Euro. Jnl of Applied Mathematics (2013), *vol.* 24, *pp.* 601–629. © Cambridge University Press 2013
doi:10.1017/S0956792513000016

A Survey in Mathematics for Industry

Interpolation of spatial data – A stochastic or a deterministic problem?

M. SCHEUERER[1], R. SCHABACK[2] and M. SCHLATHER[3]

[1]*Institut für Angewandte Mathematik, Ruprecht-Karls-Universität Heidelberg, Im Neuenheimer Feld 294, D-69120 Heidelberg, Germany*
email: michael.scheuerer@uni-heidelberg.de
[2]*Institut für Numerische und Angewandte Mathematik, Georg-August-Universität Göttingen, Lotzestr, 16-18, D-37083 Göttingen, Germany*
email: schaback@math.uni-goettingen.de
[3]*Institut für Mathematik, Universität Mannheim, A5, 6, D-68131 Mannheim, Germany*
email: schlather@math.uni-mannheim.de

(*Received 28 November 2011; revised 3 January 2013; accepted 4 January 2013; first published online 7 February 2013*)

Interpolation of spatial data is a very general mathematical problem with various applications. In geostatistics, it is assumed that the underlying structure of the data is a stochastic process which leads to an interpolation procedure known as kriging. This method is mathematically equivalent to kernel interpolation, a method used in numerical analysis for the same problem, but derived under completely different modelling assumptions. In this paper we present the two approaches and discuss their modelling assumptions, notions of optimality and different concepts to quantify the interpolation accuracy. Their relation is much closer than has been appreciated so far, and even results on convergence rates of kernel interpolants can be translated to the geostatistical framework. We sketch different answers obtained in the two fields concerning the issue of kernel misspecification, present some methods for kernel selection and discuss the scope of these methods with a data example from the computer experiments literature.

Key words: spatial interpolation, geostatistics, kernel interpolation

1 Introduction

1.1 A survey in mathematics for industry

Interpolation of spatial data is a very general mathematical problem with various applications, such as surface reconstruction, the numerical solution of partial differential equations, learning theory, computer experiments and the prediction of environmental variables, to name a few. Specific instances from different fields of application can be found in [9] and [85]. The precise mathematical formulation of the problem is as follows:

Reconstruct a function $f : T \to \mathbb{R}$, where T is a domain in $\mathbb{R}^d$, based on its values at a finite set of data points $X := \{x_1, \ldots, x_n\} \subset T$ (usually called 'sampling locations' in geostatistics and 'centres' in kernel interpolation).

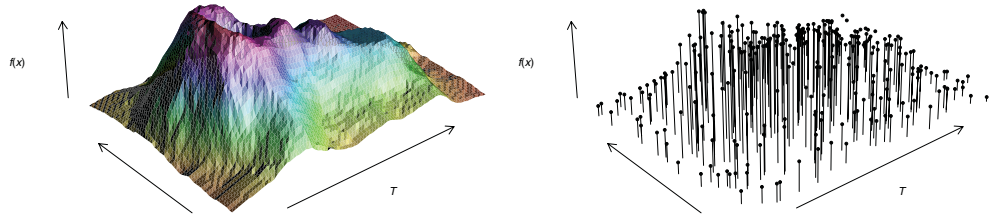

FIGURE 1. (Colour online) Perspective plot of Mount Eden (left), and 3D scatterplot of the 300 given data points (right).

This situation is illustrated in Figure 1 with topographic data from Mount Eden, New Zealand,[1] with X being a randomly chosen set of 300 sampling locations. In order to derive optimal procedures for reconstruction and to provide *a priori* estimates of their precision, it is necessary to make assumptions about f. We focus on two different approaches that deal with the above problem in different ways: kernel interpolation and kriging. The former assumes that f belongs to some Hilbert space $\mathscr{H}$ of functions of certain smoothness. This allows one to use Taylor approximation techniques to derive bounds for the approximation error in terms of the density of the data points. Smoothness is a comparatively weak and flexible assumption, and the error bounds permit control of the precision whenever it is possible to control the sampling. By construction, the kernel interpolation approach yields minimal approximation errors with respect to the norm $\|\cdot\|_{\mathscr{H}}$ on $\mathscr{H}$.

In some applications there is only limited or no control over the sampling and one has to get by with the (sometimes very sparse) data that are available. Typical examples are environmental modelling or mining where sampling involves high costs or is limited by lack of accessibility of the variable of interest. Moreover, in these applications the variable of interest is often a very rough function, and together with the sparsity of data this implies that error bounds obtained on the basis of Taylor approximation are only of limited use. A way out is possible if the stronger modelling assumption that comes with a statistical modelling approach is adequate: the assumption that f is a realization of a random field. Then again optimal approximation procedures can be derived, and a satisfactory stochastic description of the approximation error is available.

It is quite remarkable that both approaches finally come up with the same type of approximant despite different model assumptions and motivations of its construction. Several authors, including [10, 42, 51, 57], have already pointed out this connection, and a comprehensive overview over both approaches is given by [4]. While the authors of [4] also establish the link between stochastic processes and reproducing kernel Hilbert spaces (RKHS), the equivalence of kernel interpolation and kriging is shown by the usual algebraic arguments. In this paper we introduce kernel interpolation and the underlying RKHS model in a different way than that usually taken in the spline literature. This will make it clear that the derivation of optimal interpolation procedures follows the same principles in the stochastic and deterministic frameworks, and reveal that the connection between these frameworks goes much further than algebraic equivalence of respective interpolants.

In Sections 2 and 3 we describe kernel interpolation and kriging respectively along with their modelling assumptions and concepts of optimality. Some problems closely related to spatial interpolation and generalizations are discussed in Section 4. The presentation

[1] Available as dataset 'volcano' in the package 'datasets' of R [61].

https://doi.org/10.1017/S0956792513000016 Published online by Cambridge University Press

in Section 2 and 3 will show that the function that characterizes the magnitude of the pointwise approximation error appears – with different interpretations – in both frameworks. This will be used in Section 5 to apply theorems on the convergence rates of kernel interpolants to the stochastic framework, where statements of comparable generality have not been available so far. In Section 6, the issue of kernel misspecification is addressed, and we give an overview of the answers given in both communities to the question about the consequences of using an 'incorrect' kernel for the construction of the interpolant. In Section 7 we turn to the issue of parameter estimation and describe some of the procedures used to select a kernel based on the available data. The scope of these methods is briefly discussed and illustrated with a data example.

The main focus of this paper is to point out the interconnections between the two approaches to spatial interpolation. Some topics which receive considerable attention in one of the two frameworks but are (from our current perspective) hardly relevant for the respective other, will be briefly addressed in the final discussion.

2 Kernel interpolation

2.1 Positive definite kernels

In the kernel interpolation framework, f is assumed to belong to some Hilbert space $\mathscr{H}$ of real-valued functions on T with inner product $(\cdot,\cdot)_{\mathscr{H}}$. It is further assumed that for all $x \in T$ the point evaluation functional $\delta_x : f \mapsto f(x)$ is continuous in $\mathscr{H}$, i.e.

$$\delta_x \in \mathscr{H}^* \quad \text{for all} \quad x \in T, \tag{2.1}$$

where $\mathscr{H}^*$ denotes the dual of $\mathscr{H}$ with dual norm

$$\|\lambda\|_{\mathscr{H}^*} = \sup_{g \in \mathscr{H} : \|g\|_{\mathscr{H}} \leqslant 1} |\lambda(g)|, \qquad \lambda \in \mathscr{H}^*.$$

Then by the theory of reproducing kernel Hilbert spaces (see [67] and references therein), a unique symmetric function $K : T \times T \to \mathbb{R}$ ('reproducing kernel') exists with $K(\cdot, x) \in \mathscr{H}$ and

$$\begin{aligned} g(x) &= (g, K(\cdot, x))_{\mathscr{H}} \\ K(x, y) &= (K(\cdot, x), K(\cdot, y))_{\mathscr{H}} \\ K(x, y) &= (\delta_x, \delta_y)_{\mathscr{H}^*} \end{aligned} \tag{2.2}$$

for all $x, y \in T$, $g \in \mathscr{H}$. Note that either of the last two equations imply

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(x_i, x_j) \geqslant 0 \tag{2.3}$$

for any choice of points $x_1, \ldots, x_n \in T, n \in \mathbb{N}$ and coefficients $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n \setminus \{0\}$. If the functionals $\delta_{x_1}, \ldots, \delta_{x_n}$ are linear independent when based on distinct points, we even have strict inequality in (2.3). In this case K is called a positive definite kernel and $\mathscr{H} = \mathscr{H}_K$ is called the *native space* for K. The important role of positive definiteness for kernel interpolation was pointed out by Micchelli [55].

https://doi.org/10.1017/S0956792513000016 Published online by Cambridge University Press

Table 1. *Some positive definite functions $\Phi(h)$ normalized such that $\Phi(0) = 1$. Where closed forms are available, the corresponding Fourier transforms $\widehat{\Phi}(\xi)$ are also given, otherwise we give an expression indicating their rate of decay*

Gaussians	$e^{-\frac{1}{2}\|h\|^2}$	$e^{-\frac{1}{2}\|\xi\|^2}$
Inverse multiquadrics, $\beta > 0$	$(1 + \|h\|^2)^{-\beta}$	$\frac{\|\xi\|^{\beta-\frac{d}{2}}}{2^{\beta-1}\Gamma(\beta)} \mathscr{K}_{\beta-\frac{d}{2}}(\|\xi\|)$
Matérn class, $\nu > 0$	$\frac{\|h\|^\nu}{2^{\nu-1}\Gamma(\nu)} \mathscr{K}_\nu(\|h\|)$	$\frac{\Gamma(\nu+\frac{d}{2})}{\Gamma(\nu)} (1 + \|\xi\|^2)^{-\nu-\frac{d}{2}}$
Cauchy class, $\alpha = 2$, $\beta > 0$	See inverse multiquadrics	
$\alpha \in (0, 2)$, $\beta > 0$	$(1 + \|h\|^\alpha)^{-\beta/\alpha}$	$O((1 + \|\xi\|^2)^{-\frac{\alpha}{2}-\frac{d}{2}})$
Wendland functions, $d \leqslant 3$	$(1 - \|h\|)_+^2$	$O((1 + \|\xi\|^2)^{-\frac{1}{2}-\frac{d}{2}})$
cf. [68, 84] for further examples	$(1 - \|h\|)_+^4(4\|h\| + 1)$	$O((1 + \|\xi\|^2)^{-\frac{3}{2}-\frac{d}{2}})$
	$(1 - \|h\|)_+^6(\frac{35}{3}\|h\|^2 + 6\|h\| + 1)$	$O((1 + \|\xi\|^2)^{-\frac{5}{2}-\frac{d}{2}})$

The native space is an abstract concept, but for some important function spaces an explicit link can be made to translation-invariant kernels where $K(x, y) = \Phi(y - x)$ for some function $\Phi : \mathbb{R}^d \to \mathbb{R}$. Consider, for example, the Sobolev spaces $W_2^\tau(\mathbb{R}^d)$, which are widely used in numerical analysis, in particular in the context of Partial Differential Equations (PDEs). They not only have a rich mathematical structure but also characterize the degree of smoothness of the functions belonging to them. For $\tau \in \mathbb{N}$, this can be seen directly from their definition

$$W_2^\tau(\mathbb{R}^d) = \{f \in L_2(\mathbb{R}^d) : D^\alpha f \in L_2(\mathbb{R}^d) \quad \text{for all} \quad |\alpha| \leqslant \tau, \alpha \in \mathbb{N}^d\}$$

where D^α denotes an αth weak partial derivative [22, Section 5.2]. An equivalent definition exists in terms of Fourier transforms, and this definition has a straightforward generalization to non-integer orders, $\tau > 0$. In the remainder of this paper, we always have $\tau > \frac{d}{2}$, which implies, by the Sobolev embedding theorem, that every equivalence class in $W_2^\tau(\mathbb{R}^d)$ contains a continuous representer. We will interpret $W_2^\tau(\mathbb{R}^d)$ as a set of continuous functions in this way. The following theorem [85, Cor. 10.13] shows that it constitutes the native space of certain translation-invariant kernels Φ which have a degree of smoothness that depends on τ.

