

VILNIUS UNIVERSITY

Lina  
DREIŽIENĖ

# Classification risk of Gaussian spatial data using linear discriminant functions

**DOCTORAL DISSERTATION**

Physical sciences,  
mathematics 01P

---

VILNIUS 2019

This dissertation was written between 2014 and 2018 at Vilnius University.

**Academic supervisors:**

Prof. Dr. Marijus Radavičius (Vilnius University, Physical sciences, Mathematics, 01P). From 01/10/2014 to 27/09/2017;

Prof. Dr. Kęstutis Dučinskas (Vilnius University, Physical sciences, Mathematics, 01P). From 28/09/2017 to 30/09/2018.

**Academic consultant:** Prof. Dr. Marijus Radavičius (Vilnius University, Physical sciences, Mathematics, 01P). From 28/09/2017 to 30/09/2018.

VILNIAUS UNIVERSITETAS

Lina  
DREIŽIENĖ

Erdvinių Gauso duomenų klasifikavimo  
rizika naudojant tiesines  
diskriminantines funkcijas

**DAKTARO DISERTACIJA**

Fiziniai mokslai,  
matematika 01P

---

VILNIUS 2019

Disertacija rengta 2014–2018 metais Vilniaus universitete

Moksliniai vadovai:

prof. dr. Marijus Radavičius (Vilniaus universitetas, fiziniai mokslai, matematika, 01P). Nuo 2014-10-01 iki 2017-09-27;

prof. dr. Kęstutis Dučinskas (Vilniaus universitetas, fiziniai mokslai, matematika, 01P). Nuo 2017-09-28 iki 2018-09-30.

Mokslinis konsultantas: prof. dr. Marijus Radavičius (Vilniaus universitetas, fiziniai mokslai, matematika, 01P). Nuo 2017-09-28 iki 2018-09-30.

## Acknowledgements

First and foremost, I would like to express my sincere gratitude to my Scientific supervisor Professor Kęstutis Dučinskas for the continuous support, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis.

Besides my advisor, I would like to thank my Scientific advisor Professor Marijus Radavičius for his insightful comments and suggestions.

I also thank the language editors Zita Šakienė and Salomėja Šatienė for their contribution editing the thesis.

I would also like to thank all the members of staff at Vilnius University who helped me to avoid long journeys from Klaipėda to Vilnius.

Last but not least, I would like to thank my family, especially my daughter Gustė who kindly accepted to spend less time together for the sake of my extra work on the thesis.

# Contents

<b>INTRODUCTION .....</b>	<b>9</b>
PROBLEM FORMULATION AND ITS TOPICALITY .....	9
AIM AND OBJECTIVES .....	12
METHODS .....	13
SCIENTIFIC NOVELTY .....	13
STRUCTURE OF THE DISSERTATION .....	14
DISSEMINATION OF THE RESULTS.....	14
<b>GAUSSIAN MODELS AND THEIR CHARACTERISTICS .....</b>	<b>18</b>
1.1. MODELING SPATIAL DATA.....	18
1.2. ESTIMATORS FOR SPATIAL MODEL PARAMETERS.....	24
1.2.1. <i>ML estimators for geometrically anisotropic covariance</i> .....	28
1.3. NON-PARAMETRIC TEST FOR SPATIAL GEOMETRIC ANISOTROPY .....	30
<b>CLASSIFICATION OF SPATIAL GRF OBSERVATIONS .....</b>	<b>33</b>
2.1. ELEMENTS OF LINEAR DISCRIMINANT ANALYSIS .....	33
2.2. CLASSIFICATION OF GGRF OBSERVATION .....	39
2.2.1. <i>Univariate case</i> .....	40
2.2.2. <i>Multivariate case</i> .....	58
2.3. CLASSIFICATION OF GMRF OBSERVATION .....	64
<b>NUMERICAL EXPERIMENTS AND APPLICATIONS .....</b>	<b>73</b>
3.1. THE EFFICACY OF NON-PARAMETRIC TEST .....	73
3.2. THE ANALYSIS OF AER ACCURACY AND INFLUENCE OF STATISTICAL PARAMETERS TO AER .....	76
3.3. APPLICATION OF PBDF TO THE MAPPING OF PRESENCE AND ABSENCE OF ZEBRA MUSSELS IN THE CURONIAN LAGOON.....	85
<b>CONCLUSIONS.....</b>	<b>93</b>
<b>BIBLIOGRAPHY .....</b>	<b>95</b>

## Notations

$\mathbf{A}'$	transpose of matrix $\mathbf{A}$
$\mathbf{I}_n$	n-dimensional identity matrix
$\mathbf{A} \circ \mathbf{B}$	Hadamard product of matrices $\mathbf{A}$ and $\mathbf{B}$
$\mathbf{A} \oplus \mathbf{B}$	direct sum of matrices $\mathbf{A}$ and $\mathbf{B}$
$tr(\mathbf{A})$	trace of quadratic matrix $\mathbf{A}$
$rank(\mathbf{A})$	rank of matrix $\mathbf{A}$
$\mathbf{D}_p$	duplication matrix
$\mathbf{R}$	matrix of spatial correlations
$\mathbf{\Sigma}$	covariance matrix
$\mathbf{X}$	design matrix of training sample
$\mathbf{J}$	Fisher information matrix
$vec(\mathbf{A})$	vectorization of matrix $\mathbf{A}$
$vech(\mathbf{A})$	half-vectorization of symmetric matrix $\mathbf{A}$
$\Phi(\cdot)$	standard normal distribution function
$\varphi(\cdot)$	standard normal distribution density function
$e(\cdot)$	Heaviside step function
$\delta(\cdot)$	Dirac delta function
$sgn(\cdot)$	Signum function
$N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	m-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$AN_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	asymptotically normal distribution
$N_{m \times n}(\cdot, \cdot)$	matrix-variate normal distribution

Matrices and vectors are denoted with bold letters.

## Abbreviations

**AER** – approximation of expected risk

**AEER** – approximation of expected error rate

**BDF** – Bayes discriminant function

**ER** – expected risk

**EER** – expected error rate

**GGRF** – geostatistical Gaussian random field

**GMRF** – Gaussian Markov random field

**ML** – maximum likelihood method

**PBDF** – plug-in Bayes discriminant function

**STL** – set of training locations



# Introduction

## Problem formulation and its topicality

Spatial data contain information about both the attribute of interest as well as its location. The location may be a set of coordinates, such as the latitude and longitude or it may be a small region such as a county associated with an observable feature. The observations made at different locations may be closer in value than measurements made at the locations farther apart, for example, the elevation datasets have similar elevation values close to each other; the majority of minerals have location-dependent distributions; house prices and house value assessments are established by comparisons between a house and some similar nearby houses; water (wind) pollutants generate negative consequences for those downstream (downwind) of their locations, etc. This phenomenon is called *spatial correlation (autocorrelation)*. The presence of spatial correlation means that a certain amount of information is shared and duplicated among neighboring locations. This feature violates the assumption of independent observations which is the background for many classical statistical methods. Therefore, when modelling spatial data it is important to pay sufficient attention to the modelling of spatial correlation as ignoring it may affect the accuracy of the prediction (Maity and Sherman [45]) and classification procedures.

The thesis focuses on discriminant analysis (sometimes called supervised statistical classification) for spatially correlated data. Separate sampling design (see Dučinskas [16]) for training data is used throughout the thesis. With this design, training sample is formed by the observations of feature vectors taken separately from each population. Also it is assumed that sampling design for training locations (or a set of training locations) is deterministic, and all analyses are carried out conditionally on it. It should be noted, that the particular case of spatial linear classifier with the mixture sampling design for training data is considered by Stabingienė [65]. The main purpose of the thesis is to use the Bayesian classification rule, taking

into account the spatial correlation between the Gaussian observations and using the population densities conditional on the given training data; to apply the proposed classification procedure to assign Gaussian random field observation to one of several prescribed populations and to assess the classification risk or error rates. The risk of classification is an indicator of the effectiveness of a discriminant function and could be affected by rejecting the hypothesis about existence of spatial correlation. The risk of classification could also be affected by anisotropy, the situation when the spatial correlation is stronger in one direction than in another. In the situation of geometric anisotropy two additional parameters, anisotropy ratio and anisotropy angle should be included.

Spatial statistics is a relatively young science, since it emerged in early 1980s as a hybrid discipline of mining engineering, geology, mathematics, and statistics (Cressie [10]). Therefore there are not many studies in the field of discriminant analysis of spatially correlated data. Many authors (i.e. Lawoko and McLachlan [39], Kharin [36], McLachlan [50]) have investigated the problems concerned with classification of dependent observations (equicorrelated structures, autoregressive models) but Switzer [66] was the first to use classification of spatial data. Mardia [47] extended this research by including spatial discrimination methods in forming the classification maps. Adaptive Bayesian classification for spatial data was proposed by Klein and Press [37]. The application of spatial contextual (or supervised) classification methods in the geospatial data mining is considered by Shekhar et al. [61]. Okamoto [52], McLachlan [50] have investigated the performance of the plug-in version of Bayes classification rule with independent observations, or training samples where observations are temporally dependent. Batsidis and Zografos [5] proposed the plug-in approach to discrimination for independent feature observations having elliptically contoured distributions. However, all these authors did not properly analyze classification risk and its estimators. The comprehensive analysis of classification risk associated with discriminant analysis of uncorrelated observations is presented by Dučinskas [16]. Later Šaltytė [57], Šaltytė and Dučinskas [59] proposed the approximation of expected risk (in particular expected error rate) for classification of scalar Gaussian random

field observation. The extension of these results to the multivariate spatiotemporal model is realized by Šaltytė-Benth and Dučinskas [60]. Note that here and often after geostatistical models, i.e. the models with continuous support for spatial index (e.g. Cressie [10], Diggle et al. [14]), are considered.

A comprehensive review and empirical comparison of the performances for different spatial classification rules could be found in the publications of Atkinson and Lewis [4] and Berret and Calder [7]. But all above mentioned papers hold the assumption of independence between observation to be classified and training sample, i.e. used only marginal population densities in Bayes rule. In practice, such an assumption is often not reasonable, especially if the observation of interest is close to the observations from the training sample. This restrictive assumption was eliminated by Dučinskas [17], [18], but here only the mean parameters and scale parameters of covariance function are assumed unknown for the geostatistical Gaussian random field models. Batsidis and Zografos [6] applied this approach in the analysis of errors of misclassification in discrimination of dimensional coherent elliptic random field observations. Recently, Younso [75] established the consistency of a new kernel rule for spatially dependent data.

In this thesis the extension to the complete parametric uncertainty case in classification of univariate and multivariate geostatistical Gaussian random field observation is analyzed. Also the extension to the multiclass case is performed. The multiclass problem has very rarely been addressed because most of the methods proposed for two populations do not generalize. Schervish [63] analyzed the problem of classification into one of the three known normal populations by the single linear discriminant function. Techniques for multiclass probability estimation by combining all pairwise comparisons are investigated by Wu et al. [72]. Empirical comparison of different methods of error rate estimation in multiclass linear discriminant analysis for multivariate homoscedastic Gaussian data was performed by Hirst [31]. Bayesian multiclass classification problem for correlated Gaussian observation was empirically studied by Williams and Barber [71]. The novel model-free estimation method for multiclass conditional probability based on conditional quintile regression functions was

theoretically and numerically studied by Xu and Wang [73]. Correct classification rates in multi-category classification of independent multivariate Gaussian observations were provided by Schervish [64]. However, there is a lack of theoretical results for the multiclass classification problem for the dependent observation case. In this thesis the classification procedure of univariate spatially correlated Gaussian observation is proposed and the misclassification probabilities are analyzed.

Finally, the extension to the problem of classification of Gaussian Markov random field observation is made. Gaussian Markov random fields (GMRF) are powerful and important tools for modelling spatial data. They have been widely used in different areas of spatial statistics including disease mapping, spatiotemporal modelling and image analysis (see Rue and Held [56]). The observations of GMRF are usually specified by conditionally autoregressive (CAR) model first introduced in the seminal paper by Besag [8]. The practical use of Gaussian Markov random fields (GMRF) for modelling large scale spatial phenomena has significantly increased after recent advances on the efficient simulation (see Rue and Held [56]). In the thesis the classification problem for a multivariate GMRF observation is considered.

## Aim and objectives

The main aim of the thesis is to perform linear discriminant analysis for spatial Gaussian data via plug-in Bayes discriminant function using different types of covariance and to analyze the risk of classification or error rates associated with the proposed classifier.

To accomplish the aim of the thesis, the following objectives are formulated:

- To derive formulas for classification risk and analytic expressions for its estimators for geostatistical Gaussian random field and to investigate the properties of the derived formulas.
- To derive the asymptotic approximation formulas of the expected classification risk for univariate and multivariate geostatistical Gaussian random field: the case of two classes.

- To derive the asymptotic approximation formula for the expected error rate of geostatistical Gaussian random field for a multiclass case.
- To derive the actual classification risk and the asymptotic approximation formula for univariate and multivariate Gaussian Markov random field observation.
- To implement the proposed classification methods, to analyze the influence of different parameters on the classification risk by using simulated and real data.

## Methods

Discriminant analysis of spatial data is the basis of applied research methods. Many proofs in this thesis use the properties of Gaussian distribution. Taylor series expansion is applied to get the asymptotic approximation formulas. The elements of matrix calculus, Dirac delta function and its derivatives are adapted as well. The unknown population parameters are estimated using maximum likelihood method. Numerical experiments are carried out employing statistical computing software R and its packages: geoR, gstat, INLA.

## Scientific novelty

The novelty of the results of the thesis:

- Closed-form expression of asymptotic covariance matrix for the geometrically anisotropic exponential covariance model is derived.
- Novel non-parametric test for spatial geometric anisotropy is proposed.
- Actual classification risk for complete parametric uncertainty case for the univariate and multivariate two-class classification problem is derived.

- Approximation of expected risk for complete parametric uncertainty case for the univariate and multivariate two-class classification problem is derived.
- Closed-form expression of approximation of expected risk for the geometrically anisotropic exponential covariance model is presented.
- Extension to classification problem of GMRF specified by multivariate CAR model for two-class case is explored.
- Multiclass classification problem of univariate geostatistical Gaussian random field observation is investigated.

## Structure of the dissertation

This doctoral thesis consists of the introduction, three chapters, conclusions and bibliography. The first chapter is designated for Gaussian models and their characteristics. It includes the issues of modelling spatial data, discusses the estimators for spatial models and presents a non-parametric test to detect geometric anisotropy. Chapter 2 presents the main results of the thesis concerned with discriminant analysis of spatial data. The last chapter introduces the numerical experiments and applications.

## Dissemination of the results

The results of the thesis have been presented in 19 publications and both national and international conferences.

### **Publications:**

- [A1] Dučinskas K., Dreičienė L. (2018). Risks of classification of the Gaussian Markov random field observations. *Journal of Classification*, 35:422–436.

- [A2] Dreišienė L., Dučinskas K., Šaltytė-Vaisiauskė, L. (2018). Statistical classification of multivariate conditionally autoregressive Gaussian random field observations. *Spatial Statistics*, 28:216–225.
- [A3] Dučinskas K., Dreišienė L. (2016). Expected error rates in classification of Gaussian CAR observations. *Computer data analysis and modelling: theoretical and applied stochastics: proceedings of the 11th international conference*. Publishing center of BSU, Minsk, 127–130.
- [A4] Dreišienė L., Dučinskas K. (2015). Error rates in multi-category classification of the spatial multivariate Gaussian data. *Procedia Environmental Sciences. Spatial Statistics conference 2015: Emerging Patterns*. Elsevier, Science Direct. 26:78–81.
- [A5] Dreišienė L., Dučinskas K., Paulionienė L. (2015). Correct classification rates in multi-category discriminant analysis of spatial Gaussian data. *Open Journal of Statistics*. 5(1):21–26.
- [A6] Dučinskas K., Dreišienė L., Zikarienė E. (2015). Multiclass classification of the scalar Gaussian random field observation with known spatial correlation function. *Statistics and Probability Letters*. 98:107–114.
- [A7] Dučinskas K., Zikarienė E., Dreišienė L. (2014). Comparison of performances of plug-in spatial classification rules based on Bayesian and ML estimators. In proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods. SciTePress, 161–166.
- [A8] Dučinskas K., Dreišienė L. (2013). Optimal classification of multivariate GRF observations. *Multivariate Statistics: Theory and Applications*, World Scientific, 61–72. (the chapter of book).
- [A9] Dreišienė L., Karaliutė M. (2012). The influence of training sample size on the expected error rate in spatial classification. *Proc. of the Lithuanian Mathematical Society, Ser. A*. 53:24–29.

- [A10] Dreičienė L. (2011). Linear discriminant analysis of spatial Gaussian data with estimated anisotropy ratio. *Lietuvos matematikos rinkinys. LMD darbai*. 52:315–320.
- [A11] Dučinskas K., Dreičienė L. (2011). Application of Bayes discriminant functions to classification of the spatial multivariate Gaussian data. *Procedia Environmental Sciences. Spatial Statistics 2011-Mapping Global Change*. Elsevier; ScienceDirect. 7:212–217.
- [A12] Dučinskas K., Dreičienė L. (2011). Supervised classification of the scalar Gaussian random field observations under a deterministic spatial sampling design. *Austrian Journal of Statistics*. Vienna, Austria: Austrian Statistical Society; Graz University of Technology. 40(1,2):25–36.
- [A13] Dučinskas K., Dreičienė L. (2010). Supervised classification of the scalar Gaussian random field observation. *Computer data analysis and modeling: complex stochastic data and systems. Proceedings of the 9th International Conference*. 2010 September 7-11, Minsk, Publishing center of BSU. 1:33–36.
- [A14] Dučinskas K., Dreičienė L. (2010). Nonparametric test for spatial geometric anisotropy. *Lietuvos matematikos rinkinys. LMD darbai*. Vilnius: MII. 51:397–401.
- [A15] Dreičienė L., Dučinskas K. (2009). The influence of the anisotropy ratio on the expected error rates in classification of stationary GRF observations. *Applied Stochastic Models and Data Analysis*, Vilnius: Technika. 101–105.
- [A16] Dučinskas K., Dreičienė L. (2007). Effect of anisotropy coefficient on error rates of linear discriminant functions. *Lietuvos matematikos rinkinys*. Vilnius: MII. 47(spec. nr.):359–363.
- [A17] Budrikaitė L. (2005). Modeling of zonal anisotropic semivariograms. *Lietuvos Matematikos rinkinys*. Vilnius: MII. 45(spec.nr.):339–342.



- [A18] Budrikaitė L., Dučinskas K. (2005). Modelling of geometric anisotropic spatial variation. *Mathematical Modelling and Analysis*. Vilnius: Technika. 361–366.
- [A19] Budrikaitė L., Dučinskas K. (2004). Forms of anisotropy for spatial variograms. *Lietuvos matematikos rinkinys*, Vilnius: MII. 44(spec.nr.):542–546.

### **International conferences:**

1. NORDSTAT 2018. Tartu, Estonia, June 26-29, 2018. Oral presentation: *Linear discriminant analysis of spatial Gaussian data*.
2. Spatial Statistics 2017: One World, One Health. Lancaster, UK, July 4-7, 2017. Oral presentation: *Statistical classification of multivariate conditionally autoregressive Gaussian random field observations*.
3. NORDSTAT 2016. Copenhagen, Denmark, June 27-30, 2016. Poster presentation: *Application of Bayes discriminant function to classification of Gaussian Markov random field observation*.
4. Spatial statistics 2015: Emerging Patterns. Avignon, France, June 9-12, 2015. Poster presentation: *Error rates in multi-category classification of the spatial multivariate Gaussian data*.
5. The 9-th Tartu Conference on Multivariate Statistics & the 20-th International Workshop on Matrices and Statistics. Tartu, Estonia, June 26-July 1, 2011. Oral presentation: *Optimal classification of the multivariate GRF observations*.
6. Applied Stochastic Models and Data Analysis. Vilnius, Lithuania, June 30-July 3, 2009. Oral presentation: *The influence of the anisotropy ratio on the expected error rates in classification of stationary GRF observations*.
7. Mathematical modelling and analysis. Trakai, Lithuania, June 1 – 5, 2005. Oral presentation: *Modelling of geometric anisotropic spatial variation*.

**National conferences** - conferences of *Lithuanian Mathematical Society* held in 2018, 2016, 2012, 2011, 2010, 2005 and 2004.

# Chapter 1

## Gaussian models and their characteristics

In this chapter we describe the most commonly used linear model for spatial data and its characteristics. In Section 1.1 the main definitions are presented, and the linear model and its components to be used in the thesis are described. This section also contains a description of anisotropic data and commonly found forms of anisotropy and briefly discusses the methods for determining the anisotropy. In Section 1.2 the ML estimators for spatial model parameters are discussed. The conditions which are sufficient for the asymptotic normality and weak consistency of ML estimators, established by Mardia and Marshall [46] are presented. In Subsection 1.2.1 the ML estimators for the geometrically anisotropic covariance are obtained. Finally, a non-parametric test for the geometric anisotropy is presented in Section 1.3.

### 1.1. Modeling spatial data

For spatial phenomena, the model is usually a random field. Gaussian random fields (**GRF**) have a dominant role in spatial statistics and especially in the traditional field of geostatistics (Cressie [10], Cressie and Wikle [12], Diggle and Ribeiro [13], Chiles and Delfiner [9]). Traditionally (e.g. Cressie [10]), statistical models for spatial data are divided into two broad classes: geostatistical models with continuous spatial support, and lattice models, where data occur on lattice with a countable set of nodes or locations. We will focus on these two types of Gaussian random fields in this dissertation:

geostatistical Gaussian random fields (**GGRF**) and Gaussian Markov random fields (**GMRF**), a subclass of lattice model for Gaussian data. A comprehensive description and applications of GMRF could be found in the work of Rue and Held [56]. Recall that a random field  $Z(\mathbf{s})$  is said to be Gaussian if, for any positive integer  $n$  and any set of locations  $\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n \in \mathbb{R}^d$ , the joint distribution of  $Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n)$  is multivariate Gaussian. One attractive feature of Gaussian random fields is that the first two moments determine the complete distribution. Constructing a GMRF is straightforward: it is just a finite-dimensional random vector following a multivariate Gaussian distribution with additional conditional independence properties. Gaussian Markov random fields are discrete domain GRFs equipped with Markov property. If the data do not follow the normal distribution, the Box-Cox transformation could be applied to normalize the data.

Let  $D \subset \mathbb{R}^d, d \in \mathbb{N}$  denote a spatial domain of interest and  $\mathbf{s} \in D$  represents a location where the observations of variable  $Z$  are taken. Then  $Z(\mathbf{s})$  is an observation of  $Z$  at the location  $\mathbf{s}$ . Assume that  $Z(\mathbf{s})$  is a GRF observation, then the model of  $Z(\mathbf{s})$  is given by a general linear model

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \varepsilon(\mathbf{s}). \quad (1.1)$$

Here  $\mu(\mathbf{s})$  is a deterministic mean function or spatial trend, which captures the large-scale spatial variation and  $\varepsilon(\mathbf{s})$  is an error term (small-scale variation (Haining [28]) which is generated by a zero-mean GRF  $\{\varepsilon(\mathbf{s}): \mathbf{s} \in D\}$  with covariance function defined by the model for all  $\mathbf{s}_1, \mathbf{s}_2 \in D$

$$C(\mathbf{s}_1, \mathbf{s}_2) = cov(\varepsilon(\mathbf{s}_1), \varepsilon(\mathbf{s}_2)).$$

For random fields in general and for GRFs in particular, the positive definiteness of the covariance function is a sufficient and necessary condition for establishing consistent finite-dimensional distributions.

