METMA V International Workshop on Spatio-Temporal Modelling
Santiago de Compostela
30 Jun - 2 Jul 2010

# Covariance estimation via Fourier series and application to groundwater quality indicators

P. García-Soidán[1]*, R. Menezes[2], and O. Rubiños[3]

[1] *Dept. of Statistics and Operations Research, University of Vigo, Spain; pgarcia@uvigo.es*
[2] *Dept. of Mathematics and Applications, University of Minho, Portugal; rmenezes@mct.uminho.pt*
[3] *Dept. of Signal Theory and Communications, University of Vigo, Spain; oscar@com.uvigo.es*
**Corresponding author*

**Abstract.** *The Fourier series approach is a useful tool for approximation of curves in a variety of settings. In this paper we will apply this technique to estimate the covariance function of a second-order stationary random process. Furthermore, an expansion may be constructed for approximation of the covariance estimator such that it satisfies the positive-definiteness property and, therefore, is valid for prediction using the kriging techniques. We also suggest a procedure for an optimal choice of the truncation point, which specifies the number of terms to be used in the expansion. Several studies have been conducted to illustrate the performance of this approach, for both simulated and real data.*

**Keywords.** *Covariance function; Fourier coefficient; Groundwater quality indicators; Truncation point.*

## 1 Introduction

The need to reconstruct a phenomenon over the whole observation region from a finite set of data can be found in a broad spectrum of areas, such as geostatistics, hydrology, atmospheric science, etc. The use of kriging for this purpose requires estimation of the variogram or the covariance function, depending on whether intrinsic or second-order stationarity is assumed. The class of intrinsic stationary random processes is more general than that of second-order stationary random processes and the variogram does not require estimation of the constant mean of the process, unlike the covariance function. Despite the latter arguments, an important number of practitioners prefer the use of the covariance function and the reason for this might be related to their unfamiliarity with the way of characterizing dependence through the variogram. The current paper is focussed on the covariance function estimation, although a similar approach can be proposed for estimation of the variogram.

Several procedures have been suggested in the literature for estimation of the covariance function. See [8] for a review of several approaches, which are put into comparison in a numerical study covering different spatial dependence situations. In a first step, the nonparametric estimators may be used for this purpose, such as the empirical covariance [1] or the kernel-type estimator [6]. Nevertheless, they cannot be used in kriging since the positive-definiteness condition typically fails and, consequently, they might originate a negative mean squared prediction error. We can cope with this problem by choosing a valid parametric family and then selecting that covariance function in the family considered which best may fit the data. An additional option in the isotropic setting is proposed in [9], for a broad class of models dependent on a large number of parameters. Application of this approach requires selection of nodes, which could be taken to be equispaced or as the roots of some Bessel functions, as suggested in [5]; the latter method produces an orthogonal discretization so that a very small number of nodes is necessary to obtain a good nonparametric fit. Another alternative to obtain a valid estimator, which can be applied under anisotropy, is that of first truncating and then inverting the Fourier transform of a given estimator, given in [7], although selection of the truncation term is an open issue.

In this paper we develop a procedure to obtain a valid covariance estimator by using an approximation obtained from the Fourier series. This technique has been applied in different settings, such as those concerning the density or the regression estimation, as proposed in [3]. The underlying idea is based on approximating the unknown covariance function by a finite expansion, which involves two issues: specification of the truncation point and estimation of the Fourier coefficients. To ensure that the approximating partial sum is positive-definite, only those terms corresponding to positive coefficients will be included.

To estimate the Fourier coefficients in the expansion, it is necessary to choose a prior covariance estimator, referred to as the pilot estimator. This may be either supplied to carry out a specific study or may be selected from the different alternatives existing in the geostatistics literature. In this respect, our approach may be viewed as a procedure for transformation of a given covariance estimator into a valid one.

The choice of a smoothing parameter is necessary to specify the number of terms to be used in the expansion; we will refer to it as the cutoff or the truncation point. Different methods have been proposed for the latter selection with the aim of overcoming inconsistency of the resulting estimators, as suggested in [2]. In our study, an explicit procedure for choice of the truncation point is provided, based on the minimization of the corresponding mean integrated squared error.

