proceedings

MDPI

Extended Abstract

Testing Goodness-of-Fit of Parametric Spatial Trends †

Andrea Meilán-Vila [1,*], Jean Opsomer [2], Mario Francisco-Fernández [1] and Rosa M. Crujeiras [3]

[1] Departamento de Matemáticas, Universidade da Coruña, 15071 A Coruña, Spain; mariofr@udc.es
[2] Westat Inc., Rockville, MD 20850, USA; JeanOpsomer@westat.com
[3] Departamento de Estadística, Análisis Matemático y Optimización, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain; rosa.crujeiras@usc.es
* Correspondence: andrea.meilan@udc.es
† Presented at the XoveTIC Congress, A Coruña, Spain, 27–28 September 2018.

Published: 17 September 2018

Abstract: The aim of this work is to propose and analyze the behavior of a test statistic to assess a parametric trend surface, that is, a regression model with spatially correlated errors. The asymptotic behavior under the null hypothesis, as well as the asymptotic power of the test under local alternatives will be analyzed. Finite sample performance of the test is addressed by simulation, introducing a bootstrap calibration procedure.

Keywords: model checking; spatial trend; local linear regression; least squares; bootstrap

1. Introduction

Consider a spatial stochastic process, which consists of a collection of random variables indexed on a certain domain of $\mathbb{R}^2$, with a well-defined joint distribution. In this framework, the observed data usually exhibit an important feature: close observations tend to be more similar than those which are far apart. Therefore, such observations cannot be treated as independent and the dependence structure should be taken into account in any descriptive or inferential procedure. In particular, from the perspective of spatial regression models (a trend surface plus an error term), the dependence structure should be considered and properly introduced into the model.

A common task in statistics is to determine whether a parametric model is an appropriate representation of a dataset. Under the assumption of independent errors, some authors have developed goodness-of-fit tests for parametric models that rely on a smooth alternative estimated by a nonparametric regression method, as [1] or [2].

A new proposal for testing a parametric trend surface is given in this paper. The proposed test is based on a comparison between a smooth version of a parametric fit with a nonparametric estimator of the trend (specifically, the multivariate local linear estimator will be used) in terms of a distance.

2. Statistical Model

Let $\{Z(\mathbf{s}), \mathbf{s} \in D\}$ be a random spatial process consisting of collections of random variables indexed in a domain $D \subset \mathbb{R}^2$ with a well-defined joint distribution. Consider n locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ on the region D generated from a density f. The set of random variables corresponding with those locations will be represented by $\{Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)\}$. Assume the model

$$Z(\mathbf{s}_i) = m(\mathbf{s}_i) + \varepsilon(\mathbf{s}_i), \quad i = 1, \ldots, n, \tag{1}$$

where m is an unknown smooth regression function which is supposed to be twice continuously differentiable. The ε are unobserved random variables with

$$\mathbb{E}[\varepsilon(\mathbf{s}_i)] = 0, \quad \mathrm{Cov}(\varepsilon(\mathbf{s}_i), \varepsilon(\mathbf{s}_j)) = \sigma^2 \rho_n(\mathbf{s}_i - \mathbf{s}_j), \quad i, j = 1, \ldots, n,$$

where $\sigma^2 < \infty$ and ρ_n is a continuous correlation function satisfying $\rho_n(0) = 1$, $\rho_n(\mathbf{s}) = \rho_n(-\mathbf{s})$ and $|\rho_n(\mathbf{s})| \leq 1$, $\forall \mathbf{s}$. The goal of this work is to test if the trend function belongs to a parametric family:

$$H_0 : m \in \mathcal{M}_{\boldsymbol{\beta}} = \{m_{\boldsymbol{\beta}}, \boldsymbol{\beta} \in \mathcal{B}\}, \qquad \text{vs.} \qquad H_a : m \notin \mathcal{M}_{\boldsymbol{\beta}}, \tag{2}$$

with $\mathcal{B} \subset \mathbb{R}^p$ a compact set. One of the more usual approaches is to compare a smooth version of a parametric fit with a nonparametric estimator of $m(\mathbf{s})$ and "thereafter" to reject H_0 if the distance between both fits exceeds a critical value.

