Stable Likelihood Computation for Gaussian Random Fields

Michael McCourt and Gregory E. Fasshauer

Abstract Given scattered data realized from a Gaussian random field, unobserved values of the field can be predicted through kriging. To do so accurately, the covariance of the Gaussian random field must be known or, more commonly, estimated from the data. Maximum likelihood estimation of the parameters of a family of candidates for the covariance kernel is one such strategy, but evaluating the likelihood function can be ill-conditioned for certain covariance kernels. In order to stably approximate the likelihood function we leverage the Hilbert–Schmidt SVD, a decomposition of the covariance matrix constructed directly from the eigenvalues and eigenfunctions of the Hilbert–Schmidt integral operator associated with the covariance kernel. We illustrate the effectiveness of this tool, which was previously used for positive definite kernel interpolation, with some numerical experiments. We also draw connections to numerical analysis, where one might be more interested in minimizing errors or error bounds, and introduce two further criteria that can be used for parameter estimation: one based on minimizing the kriging variance (which is closely related to the power function used in numerical analysis), and the other involving the determinant of an augmented matrix.

1 Introduction

Gaussian random fields (see Section 2 for more details) provide useful models for interpolating scattered data [6, 32], design of computer experiments [27, 28], surrogate or response surface modeling [8, 13], as well as statistical or machine learning [25, 33]. There is also a development of related numerical methods based on positive definite kernels—after all, the covariance kernel of a Gaussian random field

Michael McCourt (e-mail: mccourt@sigopt.com)
SigOpt, 244 Kearny St, San Francisco, CA, 94104

Gregory E. Fasshauer (e-mail: fasshauer@iit.edu)
Department of Applied Mathematics, Illinois Institute of Technology, Chicago, IL

is just that. These kernel-based numerical methods are applied to similar problems, but also to the numerical solution of partial differential equations [5]. For an exposition that illuminates both the stochastic and deterministic perspective of these kernel-based methods we refer the reader to [9, 30].

As the references just mentioned indicate, these kernel-based methods have been around for about three decades now and many researchers have experimented with them for their specific applications—some with more success than others. When things do not turn out as expected, the most common sources of frustration for these users have been (1) the fact that many kernel-based methods tend to suffer from numerical instability, (2) the presence of one or more free parameters in the definition of many popular kernels, and (3) the high computational cost often associated with the use of kernel-based methods. In this paper we will address mostly item (2), but in doing so we will also draw upon recent advances regarding item (1). There are other exciting developments currently under way—especially in the area of numerical methods for solving PDEs [12]—that are based on localized (finite difference-like) approximations. These methods actually address all three concerns just listed, but these local methods do not converge as rapidly (for problems with sufficiently smooth solutions) as the more commonly used global methods discussed here which give rise to concerns (1)–(3).

In this paper we will focus on *parameter estimation* in the context of the *scattered data fitting problem*. This model problem is appropriate for most of the applications mentioned above and can be viewed from a deterministic or stochastic perspective. For this problem we are given locations $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \Omega \subseteq \mathbb{R}^d$ (frequently referred to as the *design*) with associated scalar[1] values $\mathbf{y} = \begin{pmatrix} y_1 & \cdots & y_N \end{pmatrix}^T \in \mathbb{R}^N$ (usually referred to as the *data*). In the stochastic setting, we interpret the given data as a realization of the vector of random variables $\mathbf{Y} = \begin{pmatrix} Y_{\mathbf{x}_1} & \cdots & Y_{\mathbf{x}_N} \end{pmatrix}^T$. Here $Y_{\mathbf{x}_i}$ denotes the random variable associated with the values taken by the (unknown) Gaussian random field $Y = \{Y_\mathbf{x}\}_{\mathbf{x} \in \Omega}$ at the point $\mathbf{x}_i \in \Omega$. In the deterministic setting, the data is viewed as samples of an (unknown) function f.

In both settings, we make a connection to a specific, albeit unknown, positive definite kernel K. In the stochastic setting, K is the covariance kernel of the Gaussian random field Y, i.e., the vector of random variables $\mathbf{Y}$ follows a multivariate normal distribution, $\mathbf{Y} \sim \mathcal{N}(\mu, \sigma^2 \mathsf{K})$, with mean vector $\mu = \mathbb{E}[\mathbf{Y}]$ and covariance matrix $\sigma^2 \mathsf{K} = \sigma^2 \left(K(\mathbf{x}_i, \mathbf{x}_j) \right)_{i,j=1}^N$. We provide more details on Gaussian random fields and motivate our explicit use of the process variance σ^2 in Section 2 below.

In the deterministic setting the connection to K appears via the function space in which the data function f "lives". We assume that this function space is a reproducing kernel Hilbert space $\mathcal{H}_K(\Omega)$ with reproducing kernel K. A specific choice of K prescribes the covariance structure of the random field in the stochastic interpretation, and it prescribes the smoothness and the inner product (and, therefore, norm)

[1] More general types of data—such as vector-valued, or even (continuous) function-valued data—have also been investigated in the context of approximation theory [14, 23], geostatistics [20] and machine learning [16, 21]. These problems necessitate the use of matrix-valued or operator-valued kernels for which the concerns addressed in this paper also apply. In the interest of keeping our discussion transparent we limit ourselves to the scalar-valued case.

of the Hilbert function space in the deterministic setting (see Section 5.1 for more details).

Since everything hinges upon our choice of the kernel K—but this kernel usually is not known—it is common to consider a *parametrized family of kernels*. Such a family may be parametrized by one or more parameters which then need to be estimated from the given data. Some common kernel families include [9]

$$\begin{aligned}\text{Gaussians (squared exponentials):}\quad & K(\mathbf{x},\mathbf{z}) = \mathrm{e}^{-\varepsilon^2 r^2},\\ \text{generalized multiquadratics:}\quad & K(\mathbf{x},\mathbf{z}) = (1+\varepsilon^2 r^2)^\beta,\ \beta \in \mathbb{R}\setminus\mathbb{N}_0, \\ \text{Matérn kernels:}\quad & K(\mathbf{x},\mathbf{z}) = \frac{\bar{K}_{\beta-d/2}(\varepsilon r)(\varepsilon r)^{\beta-d/2}}{2^{\beta-1}\Gamma(\beta)},\ \beta > d/2.\end{aligned} \tag{1}$$

Here $r = \|\mathbf{x}-\mathbf{z}\|_2$, $\bar{K}_{\beta-d/2}$ is a modified Bessel function of the second kind, β is a smoothness parameter, and ε is a positive shape parameter that determines the locality/scale of K. Other possibilities include d-dimensional tensor product kernels made up of products of one-dimensional kernels, with possibly a different set of parameters (or even a different kernel) associated with each space dimension. Furthermore, the kernels need not be radial kernels. For example, one could use (tensor products of) the univariate periodic spline kernels, iterated Brownian bridge (IBB) kernels, or Chebyshev kernels [9]:

$$\begin{aligned}\text{periodic spline:}\quad K(x,z) &= \frac{(-1)^{\beta-1}}{(2\beta)!}B_{2\beta}(|x-z|),\ \beta\in\mathbb{N},\\ &= \sum_{n=1}^{\infty}\frac{2}{(2n\pi)^{2\beta}}\cos\left(2n\pi(x-z)\right),\\ \text{IBB:}\quad K(x,z) &= \sum_{n=1}^{\infty}\frac{2}{(n^2\pi^2+\varepsilon^2)^\beta}\sin(n\pi x)\sin(n\pi z),\ \beta\in\mathbb{N},\\ \text{Chebyshev:}\quad K(x,z) &= 1-a+2a(1-b)\frac{b(1-b^2)-2b(x^2+z^2)+(1+3b^2)xz}{(1-b^2)^2+4b\left(b(x^2+z^2)-(1+b^2)xz\right)},\\ &= 1-a+2a(1-b)\sum_{n=1}^{\infty}b^{n-1}T_n(x)T_n(z),\ a\in(0,1],\ b\in(0,1).\end{aligned} \tag{2}$$

Here $B_{2\beta}$ are Bernoulli polynomials of degree 2β and T_n are Chebyshev polynomials of degree n. The parameter b also acts like a shape parameter, causing more localization for $b\to 1$ and ill-conditioning for $b\to 0$. The parameter a is not that significant as long as $a\in(0,1)$ as it just shifts and scales the kernel vertically. However, setting $a=1$ completely eliminates the vertical shift and therefore makes it markedly more difficult to fit data with a nonzero mean.

The presence of ε (and potentially other free parameters such as β or b) allows for flexibility to choose a kernel supported by the data without having to explore the endless selection of *all* positive definite kernels. Unfortunately, this flexibility is often accompanied by the danger of severe ill-conditioning for small ε because of the increasing linear dependence of $K(\cdot,\mathbf{x}_i)$ and $K(\cdot,\mathbf{x}_j)$ even when $i \neq j$.

In Section 3 we describe the Hilbert–Schmidt SVD, a strategy developed recently to avoid this ill-conditioning. In Section 4 we discuss the use of maximum likelihood estimation to choose optimal kernel parameters for prediction, and how the unstable likelihood function can be stably approximated using the Hilbert–Schmidt SVD. In Section 5 we introduce the kriging variance as another viable parametrization criterion along with a third criterion which combines the kriging variance with the maximum likelihood criterion. All of the criteria discussed in this paper are summarized in Section 6 and the effectiveness of the Hilbert–Schmidt SVD as a tool to stabilize all of these parametrization strategies is demonstrated in the context of numerical experiments in Section 7.

Other possible approaches to dealing with ill-conditioning of the linear system associated with kernel-based approximation include (possibly iterated) Tikhonov regularization [24], alternate bases such as those for polyharmonic splines of Beatson, Billings and Light [2], the Newton bases of Müller and Schaback [22], or the (weighted) SVD-bases of De Marchi and Santin [7]. Each of these methods comes with its own list of advantages and disadvantages. To our knowledge, the accuracy of the parameter estimation results we report in Section 7 has not been achieved with any other method. However, our results are limited to those special cases for which a Hilbert–Schmidt SVD is available.

2 Gaussian Random Fields and Kriging

2.1 Gaussian Random Fields

We begin by defining a probability space $(\mathcal{W},\mathcal{A},P)$, where $\mathcal{W}$ is the sample space of all possible outcomes, $\mathcal{A}$ is a set of subsets of $\mathcal{W}$ containing all the events, and P is a probability measure. We also denote by Ω the parameter space, which for our purposes will simply be $\Omega = \mathbb{R}^d$. This means that the observations from which we want to predict come from $\mathbb{R}^d$.

