p(\epsilon_2) d\epsilon_1 d\epsilon_2\right) \mathcal{F}(\Psi)(t) dt \\
&= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{i\langle x_j - x_k, t\rangle} \left(\int_{\mathbb{R}^d} e^{i\langle \epsilon_1, t\rangle} p(\epsilon_1) d\epsilon_1\right)\left(\int_{\mathbb{R}^d} e^{i\langle -\epsilon_2, t\rangle} p(\epsilon_2) d\epsilon_2\right) \mathcal{F}(\Psi)(t) dt. \qquad (34)
\end{aligned}
$$

For any $w = (w_1, \ldots, w_n)^T$, by (34), we have

$$
\begin{aligned}
&\sum_{j,k=1}^n w_j \bar{w}_k \Psi_S(x_j - x_k) \\
&= \sum_{j,k=1}^n w_j \bar{w}_k \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{i\langle x_j - x_k, t\rangle} \left(\int_{\mathbb{R}^d} e^{i\langle \epsilon_1, t\rangle} p(\epsilon_1) d\epsilon_1\right)\left(\int_{\mathbb{R}^d} e^{i\langle -\epsilon_2, t\rangle} p(\epsilon_2) d\epsilon_2\right) \mathcal{F}(\Psi)(t) dt \\
&= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left|\sum_{j=1}^n w_j e^{i\langle x_j, t\rangle}\right|^2 \left(\int_{\mathbb{R}^d} e^{i\langle \epsilon_1, t\rangle} p(\epsilon_1) d\epsilon_1\right)\left(\int_{\mathbb{R}^d} e^{i\langle -\epsilon_2, t\rangle} p(\epsilon_2) d\epsilon_2\right) \mathcal{F}(\Psi)(t) dt. \qquad (35)
\end{aligned}
$$

Let

$$
c(t) = \left(\int_{\mathbb{R}^d} e^{i\langle \epsilon_1, t\rangle} p(\epsilon_1) d\epsilon_1\right)\left(\int_{\mathbb{R}^d} e^{i\langle -\epsilon_2, t\rangle} p(\epsilon_2) d\epsilon_2\right).
$$

Thus, $c(t) \in \mathbb{R}$ and $c(t) > 0$. Therefore, $\sum_{j,k=1}^n w_j \bar{w}_k \Psi_S(x_j - x_k) \geq 0$, and equal to zero if and only if $w = 0$, which finishes the proof.

Appendix E. Proof of Theorem 2.7.

Consider the following GP with output error,

$$y_S(x) = M_S(x) + \delta(x), \tag{36}$$

where M_S is a mean zero GP with covariance function $\sigma^2\Psi_S$, and $\delta(x)$ is an independent noise process with mean zero and variance μ. The best linear unbiased predictor of (36) is

$$\hat{f}_S(x) = r_N(x)^T(R_S + \mu I_n)^{-1}Y, \tag{37}$$

and the MSPE is

$$\text{MSPE}_S = \sigma^2\Psi_S(0) - r_N(x)^T(R_S + \mu I_n)^{-1}r_N(x). \tag{38}$$

By Lemma B.1, (38) goes to zero as the fill distance of design points X goes to zero.

Take $\mu = \sigma^2(1 - \Psi_S(0))$. It can be seen that (38) is equal to $\sigma^2\Psi_S(0) - r_N(x)R^{-1}r_N(x)$. By (9), $\mathbb{E}(y(x) - \hat{y}(x))^2 = \text{MSPE}_S + \sigma^2(1 - \Psi_S(0))$, which converges to $\sigma^2(1 - \Psi_S(0))$ as the fill distance of the design points goes to zero. This completes the proof.

