Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2010, Article ID 494070, 17 pages
doi:10.1155/2010/494070

Research Article

Asymptotic Behavior of the Likelihood Function of Covariance Matrices of Spatial Gaussian Processes

Ralf Zimmermann

German Aerospace Center (DLR), Lilienthalplatz 7, 38108 Braunschweig, Germany

Correspondence should be addressed to Ralf Zimmermann, ralf.zimmermann@dlr.de

Received 16 April 2010; Revised 3 November 2010; Accepted 28 November 2010

Academic Editor: F. Marcellan

The covariance structure of spatial Gaussian predictors (aka Kriging predictors) is generally modeled by parameterized covariance functions; the associated hyperparameters in turn are estimated via the method of maximum likelihood. In this work, the asymptotic behavior of the maximum likelihood of spatial Gaussian predictor models as a function of its hyperparameters is investigated theoretically. Asymptotic sandwich bounds for the maximum likelihood function in terms of the condition number of the associated covariance matrix are established. As a consequence, the main result is obtained: *optimally trained nondegenerate spatial Gaussian processes cannot feature arbitrary ill-conditioned correlation matrices*. The implication of this theorem on Kriging hyperparameter optimization is exposed. A nonartificial example is presented, where maximum likelihood-based Kriging model training is necessarily bound to fail.

1. Introduction

Spatial Gaussian processing, also known as *best linear unbiased prediction*, refers to a statistical data interpolation method, which is nowadays applied in a wide range of scientific fields, including computer experiments in modern engineering context; see, for example, [1–5]. As a powerful tool for geostatistics, it has been pioneered by Krige in 1951 [6], and to pay tribute to his achievements, the method is also termed *Kriging*; see [7, 8] for geostatistical background.

In practical applications, the data's covariance structure is modeled through covariance functions depending on the so-called *hyperparameters*. These, in turn, are estimated by optimizing the corresponding maximum likelihood function. It has been demonstrated by many authors that the accuracy of Kriging predictors relies both heavily on hyperparameter-based model training and, from the numerical point of view, on the condition number of the associated Kriging correlation matrix. In this regard, we relate to the following, nonexhaustive selection of papers: Warnes and Ripley [9] and Mardia and

Watkins [10] present numerical examples of difficult-to-optimize covariance model functions. Ababou et al. [11] show that likelihood-optimized hyperparameters may correspond to ill-conditioned correlation matrices. Diamond and Armstrong [12] prove error estimates under perturbation of covariance models, demonstrating a strong dependence on the correlation matrix' condition number. In the same setting, Posa [13] investigates numerically the behavior of this precise condition number for different covariance models and varying hyperparameters. An extensive experimental study of the condition number as a function of all parameters in the Kriging exercise is provided by Davis and Morris [14]. Schöttle and Werner [15] propose Kriging model training under suitable conditioning constraints. Related is the work of Ying [16] and Zhang and Zimmerman [17], who prove asymptotic results on limiting distributions of maximum likelihood estimators when the number of sample points approaches infinity. Radial basis function interpolant limits are investigated in [18]. Modern textbooks covering recent results are [19, 20].

In this paper, the connection between hyper-parameter optimization and the condition number of the correlation matrix is investigated from a theoretical point of view. The setting is as follows. All sample data is considered as fixed. An arbitrary feasible covariance model function is chosen for good, so that only the covariance models' hyperparameters are allowed to vary in order to adjust the model likelihood. This is exactly the situation as it occurs in the context of computer experiments, where, based on a fixed set of sample data, predictor models have to be trained numerically. We prove that, under weak conditions, the limit values of the quantities in the model training exercise exist. Subsequently, by establishing asymptotic sandwich bounds on the model likelihood based on the condition number of the associated correlation matrix, *it is shown that ill-conditioning eventually also decreases the model likelihood*. This result implies a strategy for choosing good starting solutions for hyperparameter-based model training. We emphasize that all covariance models applied in the papers briefly reviewed above subordinate to the theoretical setting of this work.

The paper is organized as follows. In the next section, a short review of the basic theory behind Kriging is given. The main theorem is stated and proved in Section 3. In Section 4, an example of a Kriging data set is presented, which illustrates the limitations of classical model training.

2. Kriging in a Nutshell

Kriging is a statistical approach for estimating an unknown (scalar) function

$$y : \mathbb{R}^d \supseteq U \longrightarrow \mathbb{R}, \quad x \longmapsto y(x), \tag{2.1}$$

based on a finite data set of sample locations $x^1, \ldots, x^n \in U \subset \mathbb{R}^d$ with corresponding responses $y_1 := y(x^1), \ldots, y_n := y(x^n) \in \mathbb{R}$ obtained from measurements or numerical computations. The collection of responses is denoted by

$$Y^T = (y_1, \ldots, y_n) \in \mathbb{R}^n. \tag{2.2}$$

The function $y : U \to \mathbb{R}$ to be estimated is assumed to be the realization of an underlying random process given by a regression model and a random error function $\epsilon(x)$ with zero mean. More precisely

$$y(x) = f(x)\beta + \epsilon(x) = \left(f_0(x), \dots, f_p(x)\right)\begin{pmatrix}\beta_0 \\ \vdots \\ \beta_p\end{pmatrix} + \epsilon(x), \tag{2.3}$$

where the components of the row vector function $f : \mathbb{R}^d \rightarrow \mathbb{R}^{p+1}$ are the basis functions of the regression model and $\beta = (\beta_0, \dots, \beta_p)$ is the corresponding vector of regression coefficients. By assumption,

$$E[\epsilon(x)] = 0 \quad \forall x. \tag{2.4}$$

The component functions of f can be chosen arbitrarily, yet they should form a function basis suitable to the specific application. The most common choices for practical applications are

(1) *constant regression (ordinary Kriging)*: $p = 0$, $f : \mathbb{R}^d \rightarrow \mathbb{R}, x \mapsto 1$, $f(x)\beta = \beta \in \mathbb{R}$,

(2) *linear regression (universal Kriging)*: $p = d$, $f : \mathbb{R}^d \rightarrow \mathbb{R}^{d+1}, x \mapsto (1, x_1, \dots, x_d)$, $f(x)\beta = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d \in \mathbb{R}$,

and higher-order polynomials.

