LITHUANIAN JOURNAL OF STATISTICS
Lietuvos statistikos darbai
Vol. 60, 2021, pages 1–7
https://doi.org/10.15388/LJS.2021.26442

Statistics Lithuania
Lithuanian Statistical Society
Vilnius University Press

On the classification error rates in terms of semivariograms for Gaussian universal kriging models

Kęstutis Dučinskas[a,b], **Lina Dreižienė**[a,c]

[a] *Faculty of Marine Technologies and Natural Sciences, Klaipeda University*
Herkaus Manto st. 84, LT-92294 Klaipeda

[b] *Institute of Data Science and Digital Technologies, Vilnius University*
Akademijos st. 4, LT-08412 Vilnius

[c] *Faculty of Business, Klaipeda State University of Applied Sciences*
Jaunystes st. 1, LT-91274 Klaipeda

E-mail: l.dreiziene@gmail.com

Received July 20, 2021; accepted November 6, 2021; published online December 30, 2021

Abstract. Bayes multiclass classification of spatial Gaussian data following the universal kriging model is considered. The closed-form expressions for the maximum likelihood (ML) estimator of regression parameters and the actual error rate (AER) in terms of semivariograms are derived.

Keywords: pairwise discriminant function; semivariogram; actual error rate

1 Introduction

In the geostatistics literature, second-order properties are typically characterised using semivariograms that are defined directly by increments, rather than covariance functions. It is known that for random fields (RFs) with finite variances, covariance matrices are in one-to-one correspondence to semivariogram matrices and variances (e.g. [4, Section 8.3]). That was the motivating argument for the consideration of spatial prediction and classification problems via semivariograms. Another motivating argument for using semivariograms instead of covariances for stationary RFs is the unbiasedness of the classical estimator of semivariograms introduced by Matheron (e.g. [5, Section 2.2.1]).

© 2021 Authors. Published by Vilnius University Press
This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Problems of spatial data classification with continuously varying spatial index and directly specified spatial correlation or semivariogram functions were investigated by numerous authors (see, e.g. [19, 17, 1, 7]). These spatial data models are traditionally called geostatistical models (see [8]). The error rates of classification of Gaussian random field (GRF) observation via plug-in Bayes discriminant functions were explored by Ducinskas [9] for two-class case and by Ducinskas *et al.* [12] for multiclass case. Spatial classification based on plug-in Bayes discriminant function for feature observations, having elliptically contoured distributions, was explored by Batsidis and Zografos [2] and Ducinskas and Zikariene [13]. A numerical comparison of the performances for different spatial classification rules was performed by Berrett and Calder [3]. However, in the papers mentioned above, primary attention was paid to the geostatistical models with directly specified covariance functions.

We focus on the universal kriging case when several populations are specified by different regression parameters of GRF with second-order properties expressed in terms of semivariograms and variances. Classification rule based on the plug-in Bayes discriminant function with inserted ML estimators of regression parameters is explored.

In the present study, the closed-form expressions for the ML estimators of regression parameters and the actual error rate in terms of semivariograms are derived. AER and its estimators are usually used for the classification rule performance evaluation (see, e.g. [16, 10]). This is a multiclass extension of two-class classification for analogous models investigated by Ducinskas and Dreiziene [11].

This paper is organised as follows: the problem description and definitions of Bayes discriminant function and Bayes error rate are presented in Section 2; ML estimators of regression parameters and AER are derived in Section 3; discussion is presented in the last section.

2 The main concepts and definitions

This study assesses the classification of spatial data considered as realisations of a univariate RF $\{Z(s)\colon s \in D \subset \mathbb{R}^2\}$. This category is traditionally called geostatistical processes (see [6, Chapter 4]). The goal is to classify RF observation into one of L predefined populations, denoted by $\Omega_1, \ldots, \Omega_L$. We consider the linear regression model of observation $Z(s)$ in population Ω_l of the following form:

$$Z(s) = x'(s)\beta_l + \varepsilon(s),$$

where $x'(s) = (x_1(s), \ldots, x_q(s))$ is a $q \times 1$ vector of non-random regressors and β_l is a $q \times 1$ vector of unknown parameters, $l = 1, \ldots, L$, and $\beta_l \neq \beta_k$ for $l \neq k$. The error term $\varepsilon(s)$ that comprises fine-scale measurements and other errors is assumed to be zero-mean GRF $\{\varepsilon(s) : s \in D\}$ with known stationary covariance function:

$$\sigma(s-t) = cov\big(\varepsilon(s), \varepsilon(t)\big) = \sigma^2 corr\big(\varepsilon(s), \varepsilon(t)\big)$$

or semivariogram:

$$\gamma(s-t) = Var\big(\varepsilon(s) - \varepsilon(t)\big)/2 \quad \text{for } s, t \in D.$$

