An Alternative Spectral Modelling of (Cross)-Covariance Tables

R. Tolosana-Delgado[1] , K.G. van den Boogaart[2]

[1]Dept. Informàtca i Matemàtica Aplicada, U. de Girona, E-17071 Girona (Spain);
E-mail: raimon.tolosana@udg.es
[2]Institut für Mathematik und Informatik, Ernst-Moritz-Arndt-Universität Greifswald,
D-17487 Greifswald (Germany)

1. Introduction

Most-used Geostatistical techniques need the specification of a covariance structure, *e.g.* kriging and simulation of uni- and multivariate random functions, where cross-covariance modelling is critical. Covariance models must be positive definite matrix functions, which implies that any linear combination of the observations will be attached a valid variance-covariance matrix, *i.e.* a positive variance or a positive definite symmetric matrix. This rather complicated-to-test condition is simplified by Bochner's Theorem: under certain regularity conditions, the spectral representation of a covariance function must be a positive definite matrix for every frequency (Cramèr, 1940; Bochner, 1959).

The spectral modelling of auto- and cross-covariance functions came to the geostatistical field through Rehman (1995), who approximated the Fourier Transform (FT) of cross-variograms by truncated series of Bessel or *sinc* functions. Details can be found in Yao and Journel (1998), who suggest as an alternative to use the Fast FT (FFT) algorithm to validate experimental correlation tables computed by smoothing classical experimental versions with kernel-like fans, which are finally re-scaled by independent estimates of the covariance $C(h)$ to obtain a set of covariance lookup tables. But, in our opinion, this is not adequate to treat covariances due to the strong discrete character of FFT inherited from time signals, regularly sampled and with unknown continuity properties. The precision of FFT is therefore focused in the higher part of the spectrum, linked to discontinuities (nugget effect, linear behavior at the origin), but precision in the lower frequencies (hole effects, general shape) is poor. Increasing the resolution of a covariance (denser lags) only increases the high frequencies, but does not yield a better characterization of the range or hole effects (high periods) of the covariance function: to do so, we should increase the maximum lag distance. Fig. 1 illustrates this contrast by showing the spectral density of a Gaussian and an exponential covariance models, comparing the theoretical model and the computed FFT. The fit is good, but it is also seen that most (~ 80%) of the nodes computed by FFT are identically zero. In other words, effort has been spent in determining the energy of some frequencies previously known to be zero.

This knowledge comes from some standard properties of covariance functions, which make them different from signals: whereas signals are assumed to be periodic outside the sampled range and irregular or even discontinuous between the sampled nodes, covariance functions can be reasonably assumed to be continuous, smooth, derivable almost everywhere and necessarily bounded. We propose to use this extra information to constraint the frequency spectrum to investigate, shifting the focus from the high-frequency to the low-frequency part of the spectrum. Then a numerical integration can

be performed in this frequency domain to estimate the FT: we suggest a Monte Carlo method, due to the connections between one-dimensional inverse FT and an expectation.

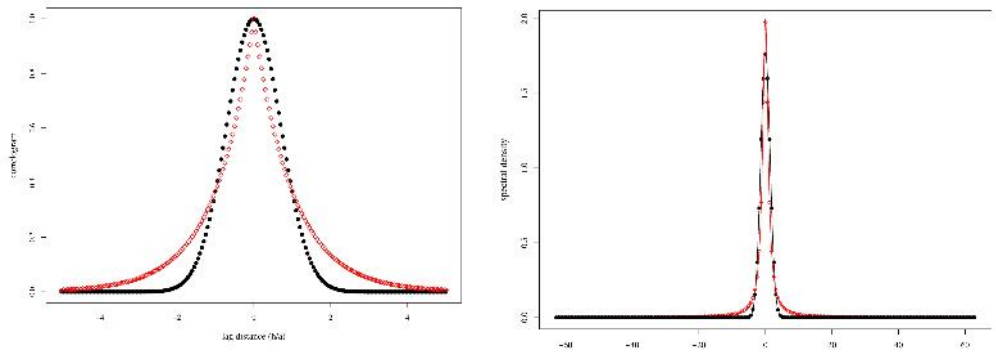

Fig. 1. Discrete correlogram and spectral density computed with FFT and with theoretical formulae (Chilès and Delfiner, 1999), for a gaussian (black dots) and an exponential (red diamonds) correlogram model, as a function of a dimensionless frequency $\nu = 2\pi a\omega$.

