A framework to understand the asymptotic properties of Kriging and splines

Eva M. Furrer and Douglas W. Nychka

Institute for Mathematics Applied to Geosciences, National Center for Atmospheric Research, P.O. Box 3000, Boulder, CO 80307

Abstract: Kriging is a nonparametric regression method used in geostatistics for estimating curves and surfaces for spatial data. It may come as a surprise that the Kriging estimator, normally derived as the best linear unbiased estimator, is also the solution of a particular variational problem. Thus, Kriging estimators can also be interpreted as generalized smoothing splines where the roughness penalty is determined by the covariance function of a spatial process. We build off the early work by Silverman (1982, 1984) and the analysis by Cox (1983, 1984), Messer (1991), Messer and Goldstein (1993) and others and develop an equivalent kernel interpretation of geostatistical estimators. Given this connection we show how a given covariance function influences the bias and variance of the Kriging estimate as well as the mean squared prediction error. Some specific asymptotic results are given for one dimensional corresponding Matérn covariances that have as their limit cubic smoothing splines.

Key words: Kriging, spline, equivalent kernel, asymptotic mean squared error

1 Introduction

A common method in the analysis of spatial data is a geostatistical estimator known as Kriging. Although Kriging is typically derived as a best linear unbiased estimator it can also be viewed as a nonparametric curve and surface estimator. Given this later perspective, it is of interest to understand Kriging in terms of the large sample properties such as the asymptotic variance and bias that are well established for kernel estimators. The key idea developed in this work is that Kriging estimators can be interpreted as generalized splines and the asymptotic techniques similar to those described in Nychka (1995) can be brought to bear on the Kriging estimators.
The problem one is faced with in nonparametric regression is to estimate an unknown function, g, on $[0,1]$, for which the observations y_i are supposed to depend on the "locations" x_i following the model:

$$y_i = g(x_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where the ε_i are iid random errors with common variance σ^2. Note that we use the interval $[0,1]$ without loss of generality. The solution to this problem by either spline or kernel methods can be written as

$$\hat{g}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \omega(x, x_i) y_i, \tag{1}$$

for a weight function ω and is therefore a linear function of the data. In contrast to kernel methods the weight function of the smoothing spline estimator is not known in closed form but the techniques presented in Nychka (1995) can be used to approximate it. In the same spirit we want to represent the Kriging estimator of a Gaussian process g at an unobserved location $\mathbf{x}$ as a weighted average of the observations at the locations $\mathbf{x}_i$ using a weight function $\omega(\mathbf{x}, \mathbf{x}_i)$, which we call the Kriging weight function.

The main contribution of this paper is the identification of a functional form approximating ω for Kriging estimators that is derived from the spatial process covariance function and is a reproducing kernel. We term this approximation the equivalent kernel because in the case of stationary covariances and uniformly distributed locations, it is similar to a kernel estimator. This paper does not give a rigorous development of specific results but rather lays out a general framework for the asymptotic theory. To this end we conjecture the large sample behavior for the mean squared error of Kriging estimators and also point to verifying a key condition that will allow the rigorous theory that has been applied to one dimensional splines to be extended to these more general estimators. As an introduction and to fix ideas, we start by discussing the structure of the classic smoothing spline problem, focusing on the details which will be important for further developments. Also, we will outline the equivalence between a Kriging estimator and a spline estimator.

1.1 Classic cubic smoothing splines and reproducing kernels

For normal errors the maximum penalized likelihood estimate is given as the minimizer of

$$\sum_{i=1}^{n} \big(y_i - \varphi(x_i)\big)^2 + \lambda \int_0^1 \varphi''(x)^2 \mathrm{d}x, \tag{2}$$

over all twice differentiable functions on $[0, 1]$. The parameter λ controls the degree of smoothness via the penalty term, which in this case penalizes roughness of the estimated function using its total curvature. The solution to this minimization is the standard cubic smoothing spline although the actual form as a piecewise cubic polynomial is not important to a general discussion. The form of the roughness penalty using second derivatives implies that linear functions will result in a penalty of zero. In order to simplify the discussion, we will decompose φ into the sum $\varphi(x) = \beta_1 + \beta_2 x + f(x)$ with $f(0) = f'(0) = 0$, then (2) is given by

$$\min_{f \in \mathcal{H}, \beta_1, \beta_2} \Big[\sum_{i=1}^{n} \Big((y_i - \beta_1 + \beta_2 x_i) - f(x_i) \Big)^2 + \lambda \int_0^1 f''(x)^2 \mathrm{d}x \Big], \tag{3}$$

$\mathcal{H}$ is a space containing all functions with square integrable second derivatives on $[0, 1]$ with $f(0) = 0$ and $f'(0) = 0$. Since only the sum of squares is affected by this complication, we can continue our discussion, which is mainly concerned with the penalty term, without further consideration of the nullspace. The full estimator can be derived by first minimizing over $\mathcal{H}$ and then over β, this results in a generalized least squares estimate for the null space parameters. We now identify the penalty term above with an inner product on $\mathcal{H}$ given as

$$\langle f, h \rangle = \int_0^1 f''(x) h''(x) \mathrm{d}x, \tag{4}$$

and we can express the minimization criterion over f in (3) as

$$\sum_{i=1}^{n} \Big((y_i^* - f(x_i) \Big)^2 + \lambda \langle f, f \rangle, \tag{5}$$

where $y_i^* = (y_i - \beta_1 + \beta_2 x_i)$.

How does one characterize this seemingly difficult minimization over a function space? The concept of a reproducing kernel provides an explicit and elegant solution to this problem. The set of functions $\mathcal{H}$ together with the inner product (4) is a Hilbert space with reproducing kernel $k(\cdot, \cdot)$ given by

$$k(u, v) = \begin{cases} \frac{1}{2}u^2 v - \frac{1}{6}u^3, & u < v, \\ \frac{1}{2}v^2 u - \frac{1}{6}v^3, & u \geq v. \end{cases}$$

The function $k(\cdot, \cdot)$ is defined as a reproducing kernel if it satisfies

$$\langle k(x, \cdot), f \rangle = f(x), \quad \text{and as a special case} \quad \langle k(u, \cdot), k(v, \cdot) \rangle = k(u, v)$$

i.e. k evaluates a function using the inner product of the Hilbert space. It will also "reproduce" itself under the inner product. In the case of the kernel given above this can be verified simply using integration by parts.