Theorem 2.1 *Suppose that the Fourier transform $\widehat{\Phi}(\xi) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{-i\,\xi' h}\, \Phi(h)\, dh$ of some positive definite function $\Phi \in L_1(\mathbb{R}^d) \cap C(\mathbb{R}^d)$ satisfies*

$$c_1 (1 + \|\xi\|^2)^{-\tau} \;\leqslant\; \widehat{\Phi}(\xi) \;\leqslant\; c_2 (1 + \|\xi\|^2)^{-\tau}, \qquad \xi \in \mathbb{R}^d, \tag{2.4}$$

for some $\tau > \frac{d}{2}$ and constants $0 < c_1 \leqslant c_2$. Then the native space of Φ coincides with the Sobolev space $W_2^\tau(\mathbb{R}^d)$ as a vector space, and the native space norm and the Sobolev norm are equivalent.

Table 1 shows examples (see [31, 46, 50, 84] for details) of positive definite functions Φ commonly used in geostatistics and the approximation theory. For $\Phi \in L_2(\mathbb{R}^d)$ their Fourier transforms are defined as stated in Theorem 2.1. We write $O((1 + \|\xi\|^2)^{-\tau})$ to denote that (2.4) is satisfied where closed forms are not available.

https://doi.org/10.1017/S0956792513000016 Published online by Cambridge University Press

While Sobolev spaces are intuitively more accessible and allow one to better understand what exactly is assumed for f, the framework of native spaces is useful to derive an optimal approximation of f based on the given data at X in the sense that the worst case approximation error is minimized pointwise. Specifically, we consider approximants of the form

$$s_{f,X}(x) = \sum_{i=1}^{n} u_i(x) f(x_i) = \underbrace{\sum_{i=1}^{n} u_i(x)\, \delta_{x_i}(f)}_{:=\lambda_{u(x)}}, \qquad x \in T, \tag{2.5}$$

which are, at each point $x \in T$, a linear combination of the given values of f. The coefficient functions $u_1, \ldots, u_n : T \to \mathbb{R}$ are defined pointwise, and for fixed $x_0 \in T$ we consider the norm of the error functional $\lambda_{\mathrm{err}} := \delta_{x_0} - \lambda_{u(x_0)}$

$$\mathcal{Q}^{1/2}(u(x_0)) := \|\delta_{x_0} - \lambda_{u(x_0)}\|_{\mathcal{H}_K^*} = \sup_{g \in \mathcal{H}_K \,:\, \|g\|_{\mathcal{H}_K} \leqslant 1} |\delta_{x_0}(g) - \lambda_{u(x_0)}(g)|\,.$$

According to (2.3), its square can be written as a quadratic form

$$\mathcal{Q}(u(x_0)) = K(x_0, x_0) - 2\sum_{i=1}^{n} u_i(x_0) K(x_0, x_i) + \sum_{i=1}^{n}\sum_{j=1}^{n} u_i(x_0) u_j(x_0) K(x_i, x_j)\,, \tag{2.6}$$

and it follows that optimal coefficients $u_1^*(x_0), \ldots, u_n^*(x_0)$ minimizing $\mathcal{Q}$ must satisfy

$$\sum_{j=1}^{n} u_j^*(x_0) K(x_i, x_j) = K(x_0, x_i) \qquad i = 1, \ldots, n. \tag{2.7}$$

If K is positive definite, then this system has a unique solution, and this in turn implies that the so-called Lagrange conditions,

$$u_i^*(x_k) = \begin{cases} 1 & \text{if } i = k \\ 0 & \text{if } i \neq k \end{cases}, \qquad i, k = 1, \ldots, n, \tag{2.8}$$

are satisfied. Hence, $s_{f,X}$ interpolates f at X, and it can be shown that it has minimal native space norm among all such interpolants [65]. This property is the usual starting point for the derivation of kernel interpolants in the spline literature.

2.2 Conditionally positive definite kernels

Some important classical interpolation schemes, such as thin-plate splines [18–20] or Hardy's multiquadrics [35], are not covered by the above theory, but can still be incorporated into the framework of kernel interpolation. To this end we must allow for kernels that are only *conditionally positive definite* with respect to some finite dimensional function space $\mathcal{P}$ (in applications we usually have $\mathcal{P} = \pi_m(T)$, the space of polynomials on T of order at most m). Let $L_{\mathcal{P}}(T)$ denote the space of all linear functionals of the form

$$\lambda_{\mathcal{X}} = \sum_{i=1}^{n} a_i \delta_{x_i}, \qquad a_1, \ldots, a_n \in \mathbb{R}, \quad \mathcal{X} := \{x_1, \ldots, x_n\} \subset T, \quad n \in \mathbb{N}$$

https://doi.org/10.1017/S0956792513000016 Published online by Cambridge University Press

Table 2. *Some conditionally positive definite functions $\Phi(h)$ together with the minimal space with respect to which they are conditionally positive definite*

Powers, $\beta \in \mathbb{R}_{>0} \setminus 2\mathbb{N}$	$\Gamma\left(-\frac{\beta}{2}\right) \|h\|^{\beta}$	$\pi_{\lceil \frac{\beta}{2} \rceil - 1}(\mathbb{R}^d)$
Thin-plate splines, $d \in \mathbb{N} \setminus 2\mathbb{N}$, $\frac{d}{2} < l \in \mathbb{N}$	$\frac{\Gamma(\frac{d}{2}-l)\, 2^{-2l}}{\pi^{d/2}(l-1)!} \|h\|^{2l-d}$	$\pi_{l-\frac{d+1}{2}}(\mathbb{R}^d)$
$d \in 2\mathbb{N}$, $\frac{d}{2} < l \in \mathbb{N}$	$\frac{(-1)^{l+1-d/2}\, 2^{1-2l}}{\pi^{d/2}(l-1)!(l-\frac{d}{2})!} \|h\|^{2l-d} \log \|h\|$	$\pi_{l-\frac{d}{2}}(\mathbb{R}^d)$
Multiquadrics, $\beta \in \mathbb{R}_{>0} \setminus \mathbb{N}$	$(-1)^{\lceil \beta \rceil} (1 + \|h\|^2)^{\beta}$	$\pi_{\lceil \beta \rceil - 1}(\mathbb{R}^d)$

that vanish on $\mathscr{P}$, i.e. $\lambda_{\mathscr{X}}(p) = 0$ for all $p \in \mathscr{P}$. This is a vector space over $\mathbb{R}$ under usual operations. A kernel K is called conditionally positive definite with respect to $\mathscr{P}$ if

$$\left(\lambda^1_{\mathscr{X}} \lambda^2_{\mathscr{X}}\right)(K) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j K(x_i, x_j) > 0 \qquad \text{for all } \lambda_{\mathscr{X}} \in L_{\mathscr{P}}(T) \setminus \{0\}, \tag{2.9}$$

where the superscripts denote the application of $\lambda_{\mathscr{X}}$ with respect to the first and second argument of K respectively. Note that such K is also conditionally positive definite with respect to any finite dimensional function space $\mathscr{P}' \supset \mathscr{P}$, in particular we can always consider a conditionally positive definite kernel with respect to $\pi_m(T)$ as conditionally positive definite with respect to $\pi_l(T)$ if $l \geqslant m$.

Assume from now that $K : T \times T \to \mathbb{R}$ is (symmetric and) conditionally positive definite with respect to $\mathscr{P}$. In analogy with (2.3) we let

$$(\lambda_{\mathscr{X}}, \mu_{\mathscr{Y}})_K := \left(\lambda^1_{\mathscr{X}} \mu^2_{\mathscr{Y}}\right)(K), \qquad \lambda_{\mathscr{X}}, \mu_{\mathscr{Y}} \in L_{\mathscr{P}}(T),$$

and due to (2.9) this defines an inner product on $L_{\mathscr{P}}(T)$. This can be used to define the native space of K as the largest space on which all functionals from $L_{\mathscr{P}}(T)$ act continuously, i.e.

$$\mathscr{H}_{K,\mathscr{P}} := \{g : T \to \mathbb{R} : |\lambda(g)| \leqslant C_g \|\lambda\|_K \text{ for all } \lambda \in L_{\mathscr{P}}(T)\}, \tag{2.10}$$

where $C_g < \infty$ is a constant depending only on g. A semi-norm on $\mathscr{H}_{K,\mathscr{P}}$ can be defined via

$$|g|_{\mathscr{H}_{K,\mathscr{P}}} := \sup_{\lambda \in L_{\mathscr{P}}(T) : \|\lambda\|_K \leqslant 1} |\lambda(g)|$$

This characterization goes back to the pioneering work of Madych and Nelson [48]. We chose it because of its striking analogy with the theory of intrinsic random fields [52], which will be further discussed in Section 3. In [67] a detailed derivation of the native space of a given conditionally positive definite kernel K is presented, showing how values $g(x), x \in T$ can be assigned to the abstract function g which *a priori* can be evaluated only by functionals from $L_{\mathscr{P}}(T)$ which does not include the point evaluation functionals δ_x. Note that the positive definite case discussed above corresponds to $\mathscr{P} = \{0\}$, and the continuity of any $\lambda \in L_{\{0\}}(T)$ is a consequence of assumption (2.1) and the Riesz representation theorem.

Examples of conditionally positive definite functions are given in Table 2. For the thin-plate splines, from now denoted by $\Phi_{d,l}$, a counterpart of Theorem 2.1 will provide

https://doi.org/10.1017/S0956792513000016 Published online by Cambridge University Press

some intuitive understanding of the corresponding native space. $\Phi_{d,l}$ will be considered as conditionally positive definite with respect to $\pi_{l-1}(\mathbb{R}^d)$. The corresponding function spaces are the Beppo–Levi spaces

$$BL_l(\mathbb{R}^d) = \left\{ g \in L_1^{loc}(\mathbb{R}^d) : D^\alpha g \in L_2(\mathbb{R}^d) \quad \text{for all} \quad |\alpha| = l, \alpha \in \mathbb{N}^d \right\}$$

with the semi-norm

$$|g|_{BL_l(\mathbb{R}^d)} = \left(\sum_{|\alpha|=l} \frac{l!}{\alpha_1! \cdots \alpha_d!} \|D^\alpha g\|^2_{L_2(\mathbb{R}^d)} \right)^{1/2}.$$

Beppo–Levi spaces are closely related to Sobolev spaces, and this relation can be used to show that any $BL_l(\mathbb{R}^d)$ with $l > \frac{d}{2}$ can be embedded into $C(\mathbb{R}^d)$ [19].

Theorem 2.2 *[85, Theorem 10.43] Let $\Phi_{d,l}$ be a thin-plate spline kernel from Table 2, considered as a conditionally positive definite with respect to $\pi_{l-1}(\mathbb{R}^d)$. Then the associated native space $\mathscr{H}_{K,\mathscr{P}}$ is the Beppo Levi space $BL_l(\mathbb{R}^d)$ of order l, and the semi-norms are the same.*

When it comes to deriving an optimal approximation of f, minimization of the norm of the error functional $\lambda_{\text{err}} = \delta_{x_0} - \lambda_{u(x_0)}$ for fixed $x_0 \in T$ again amounts to the minimization of the quadratic form $\mathscr{Q}$ in (2.6). In the general framework of conditionally positive definite kernels, however, λ_{err} is not automatically in $L_{\mathscr{P}}(T)$, and the additional constraint

$$\delta_{x_0}(p) = \lambda_{u(x_0)}(p) \qquad \text{for all} \quad p \in \mathscr{P} \tag{2.11}$$

must be satisfied to ensure that $\|\lambda_{\text{err}}\|_K$ is defined. Note that this constraint also implies that functions from $\mathscr{P}$ are always reproduced exactly by $s_{f,X}$. Since $\mathscr{P}$ was assumed finite dimensional, we can choose a basis $p_1, \ldots, p_q$, and the above condition becomes

$$p_k(x_0) = \sum_{i=1}^{n} u_i(x_0) p_k(x_i), \qquad k = 1, \ldots, q. \tag{2.12}$$

Minimizing (2.6) subject to (2.12) can be done using Lagrange multipliers $\eta_1(x_0), \ldots, \eta_q(x_0)$, and it follows that optimal coefficients $u_1^*(x_0), \ldots, u_n^*(x_0)$ must satisfy (2.12) and

$$\sum_{j=1}^{n} u_j^*(x_0) K(x_i, x_j) + \sum_{k=1}^{q} \eta_k^*(x_0) p_k(x_i) = K(x_0, x_i) \qquad i = 1, \ldots, n, \tag{2.13}$$

which generalizes the equation system (2.7) derived in the positive definite setup. If the set of data points $X = \{x_1, \ldots, x_n\}$ is $\mathscr{P}$-unisolvent, i.e. the zero function is the only function in $\mathscr{P}$ that vanishes on X, then the system of equations defined by (2.12) and (2.13) has a unique solution. Then again the Lagrange conditions (2.8) are satisfied, showing that $s_{f,X}$ interpolates the data. Criteria for $\pi_m(\mathbb{R}^d)$-unisolvency are discussed in [85, Section 2.2] and [23, Section 6.1].