Table 1 Correlogram models and their respective cumulative spectral densities, with some dimensionless angular frequency intervals, according to the degree of lost information. Notation: *a* a range parameter, *t* a period parameter, $\tau = \frac{2\pi a}{t}$ its dimensionless version, and the dimensionless frequency $\nu = 2\pi a\omega$. The intervals are symmetric, and the table shows only their upper bound.

model	cumulative density	95% interval	99% interval
spherical	(no analytical form)	7.90	22.12
exponential	$\frac{1}{2}+\frac{1}{\pi}\arctan(\nu)$	12.71	31.82
gaussian	$\varphi\left(\frac{\nu}{\sqrt{2}}\right)$	2.77	3.29
hole effect	$\frac{\pi+\arctan(\nu+\tau)+\arctan(\nu-\tau)}{2\pi}$	(Fig. 2)	
delayed exponential	$\frac{1}{2}+\frac{1}{\pi}\arctan(\nu)$	12.71	31.82

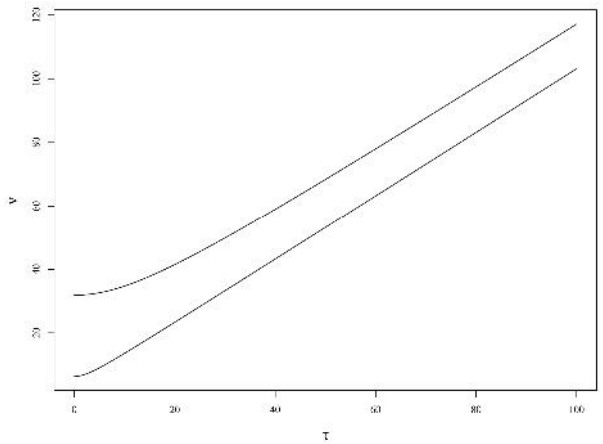

Fig. 2. Upper bound for 95% (lower curve) and 99% (upper curve) symmetric intervals of dimensionless frequency $\nu = 2\pi a\omega$ for hole effect models as a function of the dimensionless parameter $\tau = \frac{2\pi a}{t}$, with dampening range *a*, period *t* and angular frequency ω.

2. Monte Carlo Modelling

To perfectly describe a correlation function $C(h)$, being (h) a lag distance, we must know its spectral density $f^*(\omega)$ in its whole domain. If we trim the spectral density, we incur in a loss of information regarding the correlation: the degree to which this lost information will significantly alter its shape depends on the probability outside an interval of $\bar{\omega}$. Table 1 contains the spectral distribution functions for some correlation models and some exam-ples of dimensionless angular frequency intervals needed to describe the correlogram with a small loss of information. Compare these intervals and those forced by the FFT formalism (Fig. 1). By choosing a model the user has some control over the final shape of the covariance function, *e.g.* its possible range, behavior at the origin or hole effects, but the estimation procedure we propose is still non-parametric, since this choice only conditions the interval of frequencies to explore. The whole covariance estimation procedure is summarized as follows:

1. compute the experimental covariance at user- or data-defined lag distances, by using any classical standard procedure, without smoothing; it is possible to use uneven spacing if this is desired,
2. define the desired spectral frequency grid, by fixing maximum values of frequency (table 1) in each direction, and number of nodes;
3. compute a numerical approximation to the spectral density associated with the experimental covariance function; to do so, we suggest to follow (see step 6):
   (a) draw $h_n, n = 1, 2, \mathbf{i}, N$ random lag distances from a uniform distribution inside the limits of the computed covariance table,
   (b) interpolate the value of the covariance $\hat{\mathbf{C}}(h_n)$ for each h_n, by using any smooth interpolator, *e.g.* a piece-wise linear spline (1D);
   (c) for each frequency $\bar{\omega}_k$ in the user-defined spectral grid estimate the spectral density as an average of the complex exponentials of the FT, giving to each one a weight proportional to the interpolated covariance
   $$\hat{\mathbf{f}}^*(\bar{\omega}_k) \approx \sum_{n-1}^{N} \frac{\hat{\mathbf{C}}(h_n)}{N} (\cos(2\pi \langle \mathring{h}_n, \bar{\omega}_k \rangle) - i \sin(2\pi \langle \mathring{h}_n, \bar{\omega}_k \rangle));$$
   notice that if $\hat{\mathbf{C}}(h_n)$ is a matrix of auto- and cross-covariances, so will be the spectral density $f^*(\omega)$;
4. at each frequency node ω_k ensure the validity of the computed spectral density: in univariate cases, this means that this density must be always positive, while in multivariate cases, the spectral density matrix must be a positive semi-definite one; we suggest to compute the singular value decomposition of the spectral density matrix, trim its negative eigenvalues to zero, and recover a spectral density matrix by using the trimmed eigenvalues with the original eigenvectors;
5. define an output grid for the lag distances needed, *e.g.* for kriging; usually, this lag output grid will be much denser than the input lag grid of the first step;
6. compute a numerical approximation to the smoothed valid covariance associated with the spectral density, by using a Monte Carlo procedure, which estimates the covariance as the expectation of the FT with respect to the random $\bar{\omega}$:

(a) draw $\omega_n, n = 1, 2, \mathbf{i}, N$ random frequencies from a uniform distribution inside the limits of the spectral grid;

(b) interpolate the value of the spectral density $\hat{f}^*(\omega)$ for each ω_n, by using any smooth interpolator, e.g. a piece-wise linear spline (in 1D applications);

(c) for each lag distance h_k in the output lag grid, estimate the covariance as an average of the complex exponentials of the inverse FT, giving to each one a weight proportional to the interpolated spectral density

$$C(\mathbf{h}_k) \approx \sum_{n-1}^{N} \frac{\hat{f}^*(\omega_n)}{N} (\cos(2\pi \langle \mathbf{h}_k, \boldsymbol{\omega}_n \rangle) + i \sin(2\pi \langle \mathbf{h}_k, \boldsymbol{\omega}_n \rangle));$$

when estimating a correlation function, another possibility would be to use a strict Monte Carlo method, by simulating values of $\vec{\omega}$ from its spectral density an directly averaging the complex exponentials.

The suggested procedure still suffers from some problems in common with that of Yao and Journel (1998): if the trimming represents an important amount of energy, the final estimate will have an over-estimated nugget effect. However, this problem is milder in our approach, due to the lower amount of high frequencies under study. Also, we could avoid the separate estimation of the covariance at the origin, which in the FFT modelling allowed to scale the estimated correlation lookup tables to covariance tables: our approach might model directly covariances, although it is not recommended if trimming is strong.