Appendix F. Proof of Theorem 3.1. Without loss of generality, assume $\sigma = 1$. First, we consider there is noise at point x. For any $u = (u_1, \ldots, u_n)^T$, it can be shown that the MSPE of predictor u^TY is

$$\begin{aligned}
&\mathbb{E}\left\|\Psi(\cdot - (x+\epsilon)) - \sum_{j=1}^n u_i\Psi(\cdot - (x_j + \epsilon_j))\right\|^2_{\mathcal{N}_\Psi(\Omega)} \\
=&\mathbb{E}\left(1 - 2\sum_{j=1}^n u_i\Psi((x_j+\epsilon_j) - (x+\epsilon)) + \sum_{j,k=1}^n u_ju_k\Psi((x_j+\epsilon_j) - (x_k+\epsilon_k))\right) \\
=&1 - 2\sum_{j=1}^n u_j\Psi_S(x - x_j) + \sum_{j,k=1}^n u_ju_k\Psi_S(x_j - x_k) + a\|u\|_2^2, \qquad (39)
\end{aligned}$$

where $\|\cdot\|_{\mathcal{N}_\Psi(\Omega)}$ is the norm of the reproducing kernel Hilbert space $\mathcal{N}_\Psi(\Omega)$ and $a = 1 - \Psi_S(0)$, and the last equality follows from (10). Notice that

$$\Psi_S(x_j - x_k) = \frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d} e^{i\langle x_j - x_k, t\rangle}c(t)\mathcal{F}(\Psi)(t)dt,$$

where

$$c(t) = \left(\int_{\mathbb{R}^d} e^{i\langle\epsilon_j, t\rangle}p(\epsilon_j)d\epsilon_j\right)\left(\int_{\mathbb{R}^d} e^{i\langle-\epsilon_k, t\rangle}p(\epsilon_k)d\epsilon_k\right).$$

Since $|e^{i\langle-\epsilon_j, t\rangle}| \le 1$, $c(t) \le 1$. Therefore, (39) can be bounded by

$$1 - 2\sum_{j=1}^n u_j\Psi_S(x - x_j) + \sum_{j,k=1}^n u_ju_k\Psi_S(x_j - x_k) + a\|u\|_2^2$$

$$
\begin{aligned}
&= u^T R_S u - 2u^T r_S(x) + \Psi_S(x-x) + a\|u\|_2^2 + a \\
&= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left| \sum_{j=1}^n u_j e^{i\langle x_j, t\rangle} - e^{i\langle x, t\rangle} \right|^2 c(t)\mathcal{F}(\Psi)(t)dt + a\|u\|_2^2 + a \\
&\leq \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left| \sum_{j=1}^n u_j e^{i\langle x_j, t\rangle} - e^{i\langle x, t\rangle} \right|^2 \mathcal{F}(\Psi)(t)dt + a\|u\|_2^2 + a \\
&= u^T R_\Psi u - 2u^T r_\Psi(x) + 1 + a\|u\|_2^2 + a \\
&\leq \max\{1, a/\mu\}(u^T R_\Psi u - 2u^T r_\Psi(x) + 1 + \mu\|u\|_2^2) + a, \qquad (40)
\end{aligned}
$$

where $r_S(x) = (\Psi(x-x_1), ..., \Psi(x-x_n))^T$ and the second equality follows from (35). Plugging

$$u = (R_\Psi + \mu I_n)^{-1} r_\Psi(x),$$

into (39) and (40), we have the MSPE of predictor (13) upper bounded by

$$\max\{1, a/\mu\}(1 - r_\Psi(x)^T (R_\Psi + \mu I_n)^{-1} r_\Psi(x)) + a.$$

By Lemma B.1, $1 - r_\Psi(x)^T (R_\Psi + \mu I_n)^{-1} r_\Psi(x)$ converges to zero as the fill distance goes to zero since μ is a constant, which completes the proof in this case.