Representation (2.5) of $s_{f,X}$ is a good starting point to derive a pointwise optimal approximation of f, but it is quite inefficient from a computational point of view.

https://doi.org/10.1017/S0956792513000016 Published online by Cambridge University Press

Some algebraic manipulations of (2.5), (2.12) and (2.13) however yield the alternative representation

$$s_{f,X} = \sum_{j=1}^{n} \alpha_j K(\cdot, x_j) + \sum_{k=1}^{q} \beta_k p_k, \tag{2.14}$$

where the coefficients $\alpha_1, \ldots, \alpha_n$ and $\beta_1, \ldots, \beta_q$ are defined by the system

$$\begin{aligned} \sum_{j=1}^{n} \alpha_j K(x_i, x_j) + \sum_{k=1}^{q} \beta_k p_k(x_i) &= f(x_i), \qquad i = 1, \ldots, n \\ \sum_{i=1}^{n} \alpha_j p_k(x_j) &= 0, \qquad k = 1, \ldots, q \end{aligned} \tag{2.15}$$

which is again uniquely solvable if K is conditionally positive definite and X is $\mathscr{P}$-unisolvent. Note that the first set of equations simply forces $s_{f,X}$ to interpolate the data, while the second set is necessary to ensure a unique decomposition into two terms in (2.14). This system needs to be solved only once and then yields an expression for $s_{f,X}$ in closed form, valid on the whole of T. Its solution requires $O(n^3)$ floating point operations which may still be too expensive for large spatial data sets. We refer to [85, Chapter 15] for an overview over some algorithms that compute an approximate solution to reduce the computational cost to a practically manageable level.

3 Kriging

The statistical counterpart to kernel interpolation is known as kriging, the geostatistical term for optimal linear prediction of spatial processes. Kriging is based on the modelling assumption that f is a realization of a *random field* Z, which is a collection $\{Z(x) : x \in T\}$ of random variables over the same probability space $(\Omega, \mathscr{A}, P)$, indexed over T. The observations $f(x_1), \ldots, f(x_n)$ are then realizations of the random variables $Z(x_1), \ldots, Z(x_n)$. To predict Z at some (unobserved) location $x_0 \in T$, one considers all *linear predictors* of the form

$$Z_u(x_0) = \sum_{i=1}^{n} u_i(x_0) Z(x_i) \tag{3.1}$$

which are themselves random variables. The prediction of $f(x_0)$ given $f(x_1), \ldots, f(x_n)$ is then as for kernel interpolation

$$s_{f,X}(x_0) = \sum_{i=1}^{n} u_i(x_0) f(x_i).$$

To determine optimal weights $u_1^*(x_0), \ldots, u_n^*(x_0)$, additional structural assumptions on Z are needed, and depending on these assumptions one distinguishes simple, ordinary, universal and intrinsic kriging. There are still quite other forms (such as complex kriging [45], indicator kriging [41] or disjunctive kriging [53, 54]) but we shall only discuss the aforementioned ones due to their close connection to kernel interpolation.

https://doi.org/10.1017/S0956792513000016 Published online by Cambridge University Press

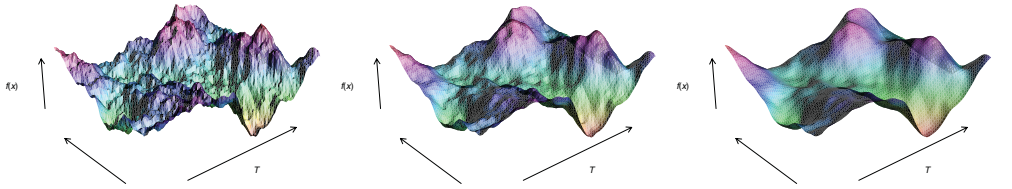

FIGURE 2. (Colour online) Perspective plots of one realization of Gaussian random fields with different Matérn covariance functions Φ_ν with $\nu = 1.0$ (left), $\nu = 1.5$ (middle) and $\nu = 2.0$ (right). We use the parametrization of Φ_ν proposed by Handcock and Wallis [34], where the argument is rescaled such that the value of ν influences the shape of $\Phi_\nu(h)$ only near $h = 0$.

3.1 Simple kriging

The additional assumption with simple kriging is that $Z(x)$ is centred, i.e. $\mathrm{E}(Z(x)) = 0$, and that the second moments exist for every $x \in T$. Then the covariance function

$$K(x, y) := \mathrm{Cov}(Z(x), Z(y)) \overset{Z \text{ centred}}{=} \mathrm{E}(Z(x)Z(y)), \quad x, y \in T \tag{3.2}$$

can be defined and it follows from some basic properties of the (co)variance that K is always symmetric and positive semi-definite. In this framework K describes the *probabilistic* structure of Z, but certain 'deterministic properties' such as the smoothness of realizations are controlled by K as well (see Figure 2, created with [73]). Note however, that a complete characterization of the probabilistic structure of a random field requires additional assumptions on its distribution (e.g. assuming all finite dimensional distributions to be multivariate Gaussian). The covariance function K can be viewed as an inner product on the vector space

$$\mathscr{V}_Z := \left\{ \sum_{i=1}^{n} a_i Z(x_i),\ a_1, \ldots, a_n \in \mathbb{R},\ x_1, \ldots, x_n \in T,\ n \in \mathbb{N} \right\}$$

of second-order random variables. The closure of $\mathscr{V}_Z$ under this inner product yields a Hilbert space that is isomorphic to $\mathscr{H}_K$ from Section 2 (see [4]). When it comes to spatial interpolation, however, the natural counterpart of the Hilbert space generated by Z is the dual $\mathscr{H}_K^*$ rather than $\mathscr{H}_K$, as will become clear in the following.

The geostatistical notion of optimality is to consider as the 'best' linear predictor $Z_{u^*}(x_0)$ the random variable with minimal expected squared deviation from $Z(x_0)$, i.e.

$$\mathrm{E}((Z(x_0) - Z_{u^*}(x_0))^2) \leqslant \mathrm{E}((Z(x_0) - Z_u(x_0))^2) \quad \text{for all } Z_u(x_0) \text{ of the form (3.1).}$$

The optimal weights are then obtained by minimizing

$$\begin{aligned} \mathrm{E}((Z(x_0) - Z_u(x_0))^2) = {} & \underbrace{\mathrm{E}(Z(x_0)Z(x_0))}_{=K(x_0,x_0)} - 2\sum_{i=1}^{n} u_i(x_0) \underbrace{\mathrm{E}(Z(x_0)Z(x_i))}_{=K(x_0,x_i)} \\ & + \sum_{i=1}^{n}\sum_{j=1}^{n} u_i(x_0)u_j(x_0) \underbrace{\mathrm{E}(Z(x_i)Z(x_j))}_{=K(x_i,x_j)}. \end{aligned} \tag{3.3}$$

https://doi.org/10.1017/S0956792513000016 Published online by Cambridge University Press

This is, however, the same quadratic form $\mathcal{Q}$ as in (2.6), and so the simple kriging prediction coincides with the optimal approximation (2.5) with weights $u_1^*(x_0), \ldots, u_n^*(x_0)$ determined by (2.7). Briefly, in kriging the covariance function takes the role of the interpolation kernel, and $\mathcal{Q}$, originally introduced as the squared norm of the pointwise error functional at x_0, becomes the expected squared prediction error.

We briefly mention another perspective which is in a way even more probabilistic: the Bayesian approach. In the simple kriging framework, it is essentially the additional assumption of Z being Gaussian that needs to be made. As mentioned above, once K is fixed, the distribution of a Gaussian random field is completely determined, and the Bayesian approach uses it as *infinite-dimensional prior distribution* for the unknown function f. The posterior distribution of f at x_0 given the observations $f(x_1), \ldots, f(x_n)$ is then a Gaussian distribution with mean $s_{f,X}(x_0)$ (with weights $u_1^*(x_0), \ldots, u_n^*(x_0)$ as above) and variance $\mathcal{Q}(u^*(x_0))$. Unlike in simple kriging and kernel interpolation, this result is not based on a particular loss function. It is an immediate consequence of the complete specification of a prior distribution for f (see also [62, Section 24.]).

3.2 Ordinary and universal kriging

Especially the assumption that Z is centred seems inappropriate in most applications of geostatistics. The first generalization of simple kriging is therefore to allow for a non-zero mean function $m(x) := \mathrm{E}(Z(x))$ while still keeping the assumption that Z has second moments. The mean function is usually unknown in practice, but this problem can be bypassed by requiring the potential predictors Z_u of Z to be *unbiased*, i.e.

$$\mathrm{E}(Z(x)) = \mathrm{E}(Z_u(x)) \qquad \text{for all} \quad x \in T.$$

This has the additional advantage of preventing systematic over- or underestimation of $Z(x_0)$. Note that any such predictor is automatically unbiased if Z is centred. Using this unbiasedness constraint to recalculate the target function (3.3) one obtains

$$\begin{aligned}
\mathrm{E}((Z(x_0) - Z_u(x_0))^2) &= \mathrm{E}((Z(x_0) - \mathrm{E}(Z(x_0)) + \mathrm{E}(Z_u(x_0)) - Z_u(x_0))^2) \\
&= \mathrm{E}\left(\left(Z(x_0) - \mathrm{E}(Z(x_0)) - \sum_{i=1}^{n} u_i(x_0)\,(Z(x_i) - \mathrm{E}(Z(x_i)))\right)^2\right) \\
&= \underbrace{\mathrm{Cov}(Z(x_0), Z(x_0))}_{=K(x_0,x_0)} - 2\sum_{i=1}^{n} u_i(x_0)\,\underbrace{\mathrm{Cov}(Z(x_0), Z(x_i))}_{=K(x_0,x_i)} \\
&\quad + \sum_{i=1}^{n}\sum_{j=1}^{n} u_i(x_0)u_j(x_0)\,\underbrace{\mathrm{Cov}(Z(x_i), Z(x_j))}_{=K(x_i,x_j)},
\end{aligned}$$

which is again the quadratic form $\mathcal{Q}$ in (2.6), depending only on K but not on m. Its minimizer, however, is in general not the same as above, since the additional unbiasedness constraint

$$m(x) = \sum_{i=1}^{n} u_i(x)\, m(x_i) \qquad \text{for all} \quad x \in T \tag{3.4}$$

https://doi.org/10.1017/S0956792513000016 Published online by Cambridge University Press

restricts the choice of weights. To ensure that condition (3.4) can be satisfied at all, one cannot let the mean function completely general, but must assume a sufficiently simple, finite dimensional model. The simplest model is a constant (but unknown) mean function, and this assumption leads to what is called *ordinary kriging*. This is, however, just a special case of *universal kriging* where the mean function is modelled as

$$m(x) := \sum_{k=1}^{q} \beta_k p_k(x), \qquad x \in T, \tag{3.5}$$

with known and linear independent functions $p_1, \ldots, p_q$, and unknown coefficients $\beta_1, \ldots, \beta_q$. Such a mean function is also called a *trend*, and condition (3.4) becomes

$$\sum_{k=1}^{q} \beta_k p_k(x) = \sum_{k=1}^{q} \beta_k \sum_{i=1}^{n} u_i(x)\, p_k(x_i) \qquad \text{for all} \quad x \in T.$$

This condition must hold for any set of coefficients $\beta_1, \ldots, \beta_q$, and so when predicting at x_0 we are back to condition (2.12) restricting the weights $u_1(x_0), \ldots, u_n(x_0)$. It follows that the universal kriging prediction coincides with the optimal approximation (2.5) from the conditionally positive definite kernel interpolation setup, with optimal weights $u_1^*(x_0), \ldots, u_n^*(x_0)$ determined by (2.12) and (2.13). Representation (2.14) was already noted by Matheron [51], and the corresponding equation system is known as *dual kriging*.

