Bull. Aust. Math. Soc. (First published online 2025), page 1 of 3*
doi:10.1017/S0004972724001357
*Provisional—final page numbers to be inserted when paper edition is published

EXPLORING OUT-OF-SAMPLE PREDICTION AND SPATIAL DEPENDENCY FOR COMPLEX BIG DATA

ADAM BILCHOURIS

(Received 4 December 2024)

2020 *Mathematics subject classification*: primary 62H11; secondary 60G15, 60G60, 62M40.
Keywords and phrases: nonparametric estimation, covariance, correlation, autocovariance, random fields, genetic algorithms.

Specific classes of stochastic processes and their multidimensional counterparts, known as random fields, often depend on several parameters that define their behaviour. Therefore, obtaining estimates for these parameters can be valuable in modelling and analysing the properties of these stochastic objects. For example, Gaussian processes and Gaussian random fields can be defined entirely through their mean and autocovariance functions [3]. As random fields have widespread applications in numerous fields, including, but not limited to, earth sciences, machine learning and physics [6], methods to estimate these parameters accurately are important and useful not only from a theoretical perspective, but have also found numerous applications in the analysis of real data.

Nonparametric estimation methods of the autocovariance function for random fields found in the literature often have desirable theoretical properties, such as unbiasedness, positive-definiteness and robustness. Also, some estimators can be computationally slow, particularly in the context of spatial big data. Furthermore, the estimators can struggle to provide a reasonable result when the autocovariance functions are long range or cyclically dependent. For example, the standard method-of-moments estimator for the isotropic autocovariance function $C(\tau)$ is given by

$$\widehat{C}(\tau) = \frac{1}{|N(\tau)|} \sum_{\boldsymbol{s},\boldsymbol{t} \in N(\tau)} (X(\boldsymbol{s}) - \overline{X})(X(\boldsymbol{t}) - \overline{X}),$$

where $N(\tau) = \{(\boldsymbol{s},\boldsymbol{t}) : \|\boldsymbol{s} - \boldsymbol{t}\| = \tau; \boldsymbol{s},\boldsymbol{t}, \in \mathbb{R}^d, \tau \in \mathbb{R}\}$, $X(\cdot)$ is a random field and $\overline{X}$ is the sample mean. This estimator is not guaranteed to be positive-definite, meaning it does not always produce a valid isotropic autocovariance function [8] and if spatial

Masters thesis submitted to La Trobe University in June 2022; degree approved on 25 October 2024; principal supervisor Andriy Olenko, co-supervisor Nikolai Leonenko (Cardiff University, UK). This work was supported by an Australian Government Research Training Program Scholarship.

dependency is present, a bias may appear in the estimate of the mean, and thus in the autocovariance function estimate [6]. Furthermore, finding the elements of $N(\tau)$ can be computationally slow as the dimension d increases or if the number of sample locations increases, such as in the case of spatial big data.

This thesis is split into three parts. The first introduces some basic concepts of time series and estimates of the parameters of a Gegenbauer process (see [1]) which are then applied to a simulated Gegenbauer process, where the parameters are chosen such that the process has long memory.

The second part considers estimating and testing the normality of the first Minkowski functional of homogeneous isotropic Gaussian random fields under varying degrees of spatial dependency, continuing the research in [7]. The covariance functions considered are the Gaussian, Bessel and Cauchy, which have short-range, cyclical and long-range dependence, respectively. Also, we raised the random fields to various powers to see how normality results changed. It was found that under the Cauchy covariance function, for all tests, the first Minkowski functional was normal.

Several nonparametric autocovariance function estimators found in the literature were compared. These nonparametric estimators mainly considered nonstandard approaches, such as kernel regression, quantile-based estimators and estimators using linear combinations of basis functions, such as B-splines. Some attention is also paid to nonparametric estimators of the (semi-)variogram. Corrections of these estimators were also considered, such as a kernel multiplier to remove estimation artefacts, or making the estimate positive-definite and thus a valid autocovariance function. For example, the kernel regression estimator, found in [4, 5], is of the form

$$\widehat{C}_H(\boldsymbol{t}) = \sum_i \sum_j \check{X}_{ij} K((\boldsymbol{t} - \boldsymbol{t}_{ij})/b) \Big/ \sum_i \sum_j K((\boldsymbol{t} - \boldsymbol{t}_{ij})/b),$$

where $\check{X}_{ij} = (X(\boldsymbol{t}_i) - \overline{X})(X(\boldsymbol{t}_j) - \overline{X})$, $\boldsymbol{t}_{ij} = \boldsymbol{t}_i - \boldsymbol{t}_j$, $K(\cdot)$ is a kernel and b is some bandwidth.