Next, we consider the case that there is no noise at point x. For any $u = (u_1, \ldots, u_n)^T$, it can be shown that the MSPE of predictor $u^T Y$ in this case is

$$
\begin{aligned}
&\mathbb{E}\left\| \Psi(\cdot - x) - \sum_{j=1}^n u_j \Psi(\cdot - (x_j + \epsilon)) \right\|_{\mathcal{N}_\Psi}^2 \\
&= u^T R_S u - 2u^T r(x) + \Psi(x-x) + a\|u\|_2^2. \qquad (41)
\end{aligned}
$$

Let $b(t) = \int_{\mathbb{R}^d} e^{i\langle \epsilon_i, t\rangle} h(\epsilon_i) d\epsilon_i$. Thus, for any $u = (u_1, \ldots, u_n)^T$, we have

$$
\begin{aligned}
&u^T R_S u - 2u^T r(x) + \Psi(x-x) + a\|u\|_2^2 \\
=&\frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left| \sum_{j=1}^n u_j e^{i\langle x_j, t\rangle} b(t) - e^{i\langle x, t\rangle} \right|^2 \mathcal{F}(\Psi)(t)dt + a\|u\|_2^2 \\
\leq&\frac{1+C^2}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left| \sum_{j=1}^n u_j e^{i\langle x_j, t\rangle} - e^{i\langle x, t\rangle} \right|^2 |b(t)|^2 \mathcal{F}(\Psi)(t)dt + \frac{1+C^{-2}}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} |1 - |b(t)||^2 \mathcal{F}(\Psi)(t)dt + a\|u\|_2^2 \\
\leq&(1+C^2)(u^T R_\Psi u - 2u^T r_\Psi(x) + 1) + a\|u\|_2^2 + (1+C^{-2}) \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} |1 - |b(t)||^2 \mathcal{F}(\Psi)(t)dt \\
&(42) \\
\leq&\max\{(1+C^2), a/\mu\}(u^T R_\Psi u - 2u^T r_\Psi(x) + 1 + \mu\|u\|_2^2) + \frac{1+C^{-2}}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} |1 - |b(t)||^2 \mathcal{F}(\Psi)(t)dt,
\end{aligned}
$$

where we use $2\langle a, b\rangle \leq C^2|a|^2 + C^{-2}|b|^2$ in the first inequality, with C a fixed constant. Plugging

$$u = (R_\Psi + \mu I_n)^{-1} r_\Psi(x),$$

into (41) and (42), we have the MSPE of predictor (13) upper bounded by

$$\max\{(1+C^2), a/\mu\}(1 - r_\Psi(x)^T(R_\Psi + \mu I_n)^{-1} r_\Psi(x)) + \frac{1+C^{-2}}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} |1 - |b(t)||^2 \mathcal{F}(\Psi)(t) dt.$$

By Lemma B.1, $1 - r_\Psi(x)^T(R_\Psi + \mu I_n)^{-1} r_\Psi(x)$ converges to zero as the fill distance goes to zero since μ is a constant. The constant C influences the number of design points needed such that $\max\{(1+C^2), a/\mu\}(1 - r_\Psi(x)^T(R_\Psi + \mu I_n)^{-1} r_\Psi(x))$ is close to zero. For a fixed number of design points, the larger C is, the larger $\max\{(1+C^2), a/\mu\}(1 - r_\Psi(x)^T(R_\Psi + \mu I_n)^{-1} r_\Psi(x))$ is. To derive an explicit bound, we let $C^2 = 25$, which yields an asymptotic upper bound

$$\frac{1.04}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} |1 - |b(t)||^2 \mathcal{F}(\Psi)(t) dt.$$

This finishes the proof.

Appendix G. Proof of Proposition 3.6. Notice that $\mathbb{E}(e^{i\epsilon_n^T t})$ converges to 1 since ϵ_n converges to 0 in distribution and $e^{i\epsilon_n^T t}$ is bounded, and $b(t)$ is bounded for all $t \in \mathbb{R}^d$. By dominated convergence theorem, the result holds.

Appendix H. Proof of Theorem 4.1. We first present a lemma, which is a generalization of Lemma B.1.