Introducing the regression design matrix

$$F := \begin{pmatrix} f(x^1) \\ \vdots \\ f(x^n) \end{pmatrix} = \begin{pmatrix} f_0(x^1) & f_1(x^1) & \cdots & f_p(x^1) \\ \vdots & \vdots & \dots & \vdots \\ f_0(x^n) & f_1(x^n) & \cdots & f_p(x^n) \end{pmatrix} \in \mathbb{R}^{n \times (p+1)}, \tag{2.5}$$

the vector of errors at the sampled sites can be written as

$$\Sigma := \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} = \begin{pmatrix} \epsilon(x^1) \\ \vdots \\ \epsilon(x^n) \end{pmatrix} = \begin{pmatrix} y_1 - f(x^1)\beta \\ \vdots \\ y_n - f(x^n)\beta \end{pmatrix} = Y - F\beta. \tag{2.6}$$

Note that the first column of F equals $\mathbf{1} \in \mathbb{R}^n$ for all polynomial regression models.

The Kriging predictor $\hat{y}$ estimates y at an untried site x as a linear combination of the sampled data

$$\hat{y}(x) = \langle \omega(x), Y \rangle = \sum_{i=1}^{n} \omega_i(x) y_i. \tag{2.7}$$

For each $x \in \mathbb{R}^d$, the unique vector of weights $\omega(x) = (\omega_1(x), \ldots, \omega_n(x))$ that leads to an unbiased prediction minimizing the mean squared error is given by the solution of the Kriging equation system

$$\begin{pmatrix} C & F \\ F^T & 0 \end{pmatrix} \begin{pmatrix} \omega(x) \\ \frac{1}{2}\mu(x) \end{pmatrix} = \begin{pmatrix} c(x) \\ f(x)^T \end{pmatrix} \in \mathbb{R}^{n+(p+1)}. \tag{2.8}$$

Here,

$$C := \left(\mathrm{Cov}\left(\epsilon\left(x^i\right), \epsilon\left(x^j\right)\right)\right)_{i,j \le n} \in \mathbb{R}^{n \times n}, \qquad c(x) := \left(\mathrm{Cov}(\epsilon(x^i), \epsilon(x))\right)_{i \le n} \in \mathbb{R}^n \tag{2.9}$$

denote the covariance matrix and the covariance vector, respectively, and the entries of the vector $\mu = (\mu_0, \ldots, \mu_p)$ are Lagrange multipliers. Solving (2.8) by Schur matrix complement inversion and substituting in (2.7) leads to the Kriging predictor formula

$$\widehat{y}(x) = f(x)\beta + c^T(x)C^{-1}(Y - F\beta), \tag{2.10}$$

where $\beta = (F^T R^{-1} F)^{-1} F^T R^{-1} Y$ is the generalized least squares solution to the regression problem $F\beta \simeq Y$. For details, see, for example, [20, 21].

For setting up Kriging predictors, it is therefore mandatory to estimate the covariances based on the sampled data set. The two most popular approaches to tackle this problem are variogram fitting (the geostatistical literature, see [7]) and application of spatial correlation functions (computer experiments, see [4, 20]). The latter ones are usually of the form

$$\mathrm{Cov}\left(\epsilon\left(x^i\right), \epsilon\left(x^j\right)\right) = \sigma^2 r\left(\theta, x^i, x^j\right) = \sigma^2 \prod_{k=1}^{d} \mathrm{scf}_k\left(\theta, x^i, x^j\right). \tag{2.11}$$

Here $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$ is a vector of *distance weights*, which models the influence of the coordinate-wise spatial correlation on the prediction. The correlation matrix is defined by

$$R(\theta) = \left(r\left(\theta, x^i, x^j\right)\right)_{i,j} \in \mathbb{R}^{n \times n}. \tag{2.12}$$

In order to avoid ambiguity due to different parameterizations of the correlation models, we fix for the rest of the paper the following.

Convention 1

Large distance weight values correspond to weak spatial correlation, and small distance weight values correspond to strong spatial correlation. More precisely, we assume that feasible spatial correlation *functions are always parameterized such that*

$$r(\theta, p, q) \longrightarrow \begin{cases} 1, & \text{for } \|\theta\| \longrightarrow 0, \\ 0, & \text{for } \|\theta\| \longrightarrow \infty, \end{cases} \tag{2.13}$$

at distinct locations $p \neq q$.

All correlation models applied in all the papers briefly reviewed in the introduction can be parameterized accordingly. A collection of spatial correlation functions is given in several publications on Cokriging/Kriging, including [21, Table 2.1]. For example, the *Gaussian correlation function* parameterized with respect to the convention above is given by

$$\begin{aligned} \mathrm{scf}_k\left(\theta, x^i, x^j\right) &= \exp\left(-\theta_k \left|x_k^i - x_k^j\right|^2\right), \\ r\left(\theta, x^i, x^k\right) &= \exp\left(-\sum_k \theta_k \left|x_k^i - x_k^j\right|^2\right). \end{aligned} \tag{2.14}$$

The results and proofs presented below hold true without change for every admissible spatial correlation model, assuming that the sample errors are normally distributed and that the process variance is *stationary*; that is, σ is independent of the locations x^i, x^j. In this setting, hyper-parameter training for Kriging models consists of optimizing the corresponding *maximum likelihood function*

$$\mathrm{MaxL}(\sigma, \beta, \theta) = \frac{1}{2\pi\sqrt{\det(\sigma^2 R(\theta))}} \exp\left(-\frac{1}{2\sigma^2}(Y - F\beta.)^T R(\theta)^{-1}(Y - F\beta.)\right). \tag{2.15}$$

For θ fixed, optima for $\sigma = \sigma(\theta)$ and $\beta = \beta(\theta)$ can be derived analytically, see [20], so that hyper-parameter training for Kriging models is reduced to the following optimization problem:

$$\min_{\{\theta \in \mathbb{R}^d, \theta_j > 0\}} \mathrm{ML}(\theta) := \left(\sigma^2(\theta)\right)^n \det(R(\theta)), \tag{2.16}$$

where the dependency on θ is as follows:

$$\sigma^2(\theta) = \frac{1}{n}\Sigma^T(\theta) R(\theta)^{-1} \Sigma(\theta), \tag{2.17}$$

$$\Sigma(\theta) = Y - E[Y] = Y - F\beta(\theta), \tag{2.18}$$

$$\beta(\theta) = \left(F^T R^{-1}(\theta) F\right)^{-1} F^T R^{-1}(\theta) Y. \tag{2.19}$$

Because the logarithm is monotonic, this is equivalent to minimizing the so-called *condensed log-likelihood function*

$$\min_{\{\theta\in\mathbb{R}^d,\theta_j>0\}} \mathrm{LogML}(\theta) := n\ln\left(\sigma^2(\theta)\right) + \ln(\det(R(\theta))), \tag{2.20}$$

often encountered in the literature.