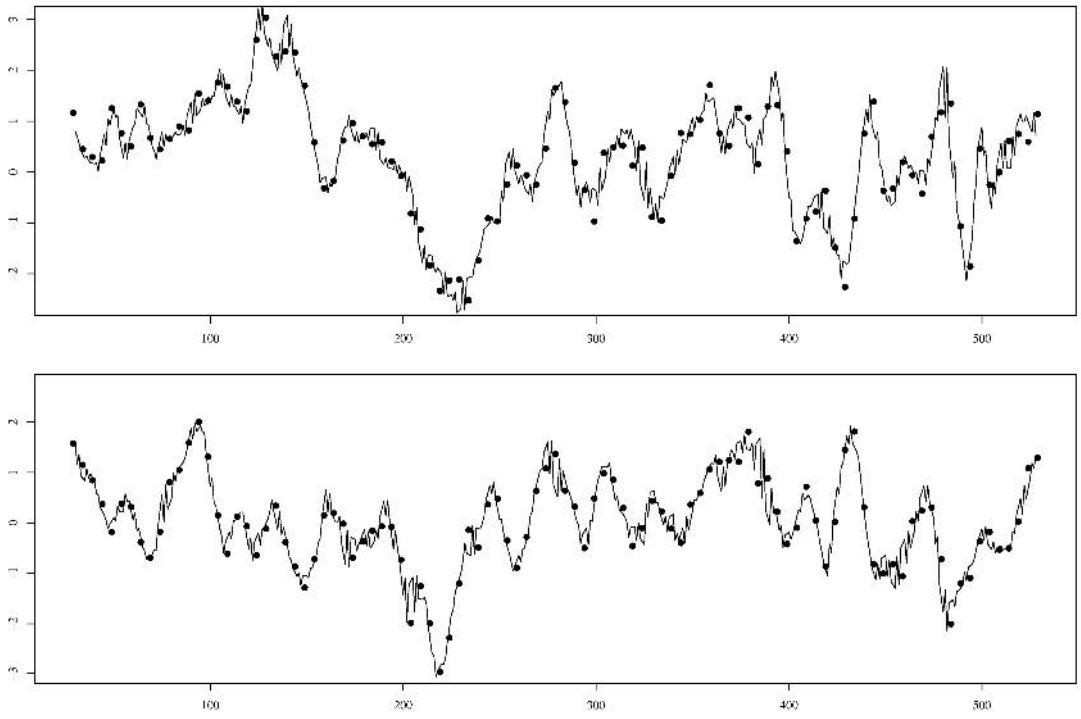

Fig. 3. Series of simulated (dots) and predicted (continuous line) values: the upper figure represents X_1, and the lower X_2.

3. A Simulated Example

To show the goodness of the method here proposed, we simulated a vectorial stochastic process, with stationary mean μ = (0,0), covariance function $C(h) - (1,1) \cdot \exp(-\frac{|h|}{10})$ and cross-covariance $C_{12}(h) - 0.5 \cdot \exp(-\frac{|h-5|}{10})$. Simulation was done using LU decomposition method, at 500 equally-spaced nodes. From them, 100 were considered as data (black dots in Fig. 3), and the rest were used for cross-validation purposes. With the data set, an experimental covariance was computed (black dots, Fig. 4, where the real covariance is also represented as a red line), and we conducted the proposed algorithm on them. The final estimated covariance function is also represented in the same figure: the real part in black, the imaginary part in blue (which should be zero, thus assessing the estimation quality). Covariance was estimated at all nodes needed in kriging ordinarily the cross-validation data set. Kriged values are shown in Fig. 3 (black line), and they are compared with the true values in Fig. 5. This last figure contains also the estimates of ordinary kriging using the true covariance model. Although the fit of the estimations with the true values and with the estimates of a classical method is good, one should notice that the values obtained with the smoothed covariance are less smooth than expected: a rather erratic small-scale fluctuation is noticeable in Fig. 3, which may be related to the residual fluctuations in the estimated covariance.

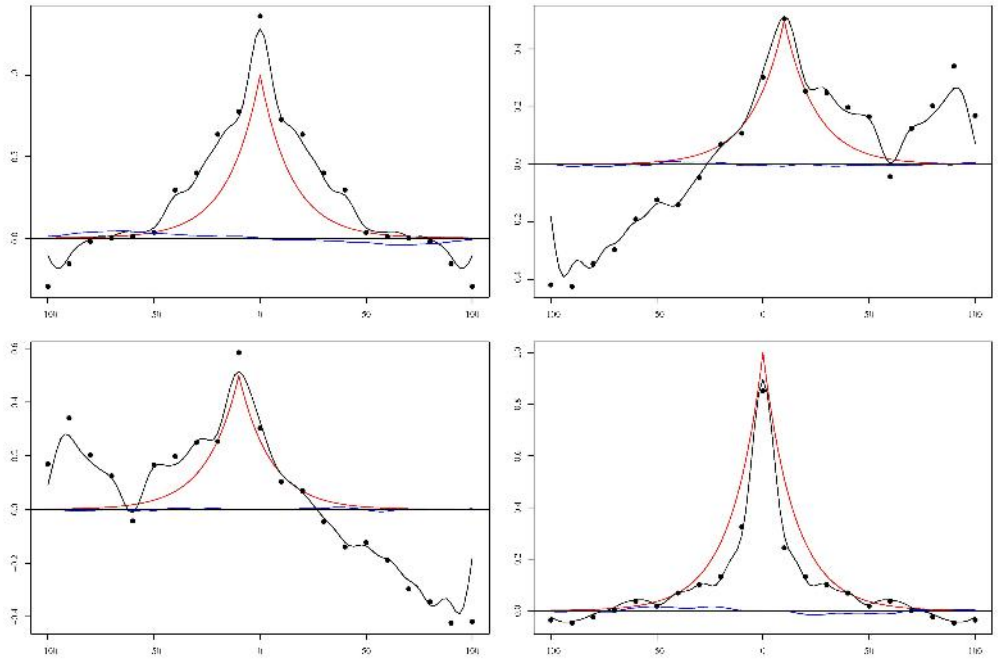

Fig. 4. Matrix of covariances of the simulated data set: true covariance (red line), estimated covariance (dots) and smoothed version (black line for the real part, blue line for the imaginary part). Left column and upper row correspond to X_1 covariances, whereas right column and lower row show X_2 covariances. Note the antisymmetric character of the cross-correlation of (X_1 , X_2,) outside the diagonal.