Lemma H.1. *Suppose the conditions of Theorem 4.1 hold. Then we have* $1 - \tilde{r}_\Psi(x)^T(\tilde{R}_\Psi + \tilde{\mu} I)^{-1} \tilde{r}_\Psi(x)$ *converges to zero as the fill distance of* X *converges to zero, where* $\tilde{\Psi} = \tilde{\Psi}_1$ *or* $\tilde{\Psi}_2$.

Proof. The proof of Lemma H.1 is similar to the proof of Lemma B.1. The only difference is that if we define $\tilde{g}(t) = \tilde{\Psi}(t-x) - \tilde{r}_\Psi(t)^T(\tilde{R}_\Psi + \tilde{\mu} I)^{-1} \tilde{r}_\Psi(x)$, then $\|\tilde{g}\|_{H^\eta(\Omega)} \leq C_2$ for all $\tilde{g}$. Thus, the result follows from the proof of Lemma B.1. ∎

Now we are ready to show the proof of Theorem 4.1. Let $\tilde{y}(x)$ be the SK predictor with parameters $(\tilde{\theta}_2, \tilde{\mu})$. Thus,

$$\tilde{y}(x) = \tilde{r}_2(x)^T(\tilde{R}_2 + \tilde{\mu} I_n)^{-1} Y, \tag{43}$$

where $\tilde{r}_2(x) = (\tilde{\Psi}_2(x, x_1), \ldots, \tilde{\Psi}_2(x, x_n))^T$ and $\tilde{R}_2 = (\tilde{\Psi}_2(x_j - x_k))_{jk}$.

Proof of Statement (i):

Direct calculation shows that the MSPE can be expressed as

$$\begin{aligned} \mathbb{E}(y(x) - \tilde{y}(x))^2 =& \sigma^2(1 - 2\tilde{r}_2(x)^T(\tilde{R}_2 + \tilde{\mu} I_n)^{-1} r_N(x) \\ &+ \tilde{r}_2(x)^T(\tilde{R}_2 + \tilde{\mu} I_n)^{-1} R(\tilde{R}_2 + \tilde{\mu} I)^{-1} \tilde{r}_2(x)), \end{aligned} \tag{44}$$

where R and r_N are as in (4) and (6), respectively. Similar to (40), we have for any $u = (u_1, \ldots, u_n)^T$,

$$1 - 2\sum_{j=1}^n u_j \Psi_S(x - x_j) + \sum_{j,k=1}^n u_j u_k \Psi_S(x_j - x_k) + a\|u\|_2^2$$

$$
\begin{aligned}
&= u^T R_S u - 2u^T r_S(x) + \Psi_S(x-x) + a\|u\|_2^2 + a \\
&= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left| \sum_{j=1}^n u_j e^{i\langle x_j, t\rangle} - e^{i\langle x, t\rangle} \right|^2 c(t)\mathcal{F}(\Psi)(t)dt + a\|u\|_2^2 + a \\
&\leq \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left| \sum_{j=1}^n u_j e^{i\langle x_j, t\rangle} - e^{i\langle x, t\rangle} \right|^2 \mathcal{F}(\Psi)(t)dt + a\|u\|_2^2 + a \\
&\leq \frac{A_1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left| \sum_{j=1}^n u_j e^{i\langle x_j, t\rangle} - e^{i\langle x, t\rangle} \right|^2 \mathcal{F}(\tilde{\Psi}_2)(t)dt + a\|u\|_2^2 + a \\
&= A_1(u^T \tilde{R}_2 u - 2u^T \tilde{r}_2(x) + \tilde{\Psi}_2(x-x)) + a\|u\|_2^2 + a \\
&\leq \max\{A_1, a/\tilde{\mu}\}(u^T \tilde{R}_2 u - 2u^T \tilde{r}_2(x) + \tilde{\Psi}_2(0) + \tilde{\mu}\|u\|_2^2) + a, \qquad (45)
\end{aligned}
$$

where

$$
c(t) = \left(\int_{\mathbb{R}^d} e^{i\langle \epsilon_j, t\rangle} p(\epsilon_j) d\epsilon_j \right) \left(\int_{\mathbb{R}^d} e^{i\langle -\epsilon_k, t\rangle} p(\epsilon_k) d\epsilon_k \right),
$$