A function $Y : \Omega \times \mathcal{W} \to \mathbb{R}$ (evaluated as $Y(\mathbf{x},\omega)$ for $\mathbf{x} \in \Omega$ and $\omega \in \mathcal{W}$) is a *random field* if, for every $\mathbf{x} \in \Omega$, Y is an $\mathcal{A}$-measurable function of ω. Our notation for this is $Y = \{Y_\mathbf{x}\}_{\mathbf{x}\in\Omega}$. Note that, for a fixed $\mathbf{x}$, $Y(\mathbf{x},\cdot) = Y_\mathbf{x}$ is a random variable, while for a fixed ω, $Y(\cdot,\omega) = y(\cdot)$ is a deterministic function of $\mathbf{x}$ referred to as a realization of the random field.

As already mentioned in Section 1, Gaussian random fields are quite popular in situations that involve the modeling of natural phenomena based on given data and one of their especially attractive features is that they are relatively easy to work with. In particular, a *Gaussian random field* is completely characterized by its first two moments, namely its mean $\mathbb{E}[Y_\mathbf{x}]$ and its covariance $\mathrm{Cov}(Y_\mathbf{x},Y_\mathbf{z}) = \sigma^2 K(\mathbf{x},\mathbf{z})$.

The mean of Y is a function μ which is defined at any point $\mathbf{x} \in \Omega$ as

$$\mu(\mathbf{x}) = \mathbb{E}[Y_{\mathbf{x}}] = \int_{\mathcal{W}} Y_{\mathbf{x}}(\omega)\,\mathrm{d}P(\omega) = \int_{\mathbb{R}} y\,\mathrm{d}F_{Y_{\mathbf{x}}}(y),$$

where $F_{Y_{\mathbf{x}}}$ is the cumulative distribution function of $Y_{\mathbf{x}}$ with respect to P. For our purposes we will assume that $Y_{\mathbf{x}}$ is continuous so that we may write

$$\mu(\mathbf{x}) = \int_{\mathbb{R}} y p_{Y_{\mathbf{x}}}(y)\mathrm{d}y$$

with density function $p_{Y_{\mathbf{x}}}$. Likewise, note that the covariance kernel K of Y satisfies

$$\sigma^2 K(\mathbf{x},\mathbf{z}) = \mathrm{Cov}(Y_{\mathbf{x}},Y_{\mathbf{z}}) = \mathbb{E}[Y_{\mathbf{x}}Y_{\mathbf{z}}] - \mu(\mathbf{x})\mu(\mathbf{z}). \tag{3}$$

Remark 1. Here the scalar parameter σ^2 is known as the *process variance* and in the statistics literature this is often included in the definition of the covariance kernel, so that, e.g., the Gaussian covariance would be, $K(\mathbf{x},\mathbf{z}) = \sigma^2 \mathrm{e}^{-\varepsilon^2\|\mathbf{x}-\mathbf{z}\|^2}$. In the approximation theory setting such an amplification factor is generally irrelevant and therefore—coming from that community—we prefer to define the Gaussian in the form $K(\mathbf{x},\mathbf{z}) = \mathrm{e}^{-\varepsilon^2\|\mathbf{x}-\mathbf{z}\|^2}$ as in (1). Having the process variance explicitly appear in our formulas will allow us to better illuminate the connection between the concepts of kriging variance (from statistics) and power function (from approximation theory) and therefore deal as precisely as possible with concepts of accuracy and error.

For the remainder of this paper we simplify the situation and assume the data was generated by a *zero-mean* Gaussian process, i.e., $\mu \equiv 0$, although a non-zero mean can also be considered. The density of the (zero-mean) multivariate normal random variable $\mathbf{Y}$ is then given by

$$p_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{\sqrt{(2\pi\sigma^2)^N \det(\mathsf{K})}} \exp\left(-\frac{1}{2\sigma^2}\mathbf{y}^T\mathsf{K}^{-1}\mathbf{y}\right), \tag{4}$$

where K is a symmetric positive semi-definite matrix since K is a positive definite (covariance) kernel. Thus, the inverse K^{-1} will exist whenever K has no zero eigenvalues. If zero eigenvalues do arise one can replace the inverse of K with its pseudoinverse (see, e.g., [18, Section 2.5.4]). The main challenge, however, is not whether K is invertible or not. Even if K is invertible it may still be numerically ill-conditioned, and we address this challenge in Section 3.

2.2 Simple Kriging

There are several different ways to arrive at the (simple) kriging predictor for $Y_{\mathbf{x}_0}$, the value of the Gaussian random field at a previously unobserved location $\mathbf{x}_0$ (see, e.g., [9, Chapter 5]). Following the *Bayesian approach*, one conditions the unobserved data at $\mathbf{x}_0$ on the observed data at all locations in $\mathcal{X} = \{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$ and, using the vector notation

$$\mathbf{k}(\mathbf{x}_0)^T = \big(K(\mathbf{x}_0,\mathbf{x}_1) \cdots K(\mathbf{x}_0,\mathbf{x}_N)\big), \tag{5}$$

one obtains

$$Y_{\mathbf{x}_0}|\mathbf{Y}=\mathbf{y} \sim \mathcal{N}\left(\, \mathbf{k}(\mathbf{x}_0)^T \mathsf{K}^{-1}\mathbf{y}\,,\, \sigma^2(K(\mathbf{x}_0,\mathbf{x}_0)-\mathbf{k}(\mathbf{x}_0)^T\mathsf{K}^{-1}\mathbf{k}(\mathbf{x}_0))\,\right), \tag{6}$$

where the posterior mean $\mathbb{E}[Y_{\mathbf{x}_0}|\mathbf{Y}=\mathbf{y}] = \mathbf{k}(\mathbf{x}_0)^T\mathsf{K}^{-1}\mathbf{y}$ is known as the *kriging prediction*. The variance

$$\mathrm{Var}(Y_{\mathbf{x}_0}|\mathbf{Y}=\mathbf{y}) = \sigma^2(K(\mathbf{x}_0,\mathbf{x}_0)-\mathbf{k}(\mathbf{x}_0)^T\mathsf{K}^{-1}\mathbf{k}(\mathbf{x}_0)) \tag{7}$$

associated with this predictor is known as the (simple) *kriging variance* and it corresponds to the minimal mean-squared error of the kriging predictor (assuming this predictor is linear). This explains the fact that the kriging prediction represents the *best linear unbiased prediction* for $Y_{\mathbf{x}_0}$ [32].

Remark 2. The reader should note that the preceding discussion—and in fact everything in this paper—assumes the data to be *noiseless*. In the presence of noise, one usually applies an additional form of regularization (such as smoothing splines or ridge regression (see, e.g., [9, Chapter 15]) and the kriging variance ends up having a more complicated form and interpretation.

3 Hilbert–Schmidt SVD

3.1 Basic Review of the Hilbert–Schmidt SVD

Positive definite kernels have an expansion in terms of their Mercer series

$$K(\mathbf{x},\mathbf{z}) = \sum_{n=1}^{\infty} \lambda_n \varphi_n(\mathbf{x})\varphi_n(\mathbf{z}),$$

where λ_n and φ_n, $n=1,2,\ldots$, are the Hilbert–Schmidt eigenvalues and eigenfunctions respectively. Three examples of positive definite kernels and their Mercer series are listed in (2). The Mercer series for the Gaussian kernel is also known (see, e.g., [9, Example 12.1] or [25, Section 4.3.1]) and it is used in later parts of this paper. As discussed in [11], the rapid decay of these eigenvalues for high smoothness kernels such as the Gaussian is the main cause of ill-conditioning in the covariance matrix K.

In [4], the authors described the vector $\mathbf{k}$ appearing in (5) using these eigenvalues and eigenfunctions,

$$\mathbf{k}(\mathbf{x})^T = \phi(\mathbf{x})^T \Lambda \Phi^T, \tag{8}$$

where $\phi(\mathbf{x})^T = \begin{pmatrix}\varphi_1(\mathbf{x}) \cdots \varphi_N(\mathbf{x}) \cdots\end{pmatrix}$ is an infinite length vector (because there are infinitely many eigenfunctions) and

$$\Lambda = \begin{pmatrix}\Lambda_1 & \\ & \Lambda_2\end{pmatrix}, \qquad \Phi = \begin{pmatrix}\phi(\mathbf{x}_1)^T \\ \vdots \\ \phi(\mathbf{x}_N)^T\end{pmatrix} = \begin{pmatrix}\Phi_1 & \Phi_2\end{pmatrix},$$

such that $\Lambda_1, \Phi_1 \in \mathbb{R}^{N\times N}$ and Λ_2 and Φ_2 are the (infinite-sized) remainders of the matrices Λ and Φ, respectively. We mention here for use in Section 4 that the eigenvalues appear in non-increasing order and that, for Gaussians, the magnitude of the smallest eigenvalue in Λ_1, λ_N, is an order of ε^2 larger than λ_{N+1}, the largest eigenvalue in Λ_2. This is true for any N, although the design of the Hilbert–Schmidt SVD in dimension $d > 1$ is too complicated for this article. It is discussed in [11] in the context of the Gaussian kernel.

Manipulations to the $\Lambda\Phi^T$ term in (8) reveal that

$$\Lambda\Phi^T = \begin{pmatrix}\Lambda_1 & \\ & \Lambda_2\end{pmatrix}\begin{pmatrix}\Phi_1^T \\ \Phi_2^T\end{pmatrix} = \begin{pmatrix}\mathsf{I}_N \\ \Lambda_2\Phi_2^T\Phi_1^{-T}\Lambda_1^{-1}\end{pmatrix}\Lambda_1\Phi_1^T,$$

which provides a way to express the vector $\mathbf{k}$ of standard basis functions in terms of a *stable basis* ψ via (8),

$$\mathbf{k}(\mathbf{x})^T = \underbrace{\phi(\mathbf{x})^T\begin{pmatrix}\mathsf{I}_N \\ \Lambda_2\Phi_2^T\Phi_1^{-T}\Lambda_1^{-1}\end{pmatrix}}_{\psi(\mathbf{x})^T}\Lambda_1\Phi_1^T = \psi(\mathbf{x})^T\Lambda_1\Phi_1^T. \tag{9}$$

The term stable basis is used because (9) isolates the swiftly decaying eigenvalues in Λ_1 which are the main source of ill-conditioning in the standard basis. Applying the same idea to K (which consists of rows of $\mathbf{k}$ evaluated at all the locations in $\mathcal{X}$) yields the *Hilbert–Schmidt SVD* (HS-SVD)

$$\mathsf{K} = \Psi\Lambda_1\Phi_1^T, \tag{10}$$

a matrix factorization of the covariance matrix K. In contrast to standard matrix decompositions which start with a matrix and produce the resulting factors, the HS-SVD is constructed from the Hilbert–Schmidt eigenvalues and eigenvectors—without the need to ever form the potentially ill-conditioned matrix K.