Lemma 1.1 (Generalized smoothing spline)**.** *If $\mathcal{H}$ is a Hilbert space with reproducing kernel k then the solution $\hat{g}$ of the minimization problem (5) is of the form*

$$\hat{g}(x) = \sum_{i=1}^{n} \theta_i \psi_i,$$

with $\psi_i = k(\cdot, x_i)$, for $i = 1, \ldots, n$ and coefficients θ_i obtained by minimizing

$$min_{\boldsymbol{\theta}} \Big[||\boldsymbol{y}^* - \boldsymbol{W\theta}||^2 + \lambda \boldsymbol{\theta}^T \boldsymbol{W\theta} \Big], \tag{6}$$

with $\boldsymbol{W} = \{W_{i,k}\} = \{k(x_i, x_k)\}$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$.

This lemma can be established using a proof by contradiction and we give a sketch of this argument. Suppose g_{other} is the minimizer but is not equal to $\hat{g}$. Let g^* be a function of the same form as $\hat{g}$ but agreeing with g_{other} at the observation points. The residual sums of squares are the same for these two estimates and using the reproducing property of k is it easy to show that $\langle g^*, g^* \rangle \leq \langle g_{\text{other}}, g_{\text{other}} \rangle$. Thus the minimization criterion evaluated at g_{other} is greater than or equal to the value at g^*. But we assumed that g_{other} was the minimizer and so one obtains a contradiction.

As discussed above, the solution $\hat{g}$ of the minimization problem (5) is a linear combination of the values $\psi_k(x_i) = k(x_i, x_k)$, i.e. $\hat{g}(x_i) = [\mathbf{W\theta}]_i$. Then restating the minimization problem in matrix notation (6) is straightforward using the reproducing property of k to simplify the penalty term. Taking derivatives with respect to $\boldsymbol{\theta}$ in (6), setting them equal to zero and solving for $\boldsymbol{\theta}$, results in

$$\hat{\boldsymbol{\theta}} = (\mathbf{W} + \lambda \mathbf{I})^{-1} \mathbf{y}^*.$$

The estimated function values at the observed points $x_1, \dots x_n$ are given by

$$\hat{\mathbf{g}} = (\hat{g}(x_1), \dots, \hat{g}(x_n))^T = \mathbf{W}\hat{\boldsymbol{\theta}} = \mathbf{W}(\mathbf{W} + \lambda\mathbf{I})^{-1}\mathbf{y}^* = \mathbf{A}(\lambda)\mathbf{y}^*,$$

i.e the smoothing spline approach leads to a linear smoothing procedure.
For completeness we now add back in the linear term that was absorbed into $\mathbf{y}^*$ and so derive the full cubic spline estimator. Substituting the minimizer over f back in (3) one obtains, after some algebra,

$$\hat{\beta} = (X^T(\mathbf{W} + \lambda\mathbf{I})^{-1})X)^{-1}X^T(\mathbf{W} + \lambda\mathbf{I})^{-1}\mathbf{y}$$

The reader can interpret this estimate as a generalized least squares regression estimate where the errors have a covariance proportional to $(\mathbf{W} + \lambda\mathbf{I})$. In particular the matrix $\mathbf{W}$, derived from the reproducing kernel is formally being manipulated as a covariance function for a random process. This observation provides a link into understanding the connection of splines with Kriging. In the next section we will see that the covariance function plays the same role as a reproducing kernel.

1.2 The classical Kriging problem

For the Kriging problem our initial context is very different, we have a spatial field

$$y_i = g(\mathbf{x}_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

where g is a zero mean Gaussian process on a domain $\mathcal{D} \subset \mathbb{R}^d$, d typically 1 or 2, and ε_i are normal errors, independent of g and distributed according to a zero mean normal distribution with variance σ^2. Without loss of generality we assume that the size of $\mathcal{D}$ is 1. The covariance of the Gaussian process g is assumed to be

$$\mathrm{Cov}(g(\mathbf{x}), g(\mathbf{x}')) = k(\mathbf{x}, \mathbf{x}'),$$

for an appropriate covariance function k. Much of the statistical craft in geostatistics is in finding good representations and estimators for k, but we will not address this aspect here. Given k, the Kriging problem is to find a prediction of the spatial field g at the locations $\mathbf{x}_1, \dots, \mathbf{x}_N$ based on the observations $y_1, \dots, y_n$. This can easily be achieved by using the properties of the multivariate normal distribution. For $\mathbf{y} = (y_1, \dots, y_n)^T$ and $\mathbf{g} = (g(\mathbf{x}_1), \dots, g(\mathbf{x}_N))^T$ we have that

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{g} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \mathrm{Cov}(\mathbf{y}, \mathbf{y}) & \mathrm{Cov}(\mathbf{y}, \mathbf{g}) \\ \mathrm{Cov}(\mathbf{g}, \mathbf{y}) & \mathrm{Cov}(\mathbf{g}, \mathbf{g}) \end{pmatrix}\right).$$

Then the prediction of g at the locations $\mathbf{x}_1, \dots, \mathbf{x}_N$ can, for example, be defined as the conditional expectation of g at those locations given the observations. The conditional density is given by:

$$[g(\mathbf{x}) \,|\, \mathbf{y}] = \mathcal{N}\big(\mathrm{Cov}(g(\mathbf{x}), \mathbf{y})\mathrm{Cov}(\mathbf{y}, \mathbf{y})^{-1}\mathbf{y}, \mathrm{Cov}(g(\mathbf{x}), g(\mathbf{x})) - \mathrm{Cov}(g(\mathbf{x}), \mathbf{y})\mathrm{Cov}(\mathbf{y}, \mathbf{y})^{-1}\mathrm{Cov}(\mathbf{y}, g(\mathbf{x}))\big), \tag{7}$$