and $a = 1 - \Psi_S(0)$. Plugging

$$
u = (\tilde{R}_2 + \tilde{\mu} I_n)^{-1} \tilde{r}_2(x),
$$

into (44) and (45), we have the MSPE of predictor (44) is upper bounded by

$$
\begin{aligned}
&\max\{A_1, a/\tilde{\mu}\}(\tilde{\Psi}_2(0) - \tilde{r}_2(x)^T (\tilde{R}_2 + \tilde{\mu} I_n)^{-1} \tilde{r}_2(x) + a \\
&\leq \max\{A_1, aC\}(\tilde{\Psi}_2(0) - \tilde{r}_2(x)^T (\tilde{R}_2 + C I_n)^{-1} \tilde{r}_2(x) + a
\end{aligned}
$$

By Lemma H.1, $\tilde{\Psi}_2(0) - \tilde{r}_2(x)^T (\tilde{R}_2 + C I_n)^{-1} \tilde{r}_2(x)$ converges to zero as the fill distance goes to zero, which indicates that $\sigma^2 a$ is an asymptotic upper bound on the MSPE of SK with parameters. Note that $\sigma^2 a$ is also the limit of KALEN with the true parameters, which is the best linear unbiased predictor. Therefore, $\sigma^2 a$ is the limit of SK with parameters.

Note that KALEN is

$$
\hat{y}(x) = \tilde{r}_N(x)^T (\tilde{R}_S + \tilde{a} I_n)^{-1} Y, \qquad (46)
$$

where $\tilde{R}_S = (\tilde{\Psi}_S(x_j - x_k))_{jk}$, $\tilde{r}_N(x) = (\tilde{\Psi}_S(x - x_1), ..., \tilde{\Psi}_S(x - x_n))$,

$$
\tilde{\Psi}_S(s-t) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \tilde{\Psi}_1(s + \epsilon_1 - (t + \epsilon_2)) \tilde{p}(\epsilon_1) \tilde{p}(\epsilon_2) d\epsilon_1 d\epsilon_2,
$$

and $\tilde{a} = \tilde{\Psi}_1(0) - \tilde{\Psi}_S(0)$. Condition (4) in Theorem 4.1 implies that $\tilde{a}$ is bounded away from zero. Thus, repeating the argument in the proof of SK completes the proof of Statement (i).

Proof of Statement (ii):

By direct calculation, it can be shown that

$$
\mathbb{E}(y(x) - \tilde{y}(x))^2 = \sigma^2 (1 - 2\tilde{r}_2(x)^T (\tilde{R}_2 + \tilde{\mu} I_n)^{-1} r(x)
$$

$$+\tilde{r}_2(x)^T(\tilde{R}_2+\tilde{\mu}I_n)^{-1}R(\tilde{R}_2+\tilde{\mu}I)^{-1}\tilde{r}_2(x), \tag{47}$$

where $r(x)$ is as in (3). Let $b(t)=\int_{\mathbb{R}^d}e^{i\langle\epsilon_j,t\rangle}p(\epsilon_j)d\epsilon_j$. For any $u=(u_1,\ldots,u_n)^T$, we have