We consider the universal kriging model for better interpretability and simplicity of obtained closed-form expressions (e.g. [8]), assuming known covariance functions and semivariograms. Relaxing this assumption has only the price of computation time and is relevant for analysing the real data.

Suppose the set of training locations $S_n = \{s_i \in D, i = 0, 1, \ldots, n\}$ is partitioned into the union of L disjoint subsets, i.e. $S_n = \bigcup_{l=1}^{L} S^{(k)}$, where $S^{(l)}$ contains n_l locations with population labels l, $\sum_{l=1}^{L} n_l = n$. For simplicity, arrange the set of training locations in the following way: $S^{(1)} = \{s_1, s_2, \ldots, s_{n_1}\}$, $S^{(2)} = \{s_{n_1+1}, \ldots, s_{n_1+n_2}\}, \ldots, S^{(L)} = \{s_{\sum_{l=1}^{L-1} n_l+1}, \ldots, s_n\}$. The location of the observation to be classified is indexed by $\{0\}$. Set $S_n^0 = S_n \cup \{0\}$.

In what follows, we use the notations $Z(s_i) = Z_i$, $x(s_i) = x_i$, $\varepsilon(s_i) = \varepsilon_i$ for $i = 0, 1, \ldots, n$, $\sigma_{ij} = cov(Z_i, Z_j)$, $r_{ij} = corr(Z_i, Z_j)$ for $i, j = 0, 1, \ldots, n$, $i \neq j$. Also, define n-dimensional vectors $Z = (Z_1, \ldots, Z_n)'$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$, $c_0 = (\sigma_{01}, \sigma_{02}, \ldots, \sigma_{0n})'$, $r_0 = (r_{01}, r_{02}, \ldots, r_{0n})'$, and $n \times n$ matrices $\Sigma = Var(Z) = (\sigma_{ij})$, $i, j = 1, \ldots, n$ and $R = (r_{ij})$, $i, j = 1, \ldots, n$. Analogously, introduce an n-dimensional vector and $n \times n$ matrix for semivariograms $\gamma_0 = (\gamma_{01}, \gamma_{02}, \ldots, \gamma_{0n})'$, $\Gamma = (\gamma_{ij})$, $i, j = 1, \ldots, n$, where $\gamma_{ij} = Var(\varepsilon_i - \varepsilon_j)/2$.

Observe that due to stationarity of error field $\Sigma = \sigma^2 R = \sigma^2 J - \Gamma$, with $J = 1_n 1_n'$ and 1_n denotes the n-dimensional vector of ones.

Let $\Sigma^{\cdot\cdot} = 1_n' \Sigma^{-1} 1_n$ and $\Gamma^{\cdot\cdot} = 1_n' \Gamma^{-1} 1_n$.

Given invertibility condition of Σ by Woodbury identity (see [15]), Pistone and Vicario [18] derived the following formulas:

$$\begin{gathered} \Gamma^{-1} = -\Sigma^{-1} + \Sigma^{-1} J \Sigma^{-1} \Big/ \left(\Gamma^{\cdot\cdot} - \frac{1}{\sigma^2} \right) \quad \text{and} \\ \Sigma^{-1} = \Gamma^{-1} + \Gamma^{-1} J \Gamma^{-1} \Big/ \left(\Sigma^{\cdot\cdot} - \frac{1}{\sigma^2} \right). \end{gathered} \tag{1}$$

These will be intensively exploited in the present study.

Put $\beta' = (\beta_1', \ldots, \beta_L')$ and denote by X_l the $n_l \times q$ matrix of regressors for observations from Ω_l, $l = 1, \ldots, L$. Then $n \times Lq$ design matrix of training sample Z is specified by $X = \bigoplus_{l=1}^{L} X_l$. Thus, the training sample Z has a multivariate Gaussian distribution $Z \sim N_n(X\beta, \Sigma)$ with $\Sigma = \sigma^2 R$.

