Geostatistical Analysis using K-splines in the Geoadditive Model

Yannick Vandendijck[1], Christel Faes[1], Niel Hens[12]

[1] Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University, Diepenbeek, Belgium
[2] Centre for Health Economic Research and Modeling Infectious Diseases, Vaccine and Infectious Disease Institute, University of Antwerp, Wilrijk, Belgium

E-mail for correspondence: yannick.vandendijck@uhasselt.be

Abstract: In geostatistics, both kriging and smoothing splines are commonly used to predict a quantity of interest. The geoadditive model proposed by Kammann and Wand (2003) represents a fusion of kriging and penalized spline additive models. The fact that the underlying spatial covariance structure is poorly estimated using geoadditive models is a drawback. We describe K-splines, an extension of geoadditive models such that estimation of the underlying spatial process parameters and predictions of the spatial map are performed with the same accuracy and precision as in kriging.

Keywords: Covariogram; Kriging; Mixed Model; Penalized Spline.

1 Introduction

The objective of geostatistics is to produce a map of a variable of interest on a specified domain based on observations which are measured with or without noise. Consider the geostatistical model $y(\mathbf{s}_i) = z(\mathbf{s}_i) + \varepsilon_i, i = 1, \ldots, n$, where the $y(\mathbf{s}_i)$ are observed data values from the underlying true values $z(\mathbf{s}_i)$. These data values are noise-corrupted by white-noise error terms ε_i. The spatial locations $\mathbf{s}_i$ belong to a specified continuous domain $D \subset \mathbb{R}^d$. The idea of geostatistics is to use the data $y(\mathbf{s}_i)$ to make predictions of $z(\mathbf{s}_0)$ where $\mathbf{s}_0 \in D$. Both kriging and spline methods can be used to handle geostatistical problems. In kriging, the values $z(\mathbf{s}_i)$ are assumed to be the realisations of an autocorrelated random process (Cressie, 1993). Smoothing splines assume that the $z(\mathbf{s}_i)$ are the values of a smooth non-parametric function (see e.g., Hutchinson and Gessler, 1994).

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

In kriging, linearity of the covariate effects is usually assumed. Kammann and Wand (2003) merged kriging and additive models to allow for non-linear relationships between covariates and the response variable in geostatistics. Their so-called geoadditive model consists of a penalized spline additive model with a geostatistical extension. The geoadditive model can be expressed as a linear mixed model which allows for estimation and inference using standard methodology. The drawback of their model is the biased estimation of the underlying spatial process.
Vandendijck et al. (2015) introduced the concept of kriging-splines, abbreviated by K-splines, which extends geoadditive models such that the estimation of the underlying spatial process and prediction of the map of interest is performed with similar accuracy and precision as in kriging. By showing a theoretical connection between kriging and K-splines, it is presented how the spatial covariance structure (covariogram) implied by K-splines is derived. K-splines are also embedded within the linear mixed model framework and the estimation uses a two-step likelihood procedure.

2 K-splines

For simplicity, suppose the data are $(y_i, \mathbf{s}_i, a_i, b_i), 1 \leq i \leq n$, where y_i is the value of the ith response, a_i and b_i are the values of two predictor variables a and b, and $\mathbf{s}_i$ represents the geographical location. Suppose the predictor a enters the model linearly and that the predictor b enters the model non-linearly. The geoadditive model is

$$y_i = \beta_0 + \beta_a a_i + f(b_i) + S(\mathbf{s}_i) + \varepsilon_i, \qquad 1 \leq i \leq n, \tag{1}$$

where f is a smooth function of b and S is the geographical component of the model. Both f and S are modelled using penalized spline functions. We use the thin plate spline family to construct f. The spatial component S is modelled through a set of radial basis functions and is of the form $S(\mathbf{s}) = \sum_{k=1}^{K_s} u_i^s g_\phi(\mathbf{s} - \boldsymbol{\kappa}_k^s)$, where g_ϕ can be any of the proper covariance or generalized covariance functions used in kriging. The subscript in g_ϕ is used to denote a possible dependence on a spatial decay parameter ϕ. An overview of the most important covariance functions g_ϕ that can be used are given in Table 1. The vector $(u_1^s, \ldots, u_{K_s}^s)$ contains the K_s unknown knot coefficients that are penalized to overcome overfitting. The K_s knots $\boldsymbol{\kappa}_1^s, \ldots, \boldsymbol{\kappa}_{K_s}^s$ are a representative subset of $(\mathbf{s}_1, \ldots, \mathbf{s}_n)$ used for the construction of the basis functions.
Kammann and Wand (2003) propose to choose ϕ via the simple rule $\hat{\phi} = \max_{1 \leq i,j \leq n} \|\mathbf{s}_i - \mathbf{s}_j\|$. We propose K-splines, a new estimation approach for geoadditive models that allows for accurate estimation of the parameter ϕ. This enables one to estimate the underlying spatial process accurately and precisely. A two-stage iterative estimation approach is proposed to estimate

TABLE 1. Some important and often used covariance functions g_ϕ.