Simulation studies, for simulated and real data, were conducted in order to assess the quality of our valid covariance estimator and to compare it with other currently existing estimators. In particular, we focus our analysis on the water quality indicators collected by the Portuguese Hydrological Resources Management System in the south litoral coast of Esposende, during 2008 and 2009. This area was recently classified as a Vulnerable Area, making it urgent to achieve a better understanding of groundwater quality over time. The data analysis requires the application of geostatistical tools to model the spatial distribution of physic-chemical variables, such as nitrates, in distinct temporal levels. This work aims to apply the proposed valid covariogram estimator to nitrates data, allowing to build groundwater quality prediction maps.

## 2 Main results

Let $\{Z(\mathrm{s}) : \mathrm{s} \in D \subset \mathbb{R}^d\}$ be a second-order stationary random process with covariance function $C$, where $D$ is a bounded observation region. Denote by $Z(\mathrm{s}_1)$, ..., $Z(\mathrm{s}_n)$, $n$ data collected at the respective spatial locations $\mathrm{s}_1$, ..., $\mathrm{s}_n \in D$. Our aim will be to estimate $C(\mathrm{t})$, for each $\mathrm{t} \in A = \{\mathrm{s} - \mathrm{s}' : \mathrm{s}, \mathrm{s}' \in D\}$, and we will address this problem by approximating the covariance function through a Fourier expansion.

We will assume, without loss of generality, that $A$ is a bounded rectangle, $A = I_1 \times ... \times I_d$, where $I_j = [0, b_j]$ and $b_j > 0$, for all $j$. Then, there exists a complete orthonormal set on $A$, which will be designed as $\{\psi_i : i \in \mathbb{N}\}$. To check the latter, it is enough to take into account that:

- $\psi_{i_1,...,i_d} = \psi_{i_1,b_1} \cdot ... \cdot \psi_{i_d,b_d}$ is a complete orthornormal basis on $A$, provided that $\{\psi_{i_j,b_j} : i_j \in \mathbb{N}\}$ is a complete orthonormal system on $I_j = [0, b_j]$, for each $j$.
- The unidimensional cosine system $\psi_{i,b}(x) = e_i \cos(i\pi x b^{-1})$ is a complete orthonormal basis on $I = [0, b]$, where $e_i$ equals $b^{-1/2}$ or $(0.5b)^{-1/2}$, for $i = 0$ or $i > 0$, respectively.
- A bijection can be established between $\mathbb{N}^d$ and $\mathbb{N}$.

Since $C$ is a bounded and positive-definite function, it can be approximated by:

$$C_m(\mathrm{t}) = \sum_{i \leq m} \theta_{C,i} \psi_i(\mathrm{t}), \text{ for all } \mathrm{t} \in A$$

with $\theta_{C,i} = \langle C, \psi_i \rangle$. Hereafter $m$ will be called the cutoff or the truncation point.

From the foregoing definition of $C_m$, it is clear that approximating the covariance function by using a Fourier series requires:

- Computing coefficients $\theta_{C,i}$, dependent on the theoretical covariance function.
- Selecting the cutoff $m$, which will specify the number of terms in the expansion.

To compute the coefficients, a pilot estimator $\hat{C}$ of the covariance function is needed. This estimator must satisfy that $\sup_{\mathrm{t} \in A} \left| \mathrm{Bias}\left[\hat{C}(\mathrm{t})\right] \right| \overset{n \to \infty}{\longrightarrow} 0$ and $\sup_{\mathrm{t} \in A} \mathrm{Var}\left[\hat{C}(\mathrm{t})\right] \overset{n \to \infty}{\longrightarrow} 0$, which would guarantee consistency of the resulting estimator, namely:

$$\tilde{C}_{1,m}(\mathrm{t}) = \sum_{i \leq m} \theta_{\hat{C},i} \psi_i(\mathrm{t}) \tag{1}$$

where $\theta_{\hat{C},i} = \langle \hat{C}, \psi_i \rangle$.