To demonstrate the usefulness of the HS-SVD, we write the kriging prediction (6) using (9) and (10):

$$\begin{aligned}\mathbb{E}[Y_{\mathbf{x}_0}|\mathbf{Y}=\mathbf{y}] &= \mathbf{k}(\mathbf{x}_0)^T\mathsf{K}^{-1}\mathbf{y} \\ &= \psi(\mathbf{x}_0)^T\Lambda_1\Phi_1^T(\Psi\Lambda_1\Phi_1^T)^{-1}\mathbf{y} \\ &= \psi(\mathbf{x}_0)^T\Psi^{-1}\mathbf{y}. \end{aligned}\tag{11}$$

Now, the ill-conditioning due to the dangerous Λ_1^{-1} term, introduced by applying $\mathsf{K}^{-1} = \Phi_1^{-T}\Lambda_1^{-1}\Psi^{-1}$, is removed analytically through the Λ_1 term present in $\mathbf{k}$.

3.2 Hilbert–Schmidt SVD for tensor product kernels

Determining, analytically, the Mercer series of a positive definite kernel is not trivial, and as of this writing we have only knowledge of a select few [9]. Tensor product kernels (also sometimes referred to as simply product kernels) are a form of positive definite kernel constructed by multiplying two or more positive definite kernels. Most often, the goal of this is to dissect a high-dimensional setting into one-dimensional kernels,

$$K(\mathbf{x},\mathbf{z}) = K_1(x_1,z_1)\cdots K_d(x_d,z_d), \tag{12}$$

although different structures of tensor product kernels are also viable. If the Mercer series of each of the component kernels $K_1,\dots,K_d$ is known, then the Hilbert–Schmidt SVD of the tensor product K in (12) can be determined. This section will work through this derivation for a product of two kernels, for which the notation is already complicated, but the same mechanism can be extended to arbitrarily many component kernels.

Define the Mercer series

$$K_1(x,z) = \sum_{m=1}^{\infty} \lambda_m \varphi_m(x)\varphi_m(z) = \phi(x)^T \Lambda \phi(z),$$

and

$$K_2(x,z) = \sum_{m=1}^{\infty} \xi_m \nu_m(x)\nu_m(z) = \mathbf{v}(x)^T \Xi \mathbf{v}(z),$$

which in turn defines the tensor product kernel

$$\begin{aligned}
K(\mathbf{x},\mathbf{z}) &= K_1(x_1,z_1)K_2(x_2,z_2)\\
&= \phi(x_1)^T\Lambda\phi(z_1)\,\mathbf{v}(x_2)^T\Xi\mathbf{v}(z_2)\\
&= \lambda_1\xi_1\varphi_1(x_1)\varphi_1(z_1)\nu_1(x_2)\nu_1(z_2)\\
&\quad + \lambda_1\xi_2\varphi_1(x_1)\varphi_1(z_1)\nu_2(x_2)\nu_2(z_2) + \lambda_2\xi_1\varphi_2(x_1)\varphi_2(z_1)\nu_1(x_2)\nu_1(z_2)\\
&\quad + \dots .
\end{aligned}$$

It may be preferable to write this as

$$K(\mathbf{x},\mathbf{z}) = (\phi(x_1)\otimes\mathbf{v}(x_2))^T(\Lambda\otimes\Xi)(\phi(z_1)\otimes\mathbf{v}(z_2)) \tag{13}$$

to more explicitly recognize the fact that, given the Mercer series of the component kernels, we know the Mercer series of a tensor product kernel.

Of course, using this tensor product runs counter to the standard strategy of sorting the eigenvalues in nonincreasing order. In order to recover that ordering we define the matrix P to be a (infinite) permutation matrix such that the diagonal of $\tilde{\Lambda} = \mathsf{P}(\Lambda \otimes \Xi)\mathsf{P}^T$ is in nonincreasing order. Then, defining $\tilde{\phi}(\mathbf{x}) = \mathsf{P}(\phi(x_1) \otimes \mathbf{v}(x_2))$, gives the more standard form of the Mercer series,

$$K(\mathbf{x},\mathbf{z}) = (\phi(x_1) \otimes \mathbf{v}(x_2))^T \mathsf{P}^T \mathsf{P}(\Lambda \otimes \Xi)\mathsf{P}^T \mathsf{P}(\phi(z_1) \otimes \mathbf{v}(z_2)) = \tilde{\phi}(\mathbf{x})^T \tilde{\Lambda} \tilde{\phi}(\mathbf{z}).$$

In summary, our ability to apply the Hilbert–Schmidt SVD change of basis to tensor product kernels relies on two points:

1. knowing the Mercer series of the component kernels, and
2. determining an appropriate permutation matrix P.

The former requirement is infinitely harder than the latter, since, if the eigenvalues are known, sorting the products of eigenvalues that form $\Lambda \otimes \Xi$ to determine P is trivial.

Remark 3. The strategy for ordering the products of eigenvalues (and their corresponding eigenfunctions) of equal magnitude—and thus the choice of permutation matrix P—is not unique. This should be done in a way that maximizes the rank of the $N \times N$ eigenfunction matrix Φ_1. The paper [11] and [9, Section 19.4.3] contain some discussion of this issue.

4 Maximum Likelihood Estimation with the HS-SVD

As mentioned in Section 1, a standard strategy for performing prediction involves choosing a preferred family of kernels to serve as the covariance of the random field Y and then, given the data, parametrizing them optimally. The kernels in (1) or (2) have one or two parameters: the shape parameter ε (or b) and the smoothness parameter β, but arbitrarily many are possible. Given that, presence of the parameter ε alone provides a sufficient challenge because small ε will cause the matrix K in (6) to become ill-conditioned. In (11) we showed how this ill-conditioning can be overcome for predictions, and in this section, we show how to overcome this for maximum likelihood estimation of the kernel parameter ε.

4.1 The Likelihood and Profile Likelihood Criteria

The *likelihood* function is related to the probability that a specific ε generated the observed data $\{\mathbf{x}_i, y_i\}_{i=1}^N$. As explained in, e.g., [19], the likelihood function for a Gaussian random variable is its joint density, $p_{\mathbf{Y}}(\mathbf{y})$ from (4). This quantity is subject to overflow and underflow, thus a modification of the likelihood,

$$\begin{aligned} \mathrm{C}_{\mathrm{MLE}}(\varepsilon,\sigma^2;\mathcal{X},\mathbf{y}) &= -2\log(p_{\mathbf{Y}}(\mathbf{y})) - N\log 2\pi \\ &= \log\left(\left(\sigma^2\right)^N \det\mathsf{K}\right) + \frac{1}{\sigma^2}\mathbf{y}^T\mathsf{K}^{-1}\mathbf{y}, \end{aligned} \tag{14}$$

is more practical for optimization. Another common modification is to solve for the σ^2 term analytically by setting the derivative with respect to σ^2,

$$\frac{\partial}{\partial\sigma^2}\mathrm{C}_{\mathrm{MLE}}(\varepsilon,\sigma^2;\mathcal{X},\mathbf{y}) = \frac{N}{\sigma^2} - \frac{\mathbf{y}^T\mathsf{K}^{-1}\mathbf{y}}{\left(\sigma^2\right)^2},$$

equal to zero. This gives the maximum likelihood estimate

$$\sigma^2_{\mathrm{mle}}(\varepsilon) = \frac{\mathbf{y}^T\mathsf{K}^{-1}\mathbf{y}}{N}, \tag{15}$$

which can be substituted back into (14) to define (minus some constants) the *profile log likelihood*,

$$\mathrm{C}_{\mathrm{MPLE}}(\varepsilon;\mathcal{X},\mathbf{y}) = \mathrm{C}_{\mathrm{MLE}}(\varepsilon,\sigma^2_{\mathrm{mle}}(\varepsilon);\mathcal{X},\mathbf{y}) = N\log(\mathbf{y}^T\mathsf{K}^{-1}\mathbf{y}) + \log\det\mathsf{K}. \tag{16}$$

The value of ε which minimizes $\mathrm{C}_{\mathrm{MPLE}}$ maximizes the likelihood (called the maximum likelihood estimator, or MLE), and thus maximizes the probability of the data having been generated over all possible shape parameters.

Remark 4. Notice that, even though the process variance σ^2 served no role in the mean of (6), its existence can play a role in the parametrization process. It is because the σ^2 term would be handled separately here that our definition of covariance in (3) separated K and σ^2.

4.2 *Computing* $\log\det\mathsf{K}$ *and* $\mathbf{y}^T\mathsf{K}^{-1}\mathbf{y}$

When K becomes ill-conditioned (such as when ε is small), computing $\mathbf{y}^T\mathsf{K}^{-1}\mathbf{y}$ and $\det\mathsf{K}$ with standard methods (e.g., Cholesky factorization) is likely inaccurate, leaving us unable to use the MLE to judge the validity of small ε for prediction purposes, despite the fact that (11) would allow us to make predictions accurately. Using the HS-SVD (10), we can follow a similar strategy as in Section 3 to approximate the value of the profile log likelihood criterion (16) for small ε.

Computing $\log\det\mathsf{K}$ is relatively straightforward using $\mathsf{K} = \Psi\Lambda_1\Phi_1^T$:

$$\log\det\mathsf{K} = \log\det\Psi + \log\det\Lambda_1 + \log\det\Phi_1^T. \tag{17}$$

First we note that Λ_1 is diagonal, and therefore the very small eigenvalues can be handled by taking their logarithms. Furthermore, because Φ_1^T was factored while forming the stable basis ψ in (9) and (assuming a prediction must also be computed) Ψ was factored while computing (11), the cost of performing (17) is negligible.

A similar strategy will allow us to compute $\mathbf{y}^T\mathsf{K}^{-1}\mathbf{y}$. In the simple kriging (or kernel interpolation) setting, the system $\mathsf{K}\mathbf{c}=\mathbf{y}$ gives rise to the best linear unbiased prediction[2] $\mathbf{k}(\mathbf{x}_0)^T\mathbf{c}=\mathbf{k}(\mathbf{x}_0)^T\mathsf{K}^{-1}\mathbf{y}$ mentioned at the end of Section 2. As demonstrated in (11), using the stable basis ψ instead of the standard basis $\mathbf{k}$, the prediction becomes $\psi(\mathbf{x}_0)^T\mathbf{b}=\psi(\mathbf{x}_0)^T\Psi^{-1}\mathbf{y}$, which corresponds to solving the system $\Psi\mathbf{b}=\mathbf{y}$. Thus we define $\mathbf{b}\in\mathbb{R}^N$ via

$$\Psi\mathbf{b}=\mathbf{y} \quad\Longleftrightarrow\quad \mathbf{b}=\Psi^{-1}\mathbf{y} \tag{18}$$

and note that the vector $\mathbf{b}$ would be available already from predicting with (11).