3. Test Statistic

A suitable test statistic in order to solve the testing problem (2) could be computed as a weighted L_2—distance between the nonparametric and parametric fits, as in [2]:

$$T_n = n|\mathbf{H}|^{1/2} \int_D (\hat{m}^{LL}_{\mathbf{H}}(\mathbf{s}) - \hat{m}^{LL}_{\mathbf{H},\hat{\boldsymbol{\beta}}}(\mathbf{s}))^2 w(\mathbf{s}) d\mathbf{s}, \tag{3}$$

where w is a weight function. A full definition of the elements of the test statistic T_n can be found in Appendix A. For the calibration of the critical values, a bootstrap procedure is considered, see Appendix B.

4. Simulations

In this section, a simulation study showing the performance of the bootstrap procedure is presented. For this purpose, 500 samples of size $n = 400$ are generated from an isotropic spatial process observed at regularly spaced locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ in the unit square, where $\mathbf{s}_i = (s_{i1}, s_{i2})$, $i = 1, \ldots, n$:

$$Z(\mathbf{s}_i) = 2 + s_{i1} + s_{i2} + c s_{i1}^3 + \varepsilon(\mathbf{s}_i), \quad 1 \leq i \leq n. \tag{4}$$

The random errors $\varepsilon(\mathbf{s}_i)$ are normally distributed with zero mean and exponential covariance function $\mathrm{Cov}(\varepsilon(\mathbf{s}_i), \varepsilon(\mathbf{s}_j)) = \sigma^2\{\exp(-\|\mathbf{s}_i - \mathbf{s}_j\|/a_e)\}$, with $\sigma = 0.4$ and $\sigma = 0.8$. Different values of parameter a_e are considered: $a_e = 0.4, 0.6, 0.8$. The bootstrap procedure has been performed using $B = 500$ replicas for each sample. The weight function used was taken as $w(\mathbf{s}) = 1$. For simplicity, the bandwidth matrix was considered $\mathbf{H} = \mathrm{diag}(h, h)$, and different bandwidth values were chosen, $h = 0.10, 0.15, 0.20$.

In Table 1, the simulated rejection probabilities obtained for T_n are presented for the significance level $\alpha = 0.05$ over the 500 trials. When c is equal to zero (under the null hypothesis of linearity of the trend), the proportion of rejections obtained is similar to the considered significance level, but this proportion depends directly on the value of the bandwidth h. When c is equal to 5 or 10, the power of the test is really good, since the proportion of rejections is close to one, in the majority of the cases. Again, this proportion depends on the value of the bandwidth.

Table 1. Proportion of rejections of the null hypothesis.

			h		
σ	a_e	c	**0.10**	**0.15**	**0.20**
0.4	0.4	0	0.052	0.047	0.042
		5	0.897	0.932	0.911
		10	0.905	0.948	0.923
0.4	0.6	0	0.054	0.042	0.034
		5	0.856	0.901	0.898
		10	0.894	0.926	0.918
0.8	0.8	0	0.068	0.048	0.038
		5	0.808	0.798	0.806
		10	0.845	0.803	0.816

Funding: This research has received financial support from the Xunta de Galicia and the European Union (European Social Fund-ESF). This research has been partially supported by MINECO grant MTM2014-52876-R, MTM2016-76969-P and MTM2017-82724-R and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 and Centro Singular de Investigación de Galicia ED431G/01), all of them through the ERDF.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Appendix A

The trend surface estimation can be performed using a parametric and a non-parametric approach. In the parametric context, an iterative estimation procedure could be used. Denoting $\mathbf{Z} = (Z(\mathbf{s}_1), \cdots, Z(\mathbf{s}_n))'$ and $\mathbf{m}_{\boldsymbol{\beta}} = (m_{\boldsymbol{\beta}}(\mathbf{s}_1), \ldots, m_{\boldsymbol{\beta}}(\mathbf{s}_n))'$, under H_0 the steps of the procedure are:

(1) Based on the sample, estimate the trend parameter $\boldsymbol{\beta}$ using the ordinary least squares estimator, ignoring the dependence structure of the errors:

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\mathbf{Z} - \mathbf{m}_{\boldsymbol{\beta}})' (\mathbf{Z} - \mathbf{m}_{\boldsymbol{\beta}}).$$