**Definition 1.1.** Let  $n$  be a positive integer and let  $\{\mathbf{s}_i, i = 1, \dots, n\}$  be a finite set of spatial locations. Then for real numbers  $\{a_i, i = 1, \dots, n\}$  the function  $C(\cdot)$  is said to be positive definite if

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j C(\mathbf{s}_i, \mathbf{s}_j) \geq 0.$$

The mean function usually is expressed as a parametric linear model  $\mu(\mathbf{s}) = \mathbf{x}'(\mathbf{s})\boldsymbol{\beta}$ , where  $\mathbf{x}'(\mathbf{s}) = (x_1(\mathbf{s}), \dots, x_q(\mathbf{s}))$  is a vector of non-random covariates, and  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_q)'$  is a vector of parameters. Here  $q$  denotes the number of covariates. In the thesis we assume that  $x_1(\mathbf{s}) = 1$  for all  $\mathbf{s} \in D$  and the following parametric linear mean models are used:

- *Constant mean model*, if  $q = 1$  and  $\beta_1 \equiv \mu = \text{const}$ .
- *Trend surface mean model*, if  $x_j(\mathbf{s})$ ,  $j = 2, \dots, q$ , are the polynomials of spatial coordinates.
- *Regression mean model*, if  $x_j(\mathbf{s})$ ,  $j = 2, \dots, q$ , are regressors (independent variables).

An essential concept related to the analysis of spatial processes is their stationarity or homogeneity (Yaglom [74]). A random field is called *strongly stationary* (or *strictly stationary*) if all its finite-dimensional distributions are invariant under the arbitrary spatial shifts. This assumption is often too strict and hard to be verified, so it is usually weakened as follows.

**Definition 1.2.** The spatial process  $\{Z(\mathbf{s}): \mathbf{s} \in D\}$  is called *stationary* or *homogeneous* if it satisfies the following properties:  $E\{|Z(\mathbf{s})|^2\} < \infty$  for all  $\mathbf{s} \in D$ ;  $E(Z(\mathbf{s})) \equiv \mu = \text{const}$  for all  $\mathbf{s} \in D$  and  $C(\mathbf{s}_1, \mathbf{s}_2) = C(\mathbf{s}_1 - \mathbf{s}_2)$  for all  $\mathbf{s}_1, \mathbf{s}_2 \in D$  is a function of distance only. These fields are also known as *second-order*, *wide-sense* or *weakly homogeneous* fields (Yaglom [74], Haining [27], Cressie [10], Cressie and Wikle [12]).

The covariance function  $C(\mathbf{s})$  of stationary random field has the following properties:

- 1)  $C(\mathbf{0}) \geq 0$ ,
- 2)  $C(-\mathbf{s}) = C(\mathbf{s})$  for all  $\mathbf{s} \in D$ ,
- 3)  $C(\mathbf{s}) \leq C(\mathbf{0})$  for all  $\mathbf{s} \in D$ .

An important characteristic of spatial data is *spatial correlation* or *autocorrelation*. Spatial correlation defines how a variable relates with itself in relation to its position in space. The correlation function of a stationary spatial process is defined as  $R(\mathbf{s}) = C(\mathbf{s})/C(\mathbf{0})$ .

A more general assumption than *stationarity* leads to modelling spatial variation using the *semivariogram* and is called an *intrinsic stationarity* (Cressie [10], Haining [28]).

**Definition 1.3.** Suppose  $\text{var}(Z(\mathbf{s}_1) - Z(\mathbf{s}_2)) = 2\gamma(\mathbf{s}_1, \mathbf{s}_2)$  for all  $\mathbf{s}_1, \mathbf{s}_2 \in D$ . The quantity  $2\gamma(\cdot)$ , which is a function of increment  $Z(\mathbf{s}_1) - Z(\mathbf{s}_2)$ , is called a *variogram* and  $\gamma(\cdot)$  is called a *semivariogram* (Cressie [10]).

A variogram  $2\gamma(\cdot)$  must satisfy a property called *conditional negative definiteness*.

**Definition 1.4.** For any finite set of spatial locations  $\{\mathbf{s}_i, i = 1, \dots, n\}$  and real numbers  $\{a_i, i = 1, \dots, n\}$  satisfying  $\sum_{i=1}^n a_i = 0$ , the function  $2\gamma(\cdot)$  is said to be *conditionally negative definite* if

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j 2\gamma(\mathbf{s}_i, \mathbf{s}_j) \leq 0.$$

**Definition 1.5.** The process  $\{Z(\mathbf{s}): \mathbf{s} \in D\}$  is called *intrinsically stationary* if  $\mu(\mathbf{s}) \equiv \mu$  for all  $\mathbf{s} \in D$  and semivariogram  $\gamma(\mathbf{s}_1, \mathbf{s}_2) = \gamma(\mathbf{s}_1 - \mathbf{s}_2)$ , for all  $\mathbf{s}_1, \mathbf{s}_2 \in D$ .

Denote the increment by  $\mathbf{h} = \mathbf{s}_1 - \mathbf{s}_2$ , then if  $\gamma(\mathbf{h})$  is a function on both the magnitude and the direction of  $\mathbf{h}$ , the semivariogram function is said to be *anisotropic* and if  $\gamma(\mathbf{h})$  depends only on the magnitude of  $\mathbf{h}$  then it is treated as an *isotropic* one. In the case of isotropic stationary processes there is a simple relationship between the semivariogram and covariance function

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}).$$

The covariance (or semivariogram) of the stationary isotropic spatial process could be described by three parameters: *nugget effect*, *sill* and *range*.

- The quantity  $\tau^2 = C(\mathbf{0}) - \lim_{|\mathbf{h}| \rightarrow +0} C(\mathbf{h})$  is called the *nugget effect*.
- The sum  $\tau^2 + \sigma^2$  is called a *sill* of covariance, where  $\sigma^2 = C(\mathbf{0})$  is the scale parameter giving the variability of the process and usually called a *partial sill*.
- The distance, at which the sill is reached, is often known as *the range* of the covariance function and will be denoted in the thesis by  $\alpha$ .

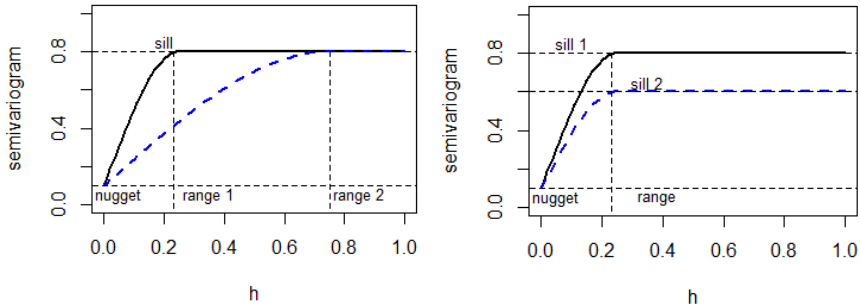
## Anisotropy

---

If the process is anisotropic, then the covariance function (or semivariogram) changes with respect to direction. A covariance/semivariogram formed by using only a certain direction-oriented pairs of observations is called directional covariance/semivariogram. Following Zimmerman [81], the anisotropy can take three forms: *sill anisotropy*, *range anisotropy* and *nugget anisotropy*. The range anisotropy is usually specified as either *geometric* (elliptical) *range anisotropy* or *non-geometric range anisotropy*. Chiles and Delfiner [9], Allard et al. [2] discuss *geometric*, *zonal* and *separable* models of anisotropy. Other authors classify anisotropies into two forms: *geometric* and *zonal* (Journel and Huijbregts [35], Goovaerts [24], Cressie [10], Cressie and Wikle [12]; Wackernagel [67]). Following these authors, the *geometric anisotropy* occurs when the range, but not the sill, of the covariance changes in different directions.

The *zonal anisotropy* exists when the sill of covariance/semivariogram function changes with the direction (Wackernagel [67]). The directional semivariograms corresponding to the different types of anisotropy are depicted in Figure 1. The forms of anisotropy are also discussed in [A19]. In this thesis we will focus on the *geometric anisotropy*. The geometric anisotropy means that the correlation is stronger in one direction than it is in the other directions. Mathematically, if one plots the directional ranges in two dimensions, they would fall on the edge of an ellipse, where the major and minor axes of ellipse correspond to the largest and shortest ranges of directional semivariograms. The geometric anisotropy can be reduced to the isotropy by a mere linear transformation of the coordinates. A description of

the linear coordinate transformation procedure can be found in the papers of Haskard [30], Sherman [62]. The issues of modelling the geometric and zonal spatial variation are also analyzed in [A18] and [A17].



**Figure 1.** Directional semivariograms in the case of the geometric and zonal anisotropy

Under the geometric anisotropy the covariance function is extended by adding two parameters: *an anisotropy ratio* and *an anisotropy angle*.

- *The anisotropy ratio*, denoted by  $\lambda$ , is equal to the ratio of the lengths of the two principal ellipse axes. This parameter is always positive and it could be greater or less than 1, with isotropy corresponding to  $\lambda = 1$ .
- *The anisotropy angle*,  $\varphi$ , is the angle of rotation which is made by the major axis of the ellipse (which is known as the direction of anisotropy) and the coordinate axis  $Ox$ . In other words it determines the orientation of the ellipse.

The assumption of the spatial isotropy is often made in practice for ease of computation and simpler interpretation. But in many applications the spatial isotropy is not a reasonable assumption, thus it is very important to verify the existence of anisotropy. There are formal and the non-formal methods (graphical techniques) to determine the anisotropy. The examples of non-formal methods are: assessing *directional semivariograms*, drawing *rose diagrams* (Isaak and Srivastava [34], Ecker and Gelfand [22]), *semivariogram* or *contour maps* (Isaak and Srivastava [34]). Despite the fact that these methods could be easily implemented, they are difficult to assess and open to interpretation.

A variety of nonparametric tests of isotropy have been proposed by Lu and Zimmerman [41], [42], Guan et al. [25], Maity and Sherman [45]. An original, simple non-parametric test statistic based on directional empirical semivariograms was proposed by Dučinskas and Dreizienė [A14] and is presented in Section 1.3. Weller and Hoeting in their paper [70] have presented a comprehensive review of different non-parametric methods for testing the isotropy. Several of the aforementioned tests were recently implemented in the R package `spTest`, available on CRAN (Weller [69]).

## 1.2. Estimators for spatial model parameters

In practical applications the true parameters are not usually known so they need to be evaluated using statistical sampling. For spatial data the adapted classical methods, such as *Maximum Likelihood* (ML) and *Least Squares* (LS) methods, may be used to estimate unknown parameters. There are also specific methods, i.e.: *Pseudo-maximum Likelihood* (PML) method (Cressie [10], Gupta and Robinson [26], Johansson [33]), *Coding method* (Besag [8], Johansson [33]), *Bayesian method* (Lu, Zhang [43], Dučinskas, Šaltytė [15]), which could be used for estimation of the unknown parameters. In this thesis the ML method for theoretical results is used.

Consider a sample with  $n$  observations ( $n > q$ ) which comes from GRF and could be described by the equation (1.1). Let  $\mathbf{Z}_n = (Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))'$  denote the vector of observations. Then the model of  $\mathbf{Z}_n$  could be specified by the equation  $\mathbf{Z}_n = \boldsymbol{\mu}_n + \boldsymbol{\varepsilon}_n$  with mean vector  $\boldsymbol{\mu}_n = (\mu(\mathbf{s}_1), \dots, \mu(\mathbf{s}_n))'$  and vector of random errors  $\boldsymbol{\varepsilon}_n = (\varepsilon(\mathbf{s}_1), \dots, \varepsilon(\mathbf{s}_n))$ .

For a non-constant (*regression* or *trend surface*) mean model  $\boldsymbol{\mu}_n$  would have the expression  $\boldsymbol{\mu}_n = \mathbf{X}\boldsymbol{\beta}$ , where  $\mathbf{X}$  is called a design matrix

$$\mathbf{X} = \begin{pmatrix} 1 & x_2(\mathbf{s}_1) & \dots & x_q(\mathbf{s}_1) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_2(\mathbf{s}_n) & \dots & x_q(\mathbf{s}_n) \end{pmatrix}.$$



If  $\text{rank}(\mathbf{X}) = q$  and  $\text{rank}(\mathbf{\Sigma}) = n$ , the term  $\boldsymbol{\varepsilon}_n$  has a multivariate Gaussian distribution  $N_n(\mathbf{0}, \mathbf{\Sigma}(\boldsymbol{\theta}))$ , where  $\mathbf{\Sigma}(\boldsymbol{\theta})$  denotes the covariance matrix which for a random vector  $\mathbf{Z}_n$  is specified as

$$\mathbf{\Sigma}(\boldsymbol{\theta}) = \begin{pmatrix} C(\mathbf{s}_1, \mathbf{s}_1; \boldsymbol{\theta}) & C(\mathbf{s}_1, \mathbf{s}_2; \boldsymbol{\theta}) & C(\mathbf{s}_1, \mathbf{s}_3; \boldsymbol{\theta}) & \cdots & C(\mathbf{s}_1, \mathbf{s}_n; \boldsymbol{\theta}) \\ C(\mathbf{s}_2, \mathbf{s}_1; \boldsymbol{\theta}) & C(\mathbf{s}_2, \mathbf{s}_2; \boldsymbol{\theta}) & C(\mathbf{s}_2, \mathbf{s}_3; \boldsymbol{\theta}) & \cdots & C(\mathbf{s}_2, \mathbf{s}_n; \boldsymbol{\theta}) \\ C(\mathbf{s}_3, \mathbf{s}_1; \boldsymbol{\theta}) & C(\mathbf{s}_3, \mathbf{s}_2; \boldsymbol{\theta}) & C(\mathbf{s}_3, \mathbf{s}_3; \boldsymbol{\theta}) & \cdots & C(\mathbf{s}_3, \mathbf{s}_n; \boldsymbol{\theta}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C(\mathbf{s}_n, \mathbf{s}_1; \boldsymbol{\theta}) & C(\mathbf{s}_n, \mathbf{s}_2; \boldsymbol{\theta}) & C(\mathbf{s}_n, \mathbf{s}_3; \boldsymbol{\theta}) & \cdots & C(\mathbf{s}_n, \mathbf{s}_n; \boldsymbol{\theta}) \end{pmatrix}.$$

The elements of this matrix are the values of parametric covariance function defined for all  $\mathbf{s}_i \in D$ ,  $i = 1, \dots, n$ .  $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k$  is a  $p \times 1$  vector of unknown covariance parameters.

**Note.** For the notational convenience we will omit  $\boldsymbol{\theta}$  if it does not play an essential role.

The covariance matrix for the geometric anisotropy case could be factorized as  $\mathbf{\Sigma} = \sigma^2 \mathbf{\Gamma}$ , where  $\sigma^2$  is a scale parameter and  $\mathbf{\Gamma} = \rho \mathbf{I}_n + \mathbf{R}$  is a standardized covariance matrix. Here  $\rho = \tau^2 / \sigma^2$  is a *relative nugget* or *noise to signal variance* (see Diggle and Ribeiro [13]) and  $\mathbf{R}$  is a spatial correlation matrix.

If the nuggetless covariance ( $\tau^2 = 0$ ) is considered then the factorized covariance matrix becomes  $\mathbf{\Sigma} = \sigma^2 \mathbf{R}$ .

The estimators of population parameters depend on the parametric uncertainty level. We will discuss three different cases of parametric uncertainty:

- The vector of mean parameters  $\boldsymbol{\beta}$  is unknown, and the vector of covariance parameters  $\boldsymbol{\theta}$  is known.
- The vector of mean parameters  $\boldsymbol{\beta}$  and the scale parameter  $\sigma^2$  are unknown, and the standardized covariance matrix  $\mathbf{\Gamma}$  is known.
- The case of complete parametric uncertainty. In this case all mean and covariance parameters,  $\boldsymbol{\beta}$  and  $\boldsymbol{\theta}$ , are unknown.

In general we will denote the combined vector of unknown parameters by  $\boldsymbol{\Psi} = (\boldsymbol{\beta}', \boldsymbol{\theta}')$ .

Since the spatial process is Gaussian then the log-likelihood function for vector  $\mathbf{Z}_n$  is

$$\ln L = (2\pi)^{-n/2} - \frac{1}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} (\mathbf{Z}_n - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} (\mathbf{Z}_n - \mathbf{X}\boldsymbol{\beta}).$$

If the vector of mean parameters  $\boldsymbol{\beta}$  is unknown then the ML estimator is

$$\widehat{\boldsymbol{\beta}}_{ML} = (\mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{Z}_n \sim N_q(\boldsymbol{\beta}, (\mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1}). \quad (1.2)$$

It is obvious that the estimator (1.2) is unbiased and efficient since  $\text{cov}(\widehat{\boldsymbol{\beta}}_{ML}) = \mathbf{J}_\beta^{-1}$ , where  $\mathbf{J}_\beta = -E\left(\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right) = \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X}$  defines the information matrix.

In the case of factorized covariance matrix  $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{\Gamma}$ , the ML estimators for  $\boldsymbol{\beta}$  and  $\sigma^2$  are

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{ML} &= (\mathbf{X}' \boldsymbol{\Gamma}^{-1} \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\Gamma}^{-1} \mathbf{Z}_n, \\ \widehat{\sigma}_{ML}^2 &= \frac{1}{n} (\mathbf{Z}_n - \mathbf{X} \widehat{\boldsymbol{\beta}}_{ML})' \boldsymbol{\Gamma}^{-1} (\mathbf{Z}_n - \mathbf{X} \widehat{\boldsymbol{\beta}}_{ML}). \end{aligned}$$

It is easy to show that

$$\widehat{\boldsymbol{\beta}}_{ML} \sim N_q(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}' \boldsymbol{\Gamma}^{-1} \mathbf{X})^{-1}).$$

The information matrix is  $\mathbf{J}_\beta = -E\left(\frac{\partial^2 \ln L(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right) = \frac{1}{\sigma^2} (\mathbf{X}' \boldsymbol{\Gamma}^{-1} \mathbf{X})$ , thus the estimator  $\widehat{\boldsymbol{\beta}}_{ML}$  is unbiased and efficient.

The ML estimator for  $\sigma^2$  is biased since  $E(\widehat{\sigma}_{ML}^2) = \sigma^2(n - q)/n$  and  $\widehat{\sigma}_{ML}^2 \sim \sigma^2 \chi_{n-q}^2/n$ .

In the thesis the unbiased estimator of  $\sigma^2$  will be used

$$\tilde{\sigma}_{ML}^2 = \widehat{\sigma}_{ML}^2 n / (n - q).$$

The most complicated situation is the case of complete parametric uncertainty. When all covariance function parameters are unknown the analytic solution does not exist. For GRF the asymptotic properties of ML estimators are established by Mardia and Marshall [47].

**Theorem 1.1.** (Mardia and Marshall [46]). Suppose  $\mathbf{Z}_n \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ ,  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$ . Let  $\lambda_1 \leq \dots \leq \lambda_n$  be the eigenvalues of the covariance matrix  $\boldsymbol{\Sigma}$

and let those of  $\Sigma_i, \Sigma_{ij}$  be  $\lambda_k^i$  and  $\lambda_k^{ij}, k = 1, \dots, n$ , respectively with  $|\lambda_1^i| \leq \dots \leq |\lambda_n^i|$  and  $|\lambda_1^{ij}| \leq \dots \leq |\lambda_n^{ij}|$  for  $i, j = 1, \dots, p$ . Here  $\Sigma_i = \partial \Sigma / \partial \theta_i, \Sigma_{ij} = \partial^2 \Sigma / \partial \theta_i \partial \theta_j$ .

Moreover, suppose that as  $n \rightarrow \infty$ :

- (a)  $\lim \lambda_n = C < \infty, \lim |\lambda_n^{ij}| = C_{ij} < \infty \forall i, j = 1, \dots, p$ ;
- (b)  $\|\Sigma_i\|^{-2} = O\left(n^{-\frac{1}{2}-\delta}\right)$  for some  $\delta > 0$ , for  $i = 1, \dots, p$ ;
- (c)  $\forall i, j = 1, \dots, p, a_{ij} = \lim \left\{ \frac{t_{ij}}{(t_{ii}t_{jj})^{1/2}} \right\}$  exists, where  
 $t_{ij} = \text{tr}(\Sigma^{-1} \Sigma_i \Sigma^{-1} \Sigma_j)$  and  $\mathbf{A} = (a_{ij})$  is a non-singular matrix;
- (d)  $\lim(\mathbf{X}'\mathbf{X})^{-1} = 0$ .

Then the ML estimator of  $\Psi = (\beta', \theta')'$  is weakly consistent and asymptotically Gaussian

$$\hat{\Psi} \sim AN(\Psi, \mathbf{J}^{-1}). \quad (1.3)$$

Here  $\mathbf{J} = \mathbf{J}_\beta \oplus \mathbf{J}_\theta$  is an information matrix which could be found as a direct sum of components:

$$\mathbf{J}_\beta = \mathbf{X}'\Sigma^{-1}\mathbf{X}, \quad (1.4)$$

$$(\mathbf{J}_\theta)_{ij} = \text{tr}(\Sigma^{-1} \Sigma_i \Sigma^{-1} \Sigma_j) / 2. \quad (1.5)$$

It should be noted that the estimators  $\hat{\beta}$  and  $\hat{\theta}$  are asymptotically independent and the asymptotic covariance matrix is simply the inverse of Fisher information matrix. This theorem is proved under the *increasing domain* asymptotics framework in which the minimum distance between sampling points is bounded away from zero and thus the spatial domain of observation is unbounded (Zhang and Zimmerman [77]). This is the spatial analogue of the asymptotics observed in time series.

There are two other asymptotic frameworks: *infill asymptotics* (or *fixed-domain asymptotics*) and *mixed domain asymptotics* (Lahiri [38]). If the *infill asymptotics* framework is considered, the spatial domain is fixed (bounded) and the locations of observations get denser, as the number of observations increases (Cressie [10]). The *mixed domain* asymptotics or *hybrid asymptotics* (Zheng, Zhu [78]) is a combination of *increasing domain* and *infill asymptotics*. Here the sampling region grows to infinity and at the same

time the distance between neighboring sampling sites goes to zero (Lahiri [38]).

The asymptotic behavior of spatial covariance parameter estimators can be different under the different asymptotic spatial frameworks. For example, Zheng and Zhu [78] showed that under each type of asymptotics the rates of convergence vary and under infill asymptotics some of the model parameter estimators are inconsistent.

In the thesis the ML estimators derived under the increasing domain asymptotic framework are considered.

**Corollary 1.1.** Consider  $\mathbf{Z}_n \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ . Then, according to the conditions formulated in Theorem 1.1, the ML estimator  $\hat{\boldsymbol{\theta}}$  satisfies the following properties (as  $n \rightarrow \infty$ ):

$$\hat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta} \text{ and } \hat{\boldsymbol{\theta}} \sim AN_p(\boldsymbol{\theta}, \mathbf{J}_{\boldsymbol{\theta}}^{-1}).$$

Similarly

$$\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta} \text{ and } \hat{\boldsymbol{\beta}} \sim AN_q(\boldsymbol{\beta}, \mathbf{J}_{\boldsymbol{\beta}}^{-1}).$$

The asymptotic covariance matrix  $\mathbf{J}_{\boldsymbol{\theta}}$  is a symmetric  $p \times p$  matrix (in the case of geometric anisotropy  $p = 5$ ) with elements defined in (1.5). The matrix  $\mathbf{J}_{\boldsymbol{\beta}}$  is defined in (1.4).

