Bandwidth selection for local linear estimation of the spatial trend

R. Fernández-Casal[1,*] and P. García-Soidán[2]

[1] *Facultad de Informática, University of A Coruña, 15071 A Coruña (Spain); ruben.fcasal@udc.es*
[2] *Facultad de CC Sociales y Comunicación, University of Vigo, 36005 Pontevedra (Spain); pgarcia@uvigo.es*
[*] *Corresponding author*

Abstract. *The estimation of the large-scale variability (spatial trend) of a geostatistical process can be accomplished by using nonparametric regression. In this work, we will focus on the local linear estimation of the trend function and, more specifically, on the selection of the bandwidth matrix involved. An overview of approaches suggested for the latter aim will be outlined and additional alternatives will be introduced, based on correcting the cross-validation selector. Numerical studies will be carried out for comparison of the different proposals and an application to a real data set will be also included to illustrate their behavior in practice.*

Keywords. *Bandwidth matrix; Local linear regression; Spatial trend.*

1 Introduction

In geostatistics, the assumption of stationarity greatly simplifies modeling of spatial data, thus enabling to make inference about a target site from those data located nearby; see, for instance, Cressie [1]. However, the aforementioned condition is quite restrictive in practice, as the underlying random process usually depart from it due to the large-scale variability. Then, estimation of the spatial trend allows the researchers to extend the validity of those techniques, originally designed for stationary data, to a more general setting. In particular, we will assume that the spatial process $\{Y(\mathbf{x}) : \mathbf{x} \in D \subset \mathbb{R}^d\}$ can be modeled as:

$$Y(\mathbf{x}) = m(\mathbf{x}) + \varepsilon(\mathbf{x}),$$

where:

- $m(\cdot)$ is the trend function.

- ε is a second-order stationary process with zero mean and covariogram C, satisfying that:

$$C(\mathbf{h}) = Cov(\varepsilon(\mathbf{x}), \varepsilon(\mathbf{x}+\mathbf{h}))$$

  for all $\mathbf{h} \in \mathbb{R}^d$.

From a nonparametric perspective, the approximation of the spatial trend can be addressed through a kernel approach, by considering either a Nadaraya-Watson estimator or a more general approach given by the local polynomial fitting, as described in Fan and Gijbels [4]. The second procedure provides an estimator with a remarkable advantage, namely, the absence of boundary effects, unlike the one derived from the Nadaraya-Watson approach. For the sake of simplicity, this work is focused on the local linear alternative. In particular we will address the selection of the bandwidth matrix, by bearing in mind that traditional methods do not perform well in the presence of correlated errors, as remarked in Opsomer *et al.* [8]. Then, the different approaches for selection of the bandwidth matrix will be developed and put into comparison in a numerical study covering a range of dependence situations. An application to a real data set will be also included to illustrate the behavior in practice of the proposed bandwidths.

2 Main results

Suppose that n data $Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_n)$ have been observed at the respective locations $\mathbf{x}_1, \ldots, \mathbf{x}_n$. The local linear trend estimator is given by:

$$\hat{m}_{\mathbf{H}}(\mathbf{x}) = \mathbf{e}_1^t \left(\mathbf{X}_{\mathbf{x}}^t \mathbf{W}_{\mathbf{x}} \mathbf{X}_{\mathbf{x}}\right)^{-1} \mathbf{X}_{\mathbf{x}}^t \mathbf{W}_{\mathbf{x}} \mathbf{Y} \equiv \mathbf{s}_{\mathbf{x}}^t \mathbf{Y},$$

where $\mathbf{e}_1 = (1,0,...,0)^t \in \mathbb{R}^{d+1}$, $\mathbf{X}_{\mathbf{x}}$ is a $n \times (d+1)$ matrix whose i-th row equals $(1, (\mathbf{x}_i - \mathbf{x})^t)$, $i = 1,...,n$, $\mathbf{W}_{\mathbf{x}} = \text{diag}\{K_{\mathbf{H}}(\mathbf{x}_1 - \mathbf{x}), ..., K_{\mathbf{H}}(\mathbf{x}_n - \mathbf{x})\}$, $K_{\mathbf{H}}(\mathbf{u}) = |\mathbf{H}|^{-1} K(\mathbf{H}^{-1}\mathbf{u})$, K is a d-variate kernel density and $\mathbf{H}$ is the bandwidth matrix. Asymptotic properties of these estimators are studied in Opsomer *et al.* [8] for correlated bidimensional data and they are generalized in Liu [7] to the multidimensional setting.

The optimal bandwidth is taken here as the minimizer of the mean averaged squared error:

$$\text{MASE}(\mathbf{H}) = \frac{1}{n} E\left((\mathbf{SY} - \mathbf{m})^t (\mathbf{SY} - \mathbf{m})\right) = \frac{1}{n} (\mathbf{Sm} - \mathbf{m})^t (\mathbf{Sm} - \mathbf{m}) + \frac{1}{n} tr\left(\mathbf{S}\boldsymbol{\Sigma}\mathbf{S}^t\right)$$

where $\mathbf{m} = (m(\mathbf{x}_1), \ldots, m(\mathbf{x}_n))^t$, $\boldsymbol{\Sigma}$ is the theoretical covariance matrix of the residuals and $\mathbf{S}$ is the smoothing matrix, namely, the $n \times n$ matrix whose i-th row equals $\mathbf{s}_{\mathbf{x}_i}^t$ (the smoother vector for $\mathbf{x} = \mathbf{x}_i$), $i = 1,...,n$. Note that this optimal bandwidth, denoted as $\mathbf{H}_{MASE}$, cannot be (normally) used in practice, due to its dependence on unknown terms.

When data are uncorrelated, different criteria have been proposed for an appropriate specification of the bandwidth from the available data. The cross validation and generalized cross validation approaches, analyzed in Craven and Wahba [3], are among the most widely used for the latter aim, which are respectively based on minimizing:

$$CV(\mathbf{H}) = \frac{1}{n} \sum_{i=1}^{n} \left(Y(\mathbf{x}_i) - \hat{m}_{-i}(\mathbf{x}_i)\right)^2$$

and:

$$GCV(\mathbf{H}) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Y(\mathbf{x}_i) - \hat{m}(\mathbf{x}_i)}{1 - \frac{1}{n} tr(\mathbf{S})}\right)^2$$

where $\hat{m}_{-i}(\mathbf{x}_i)$ is the trend estimate obtained without considering $Y(\mathbf{x}_i)$.

Under dependence, the selectors designed for independent data tend to provide wrong choices of the bandwidth, as remarked in Opsomer *et al.* [8]. To solve the aforementioned problem, different approaches can be applied, which may be obtained by adapting the cross-validation criteria, proposed in Chu and Marron [2] for the unidimensional case, to the spatial setting. Thus, for instance, a bandwidth could be taken to minimize:

$$MCV(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(Y(\mathbf{x}_i) - \hat{m}_{-N(i)}(\mathbf{x}_i)\right)^2$$

where $\hat{m}_{-N(i)}(\mathbf{x}_i)$ denotes the trend estimate obtained when ignoring observations in a neighbourhood $N(i)$ around $\mathbf{x}_i$. An alternative is suggested in Francisco-Fernández and Opsomer [6], which adjusts the generalized cross-validation criterion for the effect of spatial correlation and is based on minimizing:

$$CGCV(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y(\mathbf{x}_i) - \hat{m}(\mathbf{x}_i)}{1 - \frac{1}{n\sigma^2}tr(\mathbf{S}\boldsymbol{\Sigma})}\right)^2$$

where $\sigma^2 = C(\mathbf{0})$ is the variance (or sill).

In the current work, additional options will be derived by taking into account that:

$$E\left((\mathbf{Y}-\mathbf{SY})^t(\mathbf{Y}-\mathbf{SY})\right) = (\mathbf{Sm}-\mathbf{m})^t(\mathbf{Sm}-\mathbf{m}) + tr\left(\mathbf{S}\boldsymbol{\Sigma}\mathbf{S}^t\right) - 2tr(\mathbf{S}\boldsymbol{\Sigma}) + tr(\boldsymbol{\Sigma})$$

and, therefore:

$$E(CV(\mathbf{H})) \simeq \text{MASE}(\mathbf{H}) + \sigma^2 - \frac{2}{n}tr(\mathbf{S}_{-1}\boldsymbol{\Sigma})$$

where $\mathbf{S}_{-1}$ denotes the smoothing matrix corresponding to the estimates $(\hat{m}_{-1}(\mathbf{x}_1),\ldots,\hat{m}_{-n}(\mathbf{x}_n))^t$. Then, the selection of the bandwidth could be addressed by considering corrected versions of the cross-validation approaches, so as to minimize:

$$CCV(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}(Y(\mathbf{x}_i) - \hat{m}_{-i}(\mathbf{x}_i))^2 + \frac{2}{n}tr(\mathbf{S}_{-i}\boldsymbol{\Sigma})$$

or, alternatively:

$$CMCV(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(Y(\mathbf{x}_i) - \hat{m}_{-N(i)}(\mathbf{x}_i)\right)^2 + \frac{2}{n}tr(\mathbf{S}_{-N}\boldsymbol{\Sigma})$$

where $\mathbf{S}_{-N}$ denotes the smoothing matrix corresponding to the estimates $(\hat{m}_{-N(1)}(\mathbf{x}_1),\ldots,\hat{m}_{-N(n)}(\mathbf{x}_n))^t$

The above criteria cannot be used in practice, since the theoretical covariance matrix is usually unknown and so it must be replaced by an appropriate estimate. To deal with the latter, under non-stationary mean, the typical procedure consists in first removing the trend and then approximating the variogram (or the covariogram) from the resulting residuals. However, it is well known that the direct use of the residuals leads to a bias effect in the variogram estimation (Cressie [1], section 3.4.3). Then a different option will be explored to obtain bias-corrected nonparametric semivariogram estimates, by proceeding through a similar approach to that described in Fernández-Casal and Francisco-Fernández [5], based on the application of the iterative algorithm implemented in the `np.svariso.corr` function of the R package `npsp`.

Acknowledgments. The first author's work has been partially supported by MEC Grant MTM2008-03010. The second author acknowledges financial support from the Spanish Government by Grants CONSOLIDER-INGENIO 2010 CSD2008-00068 and TEC2011-28683-C02-02, as well as from Xunta de Galicia by Grant CN2012/279.

References

[1] Cressie, N. (1993). *Statistics for Spatial Data*. Wiley, New York.

[2] Chu, C. K. and Marron, J. S. (1991). Comparison of Two Bandwidth Selectors with Dependent Errors. *The Annals of Statistics* **19**, 1906–1918.

[3] Craven, P. and Wahba G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik* **31**, 377–403.

[4] Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*. Chapman & Hall, London.

[5] Fernández-Casal R. and Francisco-Fernández M. (2013) Nonparametric bias-corrected variogram estimation under non-constant trend. *Stoch. Environ. Res. Ris. Assess* **28**, 1247–1259.

[6] Francisco-Fernández, M. and Opsomer, J. D. (2005). Smoothing paremeter selection methods for nonparametric regression with spatially correlated errors. *The Canadian Journal of Statistics* **33**, 279–295.

[7] Liu, X. (2001). *Kernel smoothing for spatially correlated data*. Ph. D. thesis. Department of Statistics, Iowa State University.

[8] Opsomer, J. D., Wang, Y. and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science* **16**, 134–153.