The main objective of this paper is to classify a single observation of GRF $\{Z(s) : s \in D \subset \mathbb{R}^2\}$ at focal location s_0, given the stratified training sample Z.

Then the conditional distribution of Z_0, given $Z = z$, in Ω_l, $l = 1, \ldots, L$ is Gaussian with mean:

$$\begin{gathered} \mu_{lz}^0 = E(Z_0 | Z = z; \Omega_l) = x_0' \beta_l + \alpha_0'(z - X\beta) \text{ and variance} \\ \sigma_{0z}^2 = \sigma_{00} - \alpha_0' c_0 = \sigma^2 \rho_0, \end{gathered} \tag{2}$$

where $\alpha_0 = \Sigma^{-1} c_0$ and $\rho_0 = 1 - r_0' R^{-1} r_0$.

By using some matrix algebra and (1) and (2), we obtain the following formula in terms of semivariograms:

$$\rho_0 = \left(1 - \gamma_0' \Gamma^{-1} 1_n\right)^2 \big/ \left(1 - \sigma^2 \Gamma^{\cdot\cdot} + \gamma_0' \Gamma_0^{-1} \gamma_0 / \sigma^2\right). \tag{3}$$

Denote by $\pi_1^0, \ldots, \pi_L^0$ ($\sum_{i=1}^{L} \pi_i^0 = 1$) the prior probabilities of the populations $\Omega_1, \ldots, \Omega_L$, respectively, for observation at the focal location s_0, given training sample Z.

Denote the log ratio of conditional densities in populations Ω_k and Ω_l by

$$W_{kl}(Z_0, \beta) = \left(Z_0 - \frac{1}{2}\left(\mu_{kz}^0 + \mu_{lz}^0\right)\right)\left(\mu_{kz}^0 - \mu_{lz}^0\right)/\sigma_{0z}^2 + \vartheta_{kl}^0,$$

where $\vartheta_{kl}^0 = \ln(\pi_k^0/\pi_l^0)$, which will be called pairwise discriminant functions.

Then Bayes rule classifies observation Z_0, given $Z = z$, to the population Ω_k if $W_{kl}(Z_0, \beta) \geqslant 0$ for $l = 1, \dots, L$ and $l \neq k$.

The squared pairwise conditional Mahalanobis distance at location s_0, given $Z = z$, has the form

$$\Delta_{0kl}^2 = (\mu_{kz}^0 - \mu_{lz}^0)^2/\sigma_{0z}^2 = (x_0' \Delta\beta_{kl})^2/\sigma^2\rho_0,$$

where $\Delta\beta_{kl} = \beta_k - \beta_l$.

Then the misclassification probability, or Bayes error rate, for the Bayes rule specified above has the following form (see [12])

$$P(\beta) = 1 - \sum_{k=1}^{L} \pi_k^0 \int \varphi(u) H\left(u + \frac{\Delta_{0kl}^2}{2} + \vartheta_{kl}^0\right) du,$$

where $\varphi(\cdot)$ is the probability density distribution of standard Gaussian distribution and $H(\cdot)$ is the Heaviside function.

3 Estimators of parameters and actual error rates

Recall that variation in spatial models is usually represented directly by covariance functions or by semivariograms and this paper focuses on the latter.

With an insignificant loss of generality, we state the following assumption.

Intercept assumption. $x_1(s) = 1$ for $s \in D$.

This assumption implies that $Xd = 1_n$, where $d = 1_L \bigotimes e_1$ with e_1 denoting the first column of the identity matrix I_q.

It is known that ML estimator of regression parameters from training sample Z in terms of spatial covariances (correlations) hasve the explicit form

$$\hat{\beta} = \left(X' \Sigma^{-1} X\right)^{-1} X' \Sigma^{-1} Z = \left(X' R^{-1} X\right)^{-1} X' R^{-1} Z.$$

In the following, we derive the closed-form expression of the ML estimator for regression parameters through the semivariograms.

Lemma 1. *Under the Intercept assumption, the ML estimator of regression parameters from the training sample Z is*

$$\hat{\beta} = \left(X' \Gamma^{-1} X\right)^{-1} X' \Gamma^{-1} Z. \tag{4}$$

Proof. Under Intercept assumption we obtain

$$X' \Sigma^{-1} X = -X' \Gamma^{-1} X - \sigma^2 X' \Gamma^{-1} J \Gamma^{-1} X/a,$$

where $a = (1 - \sigma^2 \Gamma^{\cdot\cdot})$.