(2) Estimate the variance-covariance matrix of the errors Σ using the residuals $\tilde{\varepsilon}(\mathbf{s}_i) = Z(\mathbf{s}_i) - m_{\tilde{\boldsymbol{\beta}}}(\mathbf{s}_i)$, $i = 1, \ldots, n$, obtained from the estimator of the trend from Step (1). Note that, the entries of Σ are:

$$\Sigma(i,j) = C_{\boldsymbol{\theta}}(\mathbf{s}_i - \mathbf{s}_j), \qquad i,j = 1 \ldots, n,$$

where $C_{\boldsymbol{\theta}}(\mathbf{s}_i - \mathbf{s}_j) = \sigma^2 - \gamma_{\boldsymbol{\theta}}(\mathbf{s}_i - \mathbf{s}_j)$, being $\{2\gamma_{\boldsymbol{\theta}}(\mathbf{u}) : \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^q\}$ a valid parametric family to estimate the variogram function.

(3) Estimate the trend parameter $\boldsymbol{\beta}$ using the weighted least squares estimator, taking the dependence structure of the errors into account:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\mathbf{Z} - \mathbf{m}_{\boldsymbol{\beta}})' \tilde{\Sigma}^{-1} (\mathbf{Z} - \mathbf{m}_{\boldsymbol{\beta}}).$$

Therefore, the parametric trend estimator considered would be $m_{\hat{\boldsymbol{\beta}}}$. Note that, an estimation of Σ can be obtained from the residuals $\tilde{\varepsilon}(\mathbf{s}_i), i = 1, \ldots, n$, as follows:

$$\tilde{\Sigma}(i,j) = C_{\tilde{\boldsymbol{\theta}}_{LS}}(\mathbf{s}_i - \mathbf{s}_j) = \tilde{\sigma}^2 - \gamma_{\tilde{\boldsymbol{\theta}}_{LS}}(\mathbf{s}_i - \mathbf{s}_j), \quad i,j = 1 \ldots, n,$$

where $\gamma_{\tilde{\boldsymbol{\theta}}_{LS}}$ is the parametric least squares estimator of the variogram and $\tilde{\sigma}^2$ is an estimator of the variance. The last estimator could be obtained using a least squares procedure.

From a nonparametric point of view, model (1) has been studied by several authors. Some approaches used for this task include kernel-based methods. In this case, the trend is estimated

using the multivariate local linear estimator, see [3]. In the spatial framework, the local linear estimator for $m(\mathbf{s})$ at a location $\mathbf{s}$ can be explicitly written as

$$\hat{m}_{\mathbf{H}}^{LL}(\mathbf{s}) = \mathbf{e}_1'(X_s'W_sX_s)^{-1}X_s'W_s\mathbf{Z},$$

where $\mathbf{e}_1 = (1,0,0)'$, X_s is a $n \times 3$ matrix whose i-th row equals $(1,(\mathbf{s}_i - \mathbf{s})')$, $i = 1,\ldots,n$, $W_s = \text{diag}\{K_{\mathbf{H}}(\mathbf{s}_1 - \mathbf{s}),\ldots,K_{\mathbf{H}}(\mathbf{s}_n - \mathbf{s})\}$, where $K_{\mathbf{H}}(\mathbf{s}) = |\mathbf{H}|^{-1}K(\mathbf{H}^{-1}\mathbf{s})$ is used to assign weights. $\mathbf{H}$ is a 2×2 symmetric, positive definite matrix depending on the sample size n and K is a multivariate kernel function. Given $\mathbf{s}$, the bandwidth $\mathbf{H}$ controls the shape and the size of the local neighborhood used to estimate m.

Therefore, taking into account these estimators, the proposed test statistic is

$$T_n = n|\mathbf{H}|^{1/2}\int_D (\hat{m}_{\mathbf{H}}^{LL}(\mathbf{s}) - \hat{m}_{\mathbf{H},\hat{\boldsymbol{\beta}}}^{LL}(\mathbf{s}))^2 w(\mathbf{s})d\mathbf{s},$$

where w is a weight function and $\hat{m}_{\mathbf{H},\hat{\boldsymbol{\beta}}}^{LL}$ is a smooth version of the parametric estimator $m_{\hat{\boldsymbol{\beta}}}$, which is defined by

$$\hat{m}_{\mathbf{H},\hat{\boldsymbol{\beta}}}^{LL}(\mathbf{s}) = \mathbf{e}_1'(X_s'W_sX_s)^{-1}X_s'W_s\mathbf{m}_{\hat{\boldsymbol{\beta}}},$$

where $\mathbf{m}_{\hat{\boldsymbol{\beta}}} = (m_{\hat{\boldsymbol{\beta}}}(\mathbf{s}_1),\ldots,m_{\hat{\boldsymbol{\beta}}}(\mathbf{s}_n))'$.