For instance, the kernel covariance estimator, whose properties are derived in [4], can be taken as the pilot covariance and is given by:

$$\hat{C}_h(\mathrm{t}) = \frac{\sum_{j,k} K\left(\frac{\mathrm{t} - (\mathrm{s}_j - \mathrm{s}_k)}{h}\right) \left(Z(\mathrm{s}_j) - \bar{Z}\right)\left(Z(\mathrm{s}_k) - \bar{Z}\right)}{\sum_{j,k} K\left(\frac{\mathrm{t} - (\mathrm{s}_j - \mathrm{s}_k)}{h}\right)}$$

where $\bar{Z} = n^{-1}\sum_{j=1}^{n} Z(\mathrm{s}_j)$, $K$ denotes a $d$-variate kernel function and $h = h_n$ is the bandwidth parameter, with $h \to 0$ as $n$ tends to $\infty$.

Estimator (1) is not necessarily positive-definite, although we can overcome this problem by restricting our selection of the terms in the expansion to those involving positive Fourier coefficients $\theta_{\hat{C},i}$. This leads to the following approximating function:

$$\tilde{C}_{2,m}(\mathrm{t}) = \sum_{i \leq m} w_i \theta_{\hat{C},i} \psi_i(\mathrm{t}) \tag{2}$$

with $w_i = \mathrm{I}_{\{\theta_{\hat{C},i} > 0\}}$. The positive-definiteness of (2) can be easily checked.

With regard to the truncation point $m$, necessary for implementation of the covariance estimators (1) and (2), we propose to proceed by minimizing:

$$\mathrm{MISE}\left[\tilde{C}_{j,m}, C\right] = \sum_{i \leq m} p_i \left( \mathrm{Var}\left[\theta_{\hat{C},i}\right] + \mathrm{Bias}\left[\theta_{\hat{C},i}\right]^2 - \theta^2_{C,i} \right) + \int_A C(\mathrm{t})^2 dt = M_j(m) + \int_A C(\mathrm{s})^2 d\mathrm{s}$$

with $p_i$ equaling 1 or $w_i$, for $j = 1$ or $j = 2$, respectively.

Minimization of function $M_j$, equivalent to that of $\mathrm{MISE}[\tilde{C}_{j,m}, C]$, will provide a key idea for selection of the cutoff in each case, where the pilot covariance can be again used for approximation of the unknown terms in $M_j(m)$.

**Acknowledgments.** This work has been supported in part by grant INCITE-08-PXIB-322219-PR from Consellería de Innovación e Industria (Xunta de Galicia, Spain).

## References

[1] Cressie, N. (1993). *Statistics for spatial data*. John Wiley and Sons Inc. New York.

[2] Diggle, P. and Hall, P. (1986). The selection of terms in an orthogonal series density estimator. *Journal of the American Statistical Association* **81**, 230–233.

[3] Efromovich, S. (1999). *Nonparametric Curve Estimation: Methods, Theory and Applications*. Springer. New York.

[4] García-Soidán, P. (2007). Asymptotic normality of the Nadaraya-Watson semivariogram estimator. *TEST* **16**, 479–503.

[5] Genton, M. and Gorsich, D. J. (2002). Nonparametric variogram and covariogram estimation with Fourier-Bessel matrices. *Computational Statistics & Data Analysis* **41**, 47–57.

[6] Hall, P. and Patil, P. (1994). Properties of nonparametric estimators of autocovariance for stationary random fields. *Probability Theory and Related Fields* **99**, 399–424.

[7] Hall, P., Fisher, N. I. and Hoffman, B. (1994). On the nonparametric estimation of covariance functions. *Annals of Statistics* **22**, 2115–2134.

[8] Menezes, R. García-Soidán, P. and Febrero-Bande, M. (2005). A comparison of approaches for valid variogram achievement. *Computational Statistics* **20**, 623–642.

[9] Shapiro, A. and Botha, J. (1991). Variogram fitting with a general class of conditionally nonnegative definite functions. *Computational Statistics & Data Analysis* **11**, 87–96.