$$\begin{aligned}
&u^TR_Su-2u^Tr(x)+1+a\|u\|_2^2\\
=&\frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\left|\sum_{j=1}^n u_je^{i\langle x_j,t\rangle}b(t)-e^{i\langle x,t\rangle}\right|^2\mathcal{F}(\Psi)(t)dt+a\|u\|_2^2\\
\leq&\frac{(1+C_1^2)}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\left|\sum_{j=1}^n u_je^{i\langle x_j,t\rangle}-e^{i\langle x,t\rangle}\right|^2|b(t)|^2\mathcal{F}(\Psi)(t)dt+\frac{(1+C_1^{-2})}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}|1-|b(t)||^2\mathcal{F}(\Psi)(t)dt+a\|u\|_2^2\\
\leq&\frac{(1+C_1^2)A_1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\left|\sum_{j=1}^n u_je^{i\langle x_j,t\rangle}-e^{i\langle x,t\rangle}\right|^2|b(t)|^2\mathcal{F}(\tilde{\Psi}_2)(t)dt+\frac{(1+C_1^{-2})}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}|1-|b(t)||^2\mathcal{F}(\Psi)(t)dt+a\|u\|_2^2\\
\leq&(1+C_1^2)A_1(u^T\tilde{R}_2u-2u^T\tilde{r}_2(x)+\tilde{\Psi}_2(x-x))+a\|u\|_2^2+\frac{(1+C_1^{-2})}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}|1-|b(t)|^2|\mathcal{F}(\Psi)(t)dt\\
\leq&\max\{(1+C_1^2)A_1,a/\tilde{\mu}\}(u^T\tilde{R}_2u-2u^T\tilde{r}_2(x)+\tilde{\Psi}_2(0)+\tilde{\mu}\|u\|_2^2)\\
&+\frac{(1+C_1^{-2})}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}|1-|b(t)|^2|\mathcal{F}(\Psi)(t)dt.
\end{aligned}\tag{48}$$

Plugging $u=(\tilde{R}_2+\tilde{\mu}I)^{-1}\tilde{r}_2(x)$, into (47) and (48), we find the MSPE of predictor (13) is upper bounded by

$$\begin{aligned}
&\max\{(1+C_1^2)A_1,a/\tilde{\mu}\}(\tilde{\Psi}_2(0)-\tilde{r}_2(x)^T(\tilde{R}_2+\tilde{\mu}I_n)^{-1}\tilde{r}_2(x)\\
&+\frac{(1+C_1^{-2})}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}|1-|b(t)|^2|\mathcal{F}(\Psi)(t)dt\\
\leq&\max\{(1+C_1^2)A_1,aC\}(\tilde{\Psi}_2(0)-\tilde{r}_2(x)^T(\tilde{R}_2+CI_n)^{-1}\tilde{r}_2(x)\\
&+\frac{(1+C_1^{-2})}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}|1-|b(t)|^2|\mathcal{F}(\Psi)(t)dt.
\end{aligned}$$

We take $C_1^2=25$. By Lemma H.1, $\tilde{\Psi}_2(0)-\tilde{r}_2(x)^T(\tilde{R}_2+CI_n)^{-1}\tilde{r}_2(x)$ converges to zero as the fill distance goes to zero since C is a constant, which finishes the proof for SK.

Note that the KALE is

$$\hat{f}(x)=\tilde{r}(x)^T(\tilde{R}_S+\tilde{a}I)^{-1}Y,$$

where $\tilde{r}(x)$ is as in (3) with parameters $\tilde{\theta}_1^{(1)}$, and $\tilde{R}_S$ and $\tilde{a}$ are as in (46). Because $\tilde{\Psi}_1$ is a correlation function and $\tilde{p}(\cdot)=p(\cdot)$, we have $\tilde{\Psi}_1(0)=1$ and $\tilde{\Psi}_S(0)=\Psi_S(0)$, which imply $\tilde{a}=\tilde{1}-\tilde{\Psi}_S(0)=1-\Psi_S(0)=a$. Then for any $u=(u_1,\ldots,u_n)^T$, we have

$$u^TR_Su-2u^Tr(x)+1+a\|u\|_2^2$$

$$
\begin{aligned}
&= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left| \sum_{j=1}^{n} u_j e^{i\langle x_j, t\rangle} b(t) - e^{i\langle x, t\rangle} \right|^2 \mathcal{F}(\Psi)(t) dt + a\|u\|_2^2 \\
&\leq \frac{A_1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left| \sum_{j=1}^{n} u_j e^{i\langle x_j, t\rangle} b(t) - e^{i\langle x, t\rangle} \right|^2 \mathcal{F}(\tilde{\Psi}_1)(t) dt + a\|u\|_2^2 \\
&= A_1(u^T \tilde{R}_S u - 2u^T \tilde{r}(x) + \tilde{\Psi}_1(x - x)) + a\|u\|_2^2. \qquad (49)
\end{aligned}
$$