Applying (18) and the Hilbert–Schmidt SVD (10) to $\mathbf{y}^T\mathsf{K}^{-1}\mathbf{y}$ gives

$$\mathbf{y}^T\mathsf{K}^{-1}\mathbf{y}=(\Psi\mathbf{b})^T(\Psi\Lambda_1\Phi_1^T)^{-1}\Psi\mathbf{b}=\mathbf{b}^T\Psi^T\Phi_1^{-T}\Lambda_1^{-1}\mathbf{b}. \tag{19}$$

To evaluate $\Psi^T\Phi_1^{-T}$, we need to study ψ as defined in (9),

$$\psi(\mathbf{x})^T=\phi(\mathbf{x})^T\begin{pmatrix}\mathsf{I}_N\\ \Lambda_2\Phi_2^T\Phi_1^{-T}\Lambda_1^{-1}\end{pmatrix} \quad\Longrightarrow\quad \Psi=\Phi\begin{pmatrix}\mathsf{I}_N\\ \Lambda_2\Phi_2^T\Phi_1^{-T}\Lambda_1^{-1}\end{pmatrix}. \tag{20}$$

Using the block notation $\Phi=\begin{pmatrix}\Phi_1 & \Phi_2\end{pmatrix}$ as before, we can write

$$\Psi^T=\Phi_1^T+\Lambda_1^{-1}\Phi_1^{-1}\Phi_2\Lambda_2\Phi_2^T,$$

and using this in (19) gives

$$\mathbf{y}^T\mathsf{K}^{-1}\mathbf{y}=\mathbf{b}^T\Lambda_1^{-1}\mathbf{b}+\mathbf{b}^T\Lambda_1^{-1}\Phi_1^{-1}\Phi_2\Lambda_2\Phi_2^T\Phi_1^{-T}\Lambda_1^{-1}\mathbf{b}. \tag{21}$$

Because the *corrector matrix* $\Lambda_2\Phi_2^T\Phi_1^{-T}\Lambda_1^{-1}$ was already computed while forming Ψ with (20), computing the second term of (21) may be most efficiently done with

$$\mathbf{b}^T\Lambda_1^{-1}\Phi_1^{-1}\Phi_2\Lambda_2\Phi_2^T\Phi_1^{-T}\Lambda_1^{-1}\mathbf{b}=\left\|\Lambda_2^{-1/2}(\Lambda_2\Phi_2^T\Phi_1^{-T}\Lambda_1^{-1})\mathbf{b}\right\|_2^2. \tag{22}$$

4.3 *Approximating* $\mathbf{y}^T\mathsf{K}^{-1}\mathbf{y}$ *for small* ε

While computing (21) is possible, it may be preferable to simply exploit the quadratic (and therefore nonnegative) form of both terms on the right hand side to produce the bound

$$\mathbf{y}^T\mathsf{K}^{-1}\mathbf{y}\ge\mathbf{b}^T\Lambda_1^{-1}\mathbf{b}, \tag{23}$$

[2] In the kernel interpolation setting one instead can show that $\mathbf{k}(\mathbf{x}_0)^T\mathbf{c}=\mathbf{k}(\mathbf{x}_0)^T\mathsf{K}^{-1}\mathbf{y}$ is the *minimum norm interpolant* of the data $\mathbf{y}$ in the reproducing kernel Hilbert space $\mathcal{H}_K(\Omega)$ associated with K.

and ignore the remaining correction term (22). Note that, although $\mathsf{\Lambda}_1^{-1}$ is diagonal and $\mathbf{b}^T\mathsf{\Lambda}_1^{-1}\mathbf{b}$ is straightforward to compute, it grows unboundedly as ε shrinks to zero because of the growth in the eigenvalues[3].

Before we move on to show that it is indeed safe to ignore the correction term when ε is small, we make two comments pertinent to infinitely smooth kernels (such as Gaussians kernels), which are our primary kernels of interest:

- $\mathsf{\Psi} \to \mathsf{\Phi}_1$ for increasingly small ε which should mean that $\mathsf{\Psi}^T\mathsf{\Phi}_1^{-T} \to \mathsf{I}_N$, and, in turn,
$$\mathbf{y}^T\mathsf{K}^{-1}\mathbf{y} = \mathbf{b}^T\mathsf{\Psi}^T\mathsf{\Phi}_1^{-T}\mathsf{\Lambda}_1^{-1}\mathbf{b} \to \mathbf{b}^T\mathsf{\Lambda}_1^{-1}\mathbf{b} \quad \text{as } \varepsilon \to 0,$$
which is discussed below.
- The n^{th} value in $\mathsf{\Lambda}_1^{-1}$ is $1/\lambda_n$, which can be very large for small ε. This term is the reason that computing $\mathbf{y}^T\mathsf{K}^{-1}\mathbf{y}$ with the standard basis is unwise in the $\varepsilon \to 0$ limit.

The bound (23) is useful in the $\varepsilon \to 0$ limit so long as $\lambda_{N+1}/\lambda_N \to 0$. Because the eigenfunctions are ordered so that $\lambda_1 \geq \lambda_2 \geq \ldots$, and because the Mercer series is uniformly convergent, we know that $\lim_{\varepsilon\to 0}\mathsf{\Phi}_2\mathsf{\Lambda}_2\mathsf{\Phi}_2^T = 0$. This ordering also tells us

$$\mathsf{\Phi}_2\mathsf{\Lambda}_2\mathsf{\Phi}_2^T = \sum_{k=N+1}^{\infty}\lambda_k\hat{\varphi}_k\hat{\varphi}_k^T = \mathcal{O}(\lambda_{N+1}\hat{\varphi}_{N+1}\hat{\varphi}_{N+1}^T),$$

so that

$$\|\mathsf{\Phi}_2\mathsf{\Lambda}_2\mathsf{\Phi}_2^T\|_2 \leq \lambda_{N+1}C_{\mathsf{\Phi},N,d},$$

where $\hat{\varphi}_k^T = \begin{pmatrix}\varphi_k(\mathbf{x}_1) & \cdots & \varphi_k(\mathbf{x}_N)\end{pmatrix}$ is not to be confused with the infinite-length vector ϕ as defined in (8). Note that it can be the case that multiple eigenvalues equal λ_{N+1}, often for $d > 1$, but this will only affect the constant $C_{\mathsf{\Phi},N,d}$.

Using $\|\mathsf{\Lambda}_2^{1/2}\mathsf{\Phi}_2^T\|_2^2 = \|\mathsf{\Phi}_2\mathsf{\Lambda}_2\mathsf{\Phi}_2^T\|_2$ as $\varepsilon \to 0$ provides an upper bound for the correction term (22):

$$\begin{aligned}\left\|\mathsf{\Lambda}_2^{1/2}\mathsf{\Phi}_2^T\mathsf{\Phi}_1^{-T}\mathsf{\Lambda}_1^{-1}\mathbf{b}\right\|_2^2 &\leq \|\mathsf{\Lambda}_2^{1/2}\mathsf{\Phi}_2^T\|_2^2\|\mathsf{\Phi}_1^{-T}\|_2^2\|\mathsf{\Lambda}_1^{-1}\mathbf{b}\|_2^2\\ &\leq \lambda_{N+1}C_{\mathsf{\Phi},N,d}\|\mathsf{\Phi}_1^{-T}\|_2^2\|\mathsf{\Lambda}_1^{-1}\mathbf{b}\|_2^2\\ &\leq \frac{\lambda_{N+1}}{\lambda_N}C_{\mathsf{\Phi},N,d}\|\mathsf{\Phi}_1^{-T}\|_2^2\mathbf{b}^T\mathsf{\Lambda}_1^{-1}\mathbf{b},\end{aligned} \tag{24}$$

where we have used $\|\mathsf{\Lambda}_1^{-1}\mathbf{b}\|_2^2 \leq \left\|\mathsf{\Lambda}_1^{-1/2}\right\|\left\|\mathsf{\Lambda}_1^{-1/2}\mathbf{b}\right\|_2^2 = \mathbf{b}^T\mathsf{\Lambda}_1^{-1}\mathbf{b}/\lambda_N$. This roughly implies that

$$\mathbf{y}^T\mathsf{K}^{-1}\mathbf{y} = \mathbf{b}^T\mathsf{\Lambda}_1^{-1}\mathbf{b}\left(1 + \mathcal{O}\left(\frac{\lambda_{N+1}}{\lambda_N}\right)\right),$$

[3] We use ε here although some kernels are parametrized with other parameters, and some choices may not always approach the flat limit when that parameter → 0 (see, e.g., [9] for more details on flat limits). In that case one would have to modify the discussion accordingly.

assuming Φ_1^{-T} is well behaved in the $\varepsilon \to 0$ limit; Remark 3 mentions this and provides references discussing potential issues. We present an example that illustrates this behavior in Section 7.

5 Kriging Variance as a Parametrization Strategy

The use of profile likelihood (16) for parametrizing kernels is popular, but by no means the only viable strategy for parameter estimation. Cross-validation [15], including the leave-one-out variety [26], is also a popular parametrization strategy; the literature compares it (both favorably and unfavorably) to likelihood [17, 31, 34]. In this paper we do not discuss cross-validation.

Another tool, which parallels a strategy developed independently in numerical analysis that we discuss in Section 5.1, involves minimizing the variance of kriging predictions. Recall from (6) that the variance of a kriging prediction at a point $\mathbf{x}_0$ given data $\mathbf{y}$ observed at locations $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ is

$$\mathrm{Var}(Y_{\mathbf{x}_0}|\mathbf{Y}=\mathbf{y}) = \sigma^2 \left(K(\mathbf{x}_0,\mathbf{x}_0) - \mathbf{k}(\mathbf{x}_0)^T \mathsf{K}^{-1}\mathbf{k}(\mathbf{x}_0)\right) = \sigma^2 P_{K,\mathcal{X}}(\mathbf{x}_0)^2,$$

where we have introduced the *power function* $P_{K,\mathcal{X}}$, which is always positive, except at $\mathbf{x}_0 \in \mathcal{X}$ where $P_{K,\mathcal{X}}(\mathbf{x}_0) = 0$. As a parametrization strategy, it is important to note the presence of two distinct components which play two distinct roles: the prediction location $\mathbf{x}_0$ only appears in $P_{K,\mathcal{X}}(\mathbf{x}_0)$ and the data values $\mathbf{y}$ can only impact the process variance σ^2. Assuming the goal of using Gaussian random fields to model data is to effectively predict values at as yet unobserved locations, one strategy by which to parametrize the random field is to choose parameters that minimize the variance at the desired prediction location $\mathbf{x}_0$.

Immediately, there are some issues with this strategy. The smallest value this kriging variance can take is zero, which it will take if the process variance $\sigma^2 = 0$. Of course, $\sigma^2 = 0$ would imply that the Gaussian random field has a zero covariance kernel, and thus it would not actually be random at all. Rather than using this strategy to determine a suitable σ^2 value, we may instead refer back to (15) and use the maximum likelihood estimate $\sigma^2_{\mathrm{mle}}(\varepsilon) = \mathbf{y}^T \mathsf{K}^{-1}\mathbf{y}/N$. Doing so would produce the quantity

$$\mathrm{Var}\left(Y_{\mathbf{x}_0}|\mathbf{Y}=\mathbf{y}, \sigma^2 = \sigma^2_{\mathrm{mle}}(\varepsilon)\right) = \frac{\mathbf{y}^T \mathsf{K}^{-1}\mathbf{y}}{N} P_{K,\mathcal{X}}(\mathbf{x}_0)^2. \tag{25}$$

To convert this quantity into a parametrization strategy, we can choose ε to minimize some norm of this variance: potentially a function norm, or just the value of it at some location where a prediction is desired. For simplicity, we choose to minimize the maximum value of the variance, thus defining this parametrization objective as

$$\mathrm{C}_{\mathrm{KV}}(\varepsilon) = \log\left(\mathbf{y}^T \mathsf{K}^{-1}\mathbf{y}\right) + \max_{\mathbf{x}_0 \in \Omega \setminus \mathcal{X}} \log\left(P_{K,\mathcal{X}}(\mathbf{x}_0)^2\right), \tag{26}$$

where the constant term has been removed and the log is taken to avoid likely overflow/underflow issues during computation. In design of experiments, this might be referred to as *G*-optimality.

5.1 Numerical Analysis Connection through Reproducing Kernel Hilbert Spaces

One of the original parametrization tools developed from the numerical analysis perspective involved a bound on the error of the approximation. Recall from (6) that our prediction mechanism, given the observed data, is defined as

$$\mathbb{E}[Y_{\mathbf{x}_0}|\mathbf{Y}=\mathbf{y}]=\mathbf{k}(\mathbf{x}_0)^T\mathsf{K}^{-1}\mathbf{y}.$$

In Section 1, we described a Gaussian random field as a function of two components: the spatial component $\mathbf{x}\in\Omega\subseteq\mathbb{R}^d$ and the stochastic component $\omega\in\mathcal{W}$. If we suppose that all observations of the field occur for the same ω, then the quality of predictions would be judged against a deterministic function $y=Y(\cdot,\omega)$. One natural goal of a parametrization strategy might therefore be to choose a kernel parametrization so as to minimize

$$\left|y(\mathbf{x}_0)-\mathbf{k}(\mathbf{x}_0)^T\mathsf{K}^{-1}\mathbf{y}\right|^2, \tag{27}$$

that is, the difference between the true and predicted values, at a desired prediction location $\mathbf{x}_0$. The quantity is squared largely for cosmetic purposes.

At this point, we must recall some basic theory regarding reproducing kernel Hilbert spaces (RKHSs) from functional analysis. In particular, the *reproducing property* holds (for more details see, e.g., [10, 35]), i.e., any function $f\in\mathcal{H}_K(\Omega)$, the RKHS associated with K on Ω, satisfies $f(\mathbf{x})=\langle f,K(\cdot,\mathbf{x})\rangle_{\mathcal{H}_K}$ for $\mathbf{x}\in\Omega$, where this inner product is the RKHS inner product. Of particular consequence is that, because $K(\cdot,\mathbf{x})\in\mathcal{H}_K(\Omega)$,

$$\begin{aligned}\langle K(\cdot,\mathbf{x}),K(\cdot,\mathbf{z})\rangle_{\mathcal{H}_K}&=K(\mathbf{x},\mathbf{z}),\\ \langle K(\cdot,\mathbf{x}),\mathbf{k}(\cdot)\rangle_{\mathcal{H}_K}&=\mathbf{k}(\mathbf{x}),\\ \langle \mathbf{k}(\cdot),\mathbf{k}(\cdot)\rangle_{\mathcal{H}_K}&=\mathsf{K}.\end{aligned} \tag{28}$$

Here the latter two identities contain inner products of vectors of functions and therefore are to be considered in an elementwise sense.

Because our deterministic function y is in the RKHS $\mathcal{H}_K(\Omega)$ (see [3] for a proof) we know that $y(\mathbf{x})=\langle y,K(\cdot,\mathbf{x})\rangle_{\mathcal{H}_K}$ and therefore $\mathbf{y}=\langle y,\mathbf{k}(\cdot)\rangle_{\mathcal{H}_K}$. We can use this to express our predictions as

$$\mathbf{k}(\mathbf{x}_0)^T\mathsf{K}^{-1}\mathbf{y}=\mathbf{k}(\mathbf{x}_0)^T\mathsf{K}^{-1}\langle y,\mathbf{k}(\cdot)\rangle_{\mathcal{H}_K}=\langle y,\mathbf{k}(\mathbf{x}_0)^T\mathsf{K}^{-1}\mathbf{k}(\cdot)\rangle_{\mathcal{H}_K}.$$

Using this, we can dissect the difference (27) as

$$\begin{aligned}\left|y(\mathbf{x}_0) - \mathbf{k}(\mathbf{x}_0)^T \mathsf{K}^{-1}\mathbf{y}\right|^2 &= \left|\langle y, K(\cdot,\mathbf{x}_0)\rangle_{\mathcal{H}_K} - \langle y, \mathbf{k}(\mathbf{x}_0)^T \mathsf{K}^{-1}\mathbf{k}(\cdot)\rangle_{\mathcal{H}_K}\right|^2 \\ &= \left|\langle y, K(\cdot,\mathbf{x}_0) - \mathbf{k}(\mathbf{x}_0)^T \mathsf{K}^{-1}\mathbf{k}(\cdot)\rangle_{\mathcal{H}_K}\right|^2 \\ &\le \|y\|^2_{\mathcal{H}_K} \left\|K(\cdot,\mathbf{x}_0) - \mathbf{k}(\mathbf{x}_0)^T \mathsf{K}^{-1}\mathbf{k}(\cdot)\right\|^2_{\mathcal{H}_K}, \qquad (29)\end{aligned}$$

where the Cauchy–Schwarz inequality was invoked in the final line. Some manipulations using the identities (28) above show that

$$P_{K,\mathcal{X}}(\mathbf{x}_0)^2 = \left\|K(\mathbf{x}_0,\cdot) - \mathbf{k}(\mathbf{x}_0)^T \mathsf{K}^{-1}\mathbf{k}(\cdot)\right\|^2_{\mathcal{H}_K}.$$

We, unfortunately, do not know $\|y\|_{\mathcal{H}_K}$ (since we do not know y), but we can approximate it under the assumption that the predictions do a decent job of representing y (see [9, Chapter 9] for details). In particular, it may be reasonable to say $\|y\|_{\mathcal{H}_K} \approx \|\mathbf{k}(\cdot)^T \mathsf{K}^{-1}\mathbf{y}\|_{\mathcal{H}_K}$, and we can compute

$$\|\mathbf{k}(\cdot)^T \mathsf{K}^{-1}\mathbf{y}\|^2_{\mathcal{H}_K} = \langle \mathbf{k}(\cdot)^T \mathsf{K}^{-1}\mathbf{y}, \mathbf{k}(\cdot)^T \mathsf{K}^{-1}\mathbf{y}\rangle_{\mathcal{H}_K} = \mathbf{y}^T \mathsf{K}^{-1}\mathbf{y},$$

using (28). Substituting this into (29) gives the (approximate) bound

$$\left|y(\mathbf{x}_0) - \mathbf{k}(\mathbf{x}_0)^T \mathsf{K}^{-1}\mathbf{y}\right|^2 \le \mathbf{y}^T \mathsf{K}^{-1}\mathbf{y}\, P_{K,\mathcal{X}}(\mathbf{x}_0)^2.$$

Comparing this with (25), we see that the same logic used to minimize the kriging variance from a statistical standpoint can also have the effect of minimizing the prediction error from an approximation theory standpoint.

5.2 A Joint Profile Likelihood and Kriging Variance Objective

Interpreting the role of a parametrization objective is valuable when choosing how best to appropriately parametrize a Gaussian random field. The profile likelihood (16) measures the degree to which a chosen ε represents the data that was observed[4]. The kriging variance measures the degree to which a chosen ε is suitable for predicting unobserved values. In this section we merge these two goals into a single parametrization objective which simultaneously considers both to an appropriate degree.

To derive this joint criterion, we begin by considering the random variable

$$\widetilde{\mathbf{Y}}_{\mathbf{x}_0} = \begin{pmatrix} Y_{\mathbf{x}_0} \\ \mathbf{Y} \end{pmatrix}$$

[4] For simplicity we use only a scalar parameter ε, but everything said here holds also for a vector of kernel parameters.

which consists of the jointly distributed set of observations from the random field Y including both the observed data locations in $\mathcal{X}$ and the desired prediction location $\mathbf{x}_0$. This $\widetilde{\mathbf{Y}}$ is normally distributed with zero mean (because Y has zero mean) and covariance

$$\sigma^2 \widetilde{\mathsf{K}}(\mathbf{x}_0) = \sigma^2 \begin{pmatrix} K(\mathbf{x}_0,\mathbf{x}_0) & \mathbf{k}(\mathbf{x}_0)^T \\ \mathbf{k}(\mathbf{x}_0) & \mathsf{K} \end{pmatrix};$$

although the details were omitted, this matrix is used to determine the predictive distribution in (6).