Now examine the conditional mean in more detail:

$$\mathrm{Cov}(g(\mathbf{x}), \mathbf{y})\mathrm{Cov}(\mathbf{y}, \mathbf{y})^{-1}\mathbf{y} = \mathrm{Cov}(g(\mathbf{x}), \mathbf{y})\hat{\boldsymbol{\theta}} = \sum_i k(\mathbf{x}, \mathbf{x}_i)\hat{\theta}_i$$

where as before $\hat{\boldsymbol{\theta}} = (\mathbf{W} + \lambda \mathbf{I})^{-1}\mathbf{y}$ with $[\mathbf{W}]_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$ and $\lambda = \sigma^2$. This is exactly the form of the spline type estimator discussed in the previous section. We have now shown algebraically that if one solves a spline minimization problem where the inner product has the reproducing kernel k then the solution is the Kriging estimator.
If one considers the estimates of g at the observation points then the vector of predictions is the matrix $\mathbf{S} = \mathrm{Cov}(\mathbf{g}, \mathbf{g})\mathrm{Cov}(\mathbf{y}, \mathbf{y})^{-1}$. This is usually known as the smoothing matrix associated with Kriging and contains in its rows the so-called Kriging weights.
In summary, we have shown that the Kriging estimator is actually a smoothing spline estimator on a Hilbert space with reproducing kernel corresponding to the covariance function of the underlying Gaussian process (since the matrix $\mathbf{W}$ is determined by k). This statement will be made more rigorous in Section 2 providing the appropriate variational problem and inner product.

1.3 Outline

The previous two subsections served two purposes. On the one hand, the smoothing spline approach to function estimation has been introduced together with the use of reproducing kernel Hilbert spaces. On the other hand, it has been made intuitive why the Kriging estimator is another special case of this general approach. Section 2 will rigorously show how the Kriging problem can be stated as an analogous variational problem and how the results of Nychka (1995) can be transferred to Kriging. Section 3 gives details for some special cases of stationary covariance functions.

2 An equivalent kernel for Kriging

2.1 Variational problem and reproducing kernel of the Kriging problem

Based on the preceding discussion from Section 1.1 let

$$\mathcal{L}(f) = \sum_{i=1}^{n} \big(y_i - f(\mathbf{x}_i)\big)^2 + \lambda \langle f, f \rangle, \tag{8}$$

where the penalty inner product $\langle \cdot, \cdot \rangle$ is such that the correlation function k is the corresponding reproducing kernel. Then the estimator is the minimizer of $\mathcal{L}(f)$ over all f so that $\langle f, f \rangle < \infty$. Based on the previous discussion this has the solution $\hat{g} = \sum_{i=1}^{n} \theta_i \psi_i$, where $\psi_i = k(\cdot, \mathbf{x}_i)$, for $i = 1, \ldots, n$. Analogously to the developments of Section 1.1, this solution can then be represented in matrix notation and by construction coincides with the Kriging estimator as given in Section 1.2.
Our first objective is to define an inner product on a Hilbert space of functions having the correlation function k as reproducing kernel. If one has a positive definite kernel then one can always formally define an inner product such that the kernel is the reproducing kernel and also extend this to a Hilbert space (**?**, Theorem 2, page 117). However, in our case we would like to work with an inner product that is also interpretable with respect to integrals. Accordingly, we define the relevant inner product via the integral operator $\mathcal{K}f(\mathbf{x}) = \int_{\mathcal{D}} k(\mathbf{x}, \mathbf{x}') f(\mathbf{x}') \, \mathrm{d}\mathbf{x}'$ as

$$\langle f_1, f_2 \rangle = \int_{\mathcal{D}} (\mathcal{K}^{-1/2} f_1)(\mathbf{x}) (\mathcal{K}^{-1/2} f_2)(\mathbf{x}) \, \mathrm{d}\mathbf{x},$$

for all functions f_1 and f_2 in the space $\mathcal{H}$ of functions for which $(\mathcal{K}^{-1/2}f)$ exists and is square integrable over $\mathcal{D} \subset \mathbb{R}^d$. To derive $\mathcal{K}^{-1/2}$ we use a decomposition of k based on eigenvalues and eigenfunctions:

$$k(\mathbf{x}, \mathbf{x}') = \sum_{\nu=1}^{\infty} \lambda_\nu \varphi_\nu(\mathbf{x}) \varphi_\nu(\mathbf{x}'), \quad \text{with } \lambda_1 > \lambda_2 > \dots \text{ and } \varphi_\nu \perp \varphi_{\nu'} \text{ in } \mathrm{L}^2(\mathcal{D}).$$

Then we have that

$$\mathcal{K}f(\mathbf{x}) = \sum_{\nu=1}^{\infty} \lambda_\nu \varphi_\nu(\mathbf{x}) \int_{\mathcal{D}} \varphi_\nu(\mathbf{x}') f(\mathbf{x}') \, \mathrm{d}\mathbf{x}'$$

and by orthogonality of the φ_ν

$$\mathcal{K}^{-1/2}f = \sum_{\nu=1}^{\infty} \lambda_\nu^{-1/2} \varphi_\nu(\mathbf{x}) \int_{\mathcal{D}} \varphi_\nu(\mathbf{x}') f(\mathbf{x}') \, \mathrm{d}\mathbf{x}'.$$

Therefore, k is the reproducing kernel since

$$\begin{aligned}
\langle k(\mathbf{t}, \cdot), k(\mathbf{s}, \cdot) \rangle &= \int_{\mathcal{D}} \big(\mathcal{K}^{-1/2} k(\mathbf{t}, \cdot)\big)(\mathbf{x}) \big(\mathcal{K}^{-1/2} k(\mathbf{s}, \cdot)\big)(\mathbf{x}) \, \mathrm{d}\mathbf{x} \\
&= \sum_{\nu=1}^{\infty} \sum_{\mu=1}^{\infty} \lambda_\nu^{-1/2} \lambda_\mu^{-1/2} \Big(\int_{\mathcal{D}} \varphi_\nu(\mathbf{x}') k(\mathbf{t}, \mathbf{x}') \, \mathrm{d}\mathbf{x}' \Big) \Big(\int_{\mathcal{D}} \varphi_\mu(\mathbf{x}') k(\mathbf{s}, \mathbf{x}') \, \mathrm{d}\mathbf{x}' \Big) \Big(\int_{\mathcal{D}} \varphi_\nu(\mathbf{x}) \varphi_\mu(\mathbf{x}) \, \mathrm{d}\mathbf{x} \Big) \\
&= \sum_{\nu=1}^{\infty} \lambda_\nu \varphi_\nu(\mathbf{t}) \varphi_\nu(\mathbf{t}) = k(\mathbf{t}, \mathbf{s}),
\end{aligned}$$

where we used the orthogonality of the functions φ_ν and the decomposition of k. We have now constructed an inner product such that k is its reproducing kernel. This means that minimizing (8) is in fact equivalent to the Kriging problem.