3. Asymptotic Behavior of the Maximum Likelihood Function—Why Kriging Model Training Is Tricky: Part I

The condition number of a regular matrix $R \in \mathbb{R}^{n\times n}$ with respect to a given matrix norm $\|\cdot\|$ is defined as $\mathrm{cond}(R) := \|R\| \cdot \|R^{-1}\|$. For the matrix norm induced by the *euclidean vector norm*, one can show that

$$\mathrm{cond}(R) = \frac{\lambda_{\max}}{\lambda_{\min}}, \tag{3.1}$$

where $\lambda_{\max}$, $\lambda_{\min}$ are the largest, the smallest eigenvalue of R, respectivley. In order to prevent the solution of the Kriging equation system (2.8) from being spoiled by numerical errors, it is important to prevent the covariance matrix from being severely ill-conditioned. However, the next theorem shows that eventually, when the condition number approaches infinity, so does the associated likelihood function. (Keep in mind that we have formulated likelihood estimation as a *minimization problem;* see (2.16).) Throughout this section, we will assume that the regression design matrix F from (2.5) features the vector $\mathbf{1} \in \mathbb{R}^n$ as first column. This is the case of the highest practical relevance and, in fact, is of particular difficulty, since in this case the first column of F coincides with a limit eigenvector of the correlation matrix as will be seen in the following.

Theorem 3.1. *Let $x^1,\ldots,x^n \in \mathbb{R}^d$, $Y := (y(x^1),\ldots,y(x^n)) \in \mathbb{R}^n$ be a data set of sampled sites and responses. Let $R(\theta)$ be the associated spatial correlation matrix, and let $\Sigma(\theta)$ be the vector of errors with respect to the chosen regression model. Furthermore, let $\mathrm{cond}(R(\theta))$ be the condition number of $R(\theta)$. Suppose that the following conditions hold true:*

(1) *the eigenvalues $\lambda_i(\theta)$ of $R(\theta)$ are mutually distinct for small $0 < \|\theta\| \le \varepsilon$, $\theta \in \mathbb{R}^d$,*

(2) *the derivatives of the eigenvalues do not vanish in the limit: $(d/d\tau)|_{\tau=0}\lambda_j(\tau\mathbf{1}) = \lambda_j'(0) \neq 0$ for all $j = 2,\ldots,n$,*

(3) $\Sigma(0) \notin \mathrm{span}\{\mathbf{1}\}$, $\Sigma(0) \notin \mathbf{1}^{\perp}$.

Then,

$$\mathrm{cond}(R(\theta)) \longrightarrow \infty \quad \textit{for } \|\theta\| \longrightarrow 0, \tag{3.2}$$

and there exist constants $c_1, c_2 \in \mathbb{R}$ such that

$$c_1 \,\mathrm{cond}(R(\theta)) \le ML(\theta) \le c_2 \,\mathrm{cond}(R(\theta)), \tag{3.3}$$

for $\theta \in \mathbb{R}^d$, $0 < \|\theta\| \le \varepsilon$. The constants c_1, c_2 are independent of θ.

Remark 3.2. (1) The conditions given in the above theorem cannot be proven to hold true in general, since they depend on the data set in question. However, they hold true for nondegenerate data set. In Appendix A, a relationship between condition 2 and the regularity of $R'(0)$ is established, giving strong support that condition 2 is generally valid. Concerning the third condition, it will be shown in Lemma 3.4, that the limit $\Sigma(0)$ exists, given conditions 1 and 2. Note that the set $\text{span}\{\mathbf{1}\} \cup \mathbf{1}^{\perp}$ is of Lebesgue measure zero in $\mathbb{R}^n$. In all practical applications known to the author, these conditions were fulfilled.

(2) It holds that $\lim_{\|\theta\| \to \infty}(R(\theta))_{i,j} = I \in \mathbb{R}^{n \times n}$. Hence, the likelihood function approaches a constant limit for $\|\theta\| \to \infty$ and $\lim_{\|\theta\| \to \infty}\text{cond}(R(\theta)) = 1$. The corresponding predictor behavior is investigated in Section 4.

(3) Even though Theorem 3.1 shows that the model likelihood becomes arbitrarily bad for hyperparameters $\|\theta\| \to 0$, the optimum might lie very close to the blowup region of the condition number, leading to still quite ill-conditioned covariance matrices [11]. This fact as well as the general behavior of the likelihood function as predicted by Theorem 3.1 is illustrated in Figure 1.

(4) Figure 2(b), provides an additional illustration of Theorem 3.1.

(5) Theorem 3.1 offers a strategy for choosing starting solutions for the optimization problem (2.16): take each $\theta_k, k \in 1, \ldots, d$ as small as possible such that the corresponding correlation matrix is still (numerically) positive definite.

(6) A related investigation of interpolant limits has been performed in [18] but for standard radial basis functions.

In order to support readability, we divide the proof of Theorem 3.1 into smaller units, organized as follows. As a starting point, we establish two auxiliary lemmata on the existence of limits of eigenvalue quotients and of errors vectors. Subsequently, the proof of the main theorem is conducted relying on the lemmata.

Lemma 3.3. *In the setting of Theorem 3.1, let $\lambda_i(\theta),\ i = 1, \ldots, n$ be the eigenvalues of $R(\theta)$, ordered by size. Then,*

$$\lim_{\|\theta\| \to 0} \frac{\lambda_i(\theta)}{\lambda_j(\theta)} = \text{const.} > 0 \quad \forall i, j = 2, \ldots, n. \tag{3.4}$$