Estimates using the autocovariance function estimators were obtained in simulation studies, where the same covariance functions as above were used, each giving its own strength of spatial dependency, weak, cyclic and strong. The estimates were compared using several metrics, such as the MSPE of Kriging predictions for out-of-sample observations. It was demonstrated that certain estimators fail to capture the behaviour of the covariance function, such as cyclicality, which was shown in these simulation studies. Simulation studies were conducted for both the one-dimensional case, a Gaussian process, and the two-dimensional case, an isotropic Gaussian random field, which can be found in [2].

As these estimators can be computationally slow, particularly for large samples, this motivated the proposal of new autocovariance function estimators, aiming to be computationally faster without sacrificing too much accuracy. The proposed estimators aimed to use the isotropy and stationarity of the considered random fields, and ideas from time series analysis. For example, due to isotropy, estimates along different angles can be averaged to obtain a single estimate, and due to stationarity, estimates do not depend on spatial location, meaning, once again, many estimates can be combined

https://doi.org/10.1017/S0004972724001357 Published online by Cambridge University Press

to obtain a single estimate. Estimates are obtained and compared using simulated two-dimensional isotropic Gaussian random fields. The Bessel and Cauchy covariance functions and sample domains of varying size and density were considered when performing estimation, mimicking spatial big data. In addition to comparing the estimation errors, the time to compute the estimates was also considered.

The third part introduces an out-of-sample correction method based on scaling features of objects, using genetic algorithms for variable and regression model selection, and determining outliers through various methods. The use of a general model, that is, a model trained on all available objects, has proven to be inappropriate when performing prediction on anomalous objects. Typically, anomalous objects are manually assessed which can be costly with regards to time and resources. We aimed to create an automated correction process for predictions of anomalous objects through a nonlinear out-of-sample correction model under the assumption of scalable features. These ideas are applied to a housing data set provided by ANZ Bank and CoreLogic. A report on these results by A. Bilchouris, I. Donhauzer, A. Olenko and D. Ostapenko ['Corrected out-of-sample prediction with property pricing applications'] is in preparation.

We demonstrate that well-known autocovariance function estimators may not be suitable under certain dependency structures, when dealing with spatial big data or when computational time needs to be considered. Also, the thesis develops several methodological approaches for the analysis of big spatial data and multidimensional datasets.

We used R version 4.1.0 for all computations, and the source code can be found at https://github.com/AdamBilchouris/MastersCode. The implementation of the results obtained is currently being developed as an R package.

References

[1] H. M. Alomari, A. Ayache, M. Fradon and A. Olenko, 'Estimation of cyclic long memory parameters', *Scand. J. Stat.* **47**(1) (2019), 104–133.

[2] A. Bilchouris and A. Olenko, 'On nonparametric estimation of covariograms', *Aust. J. Stat.* **51**(1) (2025), 112–137.

[3] J. P. Chilès and P. Delfiner, *Geostatistics: Modeling Spatial Uncertainty* (Wiley, New York, 2012).

[4] P. Hall, N. I. Fisher and B. Hoffmann, 'On the nonparametric estimation of covariance functions', *Ann. Statist.* **22**(4) (1994), 2115–2134.

[5] P. Hall and P. Patil, 'Properties of nonparametric estimators of autocovariance for stationary random fields', *Probab. Theory Related Fields* **99**(3) (1994), 399–424.

[6] D. T. Hristopulos, *Random Fields for Spatial Data Modeling: A Primer for Scientists and Engineers*, Advances in Geographic Information Science (Springer, Dordrecht, 2020).

[7] N. Leonenko and A. Olenko, 'Sojourn measures of student and Fisher–Snedecor random fields', *Bernoulli* **20**(3), 1454–1483.

[8] A. M. Yaglom, *Correlation Theory of Stationary and Related Random Functions, Volume I: Basic Results* (Springer, New York, 1987).

ADAM BILCHOURIS, Mathematics and Physical Sciences,
La Trobe University, Melbourne, Victoria 3086, Australia
e-mail: a.bilchouris@latrobe.edu.au

https://doi.org/10.1017/S0004972724001357 Published online by Cambridge University Press