In closing one might ask: What exactly is the Hilbert space implied by this inner product? We know that $\mathcal{H}$ will contain finite linear combinations of the form $f = \sum_k a_k k(\cdot, x_k)$ and in fact $\mathcal{H}$ is just the completion of this class of functions. A more precise description can be inferred by considering the Fourier transform of k when the covariance is stationary and this allows one to identify the differentiability of members of $\mathcal{H}$.

2.2 The Kriging weight function

This is the main insight of this paper: we represent the minimizer of (8) as a weighted local average of the observations similar to the spline weight function representation of the smoothing spline estimate in Nychka (1995). The weight can be characterized by the first order conditions for the minimizer of $\mathcal{L}$. This idea originated in Cox (1983) and was also the basis for Nychka (1995). For any function h in some dense subset of $\mathcal{H}$, the minimizer of (8) (with smoothing parameter $\lambda \cdot n$) satisfies the equation

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon} \mathcal{L}(\hat{g} + \varepsilon h)\Big|_{\varepsilon=0} = -\frac{2}{n} \sum_{i=1}^{n} \big(y_i - \hat{g}(\mathbf{x}_i)\big) h(\mathbf{x}_i) + 2\lambda \langle \hat{g}, h \rangle = 0. \tag{9}$$

Examining (**??**) (its version for the Kriging estimator) it is trivial to see that $\omega(\cdot, \mathbf{x}_j)$ is the "estimate" for the data $y_k = n$ for $k = j$ and $y_k = 0$ otherwise. Substituting these synthetic data into (9) the exact weight function must satisfy the following identity

$$\frac{1}{n}\sum_{i=1}^{n}\omega(\mathbf{x}_i, \mathbf{x}_j)h(\mathbf{x}_k) + \lambda\langle\omega(\cdot, \mathbf{x}_j), h\rangle = h(\mathbf{x}_j).$$

Furthermore ω must be a reproducing kernel for the inner product, $\frac{1}{n}\sum_{i=1}^{n} f_1(\mathbf{x}_i)f_2(\mathbf{x}_i) + \lambda\langle f_1, f_2\rangle$ and so is a symmetric function. We assume that the empirical distribution of the $\mathbf{x}_i$ tends to a uniform distribution on $\mathcal{D}$ for $n \to \infty$. Now add and subtract the continuous integral approximation to the discrete sum and group the terms

$$\left(\frac{1}{n}\sum_{i=1}^{n}\omega(\mathbf{x}_i, \mathbf{x}_j)h(\mathbf{x}_i) - \int_{\mathcal{D}}\omega(\mathbf{t}, \mathbf{x}_j)h(\mathbf{t})\,\mathrm{d}\mathbf{t}\right) + \int_{\mathcal{D}}\omega(\mathbf{t}, \mathbf{x}_j)h(\mathbf{t})\,\mathrm{d}\mathbf{t} + \lambda\langle\omega(\cdot, \mathbf{x}_j), h\rangle = h(\mathbf{x}_j). \quad (10)$$

The final step in this derivation is to substitute a particular set of functions for h. Let G be the reproducing kernel with respect to the inner product defined by

$$\langle f_1, f_2\rangle_\lambda = \int_{\mathcal{D}} f_1(\mathbf{t})f_2(\mathbf{t})\,\mathrm{d}\mathbf{t} + \lambda\langle f_1, f_2\rangle.$$

Substituting $G(\cdot, u)$ into (10) and using its reproducing property now gives

$$\omega(\mathbf{u}, \mathbf{x}_j) = G(\mathbf{x}_j, \mathbf{u}) + \left(\int_{\mathcal{D}}\omega(\mathbf{t}, \mathbf{x}_j)G(\mathbf{t}, \mathbf{u})\,\mathrm{d}\mathbf{t} - \frac{1}{n}\sum_{i=1}^{n}\omega(\mathbf{x}_i, \mathbf{x}_j)G(\mathbf{x}_j, \mathbf{u})\right).$$

Thus provided that the difference between the integral and sum is negligible G will be a good approximation to ω. Moreover G has a simpler form as it is defined through a continuous integral rather than a sum depending on the exact distribution of the observation locations. The analysis in Nychka (1995) gives conditions under which the approximation will be accurate. Even though the final application in that work was for a specific linear spline the basic theorem establishing bounds on the error is more general. Our opinion is that the main hurdle in applying this approximation to Kriging type estimators is to establish an assumption termed the exponential envelope condition (EEC). This condition is on G and facilitates an inductive argument to infer that the bounds on the difference are asymptotically negligible relative to G. For example, in the case of m^{th} order smoothing splines the EEC requires, that there exist positive constants α, ε, $K < \infty$ such that for all $t, \tau \in [0, 1]$,

$$\begin{aligned}
\big|G(t,\tau)\big| &\le \frac{K}{\lambda^{1/2m}}\exp\Big(-(\alpha-\varepsilon)\frac{|t-\tau|}{\lambda^{1/2m}}\Big),\\
\Big|\frac{\partial}{\partial t}G(t,\tau)\Big| &\le \frac{K}{\lambda^{2/2m}}\exp\Big(-(\alpha+\varepsilon)\frac{|t-\tau|}{\lambda^{1/2m}}\Big),\\
\Big|\frac{\partial^2}{\partial t\partial\tau}G(t,\tau)\Big| &\le \frac{K}{\lambda^{3/2m}}\exp\Big(-(\alpha+\varepsilon)\frac{|t-\tau|}{\lambda^{1/2m}}\Big),
\end{aligned}$$

if the partial derivatives exist for all $t, \tau \in [0, 1]$. Otherwise, if $(\partial/\partial t)G$ is not continuous when $t = \tau$, then

$$\frac{\partial}{\partial t}G(t,\tau)\Big|_{\tau=t^-} - \frac{\partial}{\partial t}G(t,\tau)\Big|_{\tau=t^+} = \frac{1}{\lambda}.$$

In the following we will be using Fourier transform techniques, which require that integrals are taken over $\mathbb{R}^d$. Our main interest is in observations taken on a bounded domain and so we briefly sketch the link between those two. First, we observe that the penalty inner product $\langle \cdot, \cdot \rangle$ can be taken over $\mathbb{R}^d$ without changing the resulting estimator since it does not depend on the observation points. Secondly, denote by G the reproducing kernel of the inner product $\int_{\mathcal{D}} f_1(x) f_2(x) dx + \lambda \langle f_1, f_2 \rangle$ and by G^* the reproducing kernel of the inner product $\int_{\mathbb{R}^d} f_1(x) f_2(x) dx + \lambda \langle f_1, f_2 \rangle$. Suppose that q is the uniform density over the bounded domain $\mathcal{D}$ then by the properties of both G and G^* it is easy to show that

$$G^*(u, v) = G(u, v) + \Big(\int_{\mathbb{R}^d} G^*(x, u) G(x, v) \big(q(x) - 1\big) \mathrm{d}x \Big).$$

This gives an approximation for one kernel with another and the error is an integral with respect to G^*. Therefore, we can conclude that showing an EEC for G^* implies that it holds for G as well. One disadvantage of this approximation is that it will only be valid for the estimator in the interior of the domain and will not take into account edge effects in the estimator.

In summary we have identified an approximation, G^* to ω that facilitates an asymptotic analysis of the variance and bias of a Kriging estimator. This approximation is a reproducing kernel and the main challenge is to be able to verify an EEC for G^*. The next section tackles some simple covariances to illustrate how to apply these ideas. Before doing so we give an summary of how to use G to infer asymptotic properties for the estimator.

2.3 Mean squared error of a Kriging estimator

Given the approximation described in the preceding section it is of interest to derive properties of these estimators. It should be emphasized that these are still conjectured relationships but we also believe that a rigorous justification is accessible given the results for splines and also the examples in the following section.

Under the approximation

$$\hat{g}(\mathbf{x}) \approx \frac{1}{n} \sum_{i=1}^{n} G^*(\mathbf{x}, \mathbf{x}_i) y_i$$

we approximate the variance

$$\mathrm{Var}\big(\hat{g}(\mathbf{x})\big) \approx (\sigma^2/n) \int_{\mathbb{R}^d} G^*(\mathbf{u}, \mathbf{x})^2 \mathrm{d}\mathbf{u}$$

and the bias as

$$\mathrm{E}\big(\hat{g}(\mathbf{x})\big) - g(\mathbf{x}) \approx \int_{\mathbb{R}^d} G^*(\mathbf{x}, \mathbf{u}) g(\mathbf{u}) \mathrm{d}\mathbf{u} - g(\mathbf{x}).$$

Here we have approximated the discrete sum with an integral and a rigorous justification is possible. This type of analysis is supported by the same methods used to justify the equivalent kernel approximation and follows the same arguments as in the proof of Theorem 2.2 in Nychka (1995).

In the usual case of kernel estimators the variance and squared bias can be combined to give a mean squared error (MSE). However, for Kriging estimates it is the convention to also take the expectation over g and this will be the same as the conditional variance in (7):

$$\mathrm{Cov}(g(\mathbf{x}), g(\mathbf{x})) - \mathrm{Cov}(g(\mathbf{x}), \mathbf{y}) \mathrm{Cov}(\mathbf{y}, \mathbf{y})^{-1} \mathrm{Cov}(\mathbf{y}, g(\mathbf{x})).$$

Based on the discussion in Section 1.2 $\mathrm{Cov}(g(\mathbf{x}),\mathbf{y})\mathrm{Cov}(\mathbf{y},\mathbf{y})^{-1}$ is, by definition, the same as the vector of weights $\{(1/n)\omega(\mathbf{x},\mathbf{x}_i)\}$. With this identification and again substituting an integral for the discrete sum we are lead to the approximation:

$$\mathrm{MSE}(\hat{g}(\mathbf{x})) \approx k(\mathbf{x},\mathbf{x}) - \frac{1}{n}\sum_{i=1}^{n} G^*(\mathbf{x},\mathbf{x}_i)k(\mathbf{x},\mathbf{x}_i) \approx k(\mathbf{x},\mathbf{x}) - \int_{\mathbb{R}^d} G^*(\mathbf{x},\mathbf{u})k(\mathbf{x},\mathbf{u})\mathrm{d}\mathbf{u}.$$

3 Stationary covariances

We start by assuming a stationary and isotropic covariance function, i.e. $k(\mathbf{x},\mathbf{y}) = k\big(|\mathbf{x}-\mathbf{y}|\big)$. By construction we have then as well that the equivalent kernel is a function of the radial component only, $G(\mathbf{x},\mathbf{y}) = G\big(|\mathbf{x}-\mathbf{y}|\big)$. For the special cases considered in this chapter, we obtain this result by straightforward calculation. We will be applying Fourier transform techniques in order to obtain results on G, therefore we assume the domain to be all of $\mathbb{R}^d$.

3.1 Equivalent kernel

For stationary covariances the operator $\mathcal{K}f(\mathbf{x}) = \int_{\mathbb{R}^d} k(\mathbf{x}-\mathbf{y})f(\mathbf{y})\,\mathrm{d}\mathbf{y}$ is a convolution of the functions k and f. Therefore, denoting by $\mathcal{F}$ the Fourier transform, we have that $\mathcal{F}(\mathcal{K}f) = \mathcal{F}(k)\mathcal{F}(f)$. We define the operators $\mathcal{K}^{-1/2}$ and $\mathcal{K}^{1/2}$ by

$$\mathcal{K}^{-1/2}f = \mathcal{F}^{-1}\Big(\frac{\mathcal{F}(f)}{\mathcal{F}(k)^{1/2}}\Big) \quad \text{and} \quad \mathcal{K}^{1/2}g = \mathcal{F}^{-1}\big(\mathcal{F}(g)\mathcal{F}(k)^{1/2}\big),$$