### 1.2.1. ML estimators for geometrically anisotropic covariance

As it was mentioned in the case of geometric anisotropy the covariance matrix of  $\mathbf{Z}_n$  could be expressed as (see Ecker, Gelfand [22])

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \tau^2 \mathbf{I}_n + \sigma^2 \mathbf{R}, \quad (1.6)$$

here  $\tau^2$  is a nugget effect,  $\sigma^2$  is a scale parameter or partial sill,  $\mathbf{R} = \mathbf{R}(\boldsymbol{\vartheta})$  denotes the matrix of spatial correlations. The vector  $\boldsymbol{\vartheta} = (\alpha, \lambda, \varphi)'$  has three components since there are three parameters which determines spatial correlation in the case of geometric anisotropy: the range parameter  $\alpha$ , anisotropy ratio  $\lambda$ , and anisotropy angle  $\varphi$ . Then the vector of covariance parameters is

$$\boldsymbol{\theta} = (\tau^2, \sigma^2, \boldsymbol{\vartheta}')' = (\tau^2, \sigma^2, \alpha, \lambda, \varphi)'$$

In order to build the asymptotic covariance matrix  $\mathbf{J}_\theta$  for (1.6), we need to find the first order partial derivatives of  $\boldsymbol{\Sigma}(\boldsymbol{\theta})$  with respect to the  $i$ -th covariance parameter,  $i = 1, \dots, 5$ . To get the analytic expressions of  $\boldsymbol{\Sigma}_i$  we choose anisotropic exponential parametric model of covariance function

$$C(h, \boldsymbol{\theta}) = \begin{cases} \tau^2 + \sigma^2, & h = 0 \\ \sigma^2 \exp\left\{-\sqrt{(h_u)^2 + \lambda^2(h_v)^2}/\alpha\right\}, & h \neq 0 \end{cases} \quad (1.7)$$

where

$$h_u = (h_x \cos \varphi + h_y \sin \varphi)/\alpha_{max},$$

$$h_v = (-h_x \sin \varphi + h_y \cos \varphi)/\alpha_{max},$$

$$h_x = x_i - x_j, \text{ and } h_y = y_i - y_j, i, j = 1, \dots, n.$$

$x_i$  and  $y_i$  denote the X and Y coordinates of location  $\mathbf{s}_i$ . To get (1.7) the linear coordinate transformation procedure was applied. More about the coordinate transformation procedure can be found in Sherman [62], Chiles and Delfiner [9].

**Lemma 1.1.** Consider the exponential covariance model (1.7). Then the first order partial derivatives,  $\boldsymbol{\Sigma}_i = \partial \boldsymbol{\Sigma} / \partial \boldsymbol{\theta}_i$ ,  $i = 1, \dots, 5$ , are

$$\boldsymbol{\Sigma}_1 = \partial \boldsymbol{\Sigma} / \partial \tau^2 = \mathbf{I}_n, \quad (1.8)$$

$$\boldsymbol{\Sigma}_2 = \partial \boldsymbol{\Sigma} / \partial \sigma^2 = \mathbf{R}, \quad (1.9)$$

$$\boldsymbol{\Sigma}_3 = \partial \boldsymbol{\Sigma} / \partial \alpha = \frac{\sigma^2}{\alpha^2} \mathbf{R} \circ \mathbf{H}_\alpha, \quad (1.10)$$

$$\boldsymbol{\Sigma}_4 = \partial \boldsymbol{\Sigma} / \partial \lambda = -\frac{\lambda \sigma^2}{\alpha} \mathbf{R} \circ \mathbf{H}_\lambda, \quad (1.11)$$

$$\boldsymbol{\Sigma}_5 = \partial \boldsymbol{\Sigma} / \partial \varphi = \frac{\sigma^2(\lambda^2 - 1)}{\alpha} \mathbf{R} \circ \mathbf{H}_\varphi. \quad (1.12)$$

Here  $\mathbf{H}_\alpha$ ,  $\mathbf{H}_\lambda$ ,  $\mathbf{H}_\varphi$  are the  $n \times n$  matrices with diagonal elements equal to 0. The off-diagonal elements ( $i \neq j$ ) are the following

$$(\mathbf{H}_\alpha)_{ij} = \sqrt{(h_u^{ij})^2 + \lambda^2(h_v^{ij})^2}, \quad (1.13)$$

$$(\mathbf{H}_\lambda)_{ij} = (h_v^{ij})^2 / \sqrt{(h_u^{ij})^2 + \lambda^2(h_v^{ij})^2}, \quad (1.14)$$

$$(\mathbf{H}_\varphi)_{ij} = h_u^{ij} h_v^{ij} / \sqrt{(h_u^{ij})^2 + \lambda^2 (h_v^{ij})^2}, \quad (1.15)$$

$$h_u^{ij} = (h_x^{ij} \cos \varphi + h_y^{ij} \sin \varphi) / \alpha, \quad h_v^{ij} = (-h_x^{ij} \sin \varphi + h_y^{ij} \cos \varphi) / \alpha,$$

$$h_x^{ij} = x_i - x_j, \quad h_y^{ij} = y_i - y_j, \quad i, j = 1, \dots, n.$$

**Proof.** (1.8)-(1.15) were obtained differentiating (1.6) with respect to the components of the parameter vector  $\boldsymbol{\theta} = (\tau^2, \sigma^2, \alpha, \lambda, \varphi)'$  and assuming that the elements of correlation matrix  $\mathbf{R}$  are  $\exp\{-\sqrt{(h_u)^2 + \lambda^2 (h_v)^2} / \alpha\}$ .

Having the expressions of partial derivatives we can build the asymptotic covariance matrix  $\mathbf{J}_\theta$ .

The obtained expressions could be applied constructing the optimality criterion for the spatial sampling design (Zimmerman [82]). They will also be applied in the thesis for solving the classification problems of spatial data.

### 1.3. Non-parametric test for spatial geometric anisotropy

A conventional practice when checking for isotropy is to assess the plots of empirical semivariograms. However these graphical techniques are open to interpretation. Guan et al. [25] have proposed a formal approach to test the isotropy which is based on the asymptotic joint normality of empirical semivariograms for multiple directions. An  $L_2$ -consistent subsampling estimator for the asymptotic covariance matrix of the empirical semivariogram is used to construct a test statistic. But the subsampling procedure takes a large amount of computing time. We propose a simpler test statistic in Gaussian case under the assumption of independence of the classical semivariogram estimators. These results are published in Dučinskas and Dreičienė [A14].

Suppose that the spatial data are the observations of GRF modelled by the equation  $Z(\mathbf{s}) = \mu(\mathbf{s}) + \varepsilon(\mathbf{s})$  which is specified in (1.1) with constant mean model. According to Definition 1.3 recall that  $\frac{1}{2} \text{var}(Z(\mathbf{s}_1) - Z(\mathbf{s}_2)) =$

$\gamma(\mathbf{s}_1, \mathbf{s}_2)$  for all  $\mathbf{s}_1, \mathbf{s}_2 \in D$  is called a semivariogram. We consider the geometric anisotropy case which means that the semivariograms have the same nugget, same sill but different ranges in to the perpendicular directions (Wackernagel [68]).

Denote by  $\mathbf{S}_n = \{\mathbf{s}_i \in D, i = 1, \dots, n\}$  the set of locations where GRF is observed, and use the classical semivariogram estimator proposed by Matheron (1962), based on the method of moments (Cressie [10])

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{|N(\mathbf{h})|} \sum_{(\mathbf{s}_i, \mathbf{s}_j) \in N(\mathbf{h})} (Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2. \quad (1.16)$$

Here  $N(\mathbf{h})$  denotes all pairs  $(\mathbf{s}_i, \mathbf{s}_j)$  for which  $\mathbf{s}_i, \mathbf{s}_j \in \mathbf{S}_n, \mathbf{s}_i - \mathbf{s}_j = \mathbf{h}$  and  $|N(\mathbf{h})|$  denotes the cardinality of the set  $N(\mathbf{h})$ .

To assess the hypothesis of isotropy, we choose the lag set  $\mathbf{L}$  including spatial lags  $h_1, h_2, \dots, h_K$  in the direction of major axis of ellipse and spatial lags  $h_{K+1}, h_{K+2}, \dots, h_{2K}$  perpendicular to that direction, i.e.

$$\mathbf{L} = (h_1, \dots, h_K, h_{K+1}, \dots, h_{2K})'.$$

Assume that  $|h_i| = |h_{i+K}|, i = 1, \dots, K$ . Then the hypothesis of isotropy is expressed as

$$H_0: \gamma(h_i) = \gamma(h_{i+K}), \text{ for all } i = 1, \dots, K.$$

$$H_1: \exists i \in \{1, \dots, K\} \text{ such that } \gamma(h_i) \neq \gamma(h_{i+K}).$$

Rejecting the hypothesis  $H_0$  means accepting the geometric anisotropy.

Set  $\boldsymbol{\gamma}' = (\gamma(h_1), \dots, \gamma(h_{2K}))$  and let  $\hat{\boldsymbol{\gamma}}' = (\hat{\gamma}(h_1), \dots, \hat{\gamma}(h_{2K}))$  be the vector of semivariogram estimators obtained by (1.16) over  $\mathbf{S}_n$ .

In what follows, we establish the asymptotic properties of  $\hat{\boldsymbol{\gamma}}$  under an increasing domain framework proposed by Guan et al. [25]. Under some regularity conditions, it was proved that

$$\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \xrightarrow{D} N_{2K}(\mathbf{0}, \boldsymbol{\Sigma}_\gamma) \text{ as } n \rightarrow \infty,$$

where  $\boldsymbol{\Sigma}_\gamma$  is the asymptotic covariance matrix with the elements of complex structure.

Under the hypothesis of isotropy, there exists a full row rank  $r \times 2K$  matrix  $\mathbf{A}$  such that  $\mathbf{A}\boldsymbol{\gamma} = \mathbf{0}$  (Lu, Zimmerman [41]). Matrix  $\mathbf{A}$  is called a

*contrast matrix*. Then under the hypothesis of isotropy it follows from continuous mapping theorem that

$$n(\mathbf{A}\hat{\boldsymbol{\gamma}})'(\mathbf{A}\boldsymbol{\Sigma}_\gamma\mathbf{A}')^{-1}(\mathbf{A}\hat{\boldsymbol{\gamma}}) \xrightarrow{D} \chi_r^2 \text{ as } n \rightarrow \infty.$$

Following Cressie [10] we have

$$\text{var}(\hat{\boldsymbol{\gamma}}) \cong \text{diag}\left(\frac{2\gamma^2(h_1)}{|N(h_1)|}, \dots, \frac{2\gamma^2(h_{2K})}{|N(h_{2K})|}\right),$$

where the approximation yields only little loss in the estimation efficiency especially in the case of independence of the classical semivariogram estimators for different spatial lags.

We propose the following estimator of  $\boldsymbol{\Sigma}_\gamma$

$$\hat{\boldsymbol{\Sigma}}_\gamma = n \text{diag}\left(\frac{2\hat{\gamma}^2(h_1)}{|N(h_1)|}, \dots, \frac{2\hat{\gamma}^2(h_{2K})}{|N(h_{2K})|}\right)$$

and use it to form the test statistic

$$\hat{\mathbf{T}} = n(\mathbf{A}\hat{\boldsymbol{\gamma}})'(\mathbf{A}\hat{\boldsymbol{\Sigma}}_\gamma\mathbf{A}')^{-1}(\mathbf{A}\hat{\boldsymbol{\gamma}}). \quad (1.17)$$

The proposed test statistic according to the Slutsky theorem (Ferguson [23]) follows the asymptotic chi-square distribution,  $\hat{\mathbf{T}} \sim \chi_r^2$ . So the hypothesis  $H_0$  is to be rejected if  $\hat{\mathbf{T}} > \chi_{r,p}^2$ , where  $\chi_{r,p}^2$  is the  $p$ -critical value of a chi-squared distribution with  $r$  degrees of freedom. If  $\hat{\mathbf{T}} \leq \chi_{r,p}^2$ , the hypothesis  $H_0$  is to be accepted.

The simulation experiment demonstrating the efficacy of the proposed test is presented in Section 3.2.



# Chapter 2

## Classification of spatial GRF observations

This chapter contains the main results of the thesis. In Section 2.1 the general definitions related with discriminant analysis are introduced. Later the formulas for Bayes risk and actual risk as well as formulas for error rates and actual error rates for the different number of populations are presented. In Section 2.2 the problem of classification of GGRF observation is analyzed. The Bayes risk associated with Bayes discriminant function and the asymptotic approximation formula of expected risk (AER) are derived. The above mentioned results are obtained for the univariate and multivariate cases and for different number of populations. Finally, the closed-form expression of asymptotic covariance matrix for exponential covariance model is presented. Section 2.3 is designated for classification problem of GMRF observation into one of two populations. The multivariate case is considered.

The certain parts of this chapter are published in [A1]-[A4], [A6]-[A8], [A10]-[A13].

### 2.1. Elements of linear discriminant analysis

Consider a Gaussian random field  $\{Z(\mathbf{s}): \mathbf{s} \in D \subset \mathbb{R}^d\}$  defined on some probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  and taking the values in  $\mathcal{Z} = \mathbb{R}^p$ , where  $\mathcal{Z}$  is a suitable feature space. The main purpose is to solve the problem of

classification of a single GRF observation  $Z_0 = Z(\mathbf{s}_0)$ ,  $\mathbf{s}_0 \in D$ , to one of  $m$  populations  $\Omega_l$ ,  $l = 1, \dots, m$  (also known as classes or categories). Following Beret and Calder [7] the spatial location ( $\mathbf{s}_0$ ) of the observation to be classified will be called *focal location*. Based on that, we will call the above mentioned observation ( $Z_0$ ) as a *focal observation* (FO).

The model of FO  $Z_0$  in population  $\Omega_l$  is  $Z(\mathbf{s}) = \mu_l(\mathbf{s}) + \varepsilon(\mathbf{s})$ . Here  $\mu_l(\mathbf{s})$  is a mean function or spatial trend. The error term  $\varepsilon(\mathbf{s})$  is generated by the same zero-mean GRF  $\{\varepsilon(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\}$  with the covariance function defined by the model for all  $\mathbf{s}, \mathbf{s} + \mathbf{h} \in D$ ,  $C(\mathbf{h}) = cov(\varepsilon(\mathbf{s}), \varepsilon(\mathbf{s} + \mathbf{h}))$ ,  $l = 1, \dots, m$ .

Denote by  $\mathbf{S}_n = \{\mathbf{s}_i \in D, i = 1, \dots, n\}$  the set of locations where the training sample  $\mathbf{T} = (Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))'$  is taken, and call it the set of training locations (STL). It specifies the spatial sampling design or spatial framework for the training sample (Shekhar et al. [61]). The  $\mathbf{S}_n$  is partitioned into a union of  $m$  disjoint subsets, i.e.,  $\mathbf{S}_n = \mathbf{S}^{(1)} \cup \dots \cup \mathbf{S}^{(m)}$ , where  $\mathbf{S}^{(l)}$  contains  $n_l$  ( $n = \sum_{l=1}^m n_l$ ) observations of  $Z(\mathbf{s})$  from population  $\Omega_l$ ,  $l = 1, \dots, m$ .

Then the training sample  $\mathbf{T} = (\mathbf{T}'_1, \dots, \mathbf{T}'_m)'$  is a  $n$ -dimensional vector composed of observations which come from different populations and let  $\mathbf{t}$  denote the realization of training sample  $\mathbf{T}$ . Then the model of training sample in a vector form is

$$\mathbf{T} = \mathbf{M} + \mathbf{E},$$

here  $\mathbf{M} = (\mu_1(\mathbf{s}_1), \dots, \mu_1(\mathbf{s}_{n_1}), \dots, \mu_m(\mathbf{s}_{n_1+\dots+n_{m-1}+1}), \dots, \mu_m(\mathbf{s}_n))'$  and  $\mathbf{E}$  is the  $n$ -dimensional vector of random errors  $\varepsilon(\mathbf{s})$ .

The FO  $Z_0$  is assumed to be dependent on the training sample thus the conditional  $Z(\mathbf{s}_0)$  distributions will be used.

## The risk of classification: general definitions

---

Let  $p_l(Z_0|\mathbf{t})$ ,  $l = 1, \dots, m$  denote the class-conditional probability density function of  $Z_0$ , given  $\mathbf{T} = \mathbf{t}$ , in the population  $\Omega_l$ , describing the distribution of the feature vector in each population (Theodoritis [67]). The *losses of*

*classification* or the *loss function* when an object from the population  $l$  is allocated to the population  $k$ , is denoted by  $L(l, k)$ ,  $l, k = 1, \dots, m$ . The losses  $L(l, l)$  correspond to the correct decisions, in practice these are usually set equal to zero, although we have considered them for the sake of generality.

The following assumptions are made:

(A1) The prior probabilities,  $\pi_l$ ,  $l = 1, \dots, m$ ,  $\sum_{l=1}^m \pi_l = 1$ , are known and can be fixed or deterministic functions of  $\mathbf{s}_0$  and  $\mathbf{S}_n$ .

(A2) The values of *loss function*  $L(l, k)$  are non-negative and finite. Moreover, they do not depend on  $\mathbf{s}_0$  or training sample configuration.

The classification rule, given  $\mathbf{T} = \mathbf{t}$ , will be denoted as  $D_t(\bullet): \mathcal{Z} \rightarrow \{1, \dots, m\}$ . Then the *expected loss* or *conditional risk* (McLachlan [50]) of random observation  $Z_0$ , given  $\mathbf{t}$ , from the population  $\Omega_l$ , by prescribed classification rule is given by

$$R_0(l, D_t(\bullet)) = \int_{\mathcal{Z}} L(l, D_t(Z_0)) p_l(Z_0 | \mathbf{t}) dZ_0 = E_{Z_0 | \mathbf{t}, l} \{L(l, D_t(Z_0))\}.$$

Note that the sub index of  $R_0$  here represents the focal location  $\mathbf{s}_0$ .

Then the total risk is the total expected losses

$$R_0(D_t(\bullet)) = \sum_{l=1}^m \pi_l R_0(l, D_t(\bullet)). \quad (2.1)$$

The rule minimizing the total risk (2.1) is said to be *Bayes classification rule* (McLachlan [50], Anderson [3]) and will be denoted as  $D_t^B(\bullet)$  and for the FO  $Z_0$  it could be expressed as

$$D_t^B(Z_0) = \arg \min_{\{k=1, \dots, m\}} \{\sum_{l=1}^m \pi_l p_l(Z_0 | \mathbf{t}, \Psi) L(l, k)\}. \quad (2.2)$$

Recall that  $\Psi$  denotes the combined vector of unknown population parameters. Then *Bayes risk*, associated with Bayes classification rule (2.2), is

$$R_0^B = R_0(D_t^B(\bullet)) = \sum_{l=1}^m \pi_l E_{Z_0 | \mathbf{t}, l} L(l, D_t^B(Z_0)). \quad (2.3)$$

Let  $G_{ik}^B(Z_0)$  denote the *pairwise Bayes discriminant functions*

$$G_{ik}^B(Z_0, \Psi) = \sum_{j=1}^m \pi_j p_j(Z_0 | \mathbf{t}, \Psi) d(j, l, k), \quad (2.4)$$

where  $(j, l, k) = L(j, l) - L(j, k)$ ,  $l, k = 1, \dots, m$ . Also let  $e(x)$  be a Heaviside step function,

$$e(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1, & \text{if } x \geq 0 \end{cases}$$

Then the *Bayes risk*, defined in (2.3) and associated with (2.4) could be expressed as (see Dučinskas [16])

$$R_0^B(\Psi) = \sum_{l,k=1}^m \pi_l E_{Z_0|\mathbf{t},l} L(l, k) \prod_{j=1, j \neq k} e(G_{kj}^B(Z_0, \Psi)). \quad (2.5)$$

Since in practical situations the complete parametrical certainty of populations usually is not attained, the parameter estimators from the training sample should be plugged into the Bayes discriminant function (**BDF**). Plug-in BDF will be abbreviated as **PBDF**.

Let  $\hat{\Psi}$  denote the vector of parameters estimates. Thus, replacing the vector of the unknown parameters of  $G_{lk}^B(Z_0, \Psi)$  by the vector of estimates we get pairwise PBDF

$$\hat{G}_{lk}^B(Z_0) = G_{lk}^B(Z_0, \hat{\Psi}). \quad (2.6)$$

**Definition 2.1.** The *actual risk*, given  $\mathbf{T} = \mathbf{t}$ , for PBDF (2.6) is defined as

$$R_0^B(\hat{\Psi}) = \sum_{l,k=1}^m \pi_l E_{Z_0|\mathbf{t},l} L(l, k) \prod_{j=1, j \neq k} e(\hat{G}_{kj}^B(Z_0)). \quad (2.7)$$

**Definition 2.2.** The expectation of the actual risk with respect to the distribution of  $\mathbf{T}$  is called the *expected risk* and is denoted as

$$ER = E_T(R_0^B(\hat{\Psi})).$$

The ER is useful in providing a guide to the performance of the plug-in classification rule because it is actually formed from the training sample. The ER is the performance measure to the PBDF similar as the mean squared prediction error (MSPE) is the performance measure to the plug-in kriging predictor (see Diggle et al. [14]).

## The risk of classification for $m = 2$

---

The two-class case is a special case. The most of the author's publications ([A1]-[A3], [A7]-[A13], [A15], [A16]) deal with the two-class case, that is the reason this case is analyzed separately.

Suppose we have a two-class case ( $m = 2$ ). Then the Bayes classification rule is

$$D_t^B(Z_0, \Psi) = \arg \max_{\{l=1,2\}} \{g_l p_l(Z_0 | \mathbf{t}, \Psi)\},$$

where  $g_l = \pi_l(L(l, 3-l) - L(l, l))$ ,  $l = 1, 2$ .

Since replacing the discriminant function by its monotonically increasing function does not influence the decision, it is more convenient to work in terms of log-ratios. The pairwise Bayes discriminant function based on log-ratios will be denoted as  $W_{lk}^B(Z_0, \Psi)$ .

For the two-class case a single discriminant function is required, so we skip the indices from the notation of pairwise Bayes discriminant function, that is,  $W_{12}^B(Z_0, \Psi)$  is replaced by  $W^B(Z_0, \Psi)$ . Then the expression of *Bayes discriminant function* has the form

$$W^B(Z_0, \Psi) = \ln \left( \frac{p_1(Z_0 | \mathbf{t}, \Psi)}{p_2(Z_0 | \mathbf{t}, \Psi)} \right) + \gamma^*, \quad (2.8)$$

here  $\gamma^* = \ln \left( \frac{g_1}{g_2} \right)$ ,  $g_l = \pi_l(L(l, 3-l) - L(l, l))$ ,  $l = 1, 2$ . According to (2.8) the FO  $Z_0$ , given  $\mathbf{T} = \mathbf{t}$ , is allocated to the population  $\Omega_1$  if  $W^B(Z_0, \Psi) \geq 0$ , and to the population  $\Omega_2$  otherwise.

Then the *Bayes risk* could be evaluated by

$$R_0^B(\Psi) = \sum_{l=1}^2 \sum_{k=1}^2 \pi_l L(l, k) P_{lk}, \quad (2.9)$$

where for  $l, k = 1, 2$ ,  $P_{lk} = P_l((-1)^k W^B(Z_0, \Psi) < 0)$ . Here the probability measure  $P_l$  is based on the conditional distributions of  $Z_0 \in \Omega_l$ .