The universal kriging interpolant can also be derived within a Bayesian framework. As a prior distribution for f, one assumes a Gaussian random field with covariance function K and mean function m as in (3.5). For a complete specification of the prior for f, a distribution assumption needs to be made for $\beta_1, \ldots, \beta_q$ as well. Then the posterior distribution of f can be worked out, but it will depend both on K and on the prior distribution of the trend coefficients. In the special case of a flat (uninformative) prior, however, Omre and Halvorsen [59] show that the posterior of f at x_0 is a Gaussian distribution with mean $s_{f,X}(x_0)$ and variance $\mathcal{Q}(u^*(x_0))$, both calculated with the optimal weights from the universal kriging approach.

Universal kriging and kernel interpolation with conditional positive definite kernels are formally equivalent and are derived from the same loss function $\mathcal{Q}$. Nevertheless, the analogy is not yet perfect because the universal kriging assumption that $Z(x)$ has second moments for every $x \in T$ automatically entails positive (semi)definiteness of K. We therefore consider a slightly different stochastic model leading to kriging interpolants of the same form, but using a more general dependence structure that permits the use of conditionally positive definite kernels.

3.3 Intrinsic kriging

The idea with intrinsic random fields (introduced in [52]) is that one no longer specifies the full second-order structure of Z, but only the dependence structure of certain increments. More specifically, let $\mathcal{P}$ again be a finite dimensional space of functions on T, and let $L_{\mathcal{P}}(T)$ be as in Section 2, i.e. the space of functionals of the form $\lambda_{\mathcal{X}} = \sum_{i=1}^{n} a_i \delta_{x_i}$ with

https://doi.org/10.1017/S0956792513000016 Published online by Cambridge University Press

$\lambda_{\mathscr{X}}(p) = 0$ for all $p \in \mathscr{P}$. For every such $\lambda_{\mathscr{X}} \in L_{\mathscr{P}}(T)$

$$Z_{\lambda,\mathscr{X}} := \lambda_{\mathscr{X}}(Z) = \sum_{i=1}^{n} a_i Z(x_i)$$

is called an allowable linear combination of Z with respect to $\mathscr{P}$. Assume that all allowable linear combinations of Z have second moments and are centred, i.e. $\mathrm{E}(Z_{\lambda,\mathscr{X}}) = 0$. The function $K : T \times T \to \mathbb{R}$ is then called a *generalized covariance function* of Z if

$$\mathrm{E}(Z_{\lambda,\mathscr{X}} Z_{\lambda,\mathscr{X}}) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K(x_i, x_j) \qquad \text{for all } \lambda_{\mathscr{X}} \in L_{\mathscr{P}}(T).$$

We note that Z and $Z + p$ have the same generalized covariance function for any $p \in \mathscr{P}$. Moreover, since the expectation of any squared random variable is non-negative, K must be conditionally positive semi-definite with respect to $\mathscr{P}$.

The most important case in practice is the case where $\mathscr{P} = \pi_m(\mathbb{R}^d)$, the space of polynomials of order at most m, and where Z is *intrinsically (weakly) stationary of order m*, i.e. $K(x, y) = \Phi(y - x)$ for some function $\Phi : \mathbb{R}^d \to \mathbb{R}$ that is conditionally positive semi-definite with respect to $\pi_m(\mathbb{R}^d)$. Assuming K to be translation invariant is often reasonable because general dependence structures are usually too complex for reliable inference. Note that $\lambda_{\mathscr{X}} \in L_{\pi_m}(\mathbb{R}^d)$ implies $\lambda_{\mathscr{X}+x} \in L_{\pi_m}(\mathbb{R}^d)$ for all $x \in \mathbb{R}^d$, owing to the binomial formula, and hence all random variables

$$Z_{\lambda,\mathscr{X}+x} = \sum_{i=1}^{n} a_i Z(x_i + x), \qquad x \in \mathbb{R}^d$$

are allowable if only $\lambda_{\mathscr{X}} \in L_{\pi_m}(\mathbb{R}^d)$. It follows that the random field $\{Z_{\lambda,\mathscr{X}+x} : x \in \mathbb{R}^d\}$ of mth-order increments is *weakly stationary* (i.e. centred with second moments and translation-invariant covariance function) with

$$\mathrm{E}(Z_{\lambda,\mathscr{X}+x} Z_{\lambda,\mathscr{X}+y}) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \Phi(y - x + x_j - x_i) =: \Phi_{\lambda,\mathscr{X}}(y - x).$$

Whenever in practice one faces the situation that Z itself does not appear to be weakly stationary, there is still a chance that this seems plausible for some higher order increments, and this then motivates the modelling with intrinsically stationary random fields. A detailed introduction to this topic is given in [9, Chapter 4], and in particular the differences from a modelling perspective to the model underlying the universal kriging approach are illustrated excellently.

To predict $Z(x_0)$ we consider the error functional $\lambda_{\mathrm{err}} = \delta_{x_0} - \lambda_{u(x_0)}$ as in Section 2, but in the stochastic framework we are interested in the expected squared prediction error. When K is the generalized covariance function of Z, we have by definition

$$\mathrm{E}((\lambda_{\mathrm{err}}(Z))^2) = K(x_0, x_0) - 2\sum_{i=1}^{n} u_i(x_0) K(x_0, x_i) + \sum_{i=1}^{n}\sum_{j=1}^{n} u_i(x_0) u_j(x_0) K(x_i, x_j),$$

https://doi.org/10.1017/S0956792513000016 Published online by Cambridge University Press

which is again the quadratic form $\mathcal{Q}$ in (2.6). The requirement $\lambda_{\mathrm{err}} \in L_{\mathcal{P}}(T)$ again entails condition (2.12), and so it follows that the intrinsic kriging prediction is identical to the optimal approximation (2.5) in the conditionally positive definite kernel interpolation setup with optimal weights $u_1^*(x_0), \ldots, u_n^*(x_0)$ determined by the equation systems (2.13) and (2.12).

The difference to universal kriging lies in the interpretation of representation (2.14). While it is legitimate in universal kriging to interpret the second term in (2.14) as an approximate mean function and the first term as an approximate deviation from it, such an interpretation would be wrong in intrinsic kriging where a mean function is not even defined.

3.4 Comparison with the deterministic framework

We sum up what has been pointed out so far concerning the two different modelling approaches and the corresponding notions of optimality. In both frameworks we seek to minimize, at every fixed location $x_0 \in T$, the quadratic form

$$\mathcal{Q}(u(x_0)) = K(x_0, x_0) - 2\sum_{i=1}^{n} u_i(x_0)K(x_0, x_i) + \sum_{i=1}^{n}\sum_{j=1}^{n} u_i(x_0)u_j(x_0)K(x_i, x_j),$$

possibly subject to some additional restrictions. The sense in which the resulting approximation of f at x_0 is optimal then differs according to the interpretation of $\mathcal{Q}$:

(1) In the deterministic framework, $\mathcal{Q}(u(x_0))$ indicates how well δ_{x_0} can be approximated by the linear combination $\lambda_{u(x_0)}$ of the point evaluation functionals for points of X. It measures how big the approximation error can be in the *worst case* assuming only that $f \in \mathcal{H}_{K,\mathcal{P}}$.

(2) In the stochastic framework, $\mathcal{Q}(u(x_0))$ indicates how big the error for approximating a random field Z at x_0 by $\lambda_{u(x_0)}(Z)$ will be *on average*. Its calculation is based on the assumption that Z has generalized covariance function K.

Both worst-case and average-case behaviour of numerical algorithms are studied and compared by Ritter [64]. For his average-case analysis, Ritter adopts the stochastic perspective and specifies a probability measure on the space of all functions by making the geostatistical assumption of a random field Z with covariance kernel K. The average-case optimality of K-splines that is stated in this monograph is then a consequence of the equivalence of kriging and kernel interpolation. A short list with terminology used in kernel interpolation and geostatistics is provided in Table 3.

We note that in situations where both frameworks are applicable, different answers are obtained as to the question of which K should be used. Indeed, assume that f is a realization of a centred random field Z on $\mathbb{R}^d$ with translation invariant covariance function Φ_τ whose Fourier transform $\widehat{\Phi}_\tau$ satisfies (2.4). For this model the simple kriging framework applies, and states that the optimal interpolant is obtained with Φ_τ. On the other hand, Scheuerer [71] shows that the realizations of Z- and hence f- are in the Sobolev space $W_2^\mu(\mathbb{R}^d)$ if and only if $\mu < \tau - \frac{d}{2}$. The space $W_2^\mu(\mathbb{R}^d)$, however, calls for some reproducing kernel Φ_μ satisfying (2.4) with μ instead of τ (see Theorem 2.1), and so

https://doi.org/10.1017/S0956792513000016 Published online by Cambridge University Press

Table 3. *Some frequently used terms in the language of statistics (left column) and numerical analysis (right column)*

Covariance function	Symmetric and positive definite kernel
∼ of a weakly stationary random field	Translation-invariant kernel
∼ of a w. stat. and isotropic random field	Radially symmetric, translation-invariant kernel
Kriging	Kernel interpolation
Intrinsic kriging	∼ with a conditionally positive definite kernel
Kriging variance $P^2_{K,X}$	Power function $P_{K,X}$

a numerical analyst would rather use Φ_μ with $\mu \approx \tau - \frac{d}{2}$ for interpolation. In other words: if worst case optimality is aspired, then a rougher kernel is considered appropriate for the same function f than when the aim is average-case optimality.

4 Generalizations and related problems

The spatial interpolation problem discussed in this paper can be generalized by considering arbitrary functionals $\lambda_0(f), \lambda_1(f), \ldots, \lambda_n(f)$ instead of $f(x_0), f(x_1), \ldots, f(x_n)$. Such a generalization covers, for example, the situation where f is to be reconstructed from both function values and derivatives ('Hermite–Birkhoff interpolation'), or the case where the interest is in approximating integrals of f over certain sub-domains of T. The latter is an important problem in mining, where measurements (which may also effectively be integrals over small areas) are taken at certain locations in some ore deposit, and the interest is in predicting the overall content.

In numerical analysis, approximation with general functionals goes back to [87]. In the positive definite setup, the considered functionals must be in the dual space $\mathscr{H}_K^*$, the analogue requirement with simple kriging is that $\lambda_0(Z), \lambda_1(Z), \ldots, \lambda_n(Z)$ are all well-defined random variables and have second moments. If these conditions are met, the generalization

$$\mathscr{Q}(u(\lambda_0)) = \big(\lambda_0^1\lambda_0^2\big)(K) - 2\sum_{i=1}^n u_i(\lambda_0)\big(\lambda_0^1\lambda_i^2\big)(K) + \sum_{i=1}^n\sum_{j=1}^n u_i(\lambda_0)u_j(\lambda_0)\big(\lambda_i^1\lambda_j^2\big)(K),$$

of the quadratic form (2.6) is well defined, and its minimization yields optimal weights $u_1^*(\lambda_0), \ldots, u_n^*(\lambda_0)$ for the approximation of $\lambda_0(f)$ by

$$s_{f,X}(\lambda_0) = \sum_{i=1}^n u_i(\lambda_0)\lambda_i(f).$$

In the framework of conditionally positive definite kernels (intrinsic kriging), one needs to require that $(\lambda_i^1\lambda_j^2)(K)$ and $\lambda_i(p_k)$ are well defined for all $i, j \in \{0, \ldots, n\}$ and $k \in \{1, \ldots, q\}$, and condition (2.12) becomes

$$\lambda_0(p_k) = \sum_{i=1}^n u_i(\lambda_0)\lambda_i(p_k), \qquad k = 1, \ldots, q.$$

https://doi.org/10.1017/S0956792513000016 Published online by Cambridge University Press

The quadratic form $\mathcal{Q}$ from above is then again well defined, and its minimization is straightforward. Wendland [85, Section 16.2] discusses the case of Hermite–Birkhoff interpolation in more detail and shows that the resulting system of equations has a unique solution if $\lambda_1, \ldots, \lambda_n$ are linearly independent, and if $\lambda_i(p) = 0$ for all $p \in \mathcal{P}$ and $i \in \{1, \ldots, n\}$ implies that $p = 0$. The close link between approximation and integration of f and its consequences on the error analysis of these problems is discussed in [64]. In the geostatistical framework, both kriging with gradient information and kriging of block averages are described in [9]. The former method has been applied to meteorological problems involving several variables related by physical laws [7, 8].