for all f, g, functions on $\mathbb{R}^d$, for which $\mathcal{F}(f)$, $\mathcal{F}(g)$ exist and $(\mathcal{F}(f)/\mathcal{F}(k)^{1/2})$, $(\mathcal{F}(g)\mathcal{F}(k)^{1/2})$ are square-integrable. Since we are essentially concerned with covariance functions satisfying $\mathcal{F}(k) \propto |\omega|^{-\nu}$, $|\omega^{\pm\nu/2}\mathcal{F}(f)(\omega)|$ tends sufficiently quickly to zero for rapidly decreasing $f \in C^\infty(\mathbb{R}^d)$ such that the inverse transform exists for those f. The generalization to square integrable functions is by density of $C^\infty(\mathbb{R}^d)$, see, for example, Dym and McKean (1972), Chapter 2.2. These operators can actually be interpreted as the square root of $\mathcal{K}$ and its inverse, since

$$\mathcal{F}\big(\mathcal{K}^{1/2}(\mathcal{K}^{-1/2}f)\big) = \mathcal{F}(f) \quad \text{and} \quad \mathcal{F}\big(\mathcal{K}^{1/2}(\mathcal{K}^{1/2}f)\big) = \mathcal{F}(\mathcal{K}f).$$

Using Parseval's theorem, this implies that the inner product for the penalty satisfies

$$\begin{aligned}\langle f_1, f_2\rangle &= \int_{\mathbb{R}^d} (\mathcal{K}^{-1/2}f_1)(\mathbf{x})(\mathcal{K}^{-1/2}f_2)(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_{\mathbb{R}^d} \big(\mathcal{F}(\mathcal{K}^{-1/2}f_1)\big)(\boldsymbol{\omega})\big(\mathcal{F}(\mathcal{K}^{-1/2}f_2)\big)(\boldsymbol{\omega})\,\mathrm{d}\boldsymbol{\omega} \\ &= \int_{\mathbb{R}^d} \frac{\mathcal{F}(f_1)(\boldsymbol{\omega})\mathcal{F}(f_2)(\boldsymbol{\omega})}{\mathcal{F}(k)(\boldsymbol{\omega})}\,\mathrm{d}\boldsymbol{\omega},\end{aligned}$$

for all functions f_1 and f_2 in $C^\infty(\mathbb{R}^d)$. Furthermore, again using Parseval's theorem, the inner product for which we aim to determine the reproducing kernel satisfies

$$\begin{aligned}\langle f_1, f_2\rangle_\lambda &= \int_{\mathbb{R}^d} f_1(\mathbf{t})f_2(\mathbf{t})\,\mathrm{d}\mathbf{t} + \lambda\langle f_1, f_2\rangle \\ &= \int_{\mathbb{R}^d} \mathcal{F}(f_1)(\boldsymbol{\omega})\mathcal{F}(f_2)(\boldsymbol{\omega})\,\mathrm{d}\boldsymbol{\omega} + \lambda\int \frac{\mathcal{F}(f_1)(\boldsymbol{\omega})\mathcal{F}(f_2)(\boldsymbol{\omega})}{\mathcal{F}(k)(\boldsymbol{\omega})}\,\mathrm{d}\boldsymbol{\omega} \\ &= \int_{\mathbb{R}^d} \mathcal{F}(f_1)(\boldsymbol{\omega})\mathcal{F}(f_2)(\boldsymbol{\omega})\Big(1 + \lambda\frac{1}{\mathcal{F}(k)(\boldsymbol{\omega})}\Big)\,\mathrm{d}\boldsymbol{\omega}.\end{aligned}$$

Hence, the reproducing kernel G is characterized by

$$\big\langle h, G\big(|\cdot - \mathbf{x}_j|\big)\big\rangle_\lambda = \int_{\mathbb{R}^d} \mathcal{F}(h)(\boldsymbol{\omega})\mathcal{F}\big(G\big(|\cdot - \mathbf{x}_j|\big)\big)(\boldsymbol{\omega})\Big(1 + \lambda\frac{1}{\mathcal{F}(k)(\boldsymbol{\omega})}\Big)\,\mathrm{d}\boldsymbol{\omega} = h(\mathbf{x}_j)$$

for arbitrary $h \in \mathcal{C}^\infty(\mathbb{R}^d)$. On the other hand, we have by the Fourier inversion theorem that for all $h \in \mathcal{C}^\infty(\mathbb{R}^d)$

$$\frac{1}{2\pi^d}\int_{\mathbb{R}^d} \mathcal{F}(h)(\boldsymbol{\omega})\mathrm{e}^{i\boldsymbol{\omega}\mathbf{x}_j}\,\mathrm{d}\boldsymbol{\omega} = h(\mathbf{x}_j).$$

Therefore, we conclude that

$$\mathcal{F}(G\big(|\cdot - \mathbf{x}_j|\big))(\boldsymbol{\omega})\Big(1 + \frac{1}{\lambda\mathcal{F}(k)(\boldsymbol{\omega})}\Big) = \frac{1}{2\pi}\mathrm{e}^{i\boldsymbol{\omega}\mathbf{x}_j} \quad \Longleftrightarrow \quad \mathcal{F}(G)(\boldsymbol{\omega}) = \Big(1 + \lambda\frac{1}{\mathcal{F}(k)(\boldsymbol{\omega})}\Big)^{-1}, \tag{11}$$

i.e. the Fourier transform of the reproducing kernel we want to determine is essentially determined by the Fourier transform of the underlying covariance function and we will look at it in more detail in the next section under specific covariance models.

3.2 Exponential envelope condition for the Matérn class of covariances

The Matérn class of covariances is given by

$$k(r) = \phi(\alpha r)^\nu K_\nu(\alpha r), \quad \text{for } \phi > 0, \alpha > 0, \nu > 0 \text{ and } r \in \mathbb{R},$$

and the corresponding Fourier transform, i.e. the spectral density, is

$$\mathcal{F}(k)(\omega) = \frac{2^{\nu-1}\phi\Gamma(\nu + d/2)\alpha^{2\nu}}{\pi^{d/2}(\alpha^2 + \omega^2)^{\nu+d/2}}. \tag{12}$$

This parameterization is valid for any dimension d, where the arguments r and ω correspond to radial components in state and frequency space respectively. Plugging (3.2) into (11) leads to

$$\mathcal{F}(G)(\omega) = \frac{1}{1 + (\lambda/c)(\alpha^2 + \omega^2)^{\nu+d/2}}, \tag{13}$$

where $c = 2^{\nu-1}\phi\Gamma(\nu + d/2)\alpha^{2\nu}\pi^{-d/2}$ depends on ϕ, ν, α and d.