Since  $P_{lk} + P_{ll} = 1$ , it is easy to reduce (2.9) to a single-sum function

$$R_0^B(\Psi) = \sum_{l=1}^2 (\pi_l L(l, l) + g_l P M_l), \quad (2.10)$$

where  $P M_l = P_l((-1)^l W^B(Z_0, \Psi) > 0)$ .

Then the *actual risk*, given  $\mathbf{T} = \mathbf{t}$ , for PPDF is defined as

$$R_0^B(\hat{\Psi}) = \sum_{l=1}^2 \sum_{k=1}^2 \pi_l L(l, k) \hat{P}_{lk}, \quad (2.11)$$

where for  $l, k = 1, 2$ ,  $\hat{P}_{lk} = P_l((-1)^k W^B(Z_0, \hat{\Psi}) < 0)$ .

## The probability of misclassification

---

We analyze the risk of classification if the *general loss function* is considered, that is,  $L(l, k)$  is a non-negative finite function. If  $L(l, k) = 1 - \delta_{lk}$ , where  $\delta_{lk}$  is the Kronecker delta, the classification risk becomes *the probability of misclassification* or *error rate* (Dučinskas [16]). Such a loss function is called a *zero-one loss function* and is often used if there is no possibility to evaluate the losses more accurately.

For a *zero-one loss* function the pairwise Bayes discriminant functions, defined in (2.5), get the simpler expressions

$$G_{lk}^B(Z_0, \Psi) = \pi_l p_l(Z_0 | \mathbf{t}, \Psi) - \pi_k p_k(Z_0 | \mathbf{t}, \Psi).$$

In the following we will use the equivalent discriminant functions

$$W_{lk}^B(Z_0, \Psi) = \ln \left( \frac{p_l(Z_0 | \mathbf{t}, \Psi)}{p_k(Z_0 | \mathbf{t}, \Psi)} \right) + \gamma_{lk}, \quad (2.12)$$

here  $\gamma_{lk} = \ln \left( \frac{\pi_l}{\pi_k} \right)$ ,  $l, k = 1, \dots, m, k \neq l$ . According to the (2.12) the FO  $Z_0$ , given  $\mathbf{T} = \mathbf{t}$ , is allocated to the population  $\Omega_l$  if  $W_{lk}^B(Z_0, \Psi) \geq 0$ , for all  $l, k = 1, \dots, m, k \neq l$ .

Then the probability of misclassification or error rate due to aforementioned Bayes classification rule (2.3) is (see Anderson [3])

$$P_0^B(\Psi) = 1 - \sum_{l=1}^m \pi_l PC_l,$$

where  $PC_l = P_l(W_{lk}^B(Z_0, \Psi) \geq 0, l = 1, \dots, m, l \neq k)$  corresponds to the conditional probability of correct classification of the observation  $Z_0 \in \Omega_l$ , and  $P_l$  is a probability measure with a probability density function  $p_l(Z_0 | \mathbf{t}, \Psi)$ ,  $l = 1, 2$ .

Plugging the parameters estimators into (2.11) we get PBDF  $W_{lk}^B(Z_0, \hat{\Psi})$ . Then the *actual misclassification probability* or *actual error rate* could be defined.

**Definition 2.3.** The *actual misclassification probability* or *actual error rate* incurred by PBDF is

$$P_0^B(\hat{\Psi}) = 1 - \sum_{l=1}^m \pi_l P \hat{C}_l, \quad (2.13)$$

where  $P \hat{C}_l = P_l(\hat{W}_{lk}^B(Z_0, \hat{\Psi}) \geq 0, l = 1, \dots, m, l \neq k)$ .

**Definition 2.4.** The expectation of the *actual error rate* with respect to the distribution of  $\mathbf{T}$  is called the *expected error rate (EER)* and will be denoted as  $EER = E_T(P_0^B(\Psi))$ .

The probability of misclassification for  $m = 2$

---

For the *two-class case* with *zero-one loss function* the discriminant function and misclassification probability are of the following form

$$W^B(Z_0, \Psi) = \ln\left(\frac{p_1(Z_0|\mathbf{t}, \Psi)}{p_2(Z_0|\mathbf{t}, \Psi)}\right) + \gamma, \quad (2.14)$$

where  $\gamma = \ln(\pi_1/\pi_2)$ .

$$P_0^B(\Psi) = \sum_{l=1}^2 \pi_l P_l, \quad (2.15)$$

where for  $l = 1, 2, P_l((-1)^l W^B(Z_0, \Psi) \geq 0)$ .

The further theoretical results will be based on the following assumption:

**(A3)** The mean models in the populations  $\Omega_l, l = 1, \dots, m$ , are different parametric models  $\boldsymbol{\mu}_l(\mathbf{s}) = \mathbf{x}'(\mathbf{s})\boldsymbol{\beta}_l$ , defined in the Section 1.1.

## 2.2. Classification of GGRF observation

In this section we solve the classification problem of geostatistical Gaussian random field (GGRF) observation. We use the plug-in Bayes discriminant function and derive the actual classification risk and the approximation of

the expected risk for the proposed classifier. We assume a complete parametric uncertainty case, where all mean parameters and all covariance parameters are unknown and are estimated using the ML method. These results are the extension to the ones published in the papers of Dučinskas [18], [19] where the factorized nuggetless covariance function was considered and the only one covariance parameter  $\sigma^2$  was assumed to be unknown. In this section we also present the closed-form expression of AER for geometric anisotropic exponential covariance model.

## 2.2.1. Univariate case

### Two-class case

---

We will initially focus on the univariate two-class case for GGRF. The main purpose is to assign the FO  $Z_0$  to one of two populations  $\Omega_1$  or  $\Omega_2$ . Under the assumption **(A3)** the model of observation  $Z(\mathbf{s})$  in the population  $\Omega_l$  can be written as

$$Z(\mathbf{s}) = \mathbf{x}'(\mathbf{s})\boldsymbol{\beta}_l + \varepsilon(\mathbf{s}), l = 1, 2. \quad (2.16)$$

Then the model of training sample has the following matrix form

$$\mathbf{T} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}, \quad (2.17)$$

where

$$\mathbf{X} = \bigoplus_{l=1}^2 \mathbf{X}_l \quad (2.18)$$

is a  $n \times 2q$  design matrix composed by a direct sum of  $\mathbf{X}_l$ , the  $n_l \times q$  matrices of regressors for training samples  $\mathbf{T}_l$ ,  $l = 1, 2$  and  $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2)'$  is a  $2q \times 1$  vector of parameters.  $\mathbf{E}$  is the  $n \times 1$  vector of random errors that has multivariate Gaussian distribution  $N_n(\mathbf{0}, \boldsymbol{\Sigma})$ .

Let  $C(\boldsymbol{\theta})$  and  $c_0(\boldsymbol{\theta})$  denote the covariance functions between the components of the training sample  $\mathbf{T}$ , and between  $\mathbf{T}$  and  $Z_0$ , respectively. Recall that the covariance matrix between the components of  $\mathbf{T}$  was denoted by  $\boldsymbol{\Sigma}$ , then let  $\mathbf{c}_0$  represent the vector of covariance between  $\mathbf{T}$  and  $Z_0$ .

The training sample  $\mathbf{T}$  has a multivariate Gaussian distribution



$$\mathbf{T} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}).$$

Using the properties of Gaussian distribution we get the conditional probability density function of  $Z_0$  in the population  $\Omega_l, l = 1, 2$

$$p_l(Z_0|\mathbf{t}) = f(Z_0|\mu_{lt}, \sigma_t^2)$$

with conditional mean and variance

$$\mu_{lt} = E(Z_0|\mathbf{T} = \mathbf{t}; \Omega_l) = \mathbf{x}'_0 \boldsymbol{\beta}_l + \boldsymbol{\alpha}'_0 (\mathbf{t} - \mathbf{X}\boldsymbol{\beta}), l = 1, 2, \quad (2.19)$$

$$\sigma_t^2 = \text{var}(Z_0|\mathbf{T} = \mathbf{t}) = C(\mathbf{0}) - \mathbf{c}'_0 \boldsymbol{\Sigma}^{-1} \mathbf{c}_0, \quad (2.20)$$

where  $\mathbf{x}'_0 = (x_1(\mathbf{s}_0), \dots, x_q(\mathbf{s}_0))$  and  $\boldsymbol{\alpha}'_0 = \mathbf{c}'_0 \boldsymbol{\Sigma}^{-1}$ .

Since  $C(\mathbf{0}) = \sigma^2$ ,  $\mathbf{c}_0 = \sigma^2 \mathbf{r}_0$  and  $\boldsymbol{\Sigma}^{-1} = \sigma^{-2} \boldsymbol{\Gamma}^{-1}$  the conditional variance could be rewritten as

$$\sigma_t^2 = \text{var}(Z_0|\mathbf{T} = \mathbf{t}) = \sigma^2 K, \quad (2.21)$$

where  $K = 1 - \mathbf{r}'_0 \boldsymbol{\Gamma}^{-1} \mathbf{r}_0$ ,  $\boldsymbol{\Gamma}$  is the standardized covariance matrix which was already specified in Section 1.2. Here  $\mathbf{r}_0$  represents the vector of spatial correlations between the training  $\mathbf{T}$  and FO  $Z_0$ .

Then BDF specified in (2.8) is a linear  $Z_0$  function

$$W^B(Z_0, \boldsymbol{\Psi}) = (Z_0 - (\mu_{1t} + \mu_{2t})/2)(\mu_{1t} - \mu_{2t})/\sigma_t^2 + \gamma^*. \quad (2.22)$$

Replacing the conditional mean and conditional variance into (2.22) by the expressions given in (2.19) and (2.21) we get the following formula

$$W^B(Z_0, \boldsymbol{\Psi}) = (Z_0 - \boldsymbol{\alpha}'_0 (\mathbf{t} - \mathbf{X}\boldsymbol{\beta}) - \mathbf{x}'_0 \mathbf{I}_+ \boldsymbol{\beta} / 2) (\mathbf{x}'_0 \mathbf{I}_- \boldsymbol{\beta}) / K \sigma^2 + \gamma^*. \quad (2.23)$$

Here  $\mathbf{x}'_0 \mathbf{I}_+ \boldsymbol{\beta} = \mathbf{x}'_0 (\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2)$  and  $\mathbf{x}'_0 \mathbf{I}_- \boldsymbol{\beta} = \mathbf{x}'_0 (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)$ , where  $\mathbf{I}_+ = (\mathbf{I}_q, \mathbf{I}_q)$  and  $\mathbf{I}_- = (\mathbf{I}_q, -\mathbf{I}_q)$ .

Now, the Bayes risk, associated with this linear BDF specified in (2.23), will be derived but before that it is essential to introduce the Mahalanobis distance. The Mahalanobis distance is important in classification problems because it provides a way to take into account spatial correlations when computing distances between populations. Let

$$d^2 = \frac{(\mu_{1t} - \mu_{2t})^2}{\sigma_t^2} = \frac{(\mu_{1t} - \mu_{2t})^2}{\sigma^2 K}$$

be the squared *conditional Mahalanobis distance* between  $\Omega_1$  and  $\Omega_2$  at the focal location  $\mathbf{s}_0$ .  $\mu_{1t}$  and  $\mu_{2t}$  are the conditional means specified in (2.19), and  $\sigma_t^2$  is the conditional variance specified in (2.21). Then the squared *marginal Mahalanobis distance* is specified by formula

$$\Delta^2 = (\mu_1 - \mu_2)^2 / \sigma^2.$$

Here  $\mu_1$  and  $\mu_2$  are the marginal means.

These distances will be considered as the class separation measures. Using (2.19) and (2.21) it is easy to show that the conditional Mahalanobis distance  $d$  does not depend on the realizations of training sample  $\mathbf{T}$ ; it depends only on the location of training sample elements (training sample configuration)

$$d^2 = \frac{\Delta^2 \sigma^2}{\sigma_t^2} = \frac{\Delta^2}{K}.$$

**Lemma 2.1.** Suppose the assumptions **(A1)**, **(A2)**, and **(A3)** hold. Then the *Bayes risk* associated with BDF (2.23) for the two-class case is

$$R_0^B(\Psi) = \sum_{l=1}^2 \{\pi_l L(l, l) + g_l \Phi(-d/2 + (-1)^l \gamma^*/d)\}. \quad (2.24)$$

**Proof.** Since  $W^B(Z_0, \Psi)$  is a linear function of  $Z_0$  then using the properties of Gaussian distribution the conditional distribution of  $W^B(Z_0, \Psi)$  in the population  $\Omega_l$  is the conditional univariate normal distribution with mean

$$E_l(W^B(Z_0, \Psi)) = (-1)^{l+1} d^2/2 + \gamma^*, l = 1, 2,$$

and variance

$$\text{Var}(W^B(Z_0, \Psi)) = d^2,$$

i.e.,  $W^B(Z_0, \Psi) | \Omega_l \sim N((-1)^{l+1} d^2/2 + \gamma^*, d^2)$ . Then the probabilities of misclassification are

$$PM_1 = P_1(W_0^B(Z_0, \Psi) < 0) = \Phi(-d/2 - \gamma^*/d),$$

$$PM_2 = P_2(W_0^B(Z_0, \Psi) > 0) = \Phi(-d/2 + \gamma^*/d).$$

Here  $\Phi(\cdot)$  is the standard normal distribution function. Replacing these probabilities into (2.10) we obtain the expression of Bayes risk

$$R_0^B(\Psi) = \sum_{l=1}^2 \{\pi_l L(l, l) + g_l \Phi(-d/2 + (-1)^l \gamma^*/d)\}.$$

In order to get the plug-in BDF (PBDF) we have to obtain the estimators of conditional mean (2.19) and variance (2.21). For the case of complete parametric uncertainty the estimators are the following

$$\hat{\mu}_{lt} = \mathbf{x}'_0 \hat{\boldsymbol{\beta}}_l + \hat{\boldsymbol{\alpha}}'_0 (\mathbf{t} - \mathbf{X} \hat{\boldsymbol{\beta}}), \quad (2.25)$$

$$\hat{\sigma}_t^2 = \hat{\sigma}^2 \hat{K}. \quad (2.26)$$

Here  $\hat{\boldsymbol{\beta}}$  and  $\hat{\boldsymbol{\theta}}$  denote the estimators of  $\boldsymbol{\beta}$  and  $\boldsymbol{\theta}$ , thus  $\hat{\Psi} = (\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\theta}}')$  denotes the combined vector of population parameters estimators. Replacing the obtained estimators into (2.23) we get the PBDF

$$W^B(Z_0, \hat{\Psi}) = (Z_0 - \hat{\boldsymbol{\alpha}}'_0 (\mathbf{t} - \mathbf{X} \hat{\boldsymbol{\beta}}) - \mathbf{x}'_0 \mathbf{I}_+ \hat{\boldsymbol{\beta}}/2) (\mathbf{x}'_0 \mathbf{I}_- \hat{\boldsymbol{\beta}}) / \hat{K} \hat{\sigma}^2 + \gamma^*. \quad (2.27)$$

**Lemma 2.2.** The *actual risk* for  $W^B(Z_0, \hat{\Psi})$  specified in (2.27) is defined as

$$R_0^B(\hat{\Psi}) = \sum_{l=1}^2 \{\pi_l L(l, l) + g_l \Phi(\hat{Q}_l)\}. \quad (2.28)$$

Here

$$\hat{Q}_l = (-1)^l ((a_l - \hat{b}) \text{sgn}(\mathbf{x}'_0 \mathbf{I}_- \hat{\boldsymbol{\beta}}) / \sigma_t + \gamma^* \hat{\sigma}_t^2 / |\mathbf{x}'_0 \mathbf{I}_- \hat{\boldsymbol{\beta}}| \sigma_t), \quad (2.29)$$

$$a_l = \mathbf{x}'_0 \boldsymbol{\beta}_l + \boldsymbol{\alpha}'_0 (\mathbf{t} - \mathbf{X} \boldsymbol{\beta}), \quad l = 1, 2, \quad (2.30)$$

$$\hat{b} = \hat{\boldsymbol{\alpha}}'_0 (\mathbf{t} - \mathbf{X} \hat{\boldsymbol{\beta}}) + \mathbf{x}'_0 \mathbf{I}_+ \hat{\boldsymbol{\beta}}/2. \quad (2.31)$$

$\text{sgn}(\cdot)$  is the Signum function.

**Proof.** In the population  $\Omega_l$  the conditional distribution of  $W^B(Z_0, \hat{\Psi})$ , given  $\mathbf{T} = \mathbf{t}$ , is Gaussian

$$W^B(Z_0, \hat{\Psi}) | \Omega_l \sim N(\mu_l^W, \sigma_W^2).$$

The expressions of conditional mean and variance could be obtained by using the properties of Gaussian distribution. Since  $W^B(Z_0, \hat{\Psi})$  is a linear  $Z_0$  function, it could be rearranged into the form  $Z_0 \text{const}_1 + \text{const}_2$ . Then, according to the rule  $\mu_l^W = \text{const}_1 E(Z_0) + \text{const}_2$ , we get

$$\begin{aligned} \mu_l^W &= (\mathbf{x}'_0 \mathbf{I}_- \hat{\boldsymbol{\beta}}) (\mathbf{x}'_0 \boldsymbol{\beta}_l + \boldsymbol{\alpha}'_0 (\mathbf{t} - \mathbf{X} \boldsymbol{\beta})) / \hat{\sigma}_t^2 - \\ &- (\hat{\boldsymbol{\alpha}}'_0 (\mathbf{t} - \mathbf{X} \hat{\boldsymbol{\beta}}) + \mathbf{x}'_0 \mathbf{I}_+ \hat{\boldsymbol{\beta}}/2) (\mathbf{x}'_0 \mathbf{I}_- \hat{\boldsymbol{\beta}}) / \hat{\sigma}_t^2 + \gamma^*. \end{aligned}$$

After doing some algebra we obtain the following expression

$$\mu_l^W = (a_l - \hat{b})(\mathbf{x}'_0 \mathbf{I} - \hat{\boldsymbol{\beta}}) / \hat{\sigma}_t^2 + \gamma^*, l = 1, 2.$$

Similarly we get the expression of conditional variance

$$\sigma_W^2 = \frac{(\mathbf{x}'_0 \mathbf{I} - \hat{\boldsymbol{\beta}})^2 \sigma_t^2}{\hat{\sigma}_t^4}.$$

Then the probabilities of misclassification are

$$\begin{aligned} PM_1 &= P_1(\widehat{W}^B(Z_0) < 0) = \Phi\left(-\frac{\mu_1^W}{\sigma_W}\right) = \\ &= \Phi\left(-\frac{(a_1 - \hat{b})\text{sgn}(\mathbf{x}'_0 \mathbf{I} - \hat{\boldsymbol{\beta}})}{\sigma_t} - \frac{\gamma^* \hat{\sigma}_t^2}{|\mathbf{x}'_0 \mathbf{I} - \hat{\boldsymbol{\beta}}| \sigma_t}\right), \end{aligned}$$

$$\begin{aligned} PM_2 &= P_2(\widehat{W}^B(Z_0) \geq 0) = \Phi\left(\frac{\mu_2^W}{\sigma_W}\right) = \\ &= \Phi\left(\frac{(a_2 - \hat{b})\text{sgn}(\mathbf{x}'_0 \mathbf{I} - \hat{\boldsymbol{\beta}})}{\sigma_t} + \frac{\gamma^* \hat{\sigma}_t^2}{|\mathbf{x}'_0 \mathbf{I} - \hat{\boldsymbol{\beta}}| \sigma_t}\right). \end{aligned}$$

Using (2.10) we complete the proof of lemma.

Then the *expected risk* could be evaluated by

$$ER = E_T(R_0^B(\hat{\boldsymbol{\Psi}})) = E_T\{\sum_{l=1}^2(\pi_l L(l, l) + g_l \Phi(\hat{Q}_l))\}. \quad (2.32)$$

## Asymptotic expansion of the expected risk

---

As it was already mentioned, the *actual risk* and the *expected risk* are usually considered as performance measures for the plug-in BDF. Contrary to the *actual risk*, the expressions for the *expected risk* often are very cumbersome. This makes it difficult to build any qualitative conclusions. Therefore, the asymptotic approximations of the *expected risk* are especially important.

In the thesis the *approximation of expected risk* (AER) based on the asymptotic expansion is proposed. We focus on the maximum likelihood estimators, since the inverse of the information matrix associated with the likelihood function of training sample well approximates the covariance

matrix of these estimators. The asymptotic properties of ML estimators showed by Mardia and Marshall [46] under increasing domain asymptotic framework and subject to some regularity conditions are essentially exploited (see Section 1.2).

Consider the *two-class* case of *complete parametric uncertainty*, where  $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2)$  and  $\boldsymbol{\theta} = (\tau^2, \sigma^2, \alpha, \lambda, \varphi)'$  are unknown.

Let  $\mathbf{R}_\beta^{(k)}$ ,  $\mathbf{R}_\theta^{(k)}$ ,  $k = 1, 2$  denote the  $k$ -th order derivatives of  $R_0^B(\hat{\Psi})$  with respect to  $\hat{\boldsymbol{\beta}}$  and  $\hat{\boldsymbol{\theta}}$  evaluated at the point  $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}, \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ , and let  $\mathbf{R}_{\beta\theta}^{(2)}$  denote the matrix of the second order derivatives of  $R_0^B(\hat{\Psi})$  with respect to  $\hat{\boldsymbol{\beta}}$  and  $\hat{\boldsymbol{\theta}}$  evaluated at the point  $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}, \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ .

Also the following assumption is made:

**(A4)** The training sample  $\mathbf{T}$  and estimator  $\hat{\boldsymbol{\theta}}$  are statistically independent.

The restrictive assumption **(A4)** is exploited intensively by many authors (see Zimmerman [82], Zhu and Stein [79]) since Abt [1] showed that finer approximations of MSPE considering the correlation between  $\mathbf{T}$  and  $\hat{\boldsymbol{\theta}}$  do not give better results.

Let  $\mathbf{A}_\theta = \partial \hat{\boldsymbol{\alpha}}_0 / \partial \hat{\boldsymbol{\theta}}'$  be the  $n \times p$  matrix of the first order partial derivatives evaluated at the point  $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$  and let  $\varphi(\cdot)$  be the standard normal distribution density function. Denote by  $\mathbf{s}_\theta = (\hat{\sigma}_t^2)_\theta^{(1)}$  the vector of the first order partial derivatives of  $\hat{\sigma}_t^2 = \hat{C}(\mathbf{0}) - \hat{\mathbf{c}}_0' \hat{\boldsymbol{\Sigma}}^{-1} \hat{\mathbf{c}}_0$  evaluated at the point  $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ .

**Theorem 2.1.** Suppose that the FO  $Z_0$  is to be classified by PPDF (2.27) and let the Mardia and Marshall conditions (Theorem 1.1) and the assumption **(A4)** hold. Then the approximation of ER is

$$AER = \sum_{l=1}^2 g_l \Phi(Q_l) + g_1 \varphi(Q_1) d(K_\beta + K_\theta) / 2\sigma_t^2, \quad (2.33)$$

$$K_\beta = \boldsymbol{\Lambda}' \mathbf{J}_\beta^{-1} \boldsymbol{\Lambda}, \mathbf{J}_\beta = \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X}, \quad (2.34)$$

$$\boldsymbol{\Lambda}' = \boldsymbol{\alpha}'_0 \mathbf{X} - \mathbf{x}'_0 (\mathbf{I}_+ / 2 + \gamma^* \mathbf{I}_- / d^2), \quad (2.35)$$

$$K_\theta = \text{tr}(\boldsymbol{\Sigma} \mathbf{A}_\theta \mathbf{J}_\theta^{-1} \mathbf{A}'_\theta) + (\gamma^*)^2 \mathbf{s}'_\theta \mathbf{J}_\theta^{-1} \mathbf{s}_\theta / d^2 \sigma_t^2. \quad (2.36)$$

**Proof.** Expanding  $R_0^B(\widehat{\Psi})$  in the Taylor series around the point  $\widehat{\beta} = \beta, \widehat{\theta} = \theta$  we have

$$R_0^B(\widehat{\Psi}) = R_0^B(\Psi) + \mathbf{R}_\beta^{(1)} \Delta \widehat{\beta} + \mathbf{R}_\theta^{(1)} \Delta \widehat{\theta} + \frac{1}{2} \left( (\Delta \widehat{\beta})' \mathbf{R}_\beta^{(2)} \Delta \widehat{\beta} + 2(\Delta \widehat{\beta})' \mathbf{R}_{\beta\theta}^{(2)} \Delta \widehat{\theta} + (\Delta \widehat{\theta})' \mathbf{R}_\theta^{(2)} \Delta \widehat{\theta} \right) + R_3, \quad (2.37)$$

where  $\Delta \widehat{\beta} = \widehat{\beta} - \beta, \Delta \widehat{\theta} = \widehat{\theta} - \theta$  and  $R_3$  is the remainder term.

Then we have to find the partial derivatives of  $R_0^B(\widehat{\Psi}) = \sum_{l=1}^2 \{ \pi_l L(l, l) + g_l \Phi(\widehat{Q}_l) \}$  and to evaluate them at the point  $\widehat{\beta} = \beta, \widehat{\theta} = \theta$ .

The partial derivatives of standard normal distribution function with respect to  $\widehat{\beta}$  and  $\widehat{\theta}$  are

$$\frac{\partial \Phi(\widehat{Q}_l)}{\partial \widehat{\beta}} = \varphi(\widehat{Q}_l) \frac{\partial \widehat{Q}_l}{\partial \widehat{\beta}},$$

$$\frac{\partial \Phi(\widehat{Q}_l)}{\partial \widehat{\theta}} = \varphi(\widehat{Q}_l) \frac{\partial \widehat{Q}_l}{\partial \widehat{\theta}}.$$

The partial derivatives of the standard normal distribution density function with respect to  $\widehat{\beta}$  and  $\widehat{\theta}$  are

$$\frac{\partial \varphi(\widehat{Q}_l)}{\partial \widehat{\beta}} = -\varphi(\widehat{Q}_l) \widehat{Q}_l \frac{\partial \widehat{Q}_l}{\partial \widehat{\beta}},$$

$$\frac{\partial \varphi(\widehat{Q}_l)}{\partial \widehat{\theta}} = -\varphi(\widehat{Q}_l) \widehat{Q}_l \frac{\partial \widehat{Q}_l}{\partial \widehat{\theta}}.$$

Replacing the estimators  $\widehat{\beta}$  and  $\widehat{\theta}$  by the true values into  $\varphi(\widehat{Q}_l)$  it is easy to show that  $g_1 \varphi(Q_1) = g_2 \varphi(Q_2)$ , where  $Q_l = -d/2 + (-1)^l \gamma^*/d, l = 1, 2$ . Then the partial derivatives of  $R_0^B(\widehat{\Psi})$  evaluated at the point  $\widehat{\beta} = \beta, \widehat{\theta} = \theta$  are

$$\mathbf{R}_\beta^{(1)} = g_1 \varphi(Q_1) \sum_{l=1}^2 \mathbf{Q}_{l\beta}^{(1)}, \quad (2.38)$$

$$\mathbf{R}_\theta^{(1)} = g_1 \varphi(Q_1) \sum_{l=1}^2 \mathbf{Q}_{l\theta}^{(1)}, \quad (2.39)$$

$$\mathbf{R}_\beta^{(2)} = g_1 \varphi(Q_1) \sum_{l=1}^2 \left( \mathbf{Q}_{l\beta}^{(2)} - Q_l \mathbf{Q}_{l\beta}^{(1)} \left( \mathbf{Q}_{l\beta}^{(1)} \right)' \right), \quad (2.40)$$

$$\mathbf{R}_\theta^{(2)} = g_1 \varphi(Q_1) \sum_{l=1}^2 \left( \mathbf{Q}_{l\theta}^{(2)} - Q_l \mathbf{Q}_{l\theta}^{(1)} \left( \mathbf{Q}_{l\theta}^{(1)} \right)' \right). \quad (2.41)$$

$$\mathbf{R}_{\beta\theta}^{(2)} = g_1 \varphi(Q_1) \sum_{l=1}^2 \left( \mathbf{Q}_{l\beta\theta}^{(2)} - Q_l \mathbf{Q}_{l\theta}^{(1)} \left( \mathbf{Q}_{l\beta}^{(1)} \right)' \right). \quad (2.42)$$

Here  $\mathbf{Q}_{l\beta}^{(k)}$  represents the  $k$ -th order partial derivatives of  $\hat{Q}_l$  with respect to  $\hat{\beta}$  at the point  $\hat{\beta} = \beta$ ,  $\hat{\theta} = \theta$ , and  $\mathbf{Q}_{l\theta}^{(k)}$  represents the  $k$ -th order partial derivatives of  $\hat{Q}_l$  with respect to  $\hat{\theta}$  at the point  $\hat{\beta} = \beta$ ,  $\hat{\theta} = \theta$  and  $\mathbf{Q}_{l\beta\theta}^{(2)}$  represents the second order partial derivative of  $\hat{Q}_l$  with respect to  $\hat{\beta}$  and  $\hat{\theta}$  at the point  $\hat{\beta} = \beta$ ,  $\hat{\theta} = \theta$ . To obtain these derivatives at first we differentiate (2.29)-(2.31) with respect to  $\hat{\beta}$  and  $\hat{\theta}$

$$\frac{\partial \hat{Q}_l}{\partial \hat{\beta}} = (-1)^l \left\{ \frac{\partial(\hat{\alpha}'_0(\mathbf{t} - \mathbf{X}\hat{\beta}) - \mathbf{x}'_0 \mathbf{I}_+ \hat{\beta}/2)}{\partial \hat{\beta}} - \frac{\gamma^* \hat{\sigma}_t^2}{(\mathbf{x}'_0 \mathbf{I}_- \hat{\beta})^2} \frac{\partial(\mathbf{x}'_0 \mathbf{I}_- \hat{\beta})}{\partial \hat{\beta}} \right\} / \sigma_t =$$

$$= (-1)^l \{ \hat{\alpha}'_0 \mathbf{X} - \mathbf{x}_0 \mathbf{I}_+ / 2 - \gamma^* \mathbf{x}_0 \mathbf{I}_- / \hat{d}^2 \} / \sigma_t,$$

$$\frac{\partial \hat{Q}_l}{\partial \hat{\theta}} = (-1)^l \left\{ \frac{-\partial(\hat{\alpha}'_0(\mathbf{t} - \mathbf{X}\hat{\beta}) + \mathbf{x}'_0 \mathbf{I}_+ \hat{\beta}/2)}{\partial \hat{\theta}} + \frac{\gamma^*}{|\mathbf{x}'_0 \mathbf{I}_- \hat{\beta}|} \frac{\partial(\hat{\sigma}_t^2)}{\partial \hat{\theta}} \right\} / \sigma_t =$$

$$= (-1)^l \left\{ -\frac{\partial(\hat{\alpha}'_0)}{\partial \hat{\theta}} (\mathbf{t} - \mathbf{X}\hat{\beta}) + \frac{\gamma^*}{|\mathbf{x}'_0 \mathbf{I}_- \hat{\beta}|} \frac{\partial(\hat{\sigma}_t^2)}{\partial \hat{\theta}} \right\} / \sigma_t.$$

Notice that the expression  $(\mathbf{x}'_0 \mathbf{I}_- \hat{\beta})^2 / \hat{\sigma}_t^2$  corresponds to the estimator of squared conditional Mahalanobis distance, that is,  $\hat{d}^2 = (\mathbf{x}'_0 \mathbf{I}_- \hat{\beta})^2 / \hat{\sigma}_t^2$ .

Evaluating derivatives  $\partial \hat{Q}_l / \partial \hat{\beta}$  and  $\partial \hat{Q}_l / \partial \hat{\theta}$  at the point  $\hat{\beta} = \beta$ ,  $\hat{\theta} = \theta$  leads to the following expressions

$$\mathbf{Q}_{l\beta}^{(1)} = (-1)^l \Lambda / \sigma_t,$$

$$\mathbf{Q}_{l\theta}^{(1)} = (-1)^l (-\mathbf{A}'_\theta (\mathbf{t} - \mathbf{X}\beta) + \gamma^* \mathbf{s}_\theta / (d\sigma_t)) / \sigma_t.$$

It is easy to notice that

$$\sum_{l=1}^2 \mathbf{Q}_{l\beta}^{(2)} = \mathbf{0} \text{ and } \sum_{l=1}^2 \mathbf{Q}_{l\theta}^{(2)} = \mathbf{0}.$$

The application of the above formulas to (2.38)-(2.42) yields

$$\mathbf{R}_\beta^{(1)} = \mathbf{0}, \mathbf{R}_\theta^{(1)} = \mathbf{0}, \quad (2.43)$$

$$\mathbf{R}_\beta^{(2)} = g_1 d \varphi(-d/2 - \gamma^*/d) \mathbf{\Lambda} \mathbf{\Lambda}' / \sigma_t^2, \quad (2.44)$$

$$\mathbf{R}_\theta^{(2)} = \frac{g_1 d \varphi(Q_1)}{\sigma_t^2} \left( -\mathbf{A}'_\theta (\mathbf{t} - \mathbf{X}\boldsymbol{\beta}) + \frac{\gamma^* \mathbf{s}_\theta}{d \sigma_t} \right) \left( -\mathbf{A}'_\theta (\mathbf{t} - \mathbf{X}\boldsymbol{\beta}) + \frac{\gamma^* \mathbf{s}_\theta}{d \sigma_t} \right)'. \quad (2.45)$$

It is easy to show that all elements of the matrix  $\mathbf{R}_{\beta\theta}^{(2)}$  are finite, and under the assumption **(A4)** we get  $E_T(\Delta\widehat{\boldsymbol{\beta}}\Delta\widehat{\boldsymbol{\theta}}) = 0$ . Replacing  $E_T(\Delta\widehat{\boldsymbol{\beta}}\Delta\widehat{\boldsymbol{\beta}}')$  and  $E_T(\Delta\widehat{\boldsymbol{\theta}}\Delta\widehat{\boldsymbol{\theta}}')$  by their asymptotic approximations,  $\mathbf{J}_\beta^{-1}$  and  $\mathbf{J}_\theta^{-1}$ , we get the following approximations

$$E_T \left( (\Delta\widehat{\boldsymbol{\beta}})' \mathbf{R}_\beta^{(2)} (\Delta\widehat{\boldsymbol{\beta}}) \right) \cong \frac{g_1 d}{\sigma_t^2} \varphi(-d/2 - \gamma^*/d) \mathbf{\Lambda}' \mathbf{J}_\beta^{-1} \mathbf{\Lambda}, \quad (2.46)$$

$$\begin{aligned} E_T \left( (\Delta\widehat{\boldsymbol{\theta}})' \mathbf{R}_\theta^{(2)} (\Delta\widehat{\boldsymbol{\theta}}) \right) &\cong \frac{g_1 d}{\sigma_t^2} \varphi(-d/2 - \gamma^*/d) \times \\ &\times \left( \text{tr}(\boldsymbol{\Sigma} \mathbf{A}_\theta \mathbf{J}_\theta^{-1} \mathbf{A}'_\theta) + \frac{(\gamma^*)^2}{d^2 \sigma_t^2} \mathbf{s}'_\theta \mathbf{J}_\theta^{-1} \mathbf{s}_\theta \right). \end{aligned} \quad (2.47)$$

Then taking the expectation term by term of the right-hand side of (2.37), using (2.24), (2.43)-(2.47) and replacing the moments of estimators by the corresponding moments of asymptotic distributions specified in (1.3)-(1.5) we complete the proof of theorem.

**Remark 2.1.** For the nuggetless factorized covariance matrix  $\boldsymbol{\Sigma} = \sigma^2 \mathbf{R}$  and for  $\boldsymbol{\theta} = \sigma^2$ , which means that the only one covariance parameter (scale parameter or partial sill) is unknown, the approximation of ER specified in (2.33)-(2.36) coincides with the one derived by Dučinskas [18]

$$AER = \sum_{l=1}^2 g_l \Phi(Q_l) +$$

$$+ g_1 \varphi(Q_1) (\mathbf{\Lambda} (\mathbf{X}' \mathbf{R}^{-1} \mathbf{X}) \mathbf{\Lambda}' d / 2K + (\gamma^*)^2 / (n - 2q)d),$$

where  $K = 1 - \mathbf{r}'_0 \mathbf{R}^{-1} \mathbf{r}_0$ .

**Remark 2.2.** For the case  $\boldsymbol{\theta} = \lambda$  and  $\varphi = \pi/2$  the asymptotic approximation of expected risk is presented in [A10] and is of the following form



$$AER = R_0^B(\Psi) + g_1 \varphi(-d/2 - \gamma^*/d) d(K_\beta + K_\lambda) / 2\sigma_t^2,$$

$$K_\beta = \Lambda' \mathbf{J}_\beta^{-1} \Lambda, \mathbf{J}_\beta = \mathbf{X}' \Sigma^{-1} \mathbf{X},$$

$$\Lambda' = \alpha'_0 \mathbf{X} - \mathbf{x}'_0 (\mathbf{I}_+/2 + \gamma^* \mathbf{I}_-/d^2),$$

$$K_\lambda = \text{tr}(\Sigma \mathbf{A}_\lambda \mathbf{J}_\lambda^{-1} \mathbf{A}'_\lambda) + (\gamma^*)^2 \mathbf{s}'_\lambda \mathbf{J}_\lambda^{-1} \mathbf{s}_\lambda / d^2 \sigma_t^2,$$

$$\mathbf{A}_\lambda = \partial \hat{\alpha}_0 / \partial \hat{\lambda}.$$

## The closed-form expression of AER for exponential covariance model

---

In order to apply the AER formula (2.33)-(2.36) in practice there might arise difficulties evaluating the term  $K_\theta$ . This term includes the partial matrix and vector derivatives and software may fail while doing these calculations. Therefore, having the closed-form expression of AER is significant. For this reason, we need to find the closed-form expressions of  $\mathbf{A}_\theta$ ,  $\mathbf{J}_\theta^{-1}$  and  $\mathbf{s}_\theta$ .

Suppose we have the geometrically anisotropic exponential covariance function. Recall that the vector of unknown geometrically anisotropic covariance parameters is  $\boldsymbol{\theta} = (\tau^2, \sigma^2, \alpha, \lambda, \varphi)'$  and then  $\hat{\boldsymbol{\theta}} = (\hat{\tau}^2, \hat{\sigma}^2, \hat{\alpha}, \hat{\lambda}, \hat{\varphi})'$  represents the vector of parameters estimators.

The matrix  $\mathbf{A}_\theta = \partial \hat{\alpha}_0 / \partial \hat{\boldsymbol{\theta}}'$  is a  $n \times 5$  matrix composed by the first order partial derivatives of  $\hat{\alpha}_0 = \hat{\Sigma}^{-1} \hat{\mathbf{c}}_0$  with respect to the components of  $\hat{\boldsymbol{\theta}}$  and evaluated at the point  $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ . Let  $(\mathbf{c}_0)_{\theta_i}$  and  $\Sigma_{\theta_i}$  denote the first order partial derivatives of  $\hat{\mathbf{c}}_0$  and  $\hat{\Sigma}$  with respect to  $\hat{\theta}_i$  and evaluated at  $\hat{\theta}_i = \theta_i$ . Then the  $i$ -th column of matrix  $\mathbf{A}_\theta$  could be found as derivative of  $\hat{\alpha}_0$  with respect to the  $i$ -th covariance parameter ( $\hat{\theta}_i = \theta_i$ )

$$\partial \hat{\alpha}_0 / \partial \hat{\theta}_i = -\Sigma^{-1} \Sigma_{\theta_i} \alpha_0 + \Sigma^{-1} (\mathbf{c}_0)_{\theta_i}, i = 1, \dots, 5. \quad (2.48)$$

To get the closed-form expression of matrix  $\mathbf{A}_\theta$  we have to find the first order partial derivatives of  $\hat{\mathbf{c}}_0$  and  $\hat{\Sigma}$  for a selected parametric covariance model. We chose the geometrically anisotropic covariance function defined in (1.7). Then the first order partial derivatives are

$$(\mathbf{c}_0)_{\tau^2} = \mathbf{0}_n, \quad (2.49)$$

$$(\mathbf{c}_0)_{\sigma^2} = \mathbf{r}_0, \quad (2.50)$$

$$(\mathbf{c}_0)_\alpha = \frac{\sigma^2}{\alpha^2} \mathbf{r}_0 \circ \mathbf{H}_{0\alpha}, \quad (2.51)$$

$$(\mathbf{c}_0)_\lambda = -\frac{\sigma^2 \lambda}{\alpha} \mathbf{r}_0 \circ \mathbf{H}_{0\lambda}, \quad (2.52)$$

$$(\mathbf{c}_0)_\varphi = \frac{\sigma^2(\lambda^2-1)}{\alpha} \mathbf{r}_0 \circ \mathbf{H}_{0\varphi}, \quad (2.53)$$

where  $\mathbf{H}_{0\alpha}$ ,  $\mathbf{H}_{0\lambda}$  and  $\mathbf{H}_{0\varphi}$  are  $n$ -dimensional vectors with elements

$$(\mathbf{H}_{0\alpha})_j = \sqrt{(h_u^{0j})^2 + \lambda^2 (h_v^{0j})^2},$$

$$(\mathbf{H}_{0\lambda})_j = (h_v^{0j})^2 / \sqrt{(h_u^{0j})^2 + \lambda^2 (h_v^{0j})^2},$$

$$(\mathbf{H}_{0\varphi})_j = h_u^{0j} h_v^{0j} / \sqrt{(h_u^{0j})^2 + \lambda^2 (h_v^{0j})^2},$$

$$h_u^{0j} = (h_x^{0j} \cos \varphi + h_y^{0j} \sin \varphi) / \alpha, \quad h_v^{0j} = (-h_x^{0j} \sin \varphi + h_y^{0j} \cos \varphi) / \alpha,$$

$h_x^{0j} = x_0 - x_j$ ,  $h_y^{0j} = y_0 - y_j$ ,  $x_0$  and  $y_0$  are the coordinates of  $\mathbf{s}_0$  and  $x_j$  and  $y_j$ ,  $j = 1, \dots, n$ , represent the coordinates of the  $j$ -th  $\mathbf{T}$  component.

The first order partial derivatives of covariance matrix with respect to the parameter  $\hat{\theta}_i$  evaluated at  $\hat{\theta}_i = \theta_i$  coincide with the ones defined in Lemma 1.1, that is,  $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_{\theta_i}$ ,  $i = 1, \dots, 5$ . Then replacing  $(\mathbf{c}_0)_{\theta_i}$  by (2.49)-(2.53) and  $\boldsymbol{\Sigma}_{\theta_i}$  by (1.8)-(1.15) into (2.48) we get the columns of matrix  $\mathbf{A}_\theta$ :

$$\partial \hat{\boldsymbol{\alpha}}_0 / \partial \hat{\theta}_1 = \partial \hat{\boldsymbol{\alpha}}_0 / \partial \hat{t}^2 = -\boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha}_0,$$

$$\partial \hat{\boldsymbol{\alpha}}_0 / \partial \hat{\theta}_2 = \partial \hat{\boldsymbol{\alpha}}_0 / \partial \hat{\sigma}^2 = -\boldsymbol{\Sigma}^{-1} \mathbf{R} \boldsymbol{\alpha}_0 + \boldsymbol{\Sigma}^{-1} \mathbf{r}_0,$$

$$\partial \hat{\boldsymbol{\alpha}}_0 / \partial \hat{\theta}_3 = \partial \hat{\boldsymbol{\alpha}}_0 / \partial \hat{\alpha} = -\frac{\sigma^2}{\alpha^2} \boldsymbol{\Sigma}^{-1} (\mathbf{R} \circ \mathbf{H}_\alpha \boldsymbol{\alpha}_0 - \mathbf{r}_0 \circ \mathbf{H}_{0\alpha}),$$

$$\partial \hat{\boldsymbol{\alpha}}_0 / \partial \hat{\theta}_4 = \partial \hat{\boldsymbol{\alpha}}_0 / \partial \hat{\lambda} = \frac{\lambda \sigma^2}{\alpha} \boldsymbol{\Sigma}^{-1} (\mathbf{R} \circ \mathbf{H}_\lambda \boldsymbol{\alpha}_0 - \mathbf{r}_0 \circ \mathbf{H}_{0\lambda}),$$

$$\partial \hat{\boldsymbol{\alpha}}_0 / \partial \hat{\theta}_5 = \partial \hat{\boldsymbol{\alpha}}_0 / \partial \hat{\varphi} = \frac{\sigma^2(\lambda^2-1)}{\alpha} \boldsymbol{\Sigma}^{-1} (\mathbf{R} \circ \mathbf{H}_\varphi \boldsymbol{\alpha}_0 - \mathbf{r}_0 \circ \mathbf{H}_{0\varphi}).$$

To get  $\mathbf{s}_\theta$  we have to differentiate  $\hat{\sigma}_t^2$  with respect to  $\hat{\boldsymbol{\theta}}$ , and evaluate it at the point  $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ . Then

$$\mathbf{s}_\theta = \sigma_\theta^2 + \boldsymbol{\alpha}'_0 \boldsymbol{\Sigma}_\theta \boldsymbol{\alpha}_0 - 2\boldsymbol{\alpha}'_0(\mathbf{c}_0)_\theta.$$

Using (2.49)-(2.53) and (1.8)-(1.15) we obtain the elements of  $\mathbf{s}_\theta$ :

$$\begin{aligned} s_{\tau^2} &= \boldsymbol{\alpha}'_0 \boldsymbol{\alpha}_0, \\ s_{\sigma^2} &= 1 + \boldsymbol{\alpha}'_0 \mathbf{R} \boldsymbol{\alpha}_0 - 2\boldsymbol{\alpha}'_0 \mathbf{r}_0, \\ s_\alpha &= -\frac{\sigma^2}{\alpha^2} (\boldsymbol{\alpha}'_0 (\mathbf{R} \circ \mathbf{H}_\alpha) \boldsymbol{\alpha}_0 - 2\boldsymbol{\alpha}'_0 (\mathbf{r}_0 \circ \mathbf{H}_{0\alpha})), \\ s_\lambda &= -\frac{\lambda \sigma^2}{\alpha} (\boldsymbol{\alpha}'_0 (\mathbf{R} \circ \mathbf{H}_\lambda) \boldsymbol{\alpha}_0 - 2\boldsymbol{\alpha}'_0 (\mathbf{r}_0 \circ \mathbf{H}_{0\lambda})), \\ s_\varphi &= \frac{\sigma^2(\lambda^2 - 1)}{\alpha} (\boldsymbol{\alpha}'_0 (\mathbf{R} \circ \mathbf{H}_\varphi) \boldsymbol{\alpha}_0 - 2\boldsymbol{\alpha}'_0 (\mathbf{r}_0 \circ \mathbf{H}_{0\varphi})). \end{aligned}$$

**Remark 2.3.** The special case when the only one covariance parameter is unknown,  $\boldsymbol{\theta} = \lambda$ , and the angle of anisotropy is set to  $\varphi = \pi/2$  is presented in Drežiženė [A10].