Note that $\hat{f}(x)$ minimizes (49). Then repeating the proof of Theorem 3.1 gives an upper bound

$$
\frac{1.04 A_1 \sigma^2}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} |1 - |b(t)|^2| \mathcal{F}(\tilde{\Psi}_1)(t) dt.
$$

Together with $\mathcal{F}(\tilde{\Psi}_1)(t) \leq A_2 \mathcal{F}(\Psi)(t)$ for any t, we finish the proof.

REFERENCES

[1] B. Ankenman, B. L. Nelson, and J. Staum, *Stochastic kriging for simulation metamodeling*, Operations Research, 58 (2010), pp. 371–382.

[2] J. J. Barber, A. E. Gelfand, and J. A. Silander, *Modelling map positional error to infer true feature location*, Canadian Journal of Statistics, 34 (2006), pp. 659–676.

[3] M. Binois, R. B. Gramacy, and M. Ludkovski, *Practical heteroscedastic Gaussian process modeling for large simulation experiments*, Journal of Computational and Graphical Statistics, 27 (2018), pp. 808–821.

[4] B. A. Bócsi and L. Csató, *Hessian corrected input noise models*, in International Conference on Artificial Neural Networks, Springer, 2013, pp. 1–8.

[5] H. Brezis and P. Mironescu, *Where Sobolev interacts with Gagliardo–Nirenberg*, Journal of Functional Analysis, 277 (2019), pp. 2839–2864.

[6] D. Cervone and N. S. Pillai, *Gaussian process regression with location errors*, arXiv preprint arXiv:1506.08256, (2015).

[7] N. Cressie, *Statistics for Spatial Data*, John Wiley & Sons, 2015.

[8] N. Cressie and J. Kornak, *Spatial statistics in the presence of location error with an application to remote sensing of the environment*, Statistical Science, 18 (2003), pp. 436–456.

[9] P. Dallaire, C. Besse, and B. Chaib-Draa, *Learning Gaussian process models from uncertain data*, in International Conference on Neural Information Processing, Springer, 2009, pp. 433–440.

[10] M. P. Deisenroth, D. Fox, and C. E. Rasmussen, *Gaussian processes for data-efficient learning in robotics and control*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 37 (2015), pp. 408–423.

[11] R. A. DeVore and R. C. Sharpley, *Besov spaces on domains in R^d*, Transactions of the American Mathematical Society, 335 (1993), pp. 843–864.

[12] J. Du, X. Yue, J. H. Hunt, and J. Shi, *Optimal placement of actuators via sparse learning for composite fuselage shape control*, Journal of Manufacturing Science and Engineering, 141 (2019).

[13] K.-T. Fang, R. Li, and A. Sudjianto, *Design and Modeling for Computer Experiments*, CRC Press, 2005.

[14] I. I. Gihman and A. V. Skorokhod, *The Theory of Stochastic Processes I*, Springer, 1974.

[15] A. Girard, *Approximate Methods for Propagation of Uncertainty with Gaussian Process Models*, Ph.D. thesis, University of Glasgow, 2004.

[16] R. B. Gramacy and H. K. Lee, *Cases for the nugget in modeling computer experiments*, Statistics and Computing, 22 (2012), pp. 713–722.