Appendix B

Once a suitable test statistic is available, a crucial task is the calibration of critical values for a given level α, namely t_α. Usually, the estimation of these critical values t_α such that $\mathbb{P}_{H_0}(T_n \geq t_\alpha) = \alpha$ can be done by means of the asymptotic distribution. The use of asymptotic theory to calibrate the test poses some problems, such as the need to estimate some nuisance functions and a slow convergence rate to the limit distribution. Under these circumstances, calibration can be done by means of resampling procedures, such as bootstrap, see [4].

The procedure consists in generating a bootstrap sample $\{Z^*(\mathbf{s}_i), i = 1,\ldots,n\}$ and then computing a bootstrap statistic T_n^* like T_n by the squared deviation between the smooth version of the parametric fit $\hat{m}_{\hat{\boldsymbol{\beta}}^*}^{LL}$ and the nonparametric fit $\hat{m}^{*LL}$. Once the bootstrap statistic is computed, the distribution of T_n^* can be approximated by Monte Carlo. From this Monte Carlo approximation, the $(1-\alpha)$ quantile t_α^* is defined and the parametric hypothesis es rejected if $T_n > t_\alpha^*$. The specific steps for the algorithm used in this work are the following:

1. Obtain the parametric trend estimator $\hat{\boldsymbol{\beta}}$.
2. Estimate the covariance matrix of the errors $\hat{\boldsymbol{\Sigma}}$ based on the residuals $\hat{\boldsymbol{\varepsilon}} = (\hat{\varepsilon}(\mathbf{s}_1),\ldots,\hat{\varepsilon}(\mathbf{s}_n))'$, where $\hat{\varepsilon}(\mathbf{s}_i) = Z(\mathbf{s}_i) - m_{\hat{\boldsymbol{\beta}}}(\mathbf{s}_i)$, $i = 1,\ldots,n$, and find the matrix L, such that $\hat{\boldsymbol{\Sigma}} = LL'$, using Cholesky decomposition.
3. Compute the independent residuals, $\mathbf{e} = (e(\mathbf{s}_1),\ldots,e(\mathbf{s}_n))'$, given by $e(\mathbf{s}_i) = L^{-1}\hat{\varepsilon}(\mathbf{s}_i)$.
4. These independent variables are centered and, from them, we obtain an independent bootstrap sample of size n, denoted by $\mathbf{e}^* = (e^*(\mathbf{s}_1),\ldots,e^*(\mathbf{s}_n))$.
5. Finally, the bootstrap errors $\boldsymbol{\varepsilon}^* = (\varepsilon^*(\mathbf{s}_1),\ldots,\varepsilon^*(\mathbf{s}_n))$ are $\varepsilon^*(\mathbf{s}_i) = Le^*(\mathbf{s}_i)$, and the bootstrap samples are $Z^*(\mathbf{s}_i) = m_{\hat{\boldsymbol{\beta}}}(\mathbf{s}_i) + \varepsilon^*(\mathbf{s}_i)$.

References

1. Hardle, W.; Mammen, E. Comparing nonparametric versus parametric regression fits. *Ann. Stat.* **1993**, *21*, 1926–1947.
2. Alcalá, J.; Cristóbal, J.; González-Manteiga, W. Goodness-of-fit test for linear models based on local polynomials. *Stat. Probab. Lett.* **1999**, *42*, 39–46.

3. Fan, J.; Gijbels, I. *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability*; CRC Press: Boca Raton, FL, USA, 1996; Volume 66.
4. Francisco-Fernández, M.; Jurado-Expósito, M.; Opsomer, J.; López-Granados, F. A nonparametric analysis of the spatial distribution of Convolvulus arvensis in wheat-sunflower rotations. *Environmetrics* **2006**, *17*, 849–860.