## Multiclass case

---

Now consider a *multiclass case* ( $m > 2$ ) with *zero-one loss* function. The main goal is to solve the problem of classification of FO  $Z_0$ , given training sample  $\mathbf{T}$  (described in Section 2.1), into one of several populations. The model of training sample is specified in (2.17), (2.18). We consider the case of nuggetless covariance function with unknown parameter  $\sigma^2$  and known spatial correlation function.

The pairwise BDF specified in (2.12) in this case has the expression

$$W_{lk}^B(Z_0, \boldsymbol{\Psi}) = (Z_0 - (\mu_{lt} + \mu_{kt})/2)(\mu_{lt} - \mu_{kt})/\sigma_t^2 + \gamma_{lk}, \quad (2.54)$$

where  $\mu_{lt}$  and  $\sigma_t^2$  represent the conditional mean and conditional variance, respectively. Then the Bayes rule for  $l, k = 1, \dots, m, k \neq l$  is given by:

$$\text{Classify FO } Z_0 \text{ to the population } \Omega_l \text{ if } W_{lk}^B(Z_0, \boldsymbol{\Psi}) \geq 0. \quad (2.55)$$

Replacing the conditional mean and variance into (2.54) by the expressions given in (2.19) and (2.21) we get the following formula for the pairwise BDF

$$W_{lk}^B(Z_0, \Psi) = (Z_0 - \alpha'_0(\mathbf{t} - \mathbf{X}\beta) - \mu_{lk})d_{lk}/\sigma\sqrt{K} + \gamma_{lk}, \quad (2.56)$$

where  $\mu_{lk} = \mathbf{x}'_0(\beta_l + \beta_k)/2$  and  $K = 1 - \mathbf{r}'_0\mathbf{R}^{-1}\mathbf{r}_0$ . Here  $d_{lk}$  stands for conditional Mahalanobis distance

$$d_{lk} = \frac{\mu_{lt} - \mu_{kt}}{\sigma_t} = \frac{\mu_{lt} - \mu_{kt}}{\sigma\sqrt{K}}, l, k = 1, \dots, m, k \neq l.$$

$d_{lk}$  could be expressed in terms of marginal Mahalanobis distance,  $\Delta_{lk}$

$$d_{lk} = \Delta_{lk}/\sqrt{K},$$

where  $\Delta_{lk} = \frac{\mu_l - \mu_k}{\sigma}$ ,  $\mu_l = \mathbf{x}'(\mathbf{s})\beta_l$ ,  $l, k = 1, \dots, m, k \neq l$ .

Now, the probability of misclassification (Bayes misclassification probability) associated with the Bayes classification rule for  $m > 2$  will be derived. Let  $\varphi(x; \mu, \sigma^2)$  be the probability density function of the normal distribution with mean  $\mu$  and variance  $\sigma^2$  and set  $\varphi(x; 0, 1) = \varphi(x)$ .

**Lemma 2.3.** The probability of misclassification for  $m > 2$  due to the Bayes classification rule specified in (2.55) is

$$P_0^B(\Psi) = 1 - \sum_{l=1}^m \pi_l \int_{B_l} \varphi(u) du, \quad (2.57)$$

where  $B_l = \{u: u \in R^1, d_{lk}u + d_{lk}^2/2 + \gamma_{lk} \geq 0; l = 1, \dots, m, k \neq l\}$ .

**Proof.** According to the definition the probability of misclassification due to the Bayes classification rule is

$$P_0^B(\Psi) = 1 - \sum_{l=1}^m \pi_l PC_l, \quad (2.58)$$

where for  $l, k = 1, \dots, m, l \neq k$ ,  $PC_l = P_l(W_{lk}^B(Z_0, \Psi) \geq 0)$  is the probability of correct classification of  $Z_0$  when it comes from  $\Omega_l$  with mean  $\mu_{lt}$  and variance  $\sigma_t^2$ . Let  $u = \frac{Z_0 - \mu_{lt}}{\sigma_t}$  that is,  $u \sim N(0, 1)$ . Solving it for  $Z_0$  gives

$$Z_0 = \sigma_t u + \mu_{lt}, \quad (2.59)$$

After making the change of variables  $u \rightarrow Z_0$  in (2.58) and putting (2.59) into (2.54), and then replacing the expression  $\frac{\mu_{lt} - \mu_{kt}}{\sigma_t}$  by  $d_{lk}$  we get

$$\begin{aligned} W_{lk}^B(Z_0, \Psi) &= (\sigma_t u + \mu_{lt} - (\mu_{lt} + \mu_{kt})/2)(\mu_{lt} - \mu_{kt})/\sigma_t^2 + \gamma_{lk} = \\ &= u d_{lk} + d_{lk}^2/2 + \gamma_{lk} \end{aligned}$$

which completes the proof of lemma.

Recall that the case of nuggetless covariance model with known correlation function is considered, thus, the vector of unknown population parameters has two components, i.e.  $\Psi = (\beta', \sigma^2)'$ . Moreover,  $\alpha_0$  for a nuggetless covariance does not depend on  $\sigma^2$ , that is,  $\alpha_0' = \mathbf{r}_0' \mathbf{R}^{-1}$ . Based on that the estimators of conditional mean and variance are

$$\begin{aligned} \hat{\mu}_{lt} &= \mathbf{x}_0' \hat{\beta}_l + \alpha_0' (\mathbf{t} - \mathbf{X} \hat{\beta}), \\ \hat{\sigma}_t^2 &= \hat{\sigma}^2 K. \end{aligned}$$

Replacing the conditional mean and variance into BDF (2.56) by their estimators specified above we obtain the plug-in BDF,

$$W_{lk}^B(Z_0, \hat{\Psi}) = (Z_0 - \alpha_0' (\mathbf{t} - \mathbf{X} \hat{\beta}) - \hat{\mu}_{lk})' \hat{d}_{lk} / \hat{\sigma} \sqrt{K} + \gamma_{lk}, \quad (2.60)$$

where  $\hat{\mu}_{lk} = \mathbf{x}_0' (\hat{\beta}_l + \hat{\beta}_k)/2$ .

**Definition 2.5.** The *actual error rate* incurred by the plug-in Bayes classification rule associated with PBDF (2.60) is  $P_0^B(\hat{\Psi}) = 1 - \sum_{l=1}^m \pi_l P \hat{C}_l$ , where, for  $k = 1, \dots, m$ ,  $P \hat{C}_l = P_l(W_{lk}^B(Z_0, \hat{\Psi}) \geq 0, l = 1, \dots, m, k \neq l)$ .

The closed-form expression for the actual error rate is presented in the following lemma.

**Lemma 2.4.** The *actual error rate* incurred by plug-in Bayes classification rule associated with PBDF specified in (2.60) for  $m > 2$  has the following form

$$P_0^B(\hat{\Psi}) = 1 - \sum_{l=1}^m \pi_l \int_{A_l} \varphi(u) du, \quad (2.61)$$

where

$$A_l = \{u: u \in R^1, \hat{d}_{lk}u + (\mu_k + \mathbf{r}'_0 \mathbf{R}^{-1} \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \hat{\mu}_{lk})\hat{d}_{lk}/\sigma\sqrt{K} + \gamma_{lk}\hat{\sigma}/\sigma \geq 0; l = 1, \dots, m, k \neq l\}.$$

**Proof.** The proof is completed by making the transformation of random variable (2.59) in the formulas presented in Definition 2.5 and (2.60).

Then the next step is to derive the asymptotic approximation of EER (see Definition 2.4).

## The asymptotic expansion of EER for $m > 2$

---

When the spatial correlation parameters are unknown, the likelihood function is intractable. The ML estimators have not closed-form especially in multiclass case, and it is impossible to obtain the required moments of estimators. So the application of the proposed Taylor series technique is very complicate for EER approximation (the similar situation is for the MSPE of spatial prediction (see Abt [1]). That is the reason why the attention is restricted on the case of known spatial correlation function.

Let  $B(x)$  be a real linear function defined on  $R^1$ , and let  $\delta(x), \delta'(x)$  denote Dirac delta function and its derivative, respectively. The Heaviside step function  $e(x)$  is the integral of the Dirac delta function, i.e.  $e(x) = \int_{-\infty}^x \delta(t)dt$ . The following properties of the Dirac delta function will be used further:

- (d1)  $B(x)\delta(B(x)) \equiv 0$ ,
- (d2)  $B(x)\delta'(B(x)) + \delta(B(x)) \equiv 0$ ,
- (d3)  $\delta(B(x)) = \delta(x - x_0)/|B'_x(x_0)|$ , where  $x_0$  is the solution of  $B(x) = 0$  and  $B'_x(\cdot) = dB_x(\cdot)/dx$ ,
- (d4)  $d(e(x))/dx = \delta(x)$ .

Let  $\hat{\boldsymbol{\Psi}} = (\hat{\boldsymbol{\beta}}', \hat{\sigma}^2)'$  be the vector of parameters estimators. Since we deal with the nuggetless covariance function the ML estimator of  $\boldsymbol{\beta}$  has the following form:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}^{-1}\mathbf{T} \text{ and } \hat{\boldsymbol{\beta}} \sim N_{2q}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}).$$

We use the bias adjusted ML estimator of variance

$$\hat{\sigma}^2 = (\mathbf{T} - \mathbf{X}\hat{\boldsymbol{\beta}})' \mathbf{R}^{-1} (\mathbf{T} - \mathbf{X}\hat{\boldsymbol{\beta}}) / (n - mq) \sim \sigma^2 \chi_{n-mq}^2 / (n - mq).$$

Set  $\Delta\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ ,  $\Delta\hat{\sigma}^2 = \hat{\sigma}^2 - \sigma^2$ . It is easy to show that  $E(\Delta\hat{\boldsymbol{\beta}}) = E(\Delta\hat{\sigma}^2) = \mathbf{0}$ ,  $E(\Delta\hat{\sigma}^2) = 0$  and  $Var(\hat{\sigma}^2) = 2\sigma^4 / (n - mq)$ .

Since the EER approximation is based on the Taylor series, the partial derivatives of *actual error rate* are needed so the lemma presenting these results will be formulated firstly.

Let  $\nabla_{\Psi} P_t$  and  $\nabla_{\Psi}^2 P_t$  be the vector of the first order partial derivatives and the matrix of second order partial derivatives (Hessian) of  $P_0^B(\hat{\Psi})$  with respect to  $\hat{\Psi}$  evaluated at  $\Psi$ , respectively. Similarly, let  $\nabla_{\beta} P_t$  and  $\nabla_{\beta}^2 P_t$  denote the vector of the first order partial derivatives and the matrix of second order partial derivatives of  $P_0^B(\hat{\Psi})$  with respect to  $\hat{\boldsymbol{\beta}}$  evaluated at  $\boldsymbol{\beta}$ , respectively. By  $P_t^{(k)}$  we denote the  $k$ -th ( $k = 1, 2$ ) order partial derivatives of  $P_0^B(\hat{\Psi})$  with respect to  $\hat{\sigma}^2$  evaluated at  $\sigma^2$ . Finally, we denote by  $\nabla_{\beta} P_t^{(1)}$  the vector of second order partial derivatives of  $P_0^B(\hat{\Psi})$  with respect to  $\hat{\boldsymbol{\beta}}$  and  $\hat{\sigma}^2$  evaluated at the values  $\boldsymbol{\beta}$  and  $\sigma^2$ . The same notations will be used for the derivatives of other functions of the mentioned parameters. Let  $z_{lk}$  be the solution of the equation  $W_{lk}^B(z_0, \Psi) = 0$ , i.e.,  $z_{lk} = -\gamma_{lk} \sigma \sqrt{K} / d_{lk} + \mathbf{r}'_0 \mathbf{R}^{-1} (\mathbf{t} - \mathbf{X}\boldsymbol{\beta}) + \mu_{lk}$ . Set  $p_l(z) = \varphi(z; \mu_{lt}, \sigma_t^2)$ .

Since  $G_{lk}^B(z) \geq 0$  is equivalent to  $W_{lk}^B(z) \geq 0$ , the relation between  $G_{lk}^B(z)$  and  $W_{lk}^B(z)$  could be expressed as

$$\begin{aligned} G_{lk}^B(z) &= \pi_l \varphi(z; \mu_{lt}, \sigma_t^2) - \pi_k \varphi(z; \mu_{kt}, \sigma_t^2) = \\ &= \pi_l \varphi(z; \mu_{lt}, \sigma_t^2) (1 - \exp\{-W_{lk}^B(z)\}). \end{aligned} \quad (2.62)$$

**Lemma 2.5.** The *actual error rate* derivatives with respect to  $\hat{\Psi}$  evaluated at the point  $\hat{\Psi} = \Psi$ ,  $z = z_{lk}$  attain the following values

$$\nabla_{\Psi} P_t = \mathbf{0}, \quad (2.63)$$

$$\begin{aligned} \nabla_{\Psi}^2 P_t &= \\ &= \sum_{l=1}^m \sum_{k>l} \pi_l p_l(z_{lk}) \nabla_{\Psi} W_{lk}^B(z_{lk}) \nabla'_{\Psi} W_{lk}^B(z_{lk}) \prod_{j \neq l, k} e(W_{lj}^B(z_{lk})) / W'_{lk}, \end{aligned} \quad (2.64)$$

where  $W'_{lk} = dW_{lk}^B(z) / dz = d_{lk} / \sigma \sqrt{K}$ .

**Proof.** It is easy to see that

$$\begin{aligned} P_0^B(\hat{\Psi}) &= 1 - \sum_l \pi_l \int \prod_{k \neq l} e\left(\hat{W}_{lk}^B(z)\right) p_l(z) dz = \\ &= 1 - \sum_l \pi_l \int \prod_{k \neq l} e\left(\hat{G}_{lk}^B(z)\right) p_l(z) dz. \end{aligned}$$

By using the property of Dirac delta function (d4) we obtain

$$\nabla_{\Psi} P_t = - \sum_{l=1}^m \sum_{k>l} \int_R G_{lk}^B(z) \delta\left(G_{lk}^B(z)\right) \nabla_{\Psi} G_{lk}^B(z) \prod_{j \neq l, k} e(G_{lj}^B(z)) dz.$$

By using the property (d1) in the above equation we obtain (2.63). The Hessian  $\nabla_{\Psi}^2 P_t$  evaluated at  $\hat{\Psi} = \Psi$  is equal to

$$\begin{aligned} \nabla_{\Psi}^2 P_t &= \\ &= \sum_{l=1}^m \sum_{k>l} \int \left[ G_{kl}^B(z) \nabla_{\Psi} G_{lk}^B(z) \delta'\left(G_{lk}^B(z)\right) \nabla'_{\Psi} G_{lk}^B(z) \prod_j e\left(G_{lj}^B(z)\right) \right. \\ &+ G_{kl}^B(z) \nabla_{\Psi}^2 G_{kl}^B(z) \delta\left(G_{kl}^B(z)\right) \prod_j e\left(G_{lj}^B(z)\right) \\ &+ G_{kl}^B(z) \delta\left(G_{lk}^B(z)\right) \nabla_{\Psi} G_{lk}^B(z) \left. \left( \sum_{j \neq l, k} \delta\left(G_{lj}^B(z)\right) \nabla'_{\Psi} G_{lj}^B(z) \prod_{v \neq j, l, k} e\left(G_{lv}^B(z)\right) \right) \right] dz. \end{aligned}$$

It follows from (d1) and (d2) that the integral of the second and third terms in the above square brackets are equal to 0. According to (d3) we have  $\delta(G_{lk}^B) = \delta(z - z_{lk}) \frac{\pi_l p_l(z_{lk})}{|W'(z_{lk})|}$ . Then the proof of (2.64) is completed by using (d2) and (d3) to the integral of the first term.

Recall that for the nuggetless covariance model  $\alpha'_0 = \mathbf{r}'_0 \mathbf{R}^{-1}$  does not depend on  $\sigma^2$  and let  $\lambda_{max}(\mathbf{R})$  be the largest eigenvalue of  $\mathbf{R}$  and make the following assumptions:

- (B1)  $n(\mathbf{X}'\mathbf{X})^{-1} \rightarrow \mathbf{V}$ , as  $n \rightarrow \infty$ , where  $\mathbf{V}$  is a positively definite  $m q \times m q$  matrix with finite determinant.
- (B2)  $\lambda_{max}(\mathbf{R}) < v < +\infty$ , as  $n \rightarrow \infty$ .
- (B3)  $\frac{n_l}{n_k} \rightarrow v_{lk}$ , as  $n_l, n_k \rightarrow \infty$ ,  $0 < v_{lk} < \infty$ .



Put  $\mathbf{F}_{lk}$  as the  $q \times mq$  matrix which is constructed by stacking  $m$  matrices of sizes  $q \times q$ ,

$$\mathbf{F}_{lk} = \begin{pmatrix} \mathbf{0}_q & \dots & \mathbf{0}_q & f_l \mathbf{I}_q & \mathbf{0}_q & \dots & \mathbf{0}_q & f_k \mathbf{I}_q & \mathbf{0}_q & \dots & \mathbf{0}_q \\ q \times mq & 1 & \dots & l-1 & l & l+1 & \dots & k-1 & k & k+1 & \dots & m \end{pmatrix},$$

where  $f_l = \left(\frac{1}{2} + \frac{\gamma_{lk}}{d_{lk}^2}\right)$ ,  $f_k = \left(\frac{1}{2} - \frac{\gamma_{lk}}{d_{lk}^2}\right)$ ,  $\mathbf{I}_q$  and  $\mathbf{0}_q$  are identity matrix and quadratic matrix of zeros, respectively.

$$\text{Set } \mathbf{\Lambda}_{lk} = \mathbf{X}' \boldsymbol{\alpha}_0 - \mathbf{F}_{lk} \mathbf{x}_0 \text{ and } w_{lkj} = -\gamma_{lk} \left(\frac{\mu_l - \mu_j}{\mu_l - \mu_k}\right) + \gamma_{lj} + \frac{(\mu_k - \mu_j)(\mu_l - \mu_j)}{2\sigma^2 K}.$$

**Theorem 2.2.** Suppose that observation  $Z_0$  to be classified by PPDF specified in (2.60) and let assumptions **(B1)**-**(B3)** hold. Then the asymptotic expansion of EER is

$$EER = P_0^B(\boldsymbol{\Psi}) + C/2 + D + O(n^{-2}), \quad (2.65)$$

where

$$C = \sum_{l=1}^m \sum_{k>l} \pi_l \varphi \left(\frac{\gamma_{lk}}{d_{lk}} + \frac{d_{lk}}{2}\right) d_{lk} \boldsymbol{\Lambda}'_{lk} \mathbf{R}_\beta \boldsymbol{\Lambda}_{lk} \prod_{j \neq l, k} e(w_{lkj}) / K,$$

$$D = \sum_{l=1}^m \sum_{k>l} \frac{\gamma_{lk}^2}{n-2q} \pi_l \varphi \left(\frac{\gamma_{lk}}{d_{lk}} + \frac{d_{lk}}{2}\right) \prod_{j \neq l, k} e(w_{lkj}) / d_{lk}.$$

Denote by  $AEER$  the approximation of  $EER$  obtained from (2.65) by ignoring the remainder, i.e.

$$AEER = P_0^B(\boldsymbol{\Psi}) + C/2 + D. \quad (2.66)$$

**Proof.** The proof is based on the Taylor series expansion of  $P_0^B(\widehat{\boldsymbol{\Psi}})$  around  $\boldsymbol{\Psi}$  and taking the expectation with respect to the distribution of  $\mathbf{T}$  and ignoring the terms with derivatives higher than the second order.

Then using the moments of estimators specified above and Lemma 2.5 we have

$$\begin{aligned} EER &= P_0^B(\boldsymbol{\Psi}) + \text{tr} \left( \nabla_{\widehat{\boldsymbol{\Psi}}}^2 P_t \text{Var}(\widehat{\boldsymbol{\Psi}}) \right) / 2 = \\ &= P_0^B(\boldsymbol{\Psi}) \left( \text{tr} \left( \nabla_{\widehat{\boldsymbol{\beta}}}^2 P_t \text{Var}(\widehat{\boldsymbol{\beta}}) \right) + P_t^{(2)} \text{Var}(\widehat{\sigma}^2) \right) / 2 + E(R_3). \end{aligned} \quad (2.67)$$

Then by applying (2.64) we can write

$$\begin{aligned} \nabla_{\beta}^2 P_t &= \sigma\sqrt{K} \sum_{l=1}^m \sum_{k>l} \pi_l p_l(z_{lt}) \nabla_{\beta} W_{lk}^B(z_{lk}) \nabla'_{\beta} W_{lk}^B(z_{lk}) \times \\ &\quad \times \prod_{j \neq l, k} e\left(W_{lj}^B(z_{lk})\right) / d_{lk}, \end{aligned} \quad (2.68)$$

$$P_t^{(2)} = \sigma\sqrt{K} \sum_{l=1}^m \sum_{k>l} \pi_l p_l(z_{lt}) \left(W_{lk}^{(1)}\right)^2 \prod_{j \neq l, k} e\left(W_{lj}^B(z_{lk})\right) / d_{lk}. \quad (2.69)$$

Note that using  $\nabla_{\beta} W_{lk}^B = d_{lk} \mathbf{\Lambda}_{lk} / \sigma\sqrt{K}$  in (2.68) and (2.69) and inserting them into (2.67) we obtain the main term of expansion (2.65). It is obvious that all third order moments of the components of normally distributed vector  $\Delta \widehat{\boldsymbol{\beta}}$  are equal to zero, and  $(\Delta \hat{\sigma}^3) = 8/(n - mq)^2 = O(n^{-2})$ . It implies that  $E(R_3) = O(n^{-2})$ . Putting it into (2.67) we complete the proof of the theorem.

The multiclass case results presented in this section are published in [A6]. Also, the multiclass classification problem for multivariate GGRF is analyzed in [A4] and [A5].

## 2.2.2. Multivariate case

The case of multivariate GGRF with known spatial correlation function is analyzed in Dučinskas [19]. Here the error rates are investigated and the factorized covariance function is considered. In this section the extension of the above mentioned result is presented. The case of complete parametric uncertainty for classification risk of multivariate GGRF is investigated.

Thus, the main objective here is to classify a single observation of multivariate GGRF  $\{\mathbf{Z}(\mathbf{s}): \mathbf{s} \in D \subset R^2\}$  into one of two populations  $\Omega_l, l = 1, 2$ .

The model of observation  $\mathbf{Z}(\mathbf{s})$  in population  $\Omega_l$  is

$$\mathbf{Z}(\mathbf{s}) = \mathbf{B}'_l \mathbf{x}(\mathbf{s}) + \boldsymbol{\varepsilon}(\mathbf{s}),$$

where  $\mathbf{x}(\mathbf{s})$  is a  $q \times 1$  vector of non-random regressors, and  $\mathbf{B}_l$  is a  $q \times p$  matrix of parameters. The error term is generated by  $p$ -variate zero-mean GGRF  $\{\boldsymbol{\varepsilon}(\mathbf{s}): \mathbf{s} \in D\}$  with factorized nuggetless covariance function defined by the following model for all  $\mathbf{s}, \mathbf{u} \in D$

$$\text{cov}\{\boldsymbol{\varepsilon}(\mathbf{s}), \boldsymbol{\varepsilon}(\mathbf{u})\} = r(\mathbf{s} - \mathbf{u})\mathbf{S},$$

where  $r(\mathbf{s} - \mathbf{u})$  is the spatial correlation function, and  $\mathbf{S}$  is the feature variance-covariance matrix. For a given training sample  $\mathbf{T}$  consider the problem of classification of the FO  $\mathbf{Z}_0$  to one of two populations.