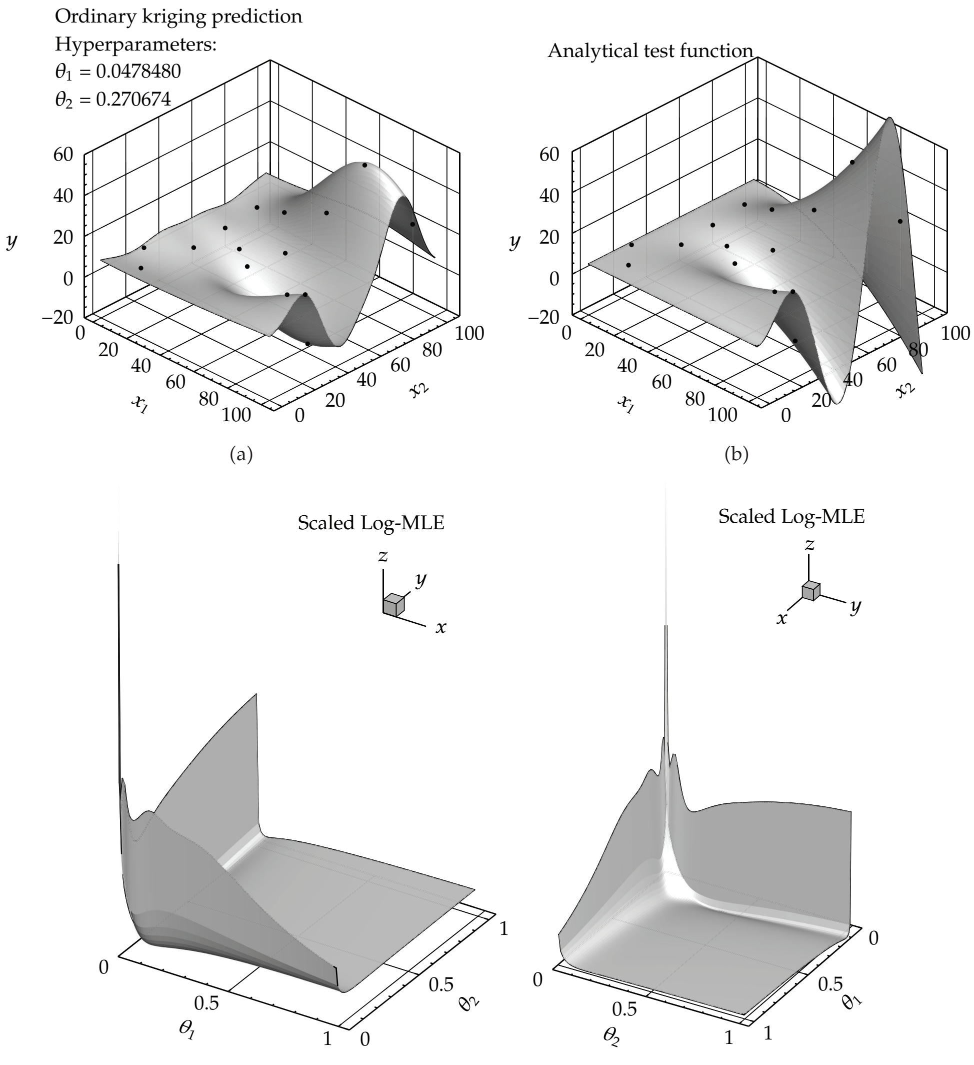
Proof. Because of (2.13), it holds that $(R(0))_{i,j} = \mathbf{1}\mathbf{1}^T \in \mathbb{R}^{n \times n}$ for every admissible spatial correlation function. Since $(\mathbf{1}\mathbf{1}^T)\mathbf{1} = n\mathbf{1} \in \mathbb{R}^n$ and $(\mathbf{1}\mathbf{1}^T)W = 0$ for all $W \in \mathbf{1}^{\perp} \subset \mathbb{R}^n$, the limit eigenvalues of the correlation matrix ordered by size are given by $\lambda_1(0) = n > 0 = \lambda_2(0) = \cdots = \lambda_n(0)$.

Under the present conditions, the eigenvalues λ_i are differentiable with respect to θ. Hence, it is sufficient to proof the lemma for $\mathbb{R} \ni \tau \mapsto \theta(\tau) = \tau\mathbf{1} \in \mathbb{R}^d$ and $\tau \to 0$. Now, condition 2 and L'Hospital's rule imply the result. □

Lemma 3.4. *In the setting of Theorem 3.1, let $\Sigma(\theta)$ be defined by* (2.18) *and* (2.19)*. Then,*

$$\lim_{\|\theta\| \to 0} \Sigma(\theta) = \Sigma(0) \tag{3.5}$$

exists.



(a)

(b)

(c)

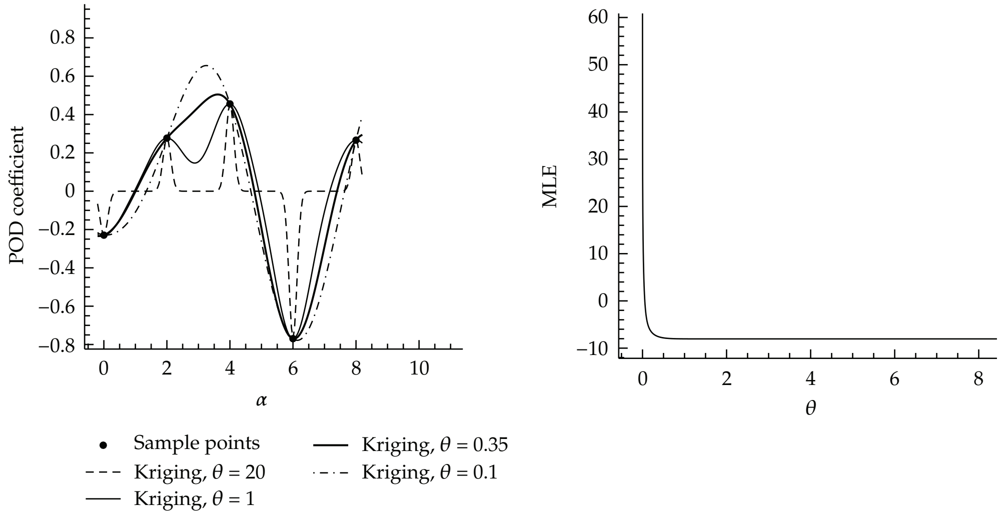
(d)

Figure 1: Ordinary Kriging estimation (a) of a two-dimensional analytical test function (b) based on 15 samples points. ((c) and (d)) Two views of the associated LogML, scaled by a factor of 1/100. This example shows that hyperparameter optima might lie very close to the blowup of the Log ML due to ill-conditioning, that is proved to occur by Theorem 3.1. Model function and sample locations are listed in Appendix B.

Proof. We prove Lemma 3.4 by showing that $\lim_{\|\theta\|\to 0}\beta(\theta)$ exists.

Remember that $\beta = (\beta_0, \ldots, \beta_p)^T \in \mathbb{R}^{p+1}$, with $p \in \mathbb{N}_0$ depending on the chosen regression model. As in the above lemma, we can restrict the considerations to the direction $\tau \mapsto \theta(\tau) = \tau\mathbf{1}$.

Let $\lambda_i(\theta),\ i = 1, \ldots, n$ be the eigenvalues of $R(\theta)$ ordered by size with corresponding eigenvector matrix $Q(\theta) = (X_1, \ldots, X_n)(\theta)$ such that $Q(\theta)R(\theta)Q^T(\theta) = \Lambda(\theta) = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. For brevity, define $X_i(\tau) := X_i(\theta(\tau)) = X_i(\tau\mathbf{1})$ and so forth.



(a)

(b)

Figure 2: Kriging of data stemming from reduced-order modeling of solutions to the Navier-Stokes equations via POD coefficient interpolation; see, for example, [23]. (a) Kriging predictors for different choices of the distance weight θ. (b) Corresponding condensed log-likelihood function, showing the limit behavior as predicted by Theorem 3.1. Numerically, the function features no local minimum. Thus, it is impossible to say which one of the estimator functions in the left picture is the most likely.