A different type of generalization of the setup in Sections 2 and 3 is required when the set X is infinite. One could think, for example, of data from moving tracking devices which measure f continuously along one-dimensional trajectories. $\lambda_{u(x_0)}$ itself is then no longer a finite linear combination of point evaluation functionals, but lies in the closure of such functionals, i.e.

$$\lambda_{u(x_0)} \in \overline{\left\{ \lambda_{\mathcal{X}}, \mathcal{X} \overset{\text{finite}}{\subset} X,\ \lambda_{\mathcal{X}}(p) = p(x_0) \text{ for all } p \in \mathcal{P} \right\}} =: \mathcal{F}_{\mathcal{P},X,x_0}.$$

The definition of $\mathcal{Q}(u(x_0)) := \|\delta_{x_0} - \lambda_{u(x_0)}\|_K^2$ remains unchanged and the functional $\lambda_{u^*(x_0)}$ defining the optimal interpolant is obtained as the projection of δ_{x_0} on $\mathcal{F}_{\mathcal{P},X,x_0}$ [69]. In the same way, the kriging approximation $Z_{u^*}(x_0)$ of $Z(x_0)$ based on the values of Z at all points of X is obtained as the projection of $Z(x_0)$ on the space

$$\overline{\left\{ \lambda_{\mathcal{X}}(Z), \mathcal{X} \overset{\text{finite}}{\subset} X,\ \lambda_{\mathcal{X}}(p) = p(x_0) \text{ for all } p \in \mathcal{P} \right\}},$$

where the closure is under the inner product induced by the generalized covariance function. While this generalization is straightforward in theory, it is not clear how a solution can be obtained in practice. Matheron [52] studies a special case where $Z_u(x_0)$ can be represented by a measure μ_{x_0} with support in X, i.e.

$$Z_u(x_0) = \int_T Z(x)\, \mu_{x_0}(dx) = \int_X Z(x)\, \mu_{x_0}(dx).$$

Such a representation is not always possible, and a counterexample in [52] shows that smooth covariance functions, such as the Gaussian model, may lead to unsolvable systems. If, however, the optimal solution can be represented in that way, then the system of equations (2.12) and (2.13) becomes a system of integral equations

$$\int_X K(x,y)\, \mu_{x_0}(dy) + \sum_{k=1}^{q} \eta_k^*(x_0) p_k(x) = K(x_0, x) \qquad \text{for all } x \in X,$$

$$\int_X p_k(x)\, \mu_{x_0}(dx) = p_k(x_0), \qquad k = 1, \ldots, q$$

that determine μ_{x_0}. At least in special cases, for example when K is such that Z has the Markov property and the geometry of X is sufficiently simple, closed-form solutions for μ_{x_0} can be obtained from these equations.

https://doi.org/10.1017/S0956792513000016 Published online by Cambridge University Press

A mathematical problem that is closely related to the spatial interpolation problem discussed in this paper arises when the function f has to be reconstructed based on data

$$y_i := f(x_i) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{4.1}$$

i.e. where f is observed with measurement errors $\epsilon_1, \ldots, \epsilon_n$. In this situation the aim is no longer interpolation of the data, but rather approximation by a function that is close to the noisy observations (4.1) on the one hand but still reasonably smooth on the other hand. This is the standard setup in machine learning, and depending on the loss function that is used to assess the fidelity to the data, different approaches turn out to be optimal. Schaback and Wendland [70] discuss the role of kernel methods in machine learning. The approach that is most closely related to the kernel interpolants ('splines') in Section 2 is the concept of *smoothing splines*, which corresponds to a quadratic loss function. More specifically, instead of minimizing the native space norm among all functions interpolating the data, smoothing splines minimize

$$\sum_{i=1}^{n} (y_i - s_{f,X}(x_i))^2 + \lambda \, \| s_{f,X} \|^2_{\mathscr{H}_{K,\mathscr{P}}}, \tag{4.2}$$

where λ is a regularization parameter that controls the fit/smoothness trade-off mentioned above. The universal kriging counterpart of (4.2) is obtained when f is assumed to be a realization of a random field Z with covariance function K, and $\epsilon_1, \ldots, \epsilon_n$ are assumed to be realizations of independent, centred random variables with variance λ. The optimal solution in that case still has the form (2.5), but in the system of equation (2.13) defining the kriging weights, $K(x_i, x_i)$ is replaced by $(K(x_i, x_i) + \lambda)$ for all $i \in \{1, \ldots, n\}$. For a detailed description of smoothing splines, their connection to RKHS on the one hand and geostatistical methods on the other, we refer to [83] and references therein.

5 Error estimates

In both frameworks the magnitude of the approximation error at $x_0 \in T$ is characterized by $\mathscr{Q}$. In the literature on kernel interpolation, the value of $Q^{1/2}$ for approximation with the optimal weights $u_1^*(x_0), \ldots, u_n^*(x_0)$ is denoted by

$$P_{K,X}(x_0) := Q^{1/2}(u^*(x_0))$$

and is called the *power function*. The definition of $Q^{1/2}$ immediately implies the error bound

$$|f(x_0) - s_{f,X}(x_0)| \leqslant P_{K,X}(x_0) \cdot |f|_{\mathscr{H}_{K,\mathscr{P}}}, \qquad x_0 \in T. \tag{5.1}$$

The function f is unknown but fixed, and so in order to control the approximation error one is interested in quantifying how $P_{K,X}$ depends on K and X. Conversely, a bound on the absolute approximation error at x_0 normalized by $|f|_{\mathscr{H}_{K,\mathscr{P}}}$ that holds for any $f \in \mathscr{H}_{K,\mathscr{P}}$ implies a bound on $P_{K,X}(x_0)$.

https://doi.org/10.1017/S0956792513000016 Published online by Cambridge University Press

In geostatistics too $P_{K,X}(x_0)$ is a well-known quantity: its square is the so-called *kriging variance*. Stein's book [77] is the main reference for an in-depth study of the asymptotic behaviour of kriging interpolants. In Section 3.6 he considers a centred, weakly stationary random field Z on $\mathbb{R}$ with covariance function Φ_τ satisfying (2.4). The simple kriging interpolant $Z_{u^*}(0)$ at $x_0 = 0$ is calculated based on the values of Z at $\pm\delta, \pm 2\delta, \ldots$ (interpolation problem) or at $-\delta, -2\delta, \ldots$ (extrapolation problem). In both cases it turns out that the kriging variance can be bounded by

$$\mathrm{E}((Z(0) - Z_{u^*}(0))^2) \leqslant C\,\delta^{2\tau-1} \qquad \text{as } \delta \to 0.$$

This result is very useful to understand the impact of the smoothness of Z on the precision of kriging approximations, but the geometric setup is rather special. In the kernel interpolation literature similar statements exist for very general alignment of points where f is observed. As a consequence of the coincidence of $P_{K,X}^2$ with the kriging variance, these results can be easily translated to the stochastic framework.

To characterize the density of locations in X without restricting to lattice data, one defines the *fill distance*

$$h_{X,T} := \sup_{x \in T} \min_{x_j \in X} \|x - x_j\|.$$

Intuitively, $h_{X,T}$ is the radius of the largest ball centred at some $x \in T$ that does not contain any of the data points. Results will be given for approximation on a bounded domain $T \subseteq \mathbb{R}^d$ with Lipschitz boundary, which means that the boundary can be thought of as locally being the graph of a Lipschitz continuous function. Moreover, T must satisfy an *interior cone condition*, i.e. there exists an angle $\theta \in (0, \frac{\pi}{2})$ and a radius $r > 0$ such that for every $x \in T$ a unit vector $\xi(x)$ exists such that the cone

$$C(x, \xi(x), \theta, r) := \{x + \lambda u : u \in \mathbb{R}^d,\ \|u\| = 1, u'\xi(t) \geqslant \cos\theta, \lambda \in [0, r]\}$$

is contained in T. This simply means that there exists a cone of fixed size that can be placed everywhere inside T, thus excluding the possibility of extremely narrow bulges of the boundary. We state two results from the kernel interpolation literature (see [85, Section 11.6] and [58, Section 4]) and formulate their consequences for kriging:

Theorem 5.1 *Suppose that $T \subset \mathbb{R}^d$ is a bounded domain, has a Lipschitz boundary, and satisfies an interior cone condition. Let $X \subset T$ be a given discrete set and $s_{f,X}$ be the kernel interpolant based on a translation invariant and positive definite kernel Φ_τ satisfying (2.4) with $\tau = k + s$, where $k > \frac{d}{2}$ is a positive integer and $0 \leqslant s < 1$. Then the error between $f \in W_2^\tau(T)$ and its interpolant $s_{f,X}$ can be bounded by*

$$|f(x) - s_{f,X}(x)| \leqslant C\, h_{X,T}^{\tau - \frac{d}{2}}\, \|f\|_{W_2^\tau(T)}, \quad x \in T,$$

for all sufficiently dense sets X.

Corollary 1 *Let Z be a centred, weakly stationary random field on a bounded domain $T \subset \mathbb{R}^d$ with covariance function Φ_τ as in Theorem 5.1. Assume that T has a Lipschitz boundary*

https://doi.org/10.1017/S0956792513000016 Published online by Cambridge University Press

and satisfies an interior cone condition. Then the kriging variance can be bounded by

$$\mathrm{E}((Z(x) - Z_{u^*}(x))^2) \leqslant C\, h_{X,T}^{2\tau-d}, \quad x \in T,$$

for all sufficiently dense sets X.

Theorem 5.2 *Suppose that $T \subset \mathbb{R}^d$ is a bounded domain that satisfies an interior cone condition. Consider the thin-plate splines $\Phi_{d,l}$ from Table 2 as conditionally positive definite with respect to $\pi_{l-1}(T)$. Then the error between $f \in W_2^l(T)$ and its thin-plate spline interpolant $s_{f,X}$ can be bounded by*

$$|f(x) - s_{f,X}(x)| \leqslant C\, h_{X,T}^{l-\frac{d}{2}}\, |f|_{BL_l(T)}, \quad x \in T,$$

for all sufficiently dense sets X.

Corollary 2 *Let Z be an intrinsically stationary random field of order l on a bounded domain $T \subset \mathbb{R}^d$ with generalized covariance function $\Phi_{d,l}$ as in Table 2. Assume that T satisfies an interior cone condition. Then the kriging variance can be bounded by*

$$\mathrm{E}((Z(x) - Z_{u^*}(x))^2) \leqslant C\, h_{X,T}^{2l-d}, \quad x \in T,$$

for all sufficiently dense sets X.

The preceding Corollaries give rates for the speed of decline of the kriging variance as the data become denser. Corollary 1 is in agreement with, but more general than the result from [77, Section 3.6] mentioned above. To our knowledge, there is no result on convergence rates of kriging predictions in the statistical literature that covers geometric setups of data points with the generality of Corollaries 1 and 2. We refer to [86] for generalizations of Theorems 5.1 and 5.2 to the situation (4.1) where f is observed with measurement error.

6 Interpolation with misspecified kernels

So far it has always been assumed that the correct K is known. In geostatistics this means that the covariance structure of the random field under study is known, in kernel interpolation it amounts to the assumption that one knows the native space in which f is contained. In practice, however, such knowledge is usually not available, and so the question arises whether the interpolation schemes discussed above are still near-optimal if an 'incorrect' kernel $\tilde{K}$ is used instead of K.

In kernel interpolation, the main interest is to ensure that the optimal rates in Theorems 5.1 and 5.2 are maintained. If this is the only goal, then rescaling the argument of Φ does not have an effect because in (2.4) rescaling only changes the constants, and thin-plate spline interpolants are invariant to rescaling of the argument of $\Phi_{d,l}$ anyway. Misspecifying the smoothness of Φ, however, does have an effect on the rate. If a kernel $\Phi_{\tilde{\tau}}$ with $\tilde{\tau} < \tau$ is used in the setup of Theorem 5.1, then the statement remains valid for the lower rate of $\tilde{\tau} - \frac{d}{2}$. For $\tilde{\tau} > \tau$ on the contrary, it cannot be guaranteed that $f \in W_2^{\tilde{\tau}}(T)$, and so

https://doi.org/10.1017/S0956792513000016 Published online by Cambridge University Press

Theorem 5.1 does not apply. Under the additional condition that the separation radius

$$q_X := \frac{1}{2} \min_{j \neq k} \|x_j - x_k\|$$

does not decline faster than $h_{X,T}$, however, it can be shown that the error rate is of the same order as if a kernel with the correct degree of smoothness was used [58]. Hence, for quasi-uniform sets X, i.e.

$$q_X \leqslant h_{X,T} \leqslant c q_X$$

for some fixed constant $c > 0$, using a very smooth kernel does not degrade the approximation accuracy of $s_{f,X}$ with respect to the error rate. The power function, however, is independent of the true smoothness of f, thus decreases with the faster rate of $\tilde{\tau} - \frac{d}{2}$, and consequently yields a false description of the magnitude of approximation errors.