One dimensional smoothers with the Matérn and cubic splines

First we consider $d = 1$ and $\nu = 1.5$, therefore the exponent in (12) is 2 and we see below that we have a correspondence to splines of order $m = 2$, i.e. roughness penalties based on second derivatives. In this case it is possible to calculate the inverse transform G by using 3.731.1 of Gradshteyn and Ryzhik (1965):

$$G(t) = \frac{\sqrt{c\pi}}{\sqrt{2(\lambda\alpha^2 + c)}}\exp(-At)\big(B\cos(Bt) + A\sin(Bt)\big), \tag{14}$$

where $A = \sqrt{\big(\sqrt{\alpha^4 + c/\lambda} + \alpha^2\big)/2}$ and $B = \sqrt{\big(\sqrt{\alpha^4 + c/\lambda} - \alpha^2\big)/2}$. The order of λ in this function is $\lambda^{-1/4}$ and an EEC for G follows straightforwardly.

It is interesting to compare this result with thin plate splines. Using $\mathcal{F}(\triangle f)(\omega) = \omega^4 \mathcal{F}(f)$, the same type of grouping of the penalty and the second term in the variational inner product as for Kriging and 3.727.1 of Gradshteyn and Ryzhik (1965), we obtain

$$\mathcal{F}(G)(\omega) = \frac{1}{1+\lambda\omega^4} \quad \text{and}$$
$$G(t) = \frac{\sqrt{\pi}}{2\lambda^{1/4}} \exp\Big(-\frac{t}{\sqrt{2}\lambda^{1/4}}\Big)\Big(\cos\Big(-\frac{t}{\sqrt{2}\lambda^{1/4}}\Big) + \sin\Big(-\frac{t}{\sqrt{2}\lambda^{1/4}}\Big)\Big). \tag{15}$$

Therefore, we consider this special case of Kriging and a thin plate spline of order 2 in $d = 1$ to be equivalent. The EEC is clear from inspection and to our knowledge this is the first analysis that draws a clear connection between the Matérn covariance estimator and a cubic spline as a limiting case.

Two dimensional thin plate spline

Naturally, the second special case should be $d = 2$ and $\nu = 1$ and this is aligned with early results by Stein (1991). Again we attempt to calculate the inverse transform in two dimensions. By switching to polar coordinates we find that the inverse transform of $\big(1 + \lambda c(\alpha^2 + \omega^2)^2\big)^{-1}$, where ω refers to the radial component in the frequency space, depends on the radial component in the state space only. It is given by

$$G(r) = \frac{c}{\lambda}\int_0^\infty \frac{J_0(r\omega)\omega}{c/\lambda + (\alpha^2+\omega^2)^2}\mathrm{d}\omega,$$

where J_0 is the Bessel function of first kind of order 0. Unfortunately, we are unable to further simplify this expression and it will be necessary to prove the EEC using the Fourier integral or to use additional theory for Bessel functions.
In this second case, it is equally interesting to look at the thin plate spline case. Similar to $d = 1$ and using the same polar coordinates substitution as for Kriging, we obtain

$$G(r) = \frac{1}{\lambda}\int_0^\infty \frac{J_0(r\omega)\omega}{1/\lambda + \omega^4}\mathrm{d}\omega = -\frac{1}{\lambda}\lambda^{-1/2}\mathrm{kei}(\lambda^{-1/4}r),$$

where kei is a Kelvin function, see Abramowitz and Stegun (1965), page 379. This result is identical to the kernel approximation for the Kriging predictor under a generalized covariance function model of Stein (1991).

3.3 Mean squared error for the Matérn class of covariances

For the Matérn covariance function in $d = 1$ with $\nu = 1.5$ we have calculated the equivalent kernel G and from its functional form it is clear that it satisfies the EEC for some $\lambda^{1/\gamma}$.
Let F_n denote the empirical distribution function of the observation points x_i, F the uniform distribution function on $\mathcal{D}$ and $D_n = \sup_{\mathcal{D}} |F_n - F|$. For discrete sums to be properly approximated by integrals we need $D_n \to 0$ at an appropriate rate but we do not require that the locations be drawn from a probability distribution. In the case of near regular grids one can expect that D_n converges to zero a rate of $1/n$.

Theorem 3.1 (MSE for the one dimensional Matérn estimator)**.** *Assume that $D_n \to 0$ as $n \to \infty$ with $\kappa D_n/\lambda^{1/\gamma} < 1$ for some constant κ and that k is the Matérn covariance function in $d = 1$ with $\nu = 1.5$. Without loss of generality assume that $k(0) = 1$ then we have*

$$MSE(\hat{g}(x)) \approx 1 - \int_{\mathbb{R}} G\big(|x-u|\big)k\big(|x-u|\big)\,du,$$

for x in the interior of $\mathcal{D}$ and λ in a range depending on D_n.

The proof of this theorem is along the lines of the proof of Theorem 2.2 of Nychka (1995). A quick argument why it holds is as follows. The following development of $\mathrm{MSE}(\hat{g}(x))$ is exact:

$$\begin{aligned}
\mathrm{MSE}(\hat{g}(x)) &= 1 - \frac{1}{n}\sum_{i=1}^{n} \omega\big(|x-x_i|\big)k\big(|x-x_i|\big) - \frac{1}{n}\sum_{i=1}^{n} G\big(|x-x_i|\big)k\big(|x-x_i|\big) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} G\big(|x-x_i|\big)k\big(|x-x_i|\big) - \int_{\mathbb{R}} G\big(|x-u|\big)k\big(|x-u|\big)\mathrm{d}u \\
&\quad + \int_{\mathbb{R}} G\big(|x-u|\big)k\big(|x-u|\big)\mathrm{d}u \\
&= \int_{\mathbb{R}} \Big(\omega\big(|x-u|\big) - G\big(|x-u|\big)\Big)k\big(|x-u|\big)\mathrm{d}F_n(u) \qquad (16) \\
&\quad + \int_{\mathbb{R}} G\big(|x-u|\big)k\big(|x-u|\big)\mathrm{d}(F_n - F)(u) \qquad (17) \\
&\quad + 1 - \int_{\mathbb{R}} G\big(|x-u|\big)k\big(|x-u|\big)\mathrm{d}u. \qquad (18)
\end{aligned}$$

The analysis of these individual terms follows very closely the arguments on page 1191 of Nychka (1995) for bounding integrals and bounding the difference between sums and integrals. (15) can be shown to be negligible using the EEC and (16) is negligible because of the assumptions on D_n. The term in (17) is the one we are interested in. By using basic properties of the Fourier transform concerning convolutions and the value of the original function at the origin, we have that

$$\mathrm{MSE}(\hat{g}(0)) \approx 1 - \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \frac{1}{\big(1+(\lambda/c)(\alpha^2+\omega^2)^2\big)}\frac{c}{(\alpha^2+\omega^2)^2}\mathrm{d}\omega.$$

By stationarity the MSE approximation is the same for all x in the interior of $\mathcal{D}$. This integral now is the basis for obtaining the asymptotic behavior of the mean squared error as a function of the parameters α and λ. As an example of the convergence rates, hold α fixed and consider the integral just as a function of λ. Using the fact that the spectral density of the Matérn covariance function integrates to $\sqrt{2\pi}$ and simplifying

$$1 - \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \frac{1}{\big(1+(\lambda/c)(\alpha^2+\omega^2)^2\big)}\frac{c}{(\alpha^2+\omega^2)^2}\mathrm{d}\omega = \frac{1}{2\pi}\int_{\mathbb{R}} \frac{\lambda}{\big(1+(\lambda/c)(\alpha^2+\omega^2)^2\big)}\mathrm{d}\omega.$$

Now make the substitution $u = \lambda^{1/4}\omega$ in this integral and we obtain

$$\frac{\lambda^{3/4}}{\sqrt{2\pi}}\int_{\mathbb{R}} \frac{1}{\big(1+(1/c)((\alpha\lambda^{1/4})^2+u^2)^2\big)}\mathrm{d}u.$$

Noting that we made an abuse of notation where λ in the equivalent kernel section is actually σ^2/n (redefined as λ/n from the introductory section), we can conclude that the pointwise MSE for this case is of order $n^{-3/4}$ based on the equivalent kernel.

4 Discussion

We have sketched a framework to understand the asymptotic properties of spatial statistics estimators based on a connection to existing theory for splines. For stationary covariances the equivalent kernel approximation has a simple form in terms of its Fourier transform and we have obtained closed form expressions in some key cases. We believe that the equivalent kernel will lead to the correct convergence rates for the Kriging estimators both under the correct covariance and also under covariance misspecification. To this end it will provide some theoretical insight in a field that is not completely developed. However, these forms are conjectured relations and we only offer a rigorous result for the one dimensional case of the Matérn covariance that is related to cubic splines. Even in this case that we have deferred including a complete proof. For future work we plan to provide a rigorous adaptation of Nychka (1995) that is streamlined for clarity and handles the more general estimators considered here. It is our opinion that this work will identify the exponential envelope condition as the key assumption and so we are comfortable emphasizing this aspect in this introduction and overview. An important detail is to understand how to verify this condition directly with the Fourier transform of the equivalent kernel and a possible avenue is through a Tauberian Theorem.

One motivation for this work is to understand the behavior of the estimator when the covariance is non-stationary. In this case the covariance is often difficult to estimate. Indeed, it can look like another function estimation problem! It is an open issue as to whether the benefits in using a more flexible estimate of a non-stationary covariance function are countered by the increased variability in using a complex estimate of the covariance. This estimated covariance may be subject to substantial sampling error and for limited sample sizes may not be an improvement over a more stable stationary model. As a first step in understanding this tradeoff, this asymptotic analysis has the potential in quantifying the improved MSE when the correct, non-stationary covariance is used versus an approximate stationary version. If the improvements are modest then this may suggest that modeling the non-stationary covariance is not important for spatial prediction. It may be important for inference.

Another area of work is modifying these techniques to account for distributions of locations that are not well approximated by a uniform distribution. A possible strategy is to modify the equivalent kernel for uniform distributions with a variable smoothing parameter that adjusts for the location density. This technique is analogous to a variable bandwidth kernel estimator and has worked for the one dimensional spline case when the densities are smooth.

In closing we note that smoothing is just one aspect of geostatistics and there are many problems where the function or surface is observed almost without error. The Kriging estimators are still very practical methods but they no longer have the property of a smoother because they may match the observations closely and provide smooth interpolations between observations. In this case of small measurement error the asymptotics discussed here break down and locations where g is not observed will have different MSE than estimates at the observed locations. In short, there are still many challenges posed in understanding these penalized estimators and advances will be useful because of the widespread use of these methods in the geosciences.

References

Abramowitz, M. and Stegun, I. (1965). *Handbook of mathematical functions*. Dover Publications Inc., New York, 9th edition.

Cox, D. D. (1983). Asymptotics for M-type smoothing splines. *The Annals of Statistics*, **11**, 530–551.

Cox, D. D. (1984). Multivariate smoothing spline functions. *SIAM Journal of Numerical Analysis*, **21**, 789–813.

Dym, H. and McKean, H. (1972). *Fourier Series and Integrals*. Acaddemic Press, New York and London.

Gradshteyn, I. and Ryzhik, I. (1965). *Table of Integrals, Series and Products*. Academic Press, New York and London, 4th edition.

Messer, K. (1991). A comparison of a spline estimate to its equivalent kernel estimate. *The Annals of Statistics*, **19**, 817–829.

Messer, K. and Goldstein, L. (1993). A new class of kernels for nonparametric curve estimation. *The Annals of Statistics*, **21**, 179–195.

Nychka, D. (1995). Splines as local smoothers. *The Annals of Statistics*, **23**, 1175–1197.

Silverman, B. (1982). The estimation of a probability density function by maximum penalized likelihood method. *The Annals of Statistics*, **10**, 795–810.

Silverman, B. (1984). Spline smoothing: The equivalent variable kernel method. *The Annals of Statistics*, **12**, 898–916.

Stein, M. (1991). A kernel approximation to the kriging predictor of a spatial process. *Annals of the Institute of Statistical Mathematics*, **43**, 61–75.