It holds that $X_1(0) = (1/\sqrt{n})\mathbf{1}$; see Lemma 3.3. Hence, $\langle \mathbf{1}, X_j(\tau)\rangle \to 0$ for $\tau \to 0$ and $j = 2, \dots, n$. In the present setting, the derivatives of eigenvalues and (normalized) eigenvectors exist and can be extended to 0; see, for example, [22]. By another application of L'Hospital's rule,

$$\lim_{\tau\to 0} \frac{\langle \mathbf{1}, X_j(\tau)\rangle}{\lambda_j(\tau)} \quad \text{exists for } j = 2, \dots, n. \tag{3.6}$$

Introducing

$$L(\tau) = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_2 \end{pmatrix}(\tau) \in \mathbb{R}^{(p+1)\times(p+1)}, \tag{3.7}$$

we can restate (2.19) as

$$\begin{aligned} \beta &= \left(F^T Q \Lambda^{-1} Q^T F\right)^{-1} \left(L^{-1} L\right) F^T Q \Lambda^{-1} Q^T Y \\ &= \left(L F^T Q \Lambda^{-1} Q^T F\right)^{-1} \left(L F^T Q \Lambda^{-1} Q^T\right) Y. \end{aligned} \tag{3.8}$$

It is sufficient to show that $\lim_{\tau\to 0}(LF^T Q \Lambda^{-1})(\tau)$ exists.

Writing columnwise $(F_0, F_1, \dots, F_p) := F$, a direct computation shows

$$\left(LF^TQ\Lambda^{-1}\right)(\tau) = \begin{pmatrix} \dfrac{\lambda_1}{\lambda_1}\langle F_0, X_1\rangle & \dfrac{\lambda_1}{\lambda_2}\langle F_0, X_2\rangle & \cdots & \dfrac{\lambda_1}{\lambda_n}\langle F_0, X_n\rangle \\ \dfrac{\lambda_2}{\lambda_1}\langle F_1, X_1\rangle & \dfrac{\lambda_2}{\lambda_2}\langle F_1, X_2\rangle & \cdots & \dfrac{\lambda_2}{\lambda_n}\langle F_1, X_n\rangle \\ \vdots & \vdots & & \vdots \\ \dfrac{\lambda_2}{\lambda_1}\langle F_p, X_1\rangle & \dfrac{\lambda_2}{\lambda_2}\langle F_p, X_2\rangle & \cdots & \dfrac{\lambda_2}{\lambda_n}\langle F_p, X_n\rangle \end{pmatrix}(\tau) \in \mathbb{R}^{(p+1)\times n}. \tag{3.9}$$

Note that $F_0 = \mathbf{1} = \sqrt{n}X_1(0)$ for the default choices of regression basis functions, such that $\langle F_0, X_1(0)\rangle \neq 0$ and $\langle F_0, X_j(0)\rangle = 0$ for $j = 2, \dots, n$. The desired result follows from (3.6) and Lemma 3.3.

Remark 3.5. Actually, one cannot prove for $(LF^TQ\Lambda^{-1})$ to be regular in general, since this matrix depends on the chosen sample locations. It might be possible to artificially choose samples such that, for example, F has not full rank. Yet if so, the whole Kriging exercise cannot be performed, since (2.19) is not well defined in this case. For constant regression, that is, $F = \mathbf{1}$, this is impossible. Note that F is independent of θ. □

Now, let us prove Theorem 3.1 using notation as introduced above.

Proof. As shown in the proof of Lemma 3.3

$$\operatorname{cond}(R(\theta)) = \frac{\lambda_{\max}(\theta)}{\lambda_{\min}(\theta)} \longrightarrow \infty \quad \text{for } \|\theta\| \longrightarrow 0. \tag{3.10}$$

Because the correlation matrix is symmetric and positive definite, a decomposition

$$R(\theta)^{-1} = Q(\theta)\Lambda(\theta)^{-1}Q^T(\theta), \tag{3.11}$$

with Q orthogonal and $\Lambda^{-1} = \operatorname{Diag}(1/\lambda_i)_{i=\{1,\dots,n\}}$, exists.

If necessary, renumber such that $\lambda_{\max} = \lambda_1 \geq \cdots \geq \lambda_n = \lambda_{\min}$. Let $W(\theta) = (W_1(\theta), \dots, W_n(\theta)) := Q(\theta)\Sigma(\theta)$. By Lemma 3.4, $W(0) := Q(0)^T\Sigma(0)$ exists. Condition 3 insures that $W_1(0) \neq 0$.

Case 1. Suppose that $W_i(0) \neq 0$ for all $i = 1, \dots, n$.

By continuity, $W_i(\theta) \neq 0$ for $0 \leq \|\theta\| \leq \varepsilon$ and $\varepsilon > 0$ small enough. Since $\{\theta \in \mathbb{R}^d,\ 0 \leq \|\theta\| \leq \varepsilon\}$ is a compact set,

$$W_m^2 := \min_{0\leq\|\theta\|\leq\varepsilon}\left\{W_i^2(\theta),\ i = 1, \dots, n\right\} \tag{3.12}$$

exists. Then, for $\|\theta\| \in [0, \epsilon]$,

$$\begin{aligned}
\mathrm{ML}(\theta) &= \frac{1}{n^n}\left(\Sigma^T(\theta)R(\theta)^{-1}\Sigma(\theta)\right)^n \det(R(\theta)) \\
&= \frac{1}{n^n}\left(\sum_i \frac{W_i^2(\theta)}{\lambda_i(\theta)}\right)^n \prod_i \lambda_i(\theta) \\
&\geq \left(\frac{W_m^2}{n}\right)^n \left(\frac{n-1}{\lambda_{\max}(\theta)} + \frac{1}{\lambda_{\min}(\theta)}\right)^n \lambda_{\min}^{n-1}(\theta)\lambda_{\max}(\theta) \\
&\geq \left(\frac{W_m^2}{n}\right)^n \left(\left(\frac{n-1}{\lambda_{\max}(\theta)}\right)^n + \left(\frac{1}{\lambda_{\min}(\theta)}\right)^n\right)\lambda_{\min}^{n-1}(\theta)\lambda_{\max}(\theta) \\
&\geq \left(\frac{W_m^2}{n}\right)^n \left(\frac{(n-1)^n}{\mathrm{cond}(R(\theta))^{n-1}} + \mathrm{cond}(R(\theta))\right) \\
&\geq \left(\frac{W_m^2}{n}\right)^n \mathrm{cond}(R(\theta)).
\end{aligned} \tag{3.13}$$

Case 2. Suppose that Case 1 does not hold true.