At this point, we can use the properties of determinants to say

$$\det \sigma^2 \widetilde{\mathsf{K}}(\mathbf{x}_0) = \det \left(\sigma^2 \begin{pmatrix} K(\mathbf{x}_0,\mathbf{x}_0) & \mathbf{k}(\mathbf{x}_0)^T \\ \mathbf{k}(\mathbf{x}_0) & \mathsf{K} \end{pmatrix} \right) = \left(\sigma^2\right)^{N+1} P_{K,\mathcal{X}}(\mathbf{x}_0)^2 \det \mathsf{K}, \quad (30)$$

using the definition of the power function $P_{K,\mathcal{X}}(\mathbf{x}_0)^2 = K(\mathbf{x}_0,\mathbf{x}_0) - \mathbf{k}(\mathbf{x}_0)^T \mathsf{K}^{-1}\mathbf{k}(\mathbf{x}_0)$. Taking the logarithm of this determinant and using $\sigma^2 = \sigma^2_{\mathrm{mle}}(\varepsilon)$ from (15) gives

$$\log\det \sigma^2 \widetilde{\mathsf{K}}(\mathbf{x}_0) = N \log\left(\mathbf{y}^T \mathsf{K}^{-1}\mathbf{y}\right) + \log\det \mathsf{K} + \log\left(\mathbf{y}^T \mathsf{K}^{-1}\mathbf{y}\right) + \log\left(P_{K,\mathcal{X}}(\mathbf{x}_0)^2\right),$$

where $\mathbf{x}_0 \notin \mathcal{X}$ is required for $P_{K,\mathcal{X}}(\mathbf{x}_0) > 0$. Following the same strategy of maximizing this determinant as was used for the kriging variance yields a new parametrization criterion,

$$\mathrm{C_{DET}}(\varepsilon) = \max_{\mathbf{x}_0 \in \Omega \setminus \mathcal{X}} \log\det \sigma^2 \widetilde{\mathsf{K}}(\mathbf{x}_0) = \mathrm{C_{MPLE}}(\varepsilon) + \mathrm{C_{KV}}(\varepsilon), \quad (31)$$

Thus minimizing $\mathrm{C_{DET}}(\varepsilon)$ has the effect of balancing the desire to minimize the prediction variance at $\mathbf{x}_0$ and maximize the fit to the existing data. Of course, trying to minimize the determinant of a matrix seems dangerous. However, since K is a positive definite covariance kernel and $\mathbf{x}_0 \notin \mathcal{X}$, we know that the matrix $\widetilde{\mathsf{K}}(\mathbf{x}_0)$ must be nonsingular. Moreover, the presence of the $\mathbf{y}^T \mathsf{K}^{-1}\mathbf{y}$ term provides a necessary counterbalance to prevent this determinant from approaching an optimum at zero.

Remark 5. Computation of the power function is subject to the same ill-conditioning as any expression involving K^{-1}; this ill-conditioning can be similarly resolved with $P_{K,\mathcal{X}}(\mathbf{x}_0)^2 = K(\mathbf{x}_0,\mathbf{x}_0) - \psi(\mathbf{x}_0)^T \Psi^{-1}\mathbf{k}(\mathbf{x}_0)$. What cannot be resolved so easily is the numerical cancelation caused by the subtraction of two close numbers which occurs in the $\varepsilon \to 0$ limit. To remedy this problem, the power function must be computed with

$$\det \widetilde{\mathsf{K}}(\mathbf{x}_0) = P_{K,\mathcal{X}}(\mathbf{x}_0)^2 \det \mathsf{K} \quad \Longleftrightarrow \quad P_{K,\mathcal{X}}(\mathbf{x}_0)^2 = \frac{\det \widetilde{\mathsf{K}}(\mathbf{x}_0)}{\det \mathsf{K}}$$

using the stable determinant computation from Section 4.2. See [9, Chapter 14.1.1] or [29] for more details.

6 Summary of Parametrization Methods

In this paper we have discussed three parametrization criteria: the maximum (profile) likelihood criterion, the kriging variance criterion, and the determinant criterion. To our knowledge, the latter has not been previously discussed. There are many other criteria that appear in the literature. Some of these, such as cross validation and a Golomb–Weinberger error criterion are discussed in [9, Chapter 14].

The three criteria of interest to us in this paper are (see (16), (26), and (31))

$$\begin{aligned} \mathrm{C}_{\mathrm{MPLE}}(\varepsilon) &= N\log(\mathbf{y}^T\mathsf{K}^{-1}\mathbf{y}) + \log\det\mathsf{K}, \\ \mathrm{C}_{\mathrm{KV}}(\varepsilon) &= \log\left(\mathbf{y}^T\mathsf{K}^{-1}\mathbf{y}\right) + \max_{\mathbf{x}_0\in\Omega\setminus\mathcal{X}} \log\left(P_{K,\mathcal{X}}(\mathbf{x}_0)^2\right), \\ \mathrm{C}_{\mathrm{DET}}(\varepsilon) &= \mathrm{C}_{\mathrm{MPLE}}(\varepsilon) + \mathrm{C}_{\mathrm{KV}}(\varepsilon). \end{aligned}$$

The discussion above addressed how to stably compute the main ingredients that appear in these criteria, namely the logarithm of the native space norm of the interpolant, $\mathbf{y}^T\mathsf{K}^{-1}\mathbf{y}$, the determinant of the covariance matrix, $\det\mathsf{K}$, and the square of the power function, $P_{K,\mathcal{X}}(\mathbf{x}_0)^2$. For the reader's convenience we summarize once more how to compute each of these quantities and include also a “standard” solution that can be used in the absence of ill-conditioning or cancelation. Whenever the (positive definite) matrix K is not severely ill-conditioned (usually this is true for kernels with low smoothness such as Matérn kernels or compactly supported Wendland kernels) it is most efficient to work with its Cholesky factorization $\mathsf{K} = \mathsf{L}\mathsf{L}^T$. This is the basis for the following “standard” approaches.

The native space norm of the interpolant can be computed either as

$$\mathbf{y}^T\mathsf{K}^{-1}\mathbf{y} = \mathbf{y}^T\mathsf{L}^{-T}\mathsf{L}^{-1}\mathbf{y} = \|\mathsf{L}^{-1}\mathbf{y}\|_2^2, \tag{32}$$

or as (see (21) and (22))

$$\mathbf{y}^T\mathsf{K}^{-1}\mathbf{y} = \mathbf{b}^T\Lambda_1^{-1}\mathbf{b} + \left\|\Lambda_2^{-1/2}(\Lambda_2\Phi_2^T\Phi_1^{-T}\Lambda_1^{-1})\mathbf{b}\right\|_2^2, \tag{33}$$

where $\mathbf{b}$ is the solution of the linear system $\Psi\mathbf{b} = \mathbf{y}$ based on the stable basis from the Hilbert–Schmidt SVD.

The logarithm of the determinant of K is either computed via the diagonal entries of the Cholesky factor L as

$$\log\det\mathsf{K} = \log\det(\mathsf{L}\mathsf{L}^T) = 2\log\det\mathsf{L} = 2\sum_{i=1}^N \log\mathsf{L}_{ii}, \tag{34}$$

or via the Hilbert–Schmidt SVD as (see (17))

$$\log\det\mathsf{K} = \log\det\Psi + \log\det\Lambda_1 + \log\det\Phi_1^T. \tag{35}$$

The standard approach to computing the square of the power function would be

$$P_{K,\mathcal{X}}(\mathbf{x}_0)^2 = K(\mathbf{x}_0,\mathbf{x}_0) - \mathbf{k}(\mathbf{x}_0)^T \mathsf{K}^{-1}\mathbf{k}(\mathbf{x}_0) = K(\mathbf{x}_0,\mathbf{x}_0) - \|\mathsf{L}^{-1}\mathbf{k}(\mathbf{x}_0)\|_2^2, \tag{36}$$

or, using the Hilbert–Schmidt SVD,

$$P_{K,\mathcal{X}}(\mathbf{x}_0)^2 = K(\mathbf{x}_0,\mathbf{x}_0) - \psi(\mathbf{x}_0)^T \Psi^{-1}\mathbf{k}(\mathbf{x}_0). \tag{37}$$

However, both of these representations can lead to severe loss of significant digits (as described in Remark 5), in which case the computation requires

$$P_{K,\mathcal{X}}(\mathbf{x}_0)^2 = \frac{\det \widetilde{\mathsf{K}}(\mathbf{x}_0)}{\det \mathsf{K}}. \tag{38}$$

Using (38) to compute $\log(P_{K,\mathcal{X}}(\mathbf{x}_0)^2)$ requires two applications of the $\log\det$ formula given above: one for the standard covariance matrix K corresponding to N points, and the other for an augmented matrix based on the locations $\widetilde{\mathcal{X}} = \mathcal{X} \cup \{\mathbf{x}_0\}$.

Based on the relative complexity of the computation required to obtain (a norm of) the power function, we can see—if one is not interested in first computing the $\mathrm{C}_{\mathrm{MPLE}}(\varepsilon)$ and $\mathrm{C}_{\mathrm{KV}}(\varepsilon)$ criteria—that it is easiest to compute the determinant criterion directly as

$$\begin{aligned}\mathrm{C}_{\mathrm{DET}}(\varepsilon) = \max_{\Omega\setminus\mathcal{X}} \log\det \sigma^2 \widetilde{\mathsf{K}}(\cdot) &= \max_{\Omega\setminus\mathcal{X}} \log\left[(\sigma^2)^{N+1} P_{K,\mathcal{X}}(\cdot)^2 \det \mathsf{K}\right] \\ &= (N+1)\log(\sigma^2) + \log\det\mathsf{K} + \max_{\Omega\setminus\mathcal{X}} \log\left(P_{K,\mathcal{X}}(\cdot)^2\right)\end{aligned}$$

where we, again, use the profile variance $\sigma^2_{\mathrm{mle}}(\varepsilon) = \mathbf{y}^T \mathsf{K}^{-1}\mathbf{y}\,/\,N$ from (15) and drop the constant term. In practice, the max is approximated by sampling at finitely many locations.

7 Numerical Experiments

The main purpose of this paper has been the development of a framework for the use of the Hilbert–Schmidt SVD as described in Section 4 to perform parameter estimation for kriging predictors (or deterministic radial basis or other kernel-based approximations) in a numerically stable way. We emphasize the profile likelihood (16) because of its popularity in the literature and numerical instabilities—especially for values of the kernel parameters that often provide highly accurate models, but lead to numerically ill-conditioned linear systems. Other criteria were also introduced and summarized in the preceding section.

We now present a series of numerical experiments that illustrate the effectiveness of our approach. In Example 1 we focus on the profile likelihood and its two components (17) and (21) as well as their approximations as discussed in Section 4.2. This example uses Gaussian kernels on a set of one-dimensional data. Example 2 demonstrates the effectiveness of the HS-SVD for scattered two-dimensional data using an anisotropic tensor product Chebyshev kernel as defined in (2). Of course, we are not limited to data fitting in Euclidean domains. Example 3 demonstrates the utility of our approach on the sphere.