Observe that

$$\left(X'\Gamma^{-1}X\right)^{-1}X'\Gamma^{-1}1_n = d \quad \text{and} \quad X'\Sigma^{-1} = -X'\Gamma^{-1}\left(I + \sigma^2 J\Gamma^{-1}/a\right). \tag{5}$$

Based on Woodbury's identity (see [14]), we get

$$\left(X'\Sigma^{-1}X\right)^{-1} = -\left(X'\Gamma^{-1}X\right)^{-1} + \sigma^2 dd'. \tag{6}$$

Consequently, from (5) and (6) follows:

$$\left(X'\Sigma^{-1}X\right)^{-1}X'\Sigma^{-1} = -\left(-\left(X'\Gamma^{-1}X\right)^{-1} + \sigma^2 dd'\right)X'\Gamma^{-1}\left(I + \frac{\sigma^2 J\Gamma^{-1}}{a}\right).$$

After simple matrix algebra, we complete the proof of lemma. □

Define the estimator of the conditional mean by:

$$\hat{\mu}_{lz}^0 = E(Z_0|Z=z;\Omega_l) = x_0'\hat{\beta}_l + \alpha_0'(z - X\hat{\beta}).$$

By using the result of Lemma 1, we observe that $\alpha_0'(z - X\hat{\beta}) = \gamma_0'\Gamma^{-1}(z - X\hat{\beta})$, then

$$\hat{\mu}_{lz}^0 = E(Z_0|Z=z;\Omega_l) = x_0'\hat{\beta}_l + \gamma_0'\Gamma^{-1}(z - X\hat{\beta}).$$

Then plug-in pairwise discriminant function has the form

$$W_{kl}(Z_0,\hat{\beta}) = \left(Z_0 - \frac{1}{2}\left(\hat{\mu}_{kz}^0 + \hat{\mu}_{lz}^0\right)\right)\left(\hat{\mu}_{kz}^0 - \hat{\mu}_{lz}^0\right)/\sigma_{0z}^2 + \vartheta_{kl}^0,$$

and plug-in Bayes rule classifies observation Z_0, given $Z = z$, to the population Ω_k if $W_{kl}(Z_0,\hat{\beta}) \geqslant 0$ for $l = 1,\dots,L$ and $l \neq k$.

Denote the estimated squared pairwise conditional Mahalanobis distance at location s_0, given $Z = z$, by

$$\hat{\Delta}_{0kl}^2 = \left(x_0'\Delta\hat{\beta}_{kl}\right)^2/\sigma_{0z}^2,$$

where $\Delta\hat{\beta}_{kl} = \hat{\beta}_k - \hat{\beta}_l$.

Lemma 2. *Under the Intercept assumption and assertion of Lemma* 1*, the actual error rate of plug-in Bayes rule has the following form*:

$$P(\beta) = 1 - \sum_{k=1}^{L}\pi_k^0\int\varphi(u)H\left(u + \frac{\hat{\Delta}_{0kl}^2}{2} + \vartheta_{kl}^0\right)du.$$

Proof. The conditional distribution of $W_{kl}(Z_0,\hat{\beta})$, given $Z = z$, in Ω_k is normal with mean $E(W_{kl}(Z_0,\hat{\beta})) = \hat{\Delta}_{0kl}^2/2 + \vartheta_{kl}^0$ and variance $Var(W_{kl}(Z_0,\hat{\beta})) = \hat{\Delta}_{0kl}^2 = (x_0'F^-\hat{\beta})^2/(\sigma^2\rho_0)$. $\hat{\beta}$ and ρ_0 included in the above formulas for error rates have their explicit expressions in terms of semivariograms presented in formulas (3) and (4), respectively. Then by using the properties of normal distribution and definition, we complete the proof of Lemma 2. □

4 Discussion

This paper contributes to the analysis of the performance of plug-in Bayes rule in the classification of spatial Gaussian data, focusing on error rates. For the Gaussian universal kriging model, the closed-form expression for the ML estimators of regression parameters in the case of stratified training sample as well as the formula of AER are derived via semivariograms with broader applicability than covariances.

We expect these findings would allow to reach broader applicability of semivariograms in designing the classifiers of spatial data based on pairwise Bayes discriminant functions.