From $\Sigma(0) \notin \mathrm{span}\{\mathbf{1}\}$, it follows that

$$W(\theta) = Q(\theta)^T\Sigma(\theta) \notin \mathrm{span}\left\{(1,0,\ldots,0)^T\right\}, \tag{3.14}$$

for $\|\theta\|$ sufficiently small. Let $J := \{i \in \{1,\ldots,n\} \mid W_i(0) \neq 0\}$. Then, $n_J := |J| \geq 2$.

Define

$$W_m^2 := \min_{0\leq\|\theta\|\leq\varepsilon, j\in J}\left\{W_j^2(\theta)\right\}. \tag{3.15}$$

For the index $\tilde{m}$ defined by $\lambda_{\tilde{m}} := \min_{j\in J}\{\lambda_j\}$, it holds that $\tilde{m} \in \{2,\ldots,n\}$.

By Lemma 3.3,

$$L := \min_{0\leq\|\theta\|\leq\varepsilon}\left\{\frac{\lambda_{\min}(\theta)}{\lambda_{\tilde{m}}(\theta)}\right\} > 0 \tag{3.16}$$

exists. Using

$$\sum_{i=1}^n \frac{W_i^2(\theta)}{\lambda_i(\theta)} \geq \sum_{j\in J}\frac{W_j^2(\theta)}{\lambda_j(\theta)} \geq W_m^2\left(\frac{n_J-1}{\lambda_{\max}(\theta)} + \frac{1}{\lambda_{\tilde{m}}(\theta)}\right), \tag{3.17}$$

together with

$$\frac{\lambda_{\max}}{\lambda_{\tilde{m}}}(\theta)\left(\frac{\lambda_{\min}}{\lambda_{\tilde{m}}}\right)^{n-1}(\theta) = \frac{\lambda_{\max}}{\lambda_{\min}}(\theta)\left(\frac{\lambda_{\min}}{\lambda_{\tilde{m}}}\right)^{n}(\theta) \geq L^{n}\operatorname{cond}(R(\theta)), \tag{3.18}$$

the result can be established as in Case 1.

The estimate of the upper bound of ML is obtained in an analogous manner. Let

$$\begin{gathered} W_M^2 := \max_{0\leq\|\theta\|\leq\varepsilon}\left\{W_i^2(\theta) \mid W_i(\theta) \neq 0,\, i = 1,\dots,n\right\} > 0, \\ L_M := \max_{0\leq\|\theta\|\leq\varepsilon}\left\{\frac{\lambda_2(\theta)}{\lambda_{\min}(\theta)}\right\} > 0. \end{gathered} \tag{3.19}$$

Then,

$$\begin{aligned} \mathrm{ML}(\theta) &= \frac{1}{n^n}\left(\Sigma^T(\theta)R(\theta)^{-1}\Sigma(\theta)\right)^n \det(R(\theta)) = \frac{1}{n^n}\left(\sum_i \frac{W_i^2(\theta)}{\lambda_i(\theta)}\right)^n \prod_i \lambda_i(\theta) \\ &\leq \left(\frac{W_M^2}{n}\right)^n \left(\frac{n-1}{\lambda_{\min}(\theta)} + \frac{1}{\lambda_{\max}(\theta)}\right)^n \lambda_{\max}(\theta)\lambda_2^{n-1}(\theta) \\ &= \left(\frac{W_M^2}{n}\right)^n \left(\frac{n-1}{\lambda_{\min}(\theta)}\right)^n \left(1 + \frac{1}{n-1}\frac{\lambda_{\min}(\theta)}{\lambda_{\max}(\theta)}\right)^n \lambda_{\max}(\theta)\lambda_2^{n-1}(\theta) \\ &\overset{(*)}{\leq} \left(W_M^2\right)^n \left(1 - \frac{1}{n}\right)^n \operatorname{cond}(R(\theta)) \left(\frac{\lambda_2(\theta)}{\lambda_{\min}(\theta)}\right)^{n-1} \left(1 + \frac{n}{(n-1)\operatorname{cond}(R(\theta))}\right) \\ &\leq \frac{W_M^{2n}}{e} L_M^{n-1}(\operatorname{cond}(R(\theta)) + 2) \leq 3\frac{W_M^{2n}}{e} L_M^{n-1}\operatorname{cond}(R(\theta)), \end{aligned} \tag{3.20}$$

where we used Bernoulli's inequality at $(*)$, Lemma 3.3, and the fact that $n/(n-1) \leq 2$. □

Remark 3.6. The extension of the main theorem to Cokriging prediction [7, 20] is a straight-forward exercise, since the limit eigenvectors of the Cokriging correlation matrix corresponding to nonzero eigenvalues can also be determined explicitly.

4. Why Kriging Model Training Is Tricky: Part II

The following simple observation illustrates Kriging predictor behavior for large-distance weights θ. Notation is to be understood as introduced in Section 3.

Observation 1. Suppose that sample locations $\{x^1,\dots,x^n\} \subset \mathbb{R}^d$ and responses $y_i = y(x^i) \in \mathbb{R}$, $i = 1,\dots,n$ are given. Let $\hat{y}$ be the corresponding Kriging predictor according to (2.10). Then, for $\mathbb{R}^d \ni x \notin \{x^1,\dots,x^n\}$ and distance weights $\|\theta\| \longrightarrow \infty$, it holds that

$$\hat{y}(x) \longrightarrow f(x)\beta. \tag{4.1}$$

Table 1: Sampled sites corresponding to the example displayed in Figure 2.

$x = \alpha$:	0.0	2.0	4.0	6.0	8.0
$y(x)$:	−0.229334	0.277018	0.455534	−0.769558	0.26634

Put in simple words: if too large distance weights are chosen, then the resulting predictor function has the shape of the regression model, $x \mapsto f(x)\beta$, with peaks at the sample sites, compare Figure 2, dashed curve.

Proof. According to (2.10) it holds that

$$\widehat{y}(x) = f(x)\beta + c_{\theta}^{T}(x)C_{\theta}^{-1}(Y - F\beta), \tag{4.2}$$

where $C = \sigma^2 R$. By (2.13), it holds that $R(\theta) \to I$, for $\|\theta\| \to \infty$ for every admissible spatial correlation model of the form of (2.11). By Cauchy-Schwartz' inequality,

$$\left|c_{\theta}^{T}(x)C_{\theta}^{-1}(Y - F\beta)\right| = \left|\left\langle c_{\theta}(x), C_{\theta}^{-1}(Y - F\beta)\right\rangle\right| \leq \|c_{\theta}(x)\|\|C_{\theta}^{-1}(Y - F\beta)\|, \tag{4.3}$$

where $\|c_{\theta}(x)\| \to 0$ and $\|C_{\theta}^{-1}(Y - F\beta)\| \to$ const. for $\|\theta\| \to \infty$. □

Remark 4.1. The same predictor behavior arises at locations far away from the sampled sites, that is, for $\operatorname{dist}(x, \{x^1, \ldots, x^n\}) \to \infty$. This has to be considered, when *extrapolating* beyond the sample data set.