4. Concluding Remarks

A relatively new way of modelling covariances and validating cross-covariance systems has been in

discussion in the last years. The exising method relies on Fast FT, which has a strong discrete nature inherited from signal characteristics. Instead, we propose to use numerical integration methods (*e.g.* Monte Carlo integration), to focus the attention on narrower parts of the spectrum, those needed to reproduce a covariance model. We show that this method allows to complement the (scarce) information provided by the raw data by some characteristics (pattern of behavior at the origin, an approximation to the range and a general shape) loosely coming from a chosen model. A simulated case example showed the possibilities of such a method to validate covariances for simple kriging, although some noisy behaviour was observed in the predictions. This calls for a better way to smooth the covariance estimates, apart from applying a validation method like this. Also, the method still lacks a generalisation to spatial problems (typically, 2D and 3D applications).

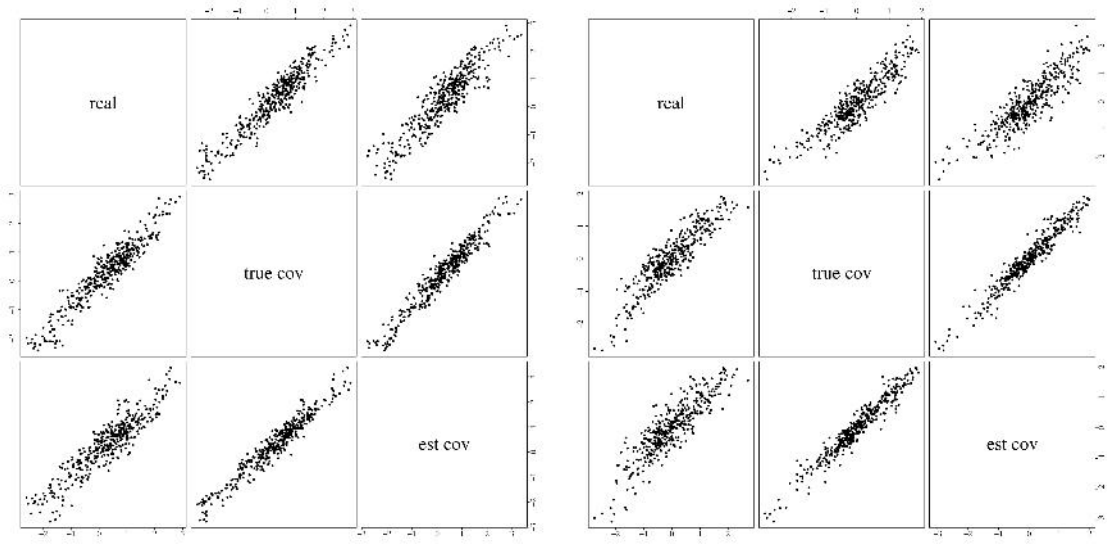

Fig. 5. Scatterplots of the predicted values, compared with the true simulations and with the predictions obtained with the true covariance model.

5. Acknowledgements

This work was elaborated during a long research stage funded by the projects A/04/33586 and 2004-BE-00147 respectively from the German academic exchange office and from the Catalan universitary grant agency, and complemented by a Student Grant of the International Association for Mathematical Geology.

6. References

Bochner, S., 1959. *Lectures on Fourier Integrals*. New Jersey (USA): Princeton University Press.

Chilès, Jean-Paul and Delfiner, Pierre ,1999. *Geostatistics—modeling spatial uncertainty*. Series in Probability and Statistics. John Wiley and Sons, Inc., New York, NY (USA) 695 p.

Cramèr H., 1940. On the theory of stationary random processes. *Annals of Mathematics* 41(1), 215–230.

Rehman, S., 1995. *Semiparametric modelling of cross-semivariograms*. Unpublished PhD disertation, Georgia Institute of Technology.

Yao, Tingting and Journel, André G., 1998. Automatic modeling of (cross) covariance tables using Fast FT. *Mathematical Geology* 30(6), 589–615.