Replacing a semivariogram with its estimator hasve only the price of computation time and is relevant for the analysis of real-world data. Investigations in that direction are in our plans.

References

[1] P. Atkinson, P. Lewis. Geostatistical classification for remote sensing: an introduction. *Comput. Geosci.*, **10**:361–371, 2000.

[2] A. Batsidis, K. Zografos. Errors of misclassification in discrimination of dimensional coherent elliptic random field observations. *Stat. Neerl.*, **65**(4):446–461, 2011.

[3] C. Berrett, C.A. Calder. Bayesian spatial binary classification. *Spat. Stat.*, **16**:72–102, 2016.

[4] R. Christensen. *Advanced Linear Modeling: Statistical Learning and Dependent Data.* Springer Texts in Statistics. Springer, 2019.

[5] N. Cressie. *Statistics for Spatial Data.* John Wiley & Sons, N.Y., 1993.

[6] N. Cressie, C.K. Wikle. *Statistics for Spatio-Temporal Data.* Wiley, Hoboken, NJ, 2011.

[7] V. De Oliveira. Bayesian prediction of clipped Gaussian random fields. *Comput. Stat. Data Anal.*, **34**:299–314, 2000.

[8] P.J. Diggle, P.J. Ribeiro, O.F. Christensen. An introduction to model-based geostatistics. In Möller J.(Ed.), *Spatial Statistics and Computational Methods*, volume 173 of *Lecture Notes in Statist.*, pp. 43–86, 2003.

[9] K. Ducinskas. Statistical classification of the observation of nugget-less spatial Gaussian process with unknown sill parameter. *Nonlinear Anal. Model. Control*, **14**(2):155–163, 2009.

[10] K. Ducinskas, L. Dreiziene. Risks of classification of the Gaussian Markov random field observations. *J. Classif.*, **35**:422–436, 2018.

[11] K. Ducinskas, L. Dreiziene. Actual error rates in linear discrimination of spatial Gaussian data in terms of semivariograms. *Commun. Stat. - Theory Methods*, **98**, 2021 (published online). https://doi.org/10.1080/03610926.2021.1968903.

[12] K. Ducinskas, L. Dreiziene, E. Zikariene. Multiclass classification of the scalar Gaussian random field observation with known spatial correlation function. *Stat. Probab. Lett.*, **98**:107–114, 2015.

[13] K. Ducinskas, E. Zikariene. Actual error rates in classification of the t-distributed random field observation based on plug-in linear discriminant function. *Informatica*, **26**(4):557–568, 2015.

[14] M.J. Heaton, A. Datta, A.O. Finley, *et al.* A case study competition among methods for analysing large spatial data. *J. Agric. Biol. Environ. Stat.*, **23**(3):398–425, 2019.

[15] H.V. Henderson, S.R. Searle. On deriving the inverse of a sum of matrices. *SIAM Rev.*, **23**(1):53–60, 1981.

[16] P.A. Lachenbruch, M.R. Mickey. Estimation of error rates in discriminant analysis. *Technometrics*, **10**:1–11, 1918.

[17] K.V. Mardia. Spatial discrimination and classification maps. *Commun. Stat. - Theory Methods*, **13**(18):2181–2197, 1984.

[18] G. Pistone, G. Vicario. A note on semivariogram. In T. Di Battista, E. Moreno, W. Racugno(Eds.), *Topics on Methodological and Applied Statistical Inference*, pp. 181–190, 2016.

[19] P. Switzer. Extensions of linear discriminant analysis for statistical classification of remotely sensed satellite imagery. *Math. Geol.*, **12**(4):367–376, 1980.

REZIUMĖ

Klasifikavimo klaidos vertinimas, pagrįstas semivariogramomis Gauso universalaus krigingo modeliui

K. Dučinskas, L. Dreižienė

Straipsnyje analizuojama erdvinių Gauso duomenų Bajeso klasifikavimo procedūra universalaus krigingo modeliui. Išvestos išreikštinės maksimalaus tikėtinumo regresijos parametrų įverčių išraiškos bei aktualioji (įvertinta) klasifikavimo paklaida, pagrįstos semivariogramomis.

Raktiniai žodžiai: porinė Bajeso diskriminantinė funkcija; semivariograma; aktualioji klasifikavimo klaida