Figure 2 shows an example data set for which the Kriging maximum likelihood function is constant over a large range of θ values. This example was *not constructed artificially* but occured in the author's daily work of computing approximate fluid flow solutions based on proper orthogonal decomposition (POD) followed by coefficient interpolation as described in [23, 24].

The sample data set is given in Table 1. The Kriging estimator given by the dashed line shows a behavior as predicted by Observation 1. Note that from the model training point of view, all distance weights $\theta > 1$ are equally likely, yet lead to quite different predictor functions. Since the ML features no local minimum, classical hyperparameter estimation is impossible.

Appendices

A. On the Validity of Condition 2 in Theorem 3.1

The next lemma strongly indicates that the second condition in the main Theorem 3.1 is given in nondegenerate cases.

Lemma A.1. *Let $\mathbb{R}^d \ni \theta \mapsto R(\theta) \in \mathbb{R}^{n\times n}$ be the correlation matrix function corresponding to a given set of Kriging data and a fixed spatial correlation model.*

Let $\lambda_i(\theta)$, $i = 1, \ldots, n$ be the eigenvalues of R ordered by size with corresponding eigenvector matrix $Q = (X_1, \ldots, X_n)$, and define $\theta : \mathbb{R} \to \mathbb{R}^d, \tau \mapsto \theta(\tau) := \tau\mathbf{1}$. Suppose that the eigenvalues are mutually distinct for $\tau > 0$ close to zero.

Denote the directional derivative in the direction $\mathbf{1}$ *with respect to* τ *by a prime* $'$, *that is,* $(d/d\tau)R(\tau\mathbf{1}) = R'(\tau)$ *and so forth. Then, it holds that*

$$\lambda_i'(0) = X_i^T(0)R'(0)X_i. \tag{A.1}$$

If $R'(0)$ *is regular, then*

$$\lambda_i'(0) \neq 0, \tag{A.2}$$

for all $i = 1, \dots, n$, *with at most one possible exception.*

Proof. For every admissible spatial spatial correlation function $r(\theta, \cdot, \cdot)$ of the form (2.11) and $x \neq z \in \mathbb{R}^d$, it holds that

$$r(\theta(\tau), x, z) \longrightarrow 1, \quad \text{for } \tau \longrightarrow 0. \tag{A.3}$$

Thus, $(R(0))_{i,j} = \mathbf{1}\mathbf{1}^T \in \mathbb{R}^{n\times n}$.

It holds that $(\mathbf{1}\mathbf{1}^T)\mathbf{1} = n\mathbf{1} \in \mathbb{R}^n$ and $(\mathbf{1}\mathbf{1}^T)W = 0$ for all $W \in \mathbf{1}^\perp \subset \mathbb{R}^n$; therefore, the limits of the eigenvalues of the correlation matrix ordered by size are given by $\lambda_1(0) = n > \lambda_2(0) = \cdots = \lambda_n(0) = 0$. The assumption, that no multiple eigenvalues occur, ensures that the eigenvalues λ_i and corresponding (normalized, oriented) eigenvectors X_i are differentiable with respect to τ. Let $Q(\tau) = (X_1(\tau), \dots, X_n(\tau)) \in \mathbb{R}^{n\times n}$ be the (orthogonal) matrix of eigenvectors, such that

$$\Lambda(\tau) := \operatorname{diag}(\lambda_i(\tau))_1^n = Q^T(\tau)R(\tau)Q(\tau). \tag{A.4}$$

Then,

$$\begin{aligned}
\Lambda'(\tau) &= Q(\tau)^T R'(\tau)Q(\tau) + Q'(\tau)^T R'(\tau)Q(\tau) + Q(\tau)^T R(\tau)Q'(\tau), \\
\Lambda'(0) &= Q(0)^T R'(0)Q(0) \\
&\quad + \begin{pmatrix} 0 & \lambda_1(0)\langle X_2'(0), X_1(0)\rangle & \cdots & \lambda_1(0)\langle X_n'(0), X_1(0)\rangle \\ \lambda_1(0)\langle X_2'(0), X_1(0)\rangle & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ \lambda_1(0)\langle X_n'(0), X_1(0)\rangle & 0 & \cdots & 0 \end{pmatrix},
\end{aligned} \tag{A.5}$$

where $Q(0)$ is the continuous extension of $Q(\tau)$; see, for example, [22].

Hence,

$$\lambda_i'(0) = e_i^T \Lambda'(0) e_i = X_i^T(0)R'(0)X_i(0). \tag{A.6}$$

Table 2

Location	x_1	x_2	$y(x_1, x_2)$
1	84.0188	39.4383	−15.0146
2	78.3099	79.844	53.5481
3	91.1647	19.7551	20.0921
4	33.5223	76.823	13.506
5	27.7775	55.397	2.10686
6	47.7397	62.8871	5.07917
7	36.4784	62.8871	−1.23344
8	95.223	51.3401	26.5839
9	63.5712	71.7297	27.5219
10	14.1603	60.6969	4.74213
11	1.63006	24.2887	5.00422
12	13.7232	80.4177	6.48784
13	15.6679	40.0944	4.24907
14	12.979	10.8809	5.16235
15	99.8925	21.8257	22.8288

Let us assume, that there exist two indices $j_0, k_0, j_0 \neq k_0$ such that $\lambda'_{k_0}(0) = 0 = \lambda'_{j_0}(0)$. Let $W := (0, \dots, -\lambda_1 \langle X'_{j_0}(0), X_1(0)\rangle, \dots, -\lambda_1 \langle X_{k_{0'}}(0), X_1(0), \dots, 0)^T \in \mathbb{R}^n$. Then,

$$\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} = \Lambda'(0) W = Q(0)^T R'(0) Q(0) W, \tag{A.7}$$

contradicting the regularity of $R'(0)$, if $W \neq 0$.

If $W = 0$, replace W by $\widetilde{W} := (0, \dots, \overset{j_0}{1}, \dots, \overset{k_0}{1}, \dots, 0)$ and repeat the above argument. □

For most correlation models, the derivative $R'(0)$ can be computed explicitly.

B. Test Setting Corresponding to Figure 1

In order to produce Figure 1, the following test function has been applied:

$$y : [0, 100] \times [0, 100] \longrightarrow \mathbb{R}, \qquad (x_1, x_2) \longmapsto 5 + \frac{x_1^2 \cdot x_2}{10,000} \sin\left(\frac{x_2}{10}\right). \tag{B.1}$$

The Kriging predictor function displayed in this figure has been constructed based on the fifteen (randomly chosen) sample points shown in Table 2.