Exponential	$g_\phi(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x}\|}{\phi}\right)$
Gaussian	$g_\phi(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x}\|^2}{\phi^2}\right)$
Spherical	$g_\phi(\mathbf{x}) = \left(1 - \frac{3}{2}\frac{\|\mathbf{x}\|}{\phi} + \frac{1}{2}\frac{\|\mathbf{x}\|^3}{\phi^3}\right)$
Matérn ($\nu = \frac{3}{2}$)	$g_\phi(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x}\|}{\phi}\right)\left(1 + \frac{\|\mathbf{x}\|}{\phi}\right)$
Circular	$g_\phi(\mathbf{x}) = 1 - \frac{2}{\pi}\left(\vartheta\sqrt{1-\vartheta^2} + \arcsin\vartheta\right)$, with $\vartheta = \min\left(\frac{\|\mathbf{x}\|}{\phi}, 1\right)$

The parameter ϕ is positive for each function.

the parameters. At the first stage, the linear mixed model representation of (1) is estimated fixing ϕ in g_ϕ at its current value, and in the second stage the parameter ϕ is optimized fixing the linear mixed model parameters. For mode details on the estimation procedure and inference using K-splines, we refer to Vandendijck et al. (2015).

3 Simulation Study

We consider as spatial domain the unit square. Data at a spatial location $\mathbf{s} = (s_x, s_y)$ on this square is simulated using the model

$$y_{\mathbf{s}} = S(\mathbf{s}) - 0.5x_{1\mathbf{s}} + \sin(2\pi x_{2\mathbf{s}}) + \varepsilon_{\mathbf{s}}, \tag{2}$$

where $\varepsilon_{\mathbf{s}} \sim \mathcal{N}(0, \sigma^2_\varepsilon = 0.10)$, $x_{1\mathbf{s}} \sim \text{Unif}[0-1]$, $x_{2\mathbf{s}} \sim \text{Unif}[0-1]$ and $S(\mathbf{s})$ is a zero-mean Gaussian Random Field (GRF) (Gelfand et al., 2010) with a Gaussian covariogram without nugget, namely $K(\mathbf{h}) = c_s \exp\left(-\frac{\|\mathbf{h}\|^2}{\tau^2}\right)$. We consider $c_s = 0.50$ and $\tau = 0.15$. We obtain 250 simulated realizations from (2). From each realization we draw a random sample of size $n = 500$. For each simulated dataset, the covariogram parameters (c_s, τ) and the measurement error parameter σ^2_ε were estimated using seven different methods: (1) Direct maximum likelihood (D-ML) parameter estimation for GRFs; (2) Direct restricted maximum likelihood (D-REML) parameter estimation for GRFs; (3) Weighted least squares (WLS) estimation of the empirical semivariogram; (4) Maximum likelihood estimation as described in Kammann and Wand (2003) (KW-ML); (5) Restricted maximum likelihood estimation as described in Kammann and Wand (2003) (KW-REML); (6) Maximum likelihood estimation using K-splines (KS-ML); and (7) Restricted maximum likelihood estimation using K-splines (KS-REML). In addition, the prediction performance at five locations on the considered spatial domain was evaluated. D-ML, D-REML and WLS are kriging approaches in which the covariates enter the mean function linearly. For KW-ML, KW-REML, KS-ML and KS-REML, we use model (1) where 150 knots are used to model the spatial component S.

Results are displayed in Table 2. It is observed that K-splines perform better for the estimation of the covariogram parameters c_s and τ. Because the covariates enter the mean function linearly in D-ML, D-REML and WLS, the measurement error parameter σ^2_ε is not well estimated. The estimated covariogram parameters for KW-ML and KW-REML are seriously biased. This can be expected since the approach of Kammann and Wand (2003) does not attempt to estimate these parameters well. In terms of prediction, we see that K-splines perform the best.

TABLE 2. MSE results over 250 simulations for the covariogram parameters and predictions with corresponding 95% confidence intervals coverage.

	c_s	τ	σ^2_ε	c_s/τ	pred.	cov.	cov.[a]
D-ML	1.67	0.02	3.76	58.97	23.65	57.2	
D-REML	1.82	0.02	3.83	62.24	23.64	57.2	
WLS	2.03	0.06	3.88	75.33	23.88	57.6	
KW-ML	$> 10^3$	148.52	3.71	$> 10^3$	14.61	63.4	
KW-REML	$> 10^3$	148.52	3.73	$> 10^3$	14.61	64.4	
K-ML	1.51	0.01	0.01	49.11	2.39	94.8	95.0
K-REML	1.55	0.01	0.01	50.06	2.40	94.8	95.3

a: based on a bootstrap procedure (see Vandendijck et al., 2015)

4 Conclusion

K-splines offer a framework wherein the covariogram parameters in a geoadditive model are estimated accurately and precisely. From simulation studies we can conclude that predictions benefit from this.

References

Cressie, N.A.C. (1993). *Statistics for spatial data*. New York: Wiley.

Gelfand, A.E., Diggle, P.J., Fuentes, M. and Guttorp, P. (Eds.) (2010). *Handbook of spatial statistics*. Boca Raton: Chapman & Hall/CRC.

Hutchinson, M.F. and Gessler, P.E. (1994). Splines - more than just a smooth interpolator. *Geoderma*, **62**, 45 – 67.

Kammann, E.E. and Wand, M.P. (2003). Geoadditive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **52**, 1 – 18.

Vandendijck, Y., Faes, C. and Hens, N. (2015). K-splines: A new method towards geostatistical analysis. *Technical Report, Hasselt University, 02/02/2015*.