The last point is not considered a big deficiency in kernel interpolation, but in geostatistics the exact quantification of the approximation error plays an important role, and a different perspective has been adopted here. A major step towards a theoretically founded answer to the kernel misspecification issue was made in [75]: If K and $\tilde{K}$ are *compatible*, then the approximation based on $\tilde{K}$ will have the same asymptotic efficiency as the optimal approximation, and the relative deviation of the true expected squared approximation error from the one calculated under the false assumption that $\tilde{K}$ is correct is asymptotically negligible. A full explanation of the concept of compatibility is beyond the scope of this paper, for details consider [38, 77, 78]. To compare with the statement above, we shall however give a sufficient condition for compatibility in an important special case where

$$K(x, y) = \frac{\sigma^2 (a \|y - x\|)^\nu}{2^{\nu-1} \Gamma(\nu)} \mathcal{K}_\nu(a \|y - x\|), \qquad x, y \in T, \quad \sigma, a, \nu > 0, \tag{6.1}$$

i.e. K is translation-invariant, radially symmetric and of the Matérn type (see Table 1). In addition to the parameter ν controlling the smoothness of K, we consider a parameter a rescaling the argument, and a variance parameter σ that does not affect the interpolant $s_{f,X}$ but scales the power function. This choice of K satisfies (2.4) with $\tau = \nu + \frac{d}{2}$ for any value of σ and a. When K has the above form and $d \leqslant 3$, compatibility of K and $\tilde{K}$ is guaranteed [88] if

$$\tilde{\nu} = \nu \quad \text{and} \quad \tilde{\sigma}^2 \tilde{a}^{2\nu} = \sigma^2 a^{2\nu}. \tag{6.2}$$

This still allows for certain deviations of $\tilde{K}$ from K, but limits the choice of $\tilde{K}$ much more than the condition $\tilde{\nu} \geqslant \nu$ that ensures optimal rates of the approximation error. Note the dependence of the above condition on the space dimension. For $d \geqslant 5$, condition (6.2) is no longer sufficient, and K and $\tilde{K}$ are compatible only in the trivial case where $\tilde{K} = K$ [1]. The case $d = 4$ is still open. To formulate the precise statement of [75], consider a random field Z on a bounded domain T with mean function of the form (3.5) and covariance function K. Let $Z_{\tilde{K}}(x_0)$ be the kriging prediction at $x_0 \in T$ based on observations of Z at some set $X_n \subset T$, derived under the (false) assumption that $\tilde{K}$ is the covariance function. Assume further that $x_0 \notin X_n$, the sequence $(X_n)_{n \in \mathbb{N}}$ of point sets is getting dense in T and

$$\mathrm{E}_K((Z_K(x_0) - Z(x_0))^2) \to 0$$

https://doi.org/10.1017/S0956792513000016 Published online by Cambridge University Press

as n tends to infinity, where E_K denotes the expectation under K. Then it holds, for any compatible covariance function $\tilde{K}$, that

$$\frac{\mathrm{E}_K((Z_{\tilde{K}}(x_0) - Z(x_0))^2)}{\mathrm{E}_K((Z_K(x_0) - Z(x_0))^2)} \longrightarrow 1 \quad \text{and} \quad \frac{\mathrm{E}_{\tilde{K}}((Z_{\tilde{K}}(x_0) - Z(x_0))^2)}{\mathrm{E}_K((Z_{\tilde{K}}(x_0) - Z(x_0))^2)} \longrightarrow 1 \tag{6.3}$$

as n tends to infinity. The convergence is even uniform on T [77]. Recall that

$$\mathrm{E}_K((Z_{\tilde{K}}(x_0) - Z(x_0))^2) = \mathcal{Q}_K(u^*_{\tilde{K}}(x_0) \quad \text{and} \quad \mathrm{E}_K((Z_K(x_0) - Z(x_0))^2) = P^2_{K,X}(x_0),$$

where the subscripts K and $\tilde{K}$ denote for $\mathcal{Q}$ that the quadratic form is calculated using K and $\tilde{K}$ respectively, and for u^* the optimal weights were obtained by minimizing $\mathcal{Q}_K$ and $\mathcal{Q}_{\tilde{K}}$ respectively. In the language of numerical analysis, (6.3) says that asymptotically the interpolant obtained with a compatible kernel $\tilde{K}$ is still optimal, and that the power function calculated with $\tilde{K}$ tends to the 'true' power function $\mathcal{Q}_K(u^*_{\tilde{K}}(x_0)$. This statement is much stronger than that of an optimal convergence rate, but it is based on more restrictive assumptions like (6.2).

An extension of this result to some conditionally positive definite covariance functions is proved in [78]. Putter and Young [60] consider the setting where $\tilde{K}$ is not fixed but may depend on n, which accommodates the situation in practice where $\tilde{K}_n$ can be estimated from the data at X_n (see Section 7) with increasing precision as n tends to infinity. This convergence is formalized by introducing the concept of *contiguity* (which replaces compatibility, see [60] for definition), and it is shown that (6.3) still holds if the stochastic models corresponding to the sequence $(\tilde{K}_n)_{n\in\mathbb{N}}$ on the one hand and the true covariance function K on the other hand are contiguous.

7 Kernel selection and parameter estimation

An immediate question to follow up the issue of kernel misspecification is how to identify the 'correct' K based on the information and data at hand. We do not intend to give a comprehensive list of all methods available, but focus on two methods that are applicable in both deterministic and stochastic frameworks.

The issue of kernel selection has received comparatively little attention in the framework of kernel interpolation. This is not surprising in the light of the preceding section where we noted that working with some smooth kernel would always guarantee optimal convergence rates whatever be the particular form (and scaling) of this kernel, provided that the sampling locations are quasi-uniform. Consequently, more emphasis was put on the study of good configurations of sampling locations [14, 16, 39] on the one hand, and edge correction strategies (see [26] for an overview) on the other hand to avoid undesired oscillations near the boundaries that often come with smooth and flat kernels. Nevertheless, several authors [6,27,28,63] have pointed out the big impact of the choice of, for example, the scaling parameter on the accuracy of interpolant. When ill-conditioning (see Section 8) is not an issue for a relevant range of parameter values, there is usually a value that minimizes interpolation errors.

In the earlier literature, the question of suitable scaling of kernel has been typically solved by ad hoc rules [25, 28, 35]. Rippa [63] was the first to propose an algorithm

https://doi.org/10.1017/S0956792513000016 Published online by Cambridge University Press

based on the idea of leave-one-out cross validation (LOOCV) which chooses the scale parameter such that some norm of the LOOCV error vector ε is minimized. In the kernel interpolation setup, the components of ε are formed by leaving out one sampling location x_i at a time, calculating the interpolant based on the remaining ones only and taking ε_i to be the difference between the true value $f(x_i)$ and the approximation $s_{f,X}(x_i)$. This procedure yields good choices of the scaling parameter and can be implemented such that the calculation of ε for a given kernel can be done with the computational cost of order $O(n^3)$, where n is the number of sampling locations. A more recent paper [24] discusses extensions of Rippa's algorithm that have been applied in the context of an iterated approximate moving least-squares approximation of function value data and RBF (radially symmetric kernels) pseudo-spectral methods for the solution of partial differential equations.

LOOCV does not make any explicit modelling assumptions and is therefore also applicable in the geostatistical framework. In the geostatistical literature, however, cross validation is mainly used as a diagnostic tool to compare the performances of geostatistical models. Traditionally, variogram-based estimation methods have been used (see e.g. [11, Sections 2.4–2.6] or [9, Chapter 2] for details) since an estimate of the variogram

$$\gamma(h) := \frac{1}{2}\mathrm{E}((Z(x+h)-Z(x))^2) \qquad \text{(assuming that } Z \text{ is stationary)}$$

usually constitutes the first step in the exploratory analysis of geostatistical data.

Here we focus on *maximum likelihood estimation* [49], which is applicable in all of the kriging setups presented above, and makes optimal use of the information contained in the data [37, Chapter 2 and Theorem 8.1]. It is usually derived under the additional modelling assumptions that Z is a Gaussian random field, and that K belongs to some parametric class $\{K_\theta : \theta \in \Theta\}$ of covariance models. In the simple case where Z has a zero mean, the log likelihood function, i.e. the logarithm of the probability density function of the random vector $(Z(x_1),\ldots,Z(x_n))'$ evaluated with the data vector $\mathrm{f} := (f(x_1),\ldots,f(x_n))'$ is then given by

$$l(\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|A_\theta|) - \frac{1}{2}\mathrm{f}'A_\theta^{-1}\mathrm{f}, \qquad \theta \in \Theta,$$

with A_θ as defined below and $|A_\theta|$ denoting its determinant. The maximum likelihood estimator then chooses the parameter that maximizes $l(\theta)$, reasoning that under the corresponding stochastic model observing the data f becomes most likely. An extension that works for both the case of a non-trivial mean of the form (3.5) and the case of a generalized covariance function was proposed by Kitanidis [43]. The idea is to use the information of $n-q$ allowable linear combinations of f only, rather than the complete data vector. In the universal kriging setup this causes the mean function to be filtered out from the data. This procedure is called *restricted maximum likelihood (REML)* estimation, and it can be shown [36] that the restricted log likelihood function can be written as

$$\begin{aligned} l(\theta) = &-\frac{n-q}{2}\log(2\pi) - \frac{1}{2}\log(|A_\theta|) - \frac{1}{2}\log\left(\left|P'A_\theta^{-1}P\right|\right) \\ &+\frac{1}{2}\log(|P'P|) - \frac{1}{2}\mathrm{f}'\left(A_\theta^{-1} - A_\theta^{-1}P\left(P'A_\theta^{-1}P\right)^{-1}P'A_\theta^{-1}\right)\mathrm{f}, \qquad \theta \in \Theta, \end{aligned}$$

https://doi.org/10.1017/S0956792513000016 Published online by Cambridge University Press

with A_θ and P by

$$A_\theta = \begin{pmatrix} K_\theta(x_1, x_1) \cdots K_\theta(x_1, x_n) \\ \vdots \quad \ddots \quad \vdots \\ K_\theta(x_n, x_1) \cdots K_\theta(x_n, x_n) \end{pmatrix}, \qquad P = \begin{pmatrix} p_1(x_1) \ \cdots p_q(x_1) \\ \vdots \quad \ddots \quad \vdots \\ p_1(x_n,) \cdots p_q(x_n) \end{pmatrix}.$$

An elementary introduction to maximum likelihood methods in spatial statistics is given in [44]. A major drawback seems to be the strong assumption that Z is Gaussian under which the maximum likelihood estimator is derived. In [72], however, an alternative derivation of REML in the framework of kernel interpretation (where much weaker modelling assumptions are made) is given, and a numerical study with several non-stochastic test cases is presented in which REML often yields very good choices of K.

Within the Bayesian paradigm, parameter selection and interpolation are not formally distinct. The full specification of a probabilistic model permits, via Bayes' Theorem, to obtain posterior distributions for the unobserved values of f, the trend parameters $\beta_1, \ldots, \beta_q$ and the covariance parameter θ. One could even step up yet another level and let the Bayesian methodology choose between different parametric model structures [62, Section 5.2]. This unified treatment of model selection and interpolation has the advantage that the additional uncertainty due to the fact that the data-generating model is unknown is reflected in posterior distributions. These distributions can, however, in general not be stated in closed form. For certain choices of the priors, some of the integrals that result from the repeated applications of Bayes' Theorem within the hierarchical model specification can be calculated analytically [17, 33], but the final posterior distributions usually require numerical approximations or Markov chain Monte Carlo (MCMC) methods.