Nomenclature

$d \in \mathbb{N}$:	Dimension of parameter space
$n \in \mathbb{N}$:	(Fixed) number of sample points
$x^i \in \mathbb{R}^d$:	ith sample location
$y_i \in \mathbb{R}$:	Sample value at sample location x^i
$I \in \mathbb{R}^{n \times n}$:	Unit matrix
$\mathbf{1} := (1, \dots, 1)^T \in \mathbb{R}^n$:	Vector with all entries equal to 1
$e_i = (0, \dots, 0, \overset{i}{1}, 0, \dots, 0)^T$:	ith standard basis vector
$V^{\perp} \subset \mathbb{R}^n$:	Subspace of all vectors orthogonal to $V \in \mathbb{R}^n$
$R \in \mathbb{R}^{n \times n}$:	Correlation matrix
$C \in \mathbb{R}^{n \times n}$:	Covariance matrix
$\mathrm{cond}(R)$:	Condition number of $R \in \mathbb{R}^{n \times n}$
$\langle \cdot, \cdot \rangle$:	Euclidean scalar product
$e = \exp(1)$:	Euler's number
$p + 1 \in \mathbb{N}$:	Dimension of regression model
$\beta \in \mathbb{R}^{p+1}$:	Vector of regression coefficients
$f : \mathbb{R}^d \to \mathbb{R}^{p+1}$:	Regression model
$\epsilon : \mathbb{R}^d \to \mathbb{R}$:	Random error function
$E[\cdot]$:	Expectation value
$\sigma = \sqrt{\mathrm{Var}}$:	Standard deviation
$\theta \in \mathbb{R}^d$:	Distance weights vector, model hyperparameters.

Acknowledgments

This research was partly sponsored by the European Regional Development Fund and Economic Development Fund of the Federal German State of Lower Saxony Contract/Grant no. W3-80026826.

References

[1] Z. H. Han, S. Gortz, and R. Zimmermann, "On improving efficiency and accuracy of variable-fidelity surrogate modeling in aero-data for loads context," in *Proceeings of European Air and Space Conference (CEAS '09)*, Manchester, UK, October 2009.

[2] M. C. Kennedy and A. O'Hagan, "Predicting the output from a complex computer code when fast approximations are available," *Biometrika*, vol. 87, no. 1, pp. 1–13, 2000.

[3] J. Laurenceau and P. Sagaut, "Building efficient response surfaces of aerodynamic functions with kriging and cokriging," *AIAA Journal*, vol. 46, no. 2, pp. 498–507, 2008.

[4] J. Sacks, J. Welch, T. J. Mitchell, and H. Wynn, "Design and analysis of computer experiments," *Statistical Science*, vol. 4, no. 4, pp. 409–423, 1989.

[5] R. Zimmermann and Z. H. Han, "Simplified cross-correlation estimation for multi-fidelity surrogate cokriging models," *Advances and Applications in Mathematical Sciences*, vol. 7, no. 2, pp. 181–202, 2010.

[6] D. Krige, "A statistical approach to some basic mine valuation problems on the Witwa-tersrand," *Journal of the Chemical, Metallurgical and Mining Engineering Society of South Africa*, vol. 52, no. 6, pp. 119–139, 1951.

[7] A. G. Journel and C. J. Huijbregts, *Mining Geostatistics*, The Blackburn Press, Caldwell, NJ, USA, 5th edition, 1991.

[8] G. Matheron, "Principles of geostatistics," *Economic Geology*, vol. 58, pp. 1246–1266, 1963.

[9] J. J. Warnes and B. D. Ripley, "Problems with likelihood estimation of covariance functions of spatial Gaussian processes," *Biometrika*, vol. 74, no. 3, pp. 640–642, 1987.

[10] K. V. Mardia and A. J. Watkins, "On multimodality of the likelihood in the spatial linear model," *Biometrika*, vol. 76, no. 2, pp. 289–295, 1989.

[11] R. Ababou, A. C. Bagtzoglou, and E. F. Wood, "On the condition number of covariance matrices in kriging, estimation, and simulation of random fields," *Mathematical Geology*, vol. 26, no. 1, pp. 99–133, 1994.

[12] P. Diamond and M. Armstrong, "Robustness of variograms and conditioning of kriging matrices," *Journal of the International Association for Mathematical Geology*, vol. 16, no. 8, pp. 809–822, 1984.

[13] D. Posa, "Conditioning of the stationary kriging matrices for some well-known covariance models," *Mathematical Geology*, vol. 21, no. 7, pp. 755–765, 1989.

[14] G. J. Davis and M. D. Morris, "Six factors which affect the condition number of matrices associated with kriging," *Mathematical Geology*, vol. 29, no. 5, pp. 669–683, 1997.

[15] K. Schöttle and R. Werner, "Improving the most general methodology to create a valid correlation matrix," *Management Information Systems*, vol. 9, pp. 701–710, 2004.

[16] Z. Ying, "Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process," *Journal of Multivariate Analysis*, vol. 36, no. 2, pp. 280–296, 1991.

[17] H. Zhang and D. L. Zimmerman, "Towards reconciling two asymptotic frameworks in spatial statistics," *Biometrika*, vol. 92, no. 4, pp. 921–936, 2005.

[18] M. D. Buhmann, S. Dinew, and E. Larsson, "A note on radial basis function interpolant limits," *IMA Journal of Numerical Analysis*, vol. 30, no. 2, pp. 543–554, 2010.

[19] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, Mass, USA, 2006.

[20] T. J. Santner, B. J. Williams, and W. I. Notz, *The Design and Analysis of Computer Experiments*, Springer, New York, NY, USA, 2003.

[21] S. Lophaven, H. B. Nielsen, and J. Søndergaard, "DACE—a MATLAB kriging tool-box, version 2.0," Tech. Rep. IMM-TR-2002-12, Technical University of Denmark, 2002.

[22] N. P. van der Aa, H. G. Ter Morsche, and R. R. M. Mattheij, "Computation of eigenvalue and eigenvector derivatives for a general complex-valued eigensystem," *Electronic Journal of Linear Algebra*, vol. 16, pp. 300–314, 2007.

[23] R. Zimmermann and S. Gortz, "Non-linear reduced order models for steady aerodynamics," *Procedia Computer Sciences*, vol. 1, no. 1, pp. 165–174, 2010.

[24] T. Bui-Thanh, M. Damadoran, and K. Willcox, "Proper orthogonal decomposition extensions for parametric applications in transonic aerodynamics," in *Proceedings of the 21th AIAA Applied Aerodynamics Conference*, Orlando Fla, USA, 2003, AIAA paper 2003-4213.