The model of training sample is

$$\mathbf{T} = \mathbf{XB} + \mathbf{E}, \quad (2.70)$$

where  $\mathbf{X}$  is the  $n \times 2q$  design matrix, defined in (2.18),  $\mathbf{B}' = (\mathbf{B}'_1, \mathbf{B}'_2)$  is a  $p \times 2q$  matrix of means parameters and  $\mathbf{E}$  represents the  $n \times p$  matrix of random errors that has matrix-variate normal distribution, i.e.

$$\mathbf{E} \sim N_{n \times p}(\mathbf{0}, \mathbf{R} \otimes \mathbf{S}).$$

Here  $\mathbf{R} = \mathbf{R}(\boldsymbol{\vartheta})$ ,  $\boldsymbol{\vartheta} = (\alpha, \lambda, \varphi)$ , has the same meaning as in the univariate case - it denotes the  $n \times n$  matrix of spatial correlations between  $\mathbf{T}$  components.  $\mathbf{S}$  is a  $p \times p$  variance-covariance matrix between features, and  $\otimes$  denotes the Kronecker product of matrices.

Notice that in the population  $\Omega_l$  the conditional distribution of  $\mathbf{Z}_0$ , given  $\mathbf{T} = \mathbf{t}$ , is Gaussian with conditional mean and variance

$$\boldsymbol{\mu}_{lt} = \mathbf{B}'_l \mathbf{x}_0 + \boldsymbol{\alpha}'_0 (\mathbf{t} - \mathbf{XB}), l = 1, 2,$$

$$\mathbf{S}_t = K\mathbf{S}, K = 1 - \mathbf{r}'_0 \mathbf{R}^{-1} \mathbf{r}_0.$$

Then BDF specified in (2.22) becomes

$$W^B(\mathbf{Z}_0, \boldsymbol{\Psi}) = (\mathbf{Z}_0 - (\boldsymbol{\mu}_{1t} + \boldsymbol{\mu}_{2t})/2)' \mathbf{S}_t^{-1} (\boldsymbol{\mu}_{1t} - \boldsymbol{\mu}_{2t}) + \gamma^*. \quad (2.71)$$

The Bayes risk for the BDF (2.71) has the same form as specified in (2.24)

$$R_0^B(\boldsymbol{\Psi}) = \sum_{l=1}^2 \{\pi_l L(l, l) + g_l \Phi(-d/2 + (-1)^l \gamma^*/d)\},$$

here the squared Mahalanobis distance between conditional distributions of  $\mathbf{Z}_0$  for given  $\mathbf{T} = \mathbf{t}$  is specified as

$$d^2 = (\boldsymbol{\mu}_{1t} - \boldsymbol{\mu}_{2t})' \mathbf{S}_t^{-1} (\boldsymbol{\mu}_{1t} + \boldsymbol{\mu}_{2t}). \quad (2.72)$$

Assume that the values of parameters  $\mathbf{B}$ ,  $\mathbf{S}$  and  $\boldsymbol{\vartheta}$  are unknown (the complete parametric uncertainty case). Replacing the parameters by their estimators in (2.71) and using the expressions of conditional mean and variance we get the following PBDF

$$W^B(\mathbf{Z}_0, \hat{\Psi}) = (\mathbf{Z}_0 - \hat{\alpha}'_0(\mathbf{t} - \mathbf{X}\hat{\mathbf{B}}) - \mathbf{x}'_0\mathbf{I}_+\hat{\mathbf{B}}/2)' \hat{\mathbf{S}}^{-1}(\mathbf{x}'_0\mathbf{I}_-\hat{\mathbf{B}})/\hat{K} + \gamma^*, \quad (2.73)$$

**Lemma 2.6.** The actual risk for  $W^B(\mathbf{Z}_0, \hat{\Psi})$  specified in (2.73) is

$$R_0^B(\hat{\Psi}) = \sum_{l=1}^2 \{\pi_l L(l, l) + g_l \Phi(\hat{Q}_l)\}, \quad (2.74)$$

where

$$\hat{Q}_l = (-1)^l \left( (\mathbf{a}_l - \hat{\mathbf{b}})' \hat{\mathbf{S}}^{-1} \hat{\mathbf{B}}' \mathbf{I}'_- \mathbf{x}_0 + \gamma^* \hat{K} \right) / \sqrt{\mathbf{x}'_0 \mathbf{I}_- \hat{\mathbf{B}} \hat{\mathbf{S}}^{-1} \hat{\mathbf{S}} \hat{\mathbf{S}}^{-1} \hat{\mathbf{B}}' \mathbf{I}'_- \mathbf{x}_0 K}, \quad (2.75)$$

$$\mathbf{a}_l = \mathbf{x}'_0 \mathbf{B}_l + \alpha'_0(\mathbf{t} - \mathbf{X}\mathbf{B}), l = 1, 2,$$

$$\mathbf{b} = \hat{\alpha}'_0(\mathbf{t} - \mathbf{X}\hat{\mathbf{B}}) + \mathbf{x}'_0 \mathbf{I}_+ \hat{\mathbf{B}}/2.$$

**Proof.** It is obvious that in the population  $\Omega_l$  the conditional distribution of PBDF, given  $\mathbf{T} = \mathbf{t}$ , is Gaussian, i.e.

$$W^B(\mathbf{Z}_0, \hat{\Psi}) | \Omega_l \sim N(\boldsymbol{\mu}_l^W, \sigma_W^2), \quad (2.76)$$

where

$$\boldsymbol{\mu}_l^W = ((\mathbf{a}_l - \hat{\mathbf{b}})' \hat{\mathbf{S}}^{-1} \hat{\mathbf{B}}' \mathbf{I}'_- \mathbf{x}_0 / \hat{K} + \gamma^*), \quad (2.77)$$

$$\sigma_W^2 = \mathbf{x}'_0 \mathbf{I}_- \hat{\mathbf{B}} \hat{\mathbf{S}}^{-1} \hat{\mathbf{S}} \hat{\mathbf{S}}^{-1} \hat{\mathbf{B}}' \mathbf{I}'_- \mathbf{x}_0 K / \hat{K}^2. \quad (2.78)$$

The proof is completed and formulas (2.74), (2.75) are obtained by using the equations (2.73), (2.76)-(2.78) and (2.11).

## The asymptotic expansion of ER for $m = 2$

---

In order to obtain the asymptotic approximation of expected risk for the multivariate two-class case we will use the maximum likelihood estimators based on the training sample  $\mathbf{T}$ . Let the conditions of Mardia and Marshall (Theorem 1.1) hold. Set

$$\boldsymbol{\beta}_v = \text{vec}(\mathbf{B}), \boldsymbol{\eta} = \text{vech}(\mathbf{S}), \mathbf{R}_g = \partial \text{vec} \mathbf{R} / \partial \boldsymbol{\theta}',$$

$$\dim \boldsymbol{\beta}_v = q_0 = 2qn, \dim \boldsymbol{\eta} = m = p(p+1)/2, \dim \boldsymbol{\theta} = r.$$

The log-likelihood function of  $\mathbf{T}$ , specified in (2.70) is

$$\Lambda(\Psi) = \text{const} - \\ -1/2(p \ln|\mathbf{R}| + n \ln|\mathbf{S}| + \text{tr}(\mathbf{R}^{-1}(\mathbf{T} - \mathbf{X}\mathbf{B})\mathbf{S}^{-1}(\mathbf{T} - \mathbf{X}\mathbf{B})')).$$

Then the information matrices for the corresponding parameters are

$$\mathbf{J}_\beta = (\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}) \otimes \mathbf{S}^{-1}, \\ \mathbf{J}_\eta = n\mathbf{D}'_p(\mathbf{S}^{-1} \otimes \mathbf{S}^{-1})\mathbf{D}_p/2, \\ \mathbf{J}_\theta = p\mathbf{R}'_\theta(\mathbf{R}^{-1} \otimes \mathbf{R}^{-1})\mathbf{R}_\theta/2,$$

where  $\mathbf{D}_p$  is a *duplication matrix* of order  $p^2 \times (p(p+1)/2)$ .

Note that  $\mathbf{J}_{\eta\theta} = E_T(\partial^2\Lambda(\Psi)/\partial\boldsymbol{\eta}\partial\boldsymbol{\theta}')$  and the above information matrices are evaluated at the true values of parameters  $\boldsymbol{\beta}_v$ ,  $\boldsymbol{\eta}$  and  $\boldsymbol{\theta}$ . It is easy to obtain

$$\mathbf{J}_{\eta\theta} = (\mathbf{D}'_p(\mathbf{S}^{-1} \otimes \mathbf{S}^{-1})\text{vec}(\mathbf{S})) \otimes (\text{vec}'\mathbf{R}(\mathbf{R}^{-1} \otimes \mathbf{R}^{-1})\mathbf{R}_\theta/2).$$

Denote by  $\mathbf{J} = \begin{pmatrix} \mathbf{J}_\eta & \mathbf{J}_{\eta\theta} \\ \mathbf{J}_{\theta\eta} & \mathbf{J}_\theta \end{pmatrix}$  and  $\mathbf{V} = \mathbf{J}^{-1} = \begin{pmatrix} \mathbf{V}_\eta & \mathbf{V}_{\eta\theta} \\ \mathbf{V}_{\theta\eta} & \mathbf{V}_\theta \end{pmatrix}$  the information matrix and inverse of information matrix, respectively.

Under some regularity condition, the matrix  $\mathbf{V}$  is an approximate covariance of the ML estimators of covariance function parameters. Using the properties of the multivariate Gaussian distribution it is easy to prove that

$$\widehat{\boldsymbol{\beta}}_v \sim AN_{q_0}(\boldsymbol{\beta}, \mathbf{V}_B), \widehat{\boldsymbol{\eta}} \sim AN_m(\boldsymbol{\eta}, \mathbf{V}_\eta), \text{ and } \widehat{\boldsymbol{\theta}} \sim AN_r(\boldsymbol{\theta}, \mathbf{V}_\theta). \quad (2.79)$$

Let  $\mathbf{R}_\beta^{(k)}$ ,  $\mathbf{R}_\eta^{(k)}$ ,  $\mathbf{R}_\theta^{(k)}$ ,  $k = 1, 2$ , denote the  $k$ -th order derivatives of  $R_0^B(\widehat{\Psi})$  with respect to  $\widehat{\boldsymbol{\beta}}$ ,  $\widehat{\boldsymbol{\eta}}$  and  $\widehat{\boldsymbol{\theta}}$  evaluated at the point  $\widehat{\boldsymbol{\beta}}_v = \boldsymbol{\beta}_v$ ,  $\widehat{\boldsymbol{\eta}} = \boldsymbol{\eta}$ ,  $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ , and let  $\mathbf{R}_{\beta\eta}^{(2)}$ ,  $\mathbf{R}_{\beta\theta}^{(2)}$  and  $\mathbf{R}_{\eta\theta}^{(2)}$  denote the matrices of the second order partial derivatives of  $R_0^B(\widehat{\Psi})$  with respect to  $\widehat{\boldsymbol{\beta}}_v$ ,  $\widehat{\boldsymbol{\eta}}$  and  $\widehat{\boldsymbol{\theta}}$  evaluated at the point  $\widehat{\boldsymbol{\beta}}_v = \boldsymbol{\beta}_v$ ,  $\widehat{\boldsymbol{\eta}} = \boldsymbol{\eta}$ ,  $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ .

Let the assumption (A4), which claims about the independence of  $\mathbf{T}$  and estimator of covariance parameters  $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\eta}}, \widehat{\boldsymbol{\theta}})$ , hold. Also let  $\mathbf{A}_\theta = \partial\widehat{\boldsymbol{\alpha}}_0/\partial\widehat{\boldsymbol{\theta}}'$  be the  $n \times k$  matrix of partial derivatives evaluated at the point  $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ ,  $\mathbf{K}_\theta = \partial K/\partial\boldsymbol{\theta}'$  be the  $k \times 1$  vector of partial derivatives evaluated at

the point  $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ , and let  $\varphi(\cdot)$  be the standard normal distribution density function.

Set  $\boldsymbol{\Lambda} = \mathbf{X}'\boldsymbol{\alpha}_0 - (\mathbf{I}'_+/2 + \gamma^*\mathbf{I}'_-/d^2)\mathbf{x}_0$ ,  $\varphi_1 = \varphi(-d/2 - \gamma^*/d)$  and  $\Delta\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ .

**Theorem 2.3.** Suppose that FO  $\mathbf{Z}_0$  is to be classified by PPDF (2.73), and let the conditions from Theorem 1.1 and assumption (A4) hold. Then the asymptotic approximation of ER is

$$\begin{aligned} AER &= R_0^B(\boldsymbol{\Psi}) + \\ &+ g_1\varphi_1 \left\{ \boldsymbol{\Lambda}'(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\boldsymbol{\Lambda}d/K + (p-1)\mathbf{x}'_0\mathbf{I}_-(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{I}'_-\mathbf{x}_0/Kd + \right. \\ &\left. + \text{tr}(\mathbf{F}_1\mathbf{V}_\eta) + \text{tr}(\mathbf{F}_2\mathbf{V}_\theta) + 2\text{tr}(\mathbf{F}_3\mathbf{V}_{\eta\theta}) \right\} / 2, \end{aligned} \quad (2.80)$$

$$\begin{aligned} \mathbf{F}_1 &= \mathbf{D}'_p \left( (\mathbf{S}^{-1}\Delta\boldsymbol{\mu}\Delta\boldsymbol{\mu}'\mathbf{S}^{-1} \otimes \mathbf{S}^{-1}\Delta\boldsymbol{\mu}\Delta\boldsymbol{\mu}'\mathbf{S}^{-1})(\gamma^*)^2K/\Delta^4 + \mathbf{S}^{-1}\Delta\boldsymbol{\mu}\Delta\boldsymbol{\mu}'\mathbf{S}^{-1} \otimes \right. \\ &\left. \otimes (\mathbf{S}^{-1} - \mathbf{S}^{-1}\Delta\boldsymbol{\mu}\Delta\boldsymbol{\mu}'\mathbf{S}^{-1}/d^2) \right) \mathbf{D}_p / (d\sqrt{K}), \end{aligned} \quad (2.81)$$

$$\mathbf{F}_2 = (\text{tr}(\mathbf{A}'_\theta\mathbf{R}\mathbf{A}_\theta\mathbf{V}_\theta)\Delta^2 + (\gamma^*)^2\mathbf{K}'_\theta\mathbf{V}_\theta\mathbf{K}_\theta) / \Delta^3\sqrt{K}, \quad (2.82)$$

$$\mathbf{F}_3 = \mathbf{D}'_p (\mathbf{S}^{-1}\Delta\boldsymbol{\mu} \otimes \mathbf{S}^{-1}\Delta\boldsymbol{\mu})(\gamma^*)^2\mathbf{K}_\theta / \Delta^4\sqrt{K}. \quad (2.83)$$

**Proof.** Expanding  $R_0^B(\widehat{\boldsymbol{\Psi}})$  in the Taylor series around the point  $\widehat{\boldsymbol{\beta}}_v = \boldsymbol{\beta}_v$ ,  $\widehat{\boldsymbol{\eta}} = \boldsymbol{\eta}$  and  $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}$  up to the second order and taking the expectation with respect to the approximate distribution specified in (2.79) we have

$$\begin{aligned} E_T \left( R_0^B(\widehat{\boldsymbol{\Psi}}) \right) &= R_0^B + \\ &+ E_T \left( (\Delta\widehat{\boldsymbol{\beta}}_v)' \mathbf{R}_\beta^{(2)} \Delta\widehat{\boldsymbol{\beta}}_v + 2(\Delta\widehat{\boldsymbol{\theta}})' \mathbf{R}_{\eta\theta}^{(2)} \Delta\widehat{\boldsymbol{\eta}} + (\Delta\widehat{\boldsymbol{\theta}})' \mathbf{R}_\theta^{(2)} (\Delta\widehat{\boldsymbol{\theta}}) + \right. \\ &\left. + (\Delta\widehat{\boldsymbol{\eta}})' \mathbf{R}_\eta^{(2)} (\Delta\widehat{\boldsymbol{\eta}}) \right) / 2 + R_3, \end{aligned} \quad (2.84)$$

$\Delta\widehat{\boldsymbol{\beta}}_v = \widehat{\boldsymbol{\beta}}_v - \boldsymbol{\beta}_v$ ,  $\Delta\widehat{\boldsymbol{\eta}} = \widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}$ ,  $\Delta\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}$  and  $R_3$  is a reminder term. After doing some matrix algebra we have

$$\begin{aligned} \mathbf{R}_\beta^{(2)} &= g_1\varphi_1 \left\{ (\mathbf{S}^{-1}\Delta\boldsymbol{\mu}\Delta\boldsymbol{\mu}'\mathbf{S}^{-1} \otimes \boldsymbol{\Lambda}\boldsymbol{\Lambda}') / K + (\mathbf{S}^{-1} - \mathbf{S}^{-1}\Delta\boldsymbol{\mu}\Delta\boldsymbol{\mu}'\mathbf{S}^{-1}/\Delta^2) \otimes \right. \\ &\left. \otimes (\mathbf{I}'_-\mathbf{x}_0\mathbf{x}'_0\mathbf{I}_-) / \Delta\sqrt{K} \right\}, \end{aligned} \quad (2.85)$$

$$\begin{aligned} \mathbf{R}_\eta^{(2)} = & g_1 \varphi_1 \mathbf{D}'_p \{ (\mathbf{S}^{-1} \Delta \boldsymbol{\mu} \Delta \boldsymbol{\mu}' \mathbf{S}^{-1} \otimes \mathbf{S}^{-1} \Delta \boldsymbol{\mu} \Delta \boldsymbol{\mu}' \mathbf{S}^{-1}) (\gamma^*)^2 K / \Delta^4 + \\ & + \mathbf{S}^{-1} \Delta \boldsymbol{\mu} \Delta \boldsymbol{\mu}' \mathbf{S}^{-1} \otimes (\mathbf{S}^{-1} - \mathbf{S}^{-1} \Delta \boldsymbol{\mu} \Delta \boldsymbol{\mu}' \mathbf{S}^{-1} / d^2) \} \mathbf{D}_p / \Delta \sqrt{K}, \end{aligned} \quad (2.86)$$

$$E \left( \mathbf{R}_\vartheta^{(2)} \right) = g_1 \varphi_1 \{ \mathbf{A}'_\vartheta \mathbf{R} \mathbf{A}_\vartheta \Delta^2 + (\gamma^*)^2 \mathbf{K}_\vartheta \mathbf{K}'_\vartheta \} / \Delta^3 \sqrt{K}, \quad (2.87)$$

$$E \left( \mathbf{R}_{\eta\vartheta}^{(2)} \right) = g_1 \varphi_1 \mathbf{D}'_m (\mathbf{S}^{-1} \Delta \boldsymbol{\mu} \otimes \mathbf{S}^{-1} \Delta \boldsymbol{\mu}) (\gamma^*)^2 \mathbf{K}_\vartheta / \Delta^4 \sqrt{K}. \quad (2.88)$$

Then by using the assumption (A4) and (2.87), (2.88) and replacing  $E_T(\Delta \hat{\boldsymbol{\theta}} \Delta \hat{\boldsymbol{\theta}}')$  and  $E_T(\Delta \hat{\boldsymbol{\eta}} \Delta \hat{\boldsymbol{\theta}}')$  by their approximations  $\mathbf{V}_\vartheta$  and  $\mathbf{V}_{\eta\vartheta}$  we get the following approximations

$$\begin{aligned} E \left( (\Delta \hat{\boldsymbol{\theta}})' \mathbf{R}_\vartheta^{(2)} (\Delta \hat{\boldsymbol{\theta}}) \right) & \cong \\ & \cong g_1 \varphi_1 \{ \text{tr}(\mathbf{A}'_\vartheta \mathbf{R} \mathbf{A}_\vartheta \mathbf{V}_\vartheta) \Delta^2 + (\gamma^*)^2 \mathbf{K}'_\vartheta \mathbf{V}_\vartheta \mathbf{K}_\vartheta \} / \Delta^3 \sqrt{K}, \end{aligned} \quad (2.89)$$

$$\begin{aligned} E \left( (\Delta \hat{\boldsymbol{\eta}})' \mathbf{R}_{\eta\vartheta}^{(2)} (\Delta \hat{\boldsymbol{\theta}}) \right) & \cong \\ & \cong g_1 \varphi_1 (\gamma^*)^2 \text{tr}(\mathbf{D}'_m (\mathbf{S}^{-1} \Delta \boldsymbol{\mu} \otimes \mathbf{S}^{-1} \Delta \boldsymbol{\mu}) \mathbf{K}_\vartheta \mathbf{V}_{\eta\vartheta}) / \Delta^4 \sqrt{K}. \end{aligned} \quad (2.90)$$

Then using (2.85), (2.86), (2.89) and (2.90) in the right-hand side of (2.84), and dropping the reminder term, we complete the proof of Theorem 2.3.

**Remark 2.4.** The problem of classification of stationary, multivariate GGRF observation to the one of two populations is presented in [A11]. Here the approximation of the *actual error rate* is derived for factorized covariance matrix when all means and covariance parameters are assumed to be unknown.

**Remark 2.5.** The problem of classification of multivariate GGRF observation into one of several populations specified by the different parametric mean models is investigated in Dreiziėnė et al. [A5]. In that paper the closed-form expressions for the *Bayes classification probability* and the *actual correct classification rate* associated with plug-in Bayes classification rule are derived.

## 2.3. Classification of GMRF observation

In this section we extend the analysis to GMRF. Geostatistical models are applied to continuous spatial processes with directly specified Matérn type or other parametric covariance function models. It is well known that the models which include covariance matrices require a large number of computer operations (see Lindgren et al. [40]). In contrast to the geostatistical models, GMRF models are based on the direct specification of sparse precision matrix. They model the data as being related to each other through a neighborhood structure. Thus, the spatial interpolation and classification problems require far fewer calculations.

The main difference comparing with GGRF here is the structure of covariance matrix. We consider the original parametric structure of covariance matrix proposed by de Oliveira and Ferreira [53] that is well-suited to the case of small samples, and ensures good frequentist properties of ML estimators of regression coefficients, spatial dependence and scale parameters. The classifier associated with PBDF is examined, and the main purpose of this chapter is to derive the closed-form expression for the actual risk and the approximation of the expected risk (AER) associated with the aforementioned classifier for the case of complete parametric uncertainty.

### Multivariate case

---

We are concerned with classification problem for a multivariate GMRF observed over a finite lattice. Mardia [48] introduced a multivariate GMRF, and more recently Jin et al. [32], Sain and Cressie [57] explored these models.

We focus on a subclass of a multivariate GMRF so-called a multivariate conditionally autoregressive (MCAR) model with parametrical structure proposed by Pettitt et al. [55]. They extended the univariate spatial model to the multivariate one, maintaining computational simplicity but modelling the essential aspects of dependence between the multivariate components and spatial dependence between sites. De Oliveira and Ferreira [53] showed that



this parametric structure is well-suited to the case of small samples and ensures good frequentist properties of ML estimators of parameters.

We consider *the two-class* case with a *zero-one loss* function. Thus, the main objective is to assign a single multivariate GMRF  $\{\mathbf{Z}(\mathbf{s}): \mathbf{s} \in D \subset R^2\}$  observation to the one of two populations if the training sample  $\mathbf{T}$  is given. The model of observation  $\mathbf{Z}(\mathbf{s})$  is

$$\mathbf{Z}(\mathbf{s}) = \mathbf{B}'_l \mathbf{x}(\mathbf{s}) + \boldsymbol{\varepsilon}(\mathbf{s}),$$

where  $\mathbf{x}(\mathbf{s})$  is a  $q \times 1$  vector of non-random regressors, and  $\mathbf{B}_l$  is a  $q \times p$  matrix of parameters,  $l = 1, 2$ . The main difference comparing with GGRF is the error term  $\boldsymbol{\varepsilon}(\mathbf{s})$  which is generated on lattice by p-variate zero-mean GMRF  $\{\boldsymbol{\varepsilon}(\mathbf{s}): \mathbf{s} \in D\}$  and is specified with respect to the neighborhood structure that will be described later.

Suppose that  $\{\mathbf{s}_i \in D; i = 0, 1, \dots, n\}$  is the set of spatial locations (nodes) where  $\mathbf{Z}(\mathbf{s})$  observations are taken. Indexing the spatial locations by integers, i.e.  $\mathbf{s}_i = i, i = 0, 1, \dots, n$  denote the set of training locations by  $\mathbf{S}_n = \mathbf{S}^{(1)} \cup \mathbf{S}^{(2)}$ , where  $\mathbf{S}^{(l)}$  are the subsets of  $\mathbf{S}_n$  that contains  $n_l$  observations of  $\mathbf{Z}(\mathbf{s})$  from  $\Omega_l, l = 1, 2, n = n_1 + n_2$ . The focal location  $\mathbf{s}_0$  is indexed by  $\{0\}$ .

Assume that the lattice  $\mathbf{S}_n^0 = \mathbf{S}_n \cup \{0\}$  is endowed with a neighborhood system  $N^0 = \{N_i^0: i = 0, 1, \dots, n\}$ , and the lattice  $\mathbf{S}_n$  is endowed with a neighborhood system  $\{N_i: i = 1, \dots, n\}$ , where  $N_i$  denotes the collection of sites that are the neighbours of site  $\mathbf{s}_i$ . Then, define the spatial weights  $w_{lk} > 0$  ( $w_{kl} = w_{lk}$ ) as a measure of similarity between the sites  $l$  and  $k$ , and let

$$\mathbf{w}'_0 = (w_{01}, \dots, w_{0n}),$$

$$\mathbf{w}'_i = (w_{i0}, \dots, w_{ii-1}, w_{ii+1}, \dots, w_{in}), i = 1, \dots, n.$$

Following de Oliveira and Ferreira [53] we construct matrices  $\mathbf{H}^0 = (h_{ij}^0: i, j = 0, 1, \dots, n)$  and  $\mathbf{H} = (h_{ij}: i, j = 1, \dots, n)$  with dimensions  $(n + 1) \times (n + 1)$  and  $(n \times n)$ , respectively. The elements of these matrices are defined as follows

$$h_{ij}^0 = \begin{cases} h_i^0 & \text{if } i = j \\ -w_{ij} & \text{if } i \in N_j^0 \\ 0 & \text{otherwise} \end{cases}, \quad h_{ij} = \begin{cases} h_i & \text{if } i = j \\ -w_{ij} & \text{if } i \in N_j \\ 0 & \text{otherwise} \end{cases}$$

where  $h_i^0 = \sum_{j \in N_i^0} w_{ij}$ ,  $i, j = 0, 1, \dots, n$ , and  $h_i = \sum_{j \in N_i} w_{ij}$ ,  $i, j = 1, \dots, n$ .

These matrices, assumed to be known, allow the modelling of different patterns of spatial correlation by the specification of different neighborhood systems and weights ( $w_{ij}$ ) (de Oliveira, Ferreira [53]).

The main objective now is to classify a single multivariate GMRF observation specified on lattice  $\mathbf{S}_n^0$  at the focal location  $\mathbf{s}_0$ .

For simplicity, we use the following notations:

$$\mathbf{Z}(\mathbf{s}_i) = \mathbf{Z}_i, \boldsymbol{\varepsilon}(\mathbf{s}_i) = \boldsymbol{\varepsilon}_i, \mathbf{x}(\mathbf{s}_i) = \mathbf{x}_i, i = 0, \dots, n.$$

$$\mathbf{Z} = (\mathbf{Z}'_0, \mathbf{Z}'_1, \dots, \mathbf{Z}'_n)', \mathbf{Z}_{-0} = (\mathbf{Z}'_1, \dots, \mathbf{Z}'_n)', \boldsymbol{\varepsilon}_{-0} = (\boldsymbol{\varepsilon}'_1, \dots, \boldsymbol{\varepsilon}'_n)'$$

Then let for  $i = 1, \dots, n$

$$\boldsymbol{\varepsilon}_{-i} = (\boldsymbol{\varepsilon}'_0, \dots, \boldsymbol{\varepsilon}'_{i-1}, \boldsymbol{\varepsilon}'_{i+1}, \dots, \boldsymbol{\varepsilon}'_n)', \mathbf{Z}_{-i} = (\mathbf{Z}'_0, \dots, \mathbf{Z}'_{i-1}, \mathbf{Z}'_{i+1}, \dots, \mathbf{Z}'_n)'$$

Suppose that  $\boldsymbol{\Lambda}$  denotes the  $p \times p$  correlation type matrix with ones on the diagonal and off-diagonal entries  $\{-\lambda_{ij}\}$  but plays a role of a precision matrix. Putting the matrix elements equal to zero gives conditional independence between the components of  $\mathbf{Z}$ . The full conditionals for  $i = 0, 1, \dots, n$  and  $l = 1, 2$  are specified as

$$\boldsymbol{\varepsilon}_i | \boldsymbol{\varepsilon}_{-i} \sim N(\boldsymbol{\mu}_i^i, \boldsymbol{\Sigma}^i),$$

$$\boldsymbol{\mu}_i^i = (\boldsymbol{\alpha}'_i \otimes \mathbf{I}_p) \boldsymbol{\varepsilon}_{-i} \text{ and } \boldsymbol{\Sigma}^i = \rho_i \boldsymbol{\Lambda}^{-1}.$$

Here  $\boldsymbol{\alpha}'_i = \alpha \mathbf{w}'_i / (1 + \alpha h_i)$  and  $\rho_i = \sigma^2 / (1 + \alpha h_i)$ .  $\alpha \geq 0$  is a spatial dependence parameter, and  $\sigma > 0$  is a scale parameter. Then the covariance matrix of  $\mathbf{Z} = (\mathbf{Z}'_0, \mathbf{Z}'_1, \dots, \mathbf{Z}'_n)'$  is

$$\text{var}(\mathbf{Z}) = \sigma^2 (\mathbf{I}_{n+1} + \alpha \mathbf{H}^0)^{-1} \otimes \boldsymbol{\Lambda}^{-1}.$$

Denote the training sample in the vector form by  $\mathbf{T} = \mathbf{Z}_{-0}$  and in the matrix form by  $\mathbf{T}^* = (\mathbf{Z}_1, \dots, \mathbf{Z}_n)'$ . The design matrix  $\mathbf{X}$  for the training sample  $\mathbf{T}$  is specified in (2.18). Then, under some regularity conditions (see Mardia [48]), the joint distribution for the training sample in the vector form is

$$\mathbf{T} \sim N_{np}(\text{vec}(\mathbf{B}'\mathbf{X}'), \sigma^2 \mathbf{V}(\alpha) \otimes \boldsymbol{\Lambda}^{-1}),$$

and in the matrix form it follows the  $n \times p$  matrix Gaussian distribution

$$\mathbf{T}^* \sim N_{p \times n}(\mathbf{XB}, \sigma^2 \mathbf{V}(\alpha) \otimes \mathbf{\Lambda}^{-1}),$$

where  $\mathbf{B}' = (\mathbf{B}'_1, \mathbf{B}'_2)$  and  $\mathbf{V}(\alpha) = (\mathbf{I}_n + \alpha \mathbf{H})^{-1}$  denotes the spatial correlation matrix for  $\mathbf{T}$ . Parameter  $\alpha$  controls the strength of correlation between the components of  $\mathbf{T}$ . When  $\alpha = 0$  the components of  $\mathbf{T}$  become independent random variables. In addition, it also controls conditional correlations among neighboring sites. In the following for brevity we will use  $\mathbf{V} = \mathbf{V}(\alpha)$ .

For a given training sample realization  $\mathbf{T} = \mathbf{t}$  (or  $\mathbf{T}^* = \mathbf{t}^*$ ), the conditional distribution of observation  $\mathbf{Z}_0$  in the population  $\Omega_l$  is p-variate Gaussian

$$(\mathbf{Z}_0 | \mathbf{T} = \mathbf{t}; \Omega_l) \sim N_p(\boldsymbol{\mu}_{lt}, \mathbf{S}_t), \quad (2.91)$$

where for  $l = 1, 2$

$$\begin{aligned} \boldsymbol{\mu}_{lt} &= \mathbf{B}'_l \mathbf{x}_0 + (\boldsymbol{\alpha}'_0 \otimes \mathbf{I}_p)(\mathbf{t} - \text{vec}(\mathbf{XB})) = \\ &= \mathbf{B}'_l \mathbf{x}_0 + (\boldsymbol{\alpha}'_0 \otimes \mathbf{I}_p)(\mathbf{t}^* - \mathbf{XB})', \end{aligned} \quad (2.92)$$

$$\mathbf{S}_t = \rho_0 \mathbf{\Lambda}^{-1}, \rho_0 = \sigma^2 / (1 + \alpha h_0). \quad (2.93)$$

In the following let  $P_{0l}$  denote the conditional distribution specified in (2.91)-(2.93), for  $l = 1, 2$ . The squared Mahalanobis distance between the populations based on the conditional distribution for the observation taken at the location  $\mathbf{s}_0$  is  $d^2 = (\boldsymbol{\mu}_{1t} - \boldsymbol{\mu}_{2t})' \mathbf{\Lambda}(\boldsymbol{\mu}_{1t} - \boldsymbol{\mu}_{2t}) / \rho_0$ . Putting the expression of conditional mean into the recent formula we get the expression for conditional Mahalanobis distance which does not depend on the realizations of training sample  $\mathbf{T}$ , but depends on the location of training sample elements

$$d^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \mathbf{\Lambda}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) / \rho_0,$$

where  $\boldsymbol{\mu}_{lt} = \mathbf{B}'_l \mathbf{x}_0, l = 1, 2$ .

Then Bayes discriminant function minimizing the probability of misclassification is

$$W^B(\mathbf{Z}_0, \boldsymbol{\Psi}) = (1 + \alpha h_0) \times$$

$$\times (\mathbf{Z}_0 - (\boldsymbol{\mu}_{1t} + \boldsymbol{\mu}_{2t})/2)' \boldsymbol{\Lambda}(\boldsymbol{\mu}_{1t} - \boldsymbol{\mu}_{2t})/\sigma^2 + \gamma, \quad (2.94)$$

where  $\gamma = \ln(\pi_1/\pi_2)$ . Using (2.92), (2.93) and replacing  $\mathbf{t}^*$  by  $\mathbf{T}^*$  in (2.94) we get

$$\begin{aligned} W^B(\mathbf{Z}_0, \boldsymbol{\Psi}) &= (1 + \alpha h_0) \times \\ &\times (\mathbf{Z}_0 - \boldsymbol{\alpha}'_0(\mathbf{T}^* - \mathbf{X}\mathbf{B}) - \mathbf{x}'_0\mathbf{I}_+\mathbf{B}/2)' \boldsymbol{\Lambda}\mathbf{x}'_0\mathbf{I}_-\mathbf{B}/\sigma^2 + \gamma. \end{aligned} \quad (2.95)$$

**Lemma 2.7.** Bayes error rate for  $W^B(\mathbf{Z}_0, \boldsymbol{\Psi})$  specified in (2.95) is

$$P_0^B(\boldsymbol{\Psi}) = \sum_{l=1}^2 \pi_l \Phi(-d/2 + (-1)^l \gamma/d).$$

**Proof.** The proof of Lemma 2.7 is analogous to the proof of Lemma 2.1. Recall, that the Bayes error rate for  $W^B(\mathbf{Z}_0, \boldsymbol{\Psi})$  is defined as

$$P_0^B(\boldsymbol{\Psi}) = \sum_{l=1}^2 \pi_l P_l((-1)^l W^B(\mathbf{Z}_0, \boldsymbol{\Psi}) \geq 0).$$

The conditional distribution of  $W^B(\mathbf{Z}_0, \boldsymbol{\Psi})$  in the population  $\Omega_l$ , given  $\mathbf{T} = \mathbf{t}$ , is Gaussian with mean and variance

$$\begin{aligned} E_l(W^B(\mathbf{Z}_0, \boldsymbol{\Psi})) &= (-1)^{l+1} d^2/2 + \gamma, l = 1, 2, \\ \text{Var}(W^B(\mathbf{Z}_0, \boldsymbol{\Psi})) &= d^2. \end{aligned}$$

Then, using the properties of normal distribution we complete the proof of lemma.

In this section we assume that the true values of parameters  $\mathbf{B}$  and  $\sigma^2$  are unknown and the ML estimators  $\hat{\mathbf{B}}$  and  $\hat{\sigma}^2$  based on  $\mathbf{T}$  are used. Then the vector of parameters that have to be estimated and the vector of their estimators are denoted by  $\boldsymbol{\Psi} = (\mathbf{B}, \sigma^2)$  and  $\hat{\boldsymbol{\Psi}} = (\hat{\mathbf{B}}, \hat{\sigma}^2)$ , respectively. After replacing  $\boldsymbol{\Psi}$  by  $\hat{\boldsymbol{\Psi}}$  in (2.95), we get the PPDF

$$W^B(\mathbf{Z}_0, \hat{\Psi}) = (1 + \alpha h_0) \times \\ \times \left( \mathbf{Z}_0 - (\mathbf{T}^* - \mathbf{X}\hat{\mathbf{B}})' \boldsymbol{\alpha}_0 - \hat{\mathbf{B}}' \mathbf{I}'_+ \mathbf{x}_0 / 2 \right)' \boldsymbol{\Lambda} \hat{\mathbf{B}}' \mathbf{I}'_- \mathbf{x}_0 / \hat{\sigma}^2 + \gamma. \quad (2.96)$$

Then the *actual error rate* for PPDF  $W^B(\mathbf{Z}_0, \hat{\Psi})$  is defined as

$$P_0^B(\hat{\Psi}) = \sum_{l=1}^2 \pi_l P_{0l}((-1)^l W^B(\mathbf{Z}_0, \hat{\Psi}) > 0). \quad (2.97)$$

**Lemma 2.8.** The *actual error rate* for PPDF specified in (2.96) is

$$P_0^B(\hat{\Psi}) = \sum_{l=1}^2 \pi_l \Phi(\hat{Q}_l),$$

where

$$\hat{Q}_l = (-1)^l \frac{\left( \mathbf{x}'_0 (\mathbf{B}_l - \mathbf{I}_+ \hat{\mathbf{B}} / 2) + \boldsymbol{\alpha}'_0 \mathbf{X} (\hat{\mathbf{B}} - \mathbf{B}) \right) \boldsymbol{\Lambda} \hat{\mathbf{B}}' \mathbf{I}'_- \mathbf{x}_0 / k_0 + \gamma \hat{\sigma}^2}{\hat{\sigma} \sqrt{\mathbf{x}'_0 \mathbf{I}_- \hat{\mathbf{B}} \boldsymbol{\Lambda} \hat{\mathbf{B}}' \mathbf{I}'_- \mathbf{x}_0 / k_0}}$$

and  $k_0 = 1/(1 + \alpha h_0)$ .

**Proof.** The proof of lemma is established by essentially exploiting the proof of Lemma 2.7, formulas (2.96), (2.97) and the properties of multivariate Gaussian distribution.

## The asymptotic expansion of expected error rate

---

In order to derive the approximation of *expected error rate* (AEER) we will use the ML estimator of matrix  $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{T}^*$ , which is a  $2q \times p$  matrix of regression coefficients. We will also use the bias adjusted ML estimator of scale parameter

$$\hat{\sigma}^2 = \left( \mathbf{T} - \text{vec} \left( (\mathbf{X}\hat{\mathbf{B}})' \right) \right)' \mathbf{V}^{-1} \otimes \boldsymbol{\Lambda}^{-1} \left( \mathbf{T} - \text{vec} \left( (\mathbf{X}\hat{\mathbf{B}})' \right) \right) / (np - 2q).$$

Using the properties of the matrix-variate normal distribution it is easy to show that

$$\hat{\mathbf{B}} \sim N_{2q \times p}(\mathbf{B}, \sigma^2 (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \otimes \boldsymbol{\Lambda}^{-1}) \quad (2.98)$$

and

$$\hat{\sigma}^2 \sim \sigma^2 \chi^2(np - 2q)/(np - 2q). \quad (2.99)$$

**Theorem 2.4.** Suppose that FO  $\mathbf{Z}_0$  to be classified by PBDF specified in (2.96). Then the approximation of EER is

$$\begin{aligned} AEER &= P_0^B(\Psi) + \pi_1 \varphi(-d/2 - \gamma/d) \times \\ &\times \{\mathbf{F}'_0 \mathbf{R}_B \mathbf{F}_0 d/k_0 + (p-1) \mathbf{x}'_0 \mathbf{I}_- \mathbf{R}_B \mathbf{I}'_- \mathbf{x}_0 / (k_0 d) + 2\gamma^2/d(np-2q)\}/2, \end{aligned}$$

where

$$\mathbf{F}_0 = \mathbf{X}' \boldsymbol{\alpha}_0 - (\mathbf{I}'_+/2 + \gamma \mathbf{I}_-/d^2) \mathbf{x}_0,$$

$$\mathbf{R}_B = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1}.$$

**Proof.** Let  $\Delta \hat{\mathbf{B}} = \hat{\mathbf{B}} - \mathbf{B}$ ,  $\Delta \hat{\sigma}^2 = \hat{\sigma}^2 - \sigma^2$ . Using (2.98)-(2.99) it is easy to show (e.g. Magnus and Neudecker [44]) that

$$E_T(\Delta \hat{\mathbf{B}}) = \mathbf{0}, E_T(\Delta \hat{\sigma}^2) = 0, E_T(\Delta \hat{\sigma}^2 \Delta \hat{\mathbf{B}}) = \mathbf{0} \quad (2.100)$$

and

$$E_T \left( \text{vec}(\Delta \hat{\mathbf{B}}') \left( \text{vec}(\Delta \hat{\mathbf{B}}') \right)' \right) = \mathbf{R}_B \otimes \boldsymbol{\Sigma}, \quad (2.101)$$

$$E_T(\Delta \hat{\sigma}^2) = \sigma^4 / (np - 2q). \quad (2.102)$$

For the proof of theorem we use the following notations. Let  $\hat{b}_{\alpha\beta}$  represent the elements of  $\hat{\mathbf{B}}$ . Denote by

$$P_B^{(1)}(\alpha, \beta) = \partial P_0^B(\Psi) / \partial \hat{b}_{\alpha\beta}, P_B^{(2)}(\alpha\beta, \gamma\delta) = \partial^2 P_0^B(\Psi) / \partial \hat{b}_{\alpha\beta} \partial \hat{b}_{\gamma\delta},$$

$$P_\sigma^{(1)} = \partial P_0^B(\Psi) / \partial \hat{\sigma}^2, P_\sigma^{(2)} = \partial^2 P_0^B(\Psi) / \partial^2 \hat{\sigma}^2,$$

$$P_{B,\sigma}^{(2)}(\alpha\beta) = \partial^2 P_0^B(\Psi) / \partial \hat{b}_{\alpha\beta} \partial \hat{\sigma}^2$$

the partial derivatives of  $P_0^B(\Psi)$  with respect to the corresponding parameters evaluated at the point  $\hat{\mathbf{B}} = \mathbf{B}$ ,  $\hat{\sigma}^2 = \sigma^2$ . The analogous notations will be used for the partial derivatives of  $\hat{Q}_l$ ,  $l = 1, 2$ .

Make a Taylor expansion of  $P_0^B(\widehat{\Psi})$  at the points  $\widehat{\mathbf{B}} = \mathbf{B}$  and  $\widehat{\sigma}^2 = \sigma^2$  up to the second order partial derivatives and use the Lagrange remainder term. Taking the expectation with respect to the distribution of  $\mathbf{T}$  and using (2.100) we get

$$E_T \left( P_0^B(\widehat{\Psi}) \right) = P_0^B(\Psi) + \left( \sum_{\alpha, \gamma=1}^{2q} \sum_{\beta, \delta=1}^p P_B^{(2)}(\alpha\beta, \gamma\delta) E_T(\Delta\widehat{b}_{\alpha\beta}\Delta\widehat{b}_{\gamma\delta}) + P_\sigma^{(2)} E_T(\Delta\widehat{\sigma}^2) \right) / 2 + R_3, \quad (2.103)$$

where  $R_3$  is the expectation of remainder term. Note that

$$\pi_1 \varphi(Q_1) = \pi_2 \varphi(Q_2). \quad (2.104)$$

By using the chain rule and equation (2.104) we have

$$P_B^{(2)}(\alpha\beta, \gamma\delta) = \pi_1 \varphi(Q_1) \sum_{l=1}^2 (-1)^l \left( Q_l Q_{lB}^{(1)}(\alpha\beta) Q_{lB}^{(1)}(\gamma\delta) - Q_{lB}^{(2)}(\alpha\beta, \gamma\delta) \right), \quad (2.105)$$

$$P_\sigma^{(2)} = \pi_1 \varphi(Q_1) \sum_{l=1}^2 (-1)^l \left( Q_l \left( Q_{l\sigma}^{(1)} \right)^2 - Q_{l\sigma}^{(2)} \right). \quad (2.106)$$

Taking the appropriate partial derivatives by elements of matrices, we have

$$Q_{lB}^{(1)}(\alpha\beta) = (-1)^l (\mathbf{X}'\alpha_0 - (\mathbf{I}'_+/2 + \gamma\mathbf{I}'_-/d^2)\mathbf{x}_0) \times \mathbf{J}^{\alpha\beta} \mathbf{A}\mathbf{B}'\mathbf{I}'_- \mathbf{x}_0 / dk_0 \sigma^2 \quad (2.107)$$

$$\sum_{l=1}^2 Q_{lB}^{(2)}(\alpha\beta, \gamma\delta) = (\mathbf{x}'_0 \mathbf{I}'_- \mathbf{J}^{\alpha\gamma} \mathbf{I}'_- \mathbf{x}_0 - \mathbf{x}'_0 \mathbf{I}'_- \mathbf{J}^{\alpha\beta} \mathbf{A}\mathbf{B}'\mathbf{I}'_- \mathbf{x}_0 \mathbf{x}'_0 \mathbf{I}'_- \mathbf{J}^{\gamma\delta} \mathbf{\Sigma}^{-1} \mathbf{B}'\mathbf{I}'_- \mathbf{x}_0 / d^2 k_0) / dk_0 \sigma^2, \quad (2.108)$$

$$Q_{l\sigma}^{(1)} = (-1)^l \gamma \mathbf{x}'_0 \mathbf{I}'_- \mathbf{B}\mathbf{\Sigma}^{-1} \mathbf{A}\mathbf{B}'\mathbf{I}'_- \mathbf{x}'_0 / d^3 k_0 \sigma^2, \quad (2.109)$$

$$\sum_{l=1}^2 Q_{l\sigma}^{(2)} = (\mathbf{x}'_0 \mathbf{I}'_- \mathbf{B}\mathbf{\Sigma}^{-1} \mathