BANDWIDTH SELECTION FOR LOCAL LINEAR ESTIMATION OF THE SPATIAL TREND

Rubén Fernández-Casal[1], Pilar García-Soidán[2]
[1] University of A Coruña (Spain); ruben.fcasal@udc.es
[2] University of Vigo (Spain); pgarcia@uvigo.es

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.

Introduction

- Let us assume that the spatial process $\{Y(\mathbf{x}) : \mathbf{x} \in D \subset \mathbb{R}^d\}$ can be modeled as:

$$Y(\mathbf{x}) = m(\mathbf{x}) + \varepsilon(\mathbf{x}), \tag{1}$$

  where $m(\cdot)$ is the trend function and $\varepsilon(\cdot)$ 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 trend can be addressed through a Nadaraya-Watson estimator or a more general approach given by the local polynomial fitting, as described in Fan and Gijbels (1996). The second procedure provides an estimator with a remarkable advantage, namely, the absence of boundary effects.

- This work is focused on the local linear alternative, where a bandwidth must be estimated and this issue will be addressed through different approaches obtained by correcting the cross-validation selector.

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

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

  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.
  - $\mathbf{H}$ is the bandwidth matrix.

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

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

  where:
  - $\mathbf{m} = (m(\mathbf{x}_1), \ldots, m(\mathbf{x}_n))^t$.
  - $\boldsymbol{\Sigma}$ is the theoretical covariance matrix of the errors.
  - $\mathbf{S}$ is the $n \times n$ matrix whose i-th row equals $\mathbf{s}_{\mathbf{x}_i}^t$

  This optimal bandwidth, denoted as $\mathbf{H}_{MASE}$, cannot be used in practice, due to its dependence on unknown terms.

- When data are uncorrelated, the cross validation and generalized cross validation approaches are among the most widely used for selection of the bandwidth, which are respectively based on minimizing:

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

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

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

Main results

- Under dependence, a bandwidth selector can be obtained by adapting the cross-validation criteria, proposed in Chu and Marron (1991). For instance, we can 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 (2005), 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 criteria will be derived by taking into account that:

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

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

- By considering corrected versions of the cross-validation approaches, the bandwidth could be taken 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 theoretical covariance matrix is usually unknown and it must be replaced by an appropriate estimate. As in the traditional geostatistical approaches, the usual dependence estimation method consists of removing the trend and estimating the variogram from the residuals $\hat{\varepsilon} = \mathbf{Y} - \mathbf{SY}$. Nevertheless, it is well-known that the direct use of the residuals in variogram estimation may produce a strong underestimation of the small-scale variability of the process (e.g. Cressie, 1993, Section 3.4.3).

- As this bias may also have a significant impact on the bandwidth selection criteria, we propose obtaining bias-corrected nonparametric semivariogram estimates, by proceeding through a similar approach to that described in Fernández-Casal and Francisco-Fernández (2014), based on the application of the iterative algorithm implemented in the `np.svariso.corr` function of the R package `npsp`.

Simulation results

- The behavior of the different criteria was evaluated in a simulation study. $N = 1,000$ samples of different sizes were generated following model (1) on a regular grid in the unit square, with mean functions $m_1(x_1, x_2) = \sin(2\pi x_1) + 4(x_2 - 0.5)^2$ and $m_2(x_1, x_2) = 4((x_1 - 0.5)^2 + (x_2 - 0.5)^2)$. The random errors ε_i were simulated through normal distributions with zero mean and isotropic exponential covariogram:

$$\gamma_\theta(\mathbf{u}) = c_0 + c_1\left(1 - \exp\left(-3\frac{\|\mathbf{u}\|}{r}\right)\right),$$

  (for $\mathbf{u} \neq \mathbf{0}$), where c_0 is the nugget effect, c_1 is the partial sill and r is the practical range.

- The values considered in the simulations were:
  - Sample sizes of $n = 10 \times 10$, 17×17 and 20×20.
  - Sill (variance) values of $\sigma^2 = 0.16$ and 1.
  - Practical ranges of $r = 0.3$, 0.6 and 0.9.
  - Nugget values of 0%, 20%, 50% and 100% of σ^2 ($c_1 = \sigma^2 - c_0$).

- Firstly, optimal bandwidths were computed through different criteria, by using the true covariance matrices. In general, it was observed a better performance of the $CGCV$ and CCV selectors. These criteria provided similar results for the different settings, with a slight improvement of the CCV selector when the correlation was strong.

- A summary of the squared errors, for $n = 20 \times 20$, $m = m_1$, $\sigma^2 = 1.0$, $r = 0.6$ and $c_0 = 0.2$, is shown in Table 1, where $MCVx$ stands for the modified cross-validation criterion when leaving out x additional observations in each direction (e.g. $CMCV2$ leaves out 25 values on a central point).

	$MASE$	GCV	CV	$MCV1$	$MCV2$	$CGCV$	CCV	$CMCV1$	$CMCV2$
Mean	.386	.630	.592	.505	.449	.423	.422	.428	.430
Median	.165	.280	.263	.222	.196	.185	.186	.186	.187
SD	.580	.914	.858	.740	.663	.625	.619	.633	.635

Table 1. Summary of squared errors of trend estimates obtained with the different criteria by using the true covariance matrices.

- Additionally, as the theoretical covariance matrix is usually unknown in practice, optimal bandwidths were computed with the $CGCV$ and CCV criteria through estimated covariance matrices. The covariances were estimated by using uncorrected and corrected residual-based nonparametric variogram estimators (Fernández-Casal and Francisco-Fernández, 2014).

- The direct use of residuals in variogram estimation seems to produce an underestimation of the small-scale variability and, as a consequence, smaller bandwidths selectors are derived from the different methods. Surprisingly, it was also observed that the use of the bias-corrected variogram estimator lead to better results than those obtained with the true covariances.

- A summary of the squared errors, for $n = 20 \times 20$, $m = m_1$, $\sigma^2 = 1.0$, $r = 0.6$ and $c_0 = 0.2$, is shown in Table 2.

	Residuals		Corrected	
	$CGCV$	CCV	$CGCV$	CCV
Mean	.448	.448	.414	.410
Median	.195	.195	.179	.177
SD	.665	.664	.617	.610

Table 2. Summary of squared errors of trend estimates obtained with estimated covariance matrices, by using the uncorrected and corrected nonparametric semivariogram estimators.

Conclusions

- As it is confirmed by the simulation results, when data are spatially correlated, the use of the $CGCV$ and CCV bandwidth selection criteria would be recommended.

- The simulation results also lead to conclude that the bias, due to the direct use of residuals in variogram estimation, may have a significant impact on the bandwidth selection method. Therefore, the use of the bias-corrected nonparametric variogram estimator would be also recommended.

- The bandwidth selection methods were implemented in the statistical environment `R` and they will be available in newer versions of the `npsp` package.

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

[4] Fernández-Casal, R. and Francisco-Fernández, M. (2014). Nonparametric bias-corrected variogram estimation under non-constant trend. *Stochastic Environmental Research and Risk Assessment* **28**, 1247–1259.

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