[17] J. H. Halton, *Algorithm 247: Radical-inverse quasi-random point sequence*, Communications of the ACM, 7 (1964), pp. 701–702.
[18] S. He, W. Lin, and S.-H. G. Chan, *Indoor localization and automatic fingerprint update with altered ap signals*, IEEE Transactions on Mobile Computing, 16 (2017), pp. 1897–1910.
[19] D. Higdon, *Space and space-time modeling using process convolutions*, in Quantitative Methods for Current Environmental Issues, Springer, 2002, pp. 37–56.
[20] V. R. Joseph and L. Kang, *Regression-based inverse distance weighting with applications to computer experiments*, Technometrics, 53 (2011), pp. 254–265.
[21] Y. B. Lim, J. Sacks, W. Studden, and W. J. Welch, *Design and analysis of computer experiments when the output is highly correlated over the input space*, Canadian Journal of Statistics, 30 (2002), pp. 109–126.
[22] G. Matheron, *Principles of geostatistics*, Economic Geology, 58 (1963), pp. 1246–1266.
[23] A. McHutchon and C. E. Rasmussen, *Gaussian process training with input noise*, in Advances in Neural Information Processing Systems, 2011, pp. 1341–1349.
[24] L. S. Muppirisetty, T. Svensson, and H. Wymeersch, *Spatial wireless channel prediction under location uncertainty*, IEEE Transactions on Wireless Communications, 15 (2016), pp. 1031–1044.
[25] C.-Y. Peng and C. J. Wu, *On the choice of nugget in kriging modeling for deterministic computer experiments*, Journal of Computational and Graphical Statistics, 23 (2014), pp. 151–168.
[26] O. Roustant, D. Ginsbourger, and Y. Deville, *DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization*, (2012).
[27] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, *Design and analysis of computer experiments*, Statistical Science, 4 (1989), pp. 409–423.
[28] T. J. Santner, B. J. Williams, and W. I. Notz, *The Design and Analysis of Computer Experiments*, Springer Science & Business Media, 2013.
[29] M. L. Stein, *Interpolation of Spatial Data: Some Theory for Kriging*, Springer Science & Business Media, 1999.
[30] R. Tuo and C. F. J. Wu, *A theoretical framework for calibration in computer models: Parametrization, estimation and convergence properties*, SIAM/ASA Journal on Uncertainty Quantification, 4 (2016), pp. 767–795.
[31] F. I. Utreras, *Convergence rates for multivariate smoothing spline functions*, Journal of approximation theory, 52 (1988), pp. 1–27.
[32] D. Veneziano and J. Van Dyck, *Statistical analysis of earthquake catalogs for seismic hazard*, in Stochastic Approaches in Earthquake Engineering, Springer, 1987, pp. 385–427.
[33] W. Wang and B. Haaland, *Controlling sources of inaccuracy in stochastic kriging*, Technometrics, 61 (2019), pp. 309–321.
[34] W. Wang, R. Tuo, and C. F. J. Wu, *On prediction properties of kriging: Uniform error bounds and robustness*, Journal of the American Statistical Association, 115 (2020), pp. 920–930.
[35] Y. Wang, X. Yue, R. Tuo, J. H. Hunt, J. Shi, et al., *Effective model calibration via sensible variable identification and adjustment with application to composite fuselage simulation*, Annals of Applied Statistics, 14 (2020), pp. 1759–1776.
[36] Y. Wen, X. Yue, J. H. Hunt, and J. Shi, *Feasibility analysis of composite fuselage shape control via finite element analysis*, Journal of Manufacturing Systems, 46 (2018), pp. 272–281.
[37] H. Wendland, *Scattered Data Approximation*, vol. 17, Cambridge University Press, 2004.
[38] C. F. J. Wu and M. S. Hamada, *Experiments: Planning, Analysis, and Optimization*, John Wiley & Sons, 2nd ed., 2009.
[39] J. K. Yamamoto, *An alternative measure of the reliability of ordinary kriging estimates*, Mathematical Geology, 32 (2000), pp. 489–509.
[40] Z. Ying, *Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process*, Journal of Multivariate Analysis, 36 (1991), pp. 280–296.
[41] X. Yue, Y. Wen, J. H. Hunt, and J. Shi, *Surrogate model based control considering uncertainty for composite fuselage assembly*, Journal of Manufacturing Science and Engineering, 140 (2018), p. 041017.
[42] H. Zhang, *Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics*, Journal of the American Statistical Association, 99 (2004), pp. 250–261.