In the situation (4.1) where only noisy observations of f are available, the main focus is on estimating the regularization parameter λ in (4.2). Wahba [82] discusses a generalized cross-validation (GCV) procedure which has the advantage over standard LOOCV that it achieves certain desirable invariance properties (see [83] for a detailed motivation and asymptotic results for GCV). While Stein [76] proves that REML is asymptotically (as the sampling locations get increasingly dense) superior to GCV when the geostatistical assumptions are true, asymptotic results from Wahba [82] suggest that REML can fail when f is a smooth deterministic function, whereas GCV chooses a good λ in all frameworks. The following example, however, shows that a rather different behaviour may be observed in our interpolation framework and finite settings.

We illustrate and compare LOOCV and REML with a test function (the 'borehole model') used by many authors (e.g. [40, 56]) to compare different methods in computer experiments. Examples from this field of application are particularly interesting in the context of the present paper because they are typically deterministic in nature but considered as realizations of Gaussian random fields. Consider the function

$$f(r_w, r, T_u, H_u, T_l, H_l, L, K_w) = \frac{2\pi T_u (H_u - H_l)}{\ln(r/r_w)\left[1 + \frac{2LT_u}{\ln(r/r_w) r_w^2 K_w} + \frac{T_u}{T_l}\right]}$$

describing the flow rate through a borehole. The eight input variables and their respective

https://doi.org/10.1017/S0956792513000016 Published online by Cambridge University Press

Table 4. *Input variables and respective ranges of interest for the borehole function*

	Variable	Range of interest
r_w	Radius of borehole	0.05 to 0.15 m
r	Radius of influence	100 to 50,000 m
T_u	Transmissivity of upper aquifer	63,070 to 115,600 m^2/year
H_u	Potentiometric head of upper aquifer	990 to 1,110 m
T_l	Transmissivity of lower aquifer	63.1 to 116 m^2/year
H_l	Potentiometric head of lower aquifer	700 to 820 m
L	Length of borehole	1,120 to 1,680 m
K_w	Hydraulic conductivity of borehole	9,855 to 12,045 m/year

Table 5. *Parameter estimates and some error statistics for the borehole test function*

	REML	LOOCV_1	LOOCV_2
σ^2	3.67×10^4	5.45×10^3	1.26×10^4
θ_1	1.27×10^{-1}	3.81×10^{-1}	2.62×10^{-1}
θ_2	1.00×10^{-6}	1.00×10^{-6}	1.00×10^{-6}
θ_3	1.00×10^{-6}	1.38×10^{-4}	1.00×10^{-6}
θ_4	7.73×10^{-3}	2.08×10^{-2}	1.21×10^{-2}
θ_5	4.96×10^{-5}	5.62×10^{-5}	2.13×10^{-5}
θ_6	1.05×10^{-2}	1.74×10^{-2}	5.80×10^{-3}
θ_7	1.40×10^{-2}	3.63×10^{-2}	2.78×10^{-2}
θ_8	1.25×10^{-3}	1.99×10^{-3}	1.23×10^{-3}
RMSE	3.96	6.73	5.03
MAE	3.08	5.02	3.77
MAStE	1.02	1.05	1.05

ranges of interest are summarized in Table 4. We rescale these variables to the range $(1, 3)$ and use the same orthogonal sampling design as Joseph *et al.* [40] with [27] locations. We now assume f to be a realization of a stationary Gaussian random field with covariance function

$$\Phi(h) = \sigma^2 e^{-\sum_{j=1}^{8} \theta_j h_j^2}$$

of the Gaussian type. Its mean function will be considered constant but unknown so that we are in the framework of ordinary kriging (see Section 3). Table 5 shows the estimates for the parameters $\theta_1, \ldots, \theta_8$ and σ^2 obtained via REML, LOOCV_1 and LOOCV_2, where the subscript indicates that either the $\|\cdot\|_1$-norm or the $\|\cdot\|_2$-norm of the cross-validation errors is minimized.

Unlike real-world applications of computer experiments, the borehole function is cheap to evaluate, and this allows us to calculate its values on the grid $\mathcal{G} := \{1, 1.5, 2, 2.5, 3\}^8$ on the space of the scaled input variables and compare them with the values predicted via ordinary kriging with the covariance functions estimated by different methods. The

https://doi.org/10.1017/S0956792513000016 Published online by Cambridge University Press

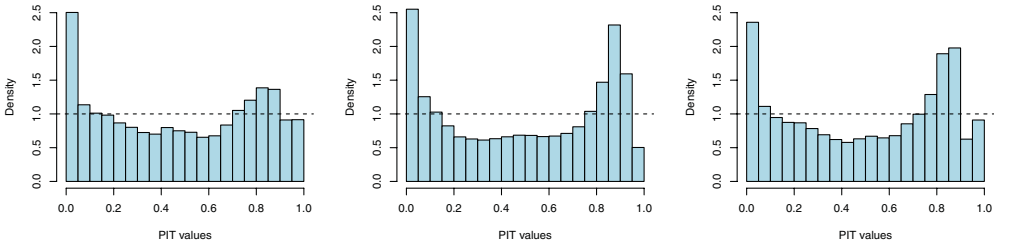

FIGURE 3. (Colour online) PIT histograms for probabilistic predictions corresponding to covariance parameters estimated via REML (left), LOOCV_1 (middle) and LOOCV_2 (right).

following error statistics are given in Table 5:

$$\mathrm{RMSE} := \sqrt{\sum_{x\in\mathscr{G}} (f(x) - s_{f,X}(x))^2}, \qquad \mathrm{MAE} := \sum_{x\in\mathscr{G}} |f(x) - s_{f,X}(x)|$$

$$\text{and} \quad \mathrm{MAStE} := \sum_{x\in\mathscr{G}} \frac{|f(x) - s_{f,X}(x)|}{P_{K,X}(x)}.$$

In the borehole example, both root mean squared error (RMSE) and mean absolute error (MAE) are the lowest for the interpolant computed with the REML estimates, but the LOOCV estimates too give good results. To judge how well the kriging variance describes the prediction uncertainty, one can look at the mean absolute standardized errors (MAStE). If the kriging variance (which also depends on the estimated parameters) has correct magnitude, the absolute standardized errors should average to 1, and indeed all three parameter choices yield an MAStE quite close to that. Since we have assumed a Gaussian random field, we can go even further and calculate, for every $x \in \mathscr{G}$, the probability integral transform (PIT) $F_x(f(x))$, where F_x is the cumulative distribution function of a Gaussian distribution with mean $s_{f,X}(x)$ and variance $P^2_{K,X}(x)$. F_x is a probabilistic forecast of $f(x)$ that automatically comes with our stochastic modelling assumptions. If it is correct, then the PIT values have a uniform distribution in $[0, 1]$, and this property can be checked by plotting them in the form of a histogram [2]. The PIT histograms in Figure 3 are quite far from uniformity, suggesting that the assumption of a Gaussian distribution is rather questionable. It is quite remarkable that REML, which is based on this assumption, does an excellent job in selecting good parameters, and we found that this is also true for many other test cases (see [72]).

A critical issue about REML estimation is the computational cost of $O(n^3)$ floating point operations for each choice of θ, which is prohibitive for large spatial data sets. When all sampling locations are on a (near-)regular lattice, spectral methods to approximate the likelihood can be used and allow to reduce the computational cost to an order of $O(n\log(n))$ [13, 29, 32]. These techniques cannot be applied to scattered data, but other approaches to approximating likelihoods [5, 47, 79, 81], covariance tapering [30] or simplified Gaussian models of low rank [3, 12, 21] have been proposed and shown to be quite effective in reducing the computational effort to an order that allows the application of REML in most practical situations.

https://doi.org/10.1017/S0956792513000016 Published online by Cambridge University Press

8 Discussion

A variety of practical problems amount to or can be linked to the mathematical problem of data interpolation. In this paper two approaches – kernel interpolation and kriging – were presented and their interconnections were pointed out. In either framework the interpolation procedure is optimal in a certain sense, but optimality is based on the assumption that the 'correct' kernel is used. Answers given by numerical analysts and statisticians to the question about the consequences on approximation accuracy of using an 'incorrect' kernel were discussed. Finally, some methods for choosing a suitable kernel based on the given data were presented.

The borehole example analysed in the preceding section poses the interesting challenge that it is not entirely clear if a stochastic modelling perspective is appropriate. While this does not matter anyway with respect to the interpolation method, it is comforting to see that with both cross validation and maximum likelihood good choices of an interpolation kernel are obtained. At first sight, this seems to contradict the asymptotic results by Wahba [82] mentioned above. It seems, however, that in this and many other examples the sample size is simply too small for asymptotic statements to hold. Moreover, in Wahba's setup the actual interpolation kernel is fixed, and only λ is estimated. Our belief is that REML is mostly competitive even in deterministic settings as long as it can choose from a sufficiently flexible class of kernels that permits, for example, adaptation to the regularity of f. Generally, when a high approximation accuracy is expected, the deterministic perspective seems more appropriate. When the data are sparse and/or f has low regularity, a random field model often yields a good description of f. The transition between the two perspectives and their respective methodologies, however, is rather smooth.

We have focused our discussion on topics that are relevant for both numerical analysts and statisticians. An important issue in kernel interpolation not mentioned so far is that of an ill-conditioned equation system (2.15). This problem frequently arises because in the deterministic framework very smooth and flat kernels are often preferred since they can achieve high convergence rates when f is very smooth (see Sections 5 and 6). Such kernels, however, inevitably lead to ill-conditioned systems which are a big challenge for numerical algorithms, and they call for special techniques such as preconditioning or changes of basis. If the standard basis $K(\cdot, x_1), \ldots, K(\cdot, x_n)$ is used as suggested by representation (2.14), ill-conditioning is tied to smoothness of kernel and small approximation errors in terms of power function [66]. But the interpolant $s_{f,X}$ in function space is not dramatically ill-conditioned [15] such that ill-conditioning is a problem of bad basis, not a problem of the reconstruction process. In geostatistics, the variables of interest in typical applications are usually very rough and call for kernels with low smoothness, and so ill-conditioning is usually not a big issue.

We shall finally mention a field of research where the methods discussed in this paper are applied in a slightly different context: the field of machine learning. The problem studied there can again be formulated as an interpolation problem, and both stochastic and deterministic modelling approaches can be used for its solution. An outline of connections to Gaussian processes and reproducing kernels is given in [62, 74]. Van der Vaart and van Zanten [80] discuss the Bayesian approach to the machine learning problem and provide – in a slightly different setting and based on a different risk function – results on convergence rates and the role of regularity of the covariance kernel similar to the

results that have been discussed in Sections 5 and 6. In machine learning too it is not always obvious if stochastic modelling assumptions are appropriate, and so understanding the implications of different assumptions and identifying the scope of the corresponding methods seem vital.

Acknowledgement

We are grateful to Tilmann Gneiting and Jon Wellner for valuable comments leading to numerous improvements in this paper.

References

[1] Anderes, E. (2010) On the consistent separation of scale and variance for Gaussian random fields. *Ann. Statist.* **38**(2), 870–893.
[2] Anderson, J. L. (1996) A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Clim.* **9**, 1518–1530.
[3] Banerjee, S., Gelfand, A. E., Finley, A. O. & Sang, H. (2008) Gaussian predictive process models for large spatial datasets. *J. R. Statist. Soc B* **70**(4), 825–848.
[4] Berlinet, A. & Thomas-Agnan, C. (2004) *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, Berlin, Germany.
[5] Caragea, P. & Smith, R. L. (2007) Asymptotic properties of computationally efficient alternative estimators for a class of multivariate normal models. *J. Multivariate Anal.* **98**(7), 1417–1440.
[6] Carlson, R. E. & Foley, T. A. (1991) The parameter R^2 in multiquadric interpolation. *Comp. Math. Appl.* **21**, 29–42.
[7] Chauvet, P., Pailleux, J. & Chilès, J.-P. (1976) Analyse objective des champs météorologiques par cokrigeage. *La Météorologie, 6ième Série* **4**, 37–54.
[8] Chilès, J.-P. (1976) How to adapt kriging to non-classical problems: three case studies. In: M. Guarascio, M. David & C. Huijbregts (editors), *Advanced Geostatistics in the Mining Industry*, D. Reidel, Dordrecht, Holland, pp. 69–89.
[9] Chilès, J.-P. & Delfiner, P. (1999) *Geostatistics. Modeling Spatial Uncertainty*, John Wiley, New York.
[10] Cressie, N. (1989) Geostatistics. *Am. Stat.* **43**(4), 197–202.
[11] Cressie, N. (1993) *Statistics for Spatial Data* (rev. ed. edition), Wiley, New York.
[12] Cressie, N. & Johannesson, G. (2008) Fixed rank kriging for very large spatial data sets. *J. R. Statist. Soc B* **70**(1), 209–226.
[13] Dahlhaus, R. & Künsch, H. R. (1987) Edge effects and efficient parameter estimation for stationary random fields. *Biometrika* **74**(4), 877–882.
[14] de Marchi, S. (2003) On optimal center locations for radial basis function interpolation: Computational aspects. *Rend. Sem. Mat. Torino* **61**(3), 343–358.
[15] de Marchi, S. & Schaback, R. (2010) Stability of kernel-based interpolation. *Adv. Comput. Math.* **32**, 155–161.
[16] de Marchi, S., Schaback, R. & Wendland, H. (2005) Near-optimal data-independent point locations for radial basis functions. *Adv. Comput. Math.* **23**(3), 317–330.
[17] Diggle, P. J. & Ribeiro, P. J. (2007) *Model-Based Geostatistics*, Springer, Berlin, Germany.
[18] Duchon, J. (1976) Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces. *Adv. RAIRO Anal. Num.* **10**, 5–12.
[19] Duchon, J. (1977) Splines minimizing rotation invariant seminorms in Sobolev spaces. In: W. Schempp & K. Zeller (editors), *Constructive Theory of Functions of Several Variables*, Springer-Verlag, Berlin, Germay, pp. 85–100.
[20] Duchon, J. (1978) Sur l'erreur d'interpolation des fonctions de plusieurs variables par les D^m-splines. *Adv. RAIRO Anal. Num.* **12**, 325–334.

[21] Eidsvik, J., Finley, S., Banerjee, S. & Rue, H. (2010) *Approximate Bayesian Inference for Large Spatial Datasets using Predictive Process Models*. Technical Report 9, Department of Mathematical Sciences, Norwegian University of Science and Technology, Norway.
[22] Evans, L. C. (2002) *Partial Differential Equations*, American Mathematical Society, Providence, RI.
[23] Fasshauer, G. E. (2007) *Meshfree Approximation Methods with Matlab*, World Scientific, Singapore.
[24] Fasshauer, G. E. & Zhang, J. G. (2007) On choosing "optimal" shape parameters for RBF approximation. *Numer. Algorithms* **45**, 345–368.
[25] Foley, T. A. (1987) Interpolation and approximation of 3-D and 4-D scattered data. *Comput. Math. Appl.* **13**, 711–740.
[26] Fornberg, B., Driscoll, T. A., Wright G. & Charles, R. (2002) Observations on the behaviour of radial basis function approximations near boundaries. *Comput. Math. Appl.* **43**, 473–490.
[27] Fornberg, B. & Wright, G. (2004) Stable computation of multiquadric interpolants for all values of the shape parameter. *Comput. Math. Appl.* **47**, 497–523.
[28] Franke, R. (1982) Scattered data interpolation: Tests of some methods. *Math. Comput.* **38**, 181–200.
[29] Fuentes, M. (2008) Approximate likelihood for large irregular spaced spatial data. *J. Am. Stat. Assoc.* **102**(477), 321–331.
[30] Furrer, R., Genton, M. G. & Nychka, D. (2006) Covariance tapering for interpolation of large spatial datasets. *J. Comput. Graph. Stat.* **15**(3), 502–523.
[31] Gneiting, T. & Schlather, M. (2004) Stochastic models that separate fractal dimension and the Hurst effect. *SIAM Rev.* **46**(2), 269–282.
[32] Guyon, X. (1982) Parameter estimation for a stationary process on a d-dimensional lattice. *Biometrika* **69**(1), 95–105.
[33] Handcock, M. S. & Stein, M. L. (1993) A Bayesian analysis of kriging. *Technometrics* **35**(4), 403–410.
[34] Handcock, M. S. & Wallis, J. R. (1994) An approach to statistical spatial-temporal modeling of meteorological fields (with discussion). *J. Am. Stat. Assoc.* **89**(7), 368–390.
[35] Hardy, R. L. (1971) Multiquadric equations of topography and other irregular surfaces. *J. Geophys. Res.* **76**, 1905–1915.
[36] Harville, D. A. (1974) Bayesian inference for variance components using only error contrasts. *Biometrika* **61**, 383–385.
[37] Heyde, C. C. (1997) *Quasi-Likelihood and Its Application*, Springer, New York.
[38] Ibragimov, I. A. & Rozanov, Y. A. (1978) *Gaussian Random Processes*, A. B. Aries (trans.), Springer, New York.
[39] Iske, A. (2000) *Optimal Distributions of Centers for Radial Basis Function Methods*. Technical Report M0004, Technische Universität München, Munich, Germany.
[40] Joseph, V. R., Hung, Y. & Sudjianto, A. (2008) Blind kriging: A new method for developing metamodels. *J. Mech. Des.* **130**(3), 1–8.
[41] Journel, A. G. (1982) The indicator approach to estimation of spatial distributions. In *Proceedings of the 17th APCOM International Symposium*, New York, pp. 793–806.
[42] Kimeldorf, G. S. & Wahba, G. (1970) A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Stat.* **41**(2), 495–502.
[43] Kitanidis, P. K. (1983) Statistical estimation of polynomial generalized covariance functions and hydrologic applications. *Water Resour. Res.* **19**(4), 909–921.
[44] Kitanidis, P. K. (1997) *Introduction to Geostatistics: Applications in Hydrology*, Cambridge University Press, New York.
[45] Lajaunie, C. & Béjaoui, R. (1991) *Sur le Krigeage des Fonctions Complexes*. Technical Report N-23/91/G, Centre de Géostatistique, Ecole des Mines de Paris, Fontainebleau.

[46] Lim, S. C. & Teo, L. P. (2010) Analytic and asymptotic properties of multivariate generalized Linniks probability densities. *J. Fourier Anal. Appl.* **16**, 715–747.
[47] Lindgren, F., Rue, H. & Lindström, J. (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J. R. Stat. Soc B* **73**(4), 423–498.
[48] Madych, W. R. & Nelson, S. A. (1988) Multivariate interpolation and conditionally positive definite functions. *Approx. Theory Appl.* **4**, 77–89.
[49] Mardia, K. V. & Marshall, R. J. (1984) Maximum likelihood estimation of models for residual covariance in spatial statistics. *Biometrika* **71**, 135–146.
[50] Matérn, B. (1986) *Spatial Variation*, 2nd ed., Lecture Notes in Statistics, Vol. 36, Springer-Verlag, Berlin, Germany.
[51] Matheron, G. (1971) *The Theory of Regionalized Variables and its Applications.* Technical Report, Cahiers du Centre de Morphologie Mathématique de Fontainebleau, Ecole des Mines de Paris.
[52] Matheron, G. (1973) The intrinsic random functions and their applications. *Adv. Appl. Prob.* **5**, 439–468.
[53] Matheron, G. (1973) *Le Krigeage Disjonctive*. Technical Report N-360, Centre de Géostatistique, Ecole des Mines de Paris.
[54] Matheron, G. (1976) A simple substitute for conditional expectation: The disjunctive kriging. In: M. Guarascio, M. David & C. Huijbregts (editors), *Advanced Geostatistics in the Mining Industry*, Reidel, Dordrecht, Netherland, pp. 221–236.
[55] Micchelli, C. A. (1986) Interpolation of scattered data: Distance matrices and conditionally positive definite functions. *Constr. Approx.* **2**, 11–22.
[56] Morris, M. D., Mitchell, T. J. & Ylvisaker, D. (1993) Bayesian design and analysis of computer experiments: Use of derivatives in surface prediction. *Technometrics* **35**(3), 243–255.
[57] Myers, D. E. (1992) Kriging, cokriging, radial basis functions and the role of positive definiteness. *Comput. Math. Appl.* **24**(12), 139–148.
[58] Narcowich, F. J., Ward, J. D. & Wendland, H. (2006) Sobolev error estimates and a Bernstein inequality for scattered data interpolation via radial basis functions. *Constr. Approx.* **24**, 175–186.
[59] Omre, H. & Halvorsen, K. B. (1989) The Bayesian bridge between simple and universal kriging. *Math. Geol.* **21**(7), 767–786.
[60] Putter, H. & Young, G. A. (2001) On the effect of covariance function estimation on the accuracy of kriging predictors. *Bernoulli* **7**(3), 421–438.
[61] R Development Core Team. (2011) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
[62] Rasmussen, C. E. & Williams, C. K. I. (2006) *Gaussian Processes for Machine Learning*, MIT Press, Boston, MA.
[63] Rippa, S. (1999) An algorithm for selecting a good value for the parameter c in radial basis function interpolation. *Adv. Comput. Math.* **11**, 193–210.
[64] Ritter, K. (2000) *Average-Case Analysis of Numerical Problems*, Lecture Notes in Mathematics, No. 1733, Springer, New York.
[65] Schaback, R. (1993) Comparison of radial basis function interpolants. In: K. Jetter & F. Utreras (editors), *Multivariate Approximation: From CAGD to Wavelets*, World Scientific, London, pp. 293–305.
[66] Schaback, R. (1995) Error estimates and condition numbers for radial basis function interpolation. *Adv. Comput. Math.* **3**, 251–264.
[67] Schaback, R. (1997) Native Hilbert spaces for radial basis functions I. In: *New Developments in Approximation Theory*, International Series of Numerical Mathematics, No. 132, Birkhauser Verlag, Berlin, Germany, pp. 255–282.
[68] Schaback, R. (2011) The missing Wendland functions. *Adv. Comput. Math.* **34**, 67–81.
[69] Schaback, R. (2011) *Kernel-Based Meshless Methods.* Technical report, Georg-August-Universität Göttingen, Göttingen, Germany.

[70] Schaback, R. & Wendland, H. (2006) Kernel techniques: From machine learning to meshless methods. *Acta Numer.* **15**, 543–639.
[71] Scheuerer, M. (2010) Regularity of the sample paths of a general second-order random field. *Stoch. Proc. Appl.* **120**, 1879–1897.
[72] Scheuerer, M. (2011) An alternative procedure for selecting a good value for the parameter c in RBF-interpolation. *Adv. Comput. Math.* **34**(1), 105–126.
[73] Schlather, M. (2001) RandomFields: Contributed extension package to R for the simulation of Gaussian and max-stable random fields. URL: cran.r-project.org.
[74] Seeger, M. (2004) Gaussian processes for machine learning. *Int. J. Neural Syst.* **14**, 1–38.
[75] Stein, M. L. (1988) Asymptotically efficient prediction of a random field with a misspecified covariance function. *Ann. Stat.* **16**, 55–63.
[76] Stein, M. L. (1990) A comparison of generalized cross validation and modified maximum likelihood for estimating the parameters of a stochastic process. *Ann. Stat.* **18**(3), 1139–1157.
[77] Stein, M. L. (1999) *Interpolation of Spatial Data*, Springer, New York.
[78] Stein, M. L. (2004) Equivalence of Gaussian measures for some nonstationary random fields. *J. Stat. Plann. Inference* **123**, 1–11.
[79] Stein, M. L., Chi, Z. & Welty, L. J. (2004) Approximating likelihoods for large spatial data sets. *J. R. Stat. Soc B* **66**(2), 275–296.
[80] van der Vaart, A. W. & van Zanten, J. H. (2011) Information rates of nonparametric Gaussian process methods. *J. Mach. Learn. Res.* **12**, 2095–2119.
[81] Vecchia, A. V. (1988) Estimation and model identification for continuous spatial processes. *J. R. Stat. Soc B* **50**, 297–312.
[82] Wahba, G. (1985) A comparison of gcv and gml for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Stat.* **13**(1), 1378–1402.
[83] Wahba, G. (1990) *Spline Models for Observational Data*, SIAM, Philadelphia, PA.
[84] Wendland, H. (1995) Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Adv. Comput. Math.* **4**, 389–396.
[85] Wendland, H. (2005) *Scattered Data Approximation*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, UK.
[86] Wendland, H. & Rieger, C. (2005) Approximate interpolation with applications to selecting smoothing parameters. *Numer. Math.* **101**, 729–748.
[87] Wu, Z. (1992) Hermite-Birkhoff interpolation of scattered data by radial basis functions. *Approx. Theory Appl.* **8**(2), 1–10.
[88] Zhang, H. (2004) Inconsistent estimation and asymptotically equivalent interpolations in model-based geostatistics. *J. Am. Stat. Assoc.* **99**, 250–261.