The last two examples, Example 4 and Example **??**, provide some comparisons amongst the additional parametrization criteria discussed in Section 5 and Section 6.

Another, separate, question regards the validity of these parameter estimation criteria for parametrizing any given kernel for predictive purposes. Unfortunately, such a question must be answered on an application-specific and kernel-specific basis, and it is far beyond the scope of this paper. As mentioned earlier, we also do not deal with randomness/noise in the data.

Example 1 (Stable computation and approximation of the profile likelihood). In this example we use data generated from the test function $f(x) = \cos(3\pi x)$. The function is sampled at $N = 24$ locations $\mathcal{X}$ sampled evenly within $[-1,1]$ to produce the vector $\mathbf{y}$ of data values. The profile likelihood criterion $\mathrm{C}_{\mathrm{MPLE}}(\varepsilon)$ (cf. (16)) is evaluated for values of ε spaced logarithmically in $[.1,10]$ using both the direct approach (labeled MLE direct in Fig. 1) based on Cholesky decomposition as laid out in Section 6, and the more elaborate formulas in (17) and (21) which provide the stable result (labeled MLE HS-SVD).

This data is then used to make predictions at $N_{\mathrm{eval}} = 100$ evenly spaced points in the domain, and the relative error compared to f is displayed in Fig. 1 with the label Error. It is apparent that the MLE direct computation loses accuracy for $\varepsilon < 3$ and suffers a complete breakdown for $\varepsilon < 1$ because K^{-1} is too ill-conditioned. By comparison, the MLE HS-SVD method suffers no ill-conditioning. The maximum likelihood estimator is near the "optimal" Error, though it does not precisely locate it.

The MLE direct curve in Fig. 1 becomes wholly unreliable for small ε because K^{-1} (computed here using the MATLAB function `pinv`) is too ill-conditioned; there appears to be a minimum value for $\varepsilon \approx .1$ which is just an artifact of the inaccurate computation. The stable method using the HS-SVD is reliable for all values of ε and clearly identifies a single region where the likelihood function is minimized. Moreover, this region is close to the "true" optimal value of ε which can be inferred from the Error graph based on the (known) function that generated the data for this test problem.

In Fig. 2 we illustrate the behavior of the components of the MLE HS-SVD: the Bound (23) and the Correction (22). As described in Section 4, the correction is only guaranteed to be negligible for $\varepsilon \to 0$, as supported by this graph. For larger values of ε, the correction may be on the same order as the bound, as indicated in the shaded strip. Because the cost of computing the full profile likelihood criterion $\mathrm{C}_{\mathrm{MPLE}}(\varepsilon)$ is negligible by comparison to solving (18), approximating the value of $\mathrm{C}_{\mathrm{MPLE}}(\varepsilon)$ by (23) is of more use from a theoretical standpoint than a computational one.

Fig. 2 also illustrates the Gap between the bound and the correction. This gap is computed as $\log_{10}(\mathsf{Bound}/\mathsf{Correction})$ and measured on the right y-axis.

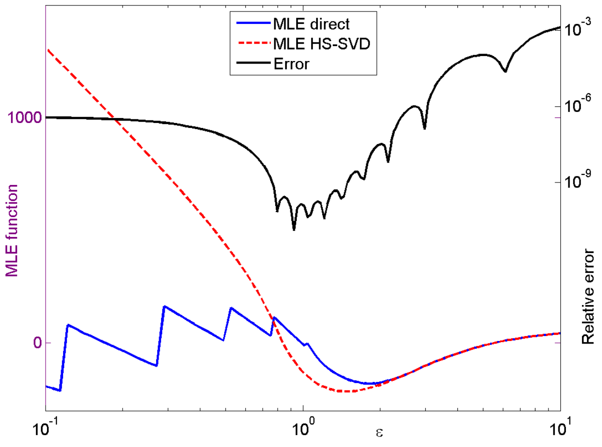

Fig. 1 Comparison of relative error based on the known test function $f(x) = \cos(3\pi x)$ and MLE estimators of the optimal shape parameter ε for Gaussian kernels computed via the (unstable) direct approach MLE direct and the stable approach MLE HS-SVD.

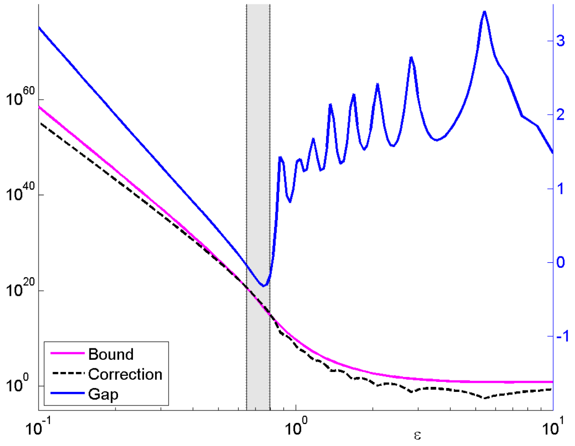

Fig. 2 Comparison of the Bound, computed with (23), and the Correction, computed with (22), using the left y-axis. The right y-axis is used to measure the Gap between these values. As $\varepsilon \to 0$, the bound dominates, but for larger values of ε no guarantee exists. In fact, the shaded gray strip denotes a region where the correction is greater than the bound.

Example 2 (2D interpolation with anisotropic tensor product Chebyshev kernel). In this example we demonstrate that the same strategy for computing the profile likelihood can also be applied in higher dimensions, with kernels different from the Gaussian kernel, and with more than one parameter.

We create data values $\mathbf{y}$ by sampling the test function

$$f(x_1,x_2) = \cos\left(\sqrt{x_1^2 + .49x_2^2}\right) + (x_1+x_2)^2 - 1 \tag{39}$$

at 81 Halton points scattered in the square $[-1,1]\times[-1,1]$. The kernel used for this example is a tensor product version of the Chebyshev kernel from (2) with a fixed value of $\mathbf{a} = [0.1, 0, 1]$ and a grid of 625 different values of the shape parameter vector $\mathbf{b} = [b_1, b_2]$ with each component spaced logarithmically in $[0.0001, 0.5012]$.

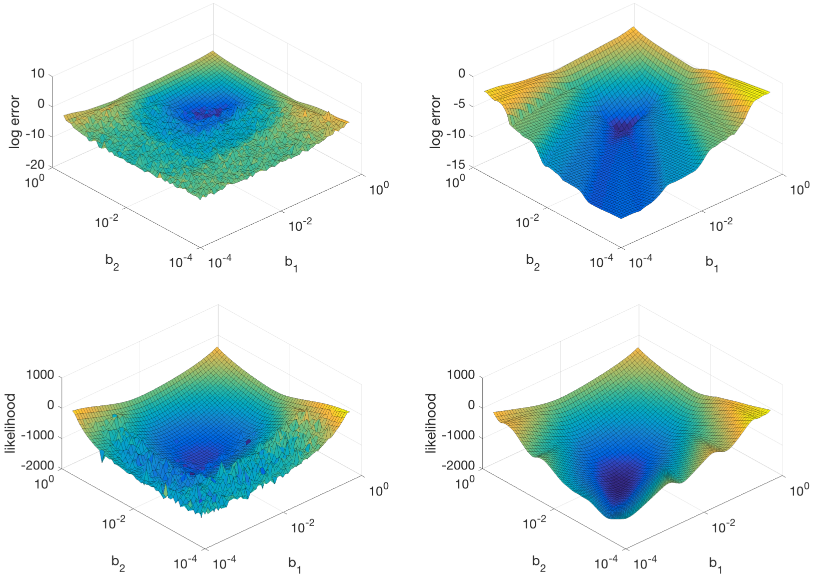

Fig. 3 Comparison of profile likelihood criterion computed without (left) and with HS-SVD (right). The top row shows the error of the kriging prediction based on Halton data sampled from the test function (39) using an anisotropic tensor product Chebyshev kernel displayed on a logarithmic scale. The bottom row displays the corresponding profile likelihood estimates.

As Fig. 3 shows, the prediction (top row) as well as the profile likelihood criterion $\mathrm{C}_{\mathrm{MPLE}}(\varepsilon)$ (bottom row) can be stably and reliably computed with the help of the HS-SVD (right column)—as compared to the direct approach, displayed in the left column, and computed using the standard basis and standard linear algebra tools such as the Cholesky decomposition and SVD. It is apparent that the stably computed profile likelihood parametrization criterion (bottom right) identifies a region for an "optimal" parameter estimate $\mathbf{b} = [b_1, b_2]$ that matches the region of smallest error (top right).

As an aside, we point out that use of a tensor product kernel in no way requires the data to be located on a grid. For more on tensor product kernels, and the Chebyshev kernel in particular, we point the reader to [9].

Example 3 (Multiquadric interpolation on a sphere). As discussed in [9, Chapter 15.3], spherical harmonics can be used to form the Hilbert-Schmidt series of zonal positive definite kernels appropriate for data from a sphere. The kernel $K(\mathbf{x},\mathbf{z}) = (1+\gamma^2 - 2\gamma\mathbf{x}^T\mathbf{z})^{-1/2}$ has the same locality for $\gamma \to 1$ and flat limit as $\gamma \to 0$ that we observed in the b parameter of the Chebyshev kernels, and as a result it is subject to the same ill-conditioning issues.

$N = 400$ points from the Womersley maximal determinant design [1] were used to sample the function

$$f(x_1,x_2,x_3) = 2\mathrm{e}^{-2x_2^2} - 3\cos(7x_1 - 2x_3)$$

at locations satisfying $x_1^2 + x_2^2 + x_3^2 = 1$. 2000 quasi random points on the sphere were chosen at which to evaluate the prediction error, which is plotted alongside the values of $\mathrm{C}_{\mathrm{MPLE}}(\gamma)$ for a range of γ values in Fig. 4.

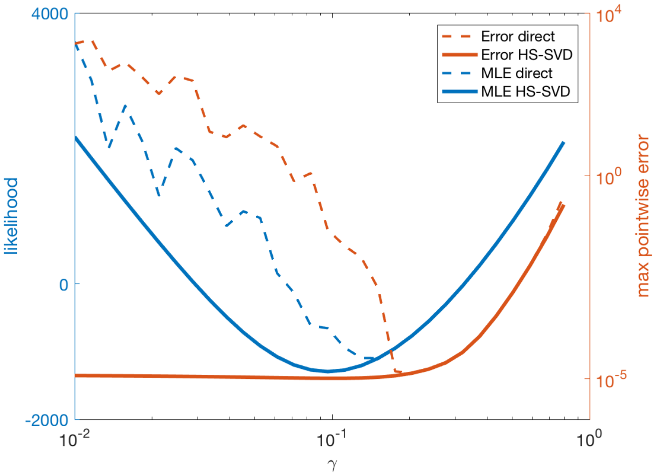

Fig. 4 The HS-SVD provides a stable strategy for covariance parametrization even for analyzing data on a sphere.

Example 4 (Comparison of various parametrization criteria for 1D interpolation with Gaussian kernel). In this example we compare the different parametrization criteria summarized in Section 6, i.e., the profile likelihood criterion $\mathrm{C}_{\mathrm{MPLE}}(\varepsilon)$, the kriging variance criterion $\mathrm{C}_{\mathrm{KV}}(\varepsilon;\mathbf{x}_0)$, and the joint determinant criterion $\mathrm{C}_{\mathrm{DET}}(\varepsilon)$. Each of these criteria is computed in different ways. For the profile likelihood criterion we have the direct approach using the Cholesky decomposition (32) and (34) (denoted as MLE direct in Fig. 5) as well as the stable version computed via the HS-SVD as in (33) and (35) (denoted as MLE HS-SVD). The kriging variance criterion is computed directly using (32) and (36) (denoted as KV direct), and computed stably either via (33) and (37) (denoted as KV HS-SVD) or via the determinant formula (38) (denoted as KV HS-SVDdet), which avoids loss of significant digits due to numerical cancelation. The determinant criterion is also computed analogously leading to the three curves denoted by DET direct, DET HS-SVD, and DET HS-SVDdet.

The test function for this example is the same as for Example 1, i.e., $f(x) = \cos(3\pi x)$ with $N = 24$ evenly spaced locations $\mathcal{X} \subset [-1,1]$. A radial Gaussian kernel with 111 values of the shape parameter ε spaced logarithmically in $[.1, 10]$. As a benchmark for the effectiveness of the various parametrization we have also added an Error curve in Fig. 5, which requires knowledge of the test function. As in Example 1, we have used two y-axes in Fig. 5 to measure the parametrization criteria on the left y-axis (some of them scaled by N so that they all evaluate over a similar range), and the fitting error on the right y-axis.

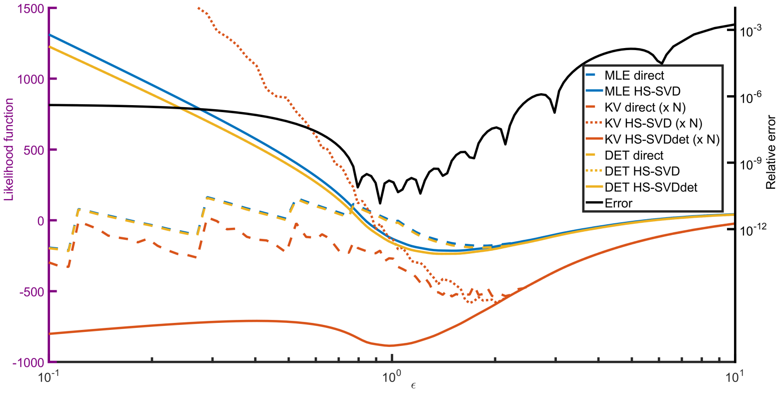

Fig. 5 Comparison of various parametrization criteria for a 1D interpolation problem with Gaussian kernels.

This example illustrates that we have derived stable and reliable versions for each of the three parametrization criteria by using the stable basis representation associated with the HS-SVD as well as the determinant formula for the computation of the power function which is not subject to loss of significant digits. For this example, all three criteria identify similar optimality regions for the shape parameter which are also similar to the optimal value from the Error graph. On the other hand, the standard/direct representations for all three criteria fail to provide any reliable estimates since they all lose accuracy for values of ε that are significantly larger than those required for the best accuracy.

8 Conclusions

Standard computations involving positive definite kernels, including prediction of unobserved values of Gaussian random fields and maximum likelihood estimation of kernel parameters, can become severely ill-conditioned for sufficiently flat kernels, such as Gaussians, with a small shape parameter ε. In earlier work, the Hilbert–Schmidt SVD was used for stable prediction when the Mercer series of the kernel is known. In this paper we have demonstrated how a similar approach allows for stable approximation of the likelihood function, which is used in determining the maximum likelihood estimator for optimal predictions. In addition, we have developed two additional parametrization criteria related to the kriging variance (and error bounds in numerical analysis) which can be stabilized using similar techniques. Numerical experiments confirm the stability for small ε, which traditional computations would be unable to achieve.

Future work should carefully investigate the advantages and disadvantages of the different criteria proposed here—depending on the specific type of application at hand, and the choice of covariance kernel used for the prediction. Further study on the relationship between the two terms present in (21) for various observed data $\mathbf{y}$ will help understand when (23) is a suitable approximation for $\mathbf{y}^T \mathsf{K}^{-1}\mathbf{y}$. Also, understanding how a Mercer series with numerically computed eigenvalues and eigenfunctions affects the quality of these computations will allow application of this strategy to a wider variety of kernels (which is currently limited by the availability of the Mercer series).

Acknowledgement The authors would like to thank Fred Hickernell for his intuition regarding Gaussian random fields, and Matthew Gliebe for his contributions as an REU student during parts of this project. The National Science Foundation grants DMS–1115392 and DMS–1522687 provided support for Matthew Gliebe and the second author.

References

1. C. An, X. Chen, I. Sloan, and R. Womersley. Well conditioned spherical designs for integration and interpolation on the two-sphere. *SIAM J. Numer. Anal.*, 48(6):2135–2157, 2010.
2. R. K. Beatson, W. A. Light, and S. Billings. Fast solution of the radial basis function interpolation equations: domain decomposition methods. *SIAM J. Sci. Comput.*, 22(5):1717–1740, 2001.
3. A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, Dordrecht, 2004.
4. R. Cavoretto, G. E. Fasshauer, and M. J. McCourt. An introduction to the Hilbert–Schmidt SVD using iterated Brownian bridge kernels. *Numer. Algorithms*, 68:393–422, 2015.
5. W. Chen, Z.-J. Fu, and C. S. Chen. *Recent Advances on Radial Basis Function Collocation Methods*. Springer Briefs in Applied Sciences and Technology. Springer, Berlin; New York, 2014.
6. N. Cressie. *Statistics for Spatial Data*. Wiley–Interscience, New York, revised edition, 1993.
7. S. De Marchi and G. Santin. A new stable basis for radial basis function interpolation. *J. Comput. Appl. Math.*, 253:1–13, 2013.

8. K. T. Fang, R. Li, and A. Sudjianto. *Design and Modeling for Computer Experiments*. Computer Science and Data Analysis. Chapman & Hall, New York, 2006.
9. G. Fasshauer and M. McCourt. *Kernel-based Approximation Methods using* MATLAB, volume 19 of *Interdisciplinary Mathematical Sciences*. World Scientific Publishing Co., Singapore, 2015.
10. G. E. Fasshauer. *Meshfree Approximation Methods with* MATLAB, volume 6 of *Interdisciplinary Mathematical Sciences*. World Scientific Publishing Co., Singapore, 2007.
11. G. E. Fasshauer and M. J. McCourt. Stable evaluation of Gaussian radial basis function interpolants. *SIAM J. Sci. Comput.*, 34(2):A737–A762, 2012.
12. B. Fornberg and N. Flyer. *A Primer on Radial Basis Functions with Applications to the Geosciences*. SIAM, Philadelphia, 2015.
13. A. I. J. Forrester, A. Sobester, and A. J. Keane. *Engineering Design via Surrogate Modelling*. Wiley, Chichester, 2008.
14. E. Fuselier. Improved stability estimates and a characterization of the native space for matrix-valued RBFs. *Adv. Comput. Math.*, 29(3):311–313, 2008.
15. F. J. Hickernell and Y. C. Hon. Radial basis function approximations as smoothing splines. *Appl. Math. Comput.*, 102(1):1–24, 1999.
16. H. Kadri, E. Duflos, P. Preux, S. Canu, A. Rakotomamonjy, and J. Audiffren. Operator-valued kernels for learning from functional response data. *arXiv:1510.08231 [cs, stat]*, 2015.
17. R. Kohn, C. F. Ansley, and D. Tharm. The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. *Journal of the American Statistical Association*, 86(416):1042–1050, 1991.
18. K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, London; New York, 1979.
19. K. V. Mardia and R. J. Marshall. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71(1):135–146, 1984.
20. Alessandra Menafoglio and Giovanni Petris. Kriging for Hilbert-space valued random fields: The operatorial point of view. *Journal of Multivariate Analysis*, 2015. Online.
21. C. A. Micchelli and M. Pontil. Kernels for multi-task learning. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 921–928. MIT Press, 2005.
22. S. Müller and R. Schaback. A Newton basis for kernel spaces. *J. Approx. Theory*, 161(2):645–655, December 2009.
23. F. J. Narcowich and J. D. Ward. Generalized Hermite interpolation via matrix-valued conditionally positive definite functions. *Math. Comp.*, 63(208):661–687, 1994.
24. A. Neumaier. Solving ill-conditioned and singular linear systems: A tutorial on regularization. *SIAM Rev.*, 40(3):636–666, 1998.
25. C. E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.
26. S. Rippa. An algorithm for selecting a good value for the parameter c in radial basis function interpolation. *Adv. Comput. Math.*, 11(2–3):193–210, 1999.
27. J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. *Stat. Sci.*, 4(4):409–423, 1989.
28. T. J. Santner, B. J. Williams, and W. I. Notz. *The Design and Analysis of Computer Experiments*. Springer, Berlin; New York, 2003.
29. R. Schaback. Multivariate interpolation by polynomials and radial basis functions. *Constr. Approx.*, 21:293–317, 2005.
30. M. Scheuerer, R. Schaback, and M. Schlather. Interpolation of spatial data — a stochastic or a deterministic problem? *Eur. J. Appl. Math.*, 24(4):601–629, 2013.
31. M. L. Stein. A comparison of generalized cross validation and modified maximum likelihood for estimating the parameters of a stochastic process. *Ann. Stat.*, 18(3):1139–1157, 1990.
32. M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, Berlin; New York, 1999.
33. I. Steinwart and A. Christmann. *Support Vector Machines*. Information Science and Statistics. Springer, Berlin; New York, 2008.

34. G. Wahba. A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics*, pages 1378–1402, 1985.
35. H. Wendland. *Scattered Data Approximation*, volume 17 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2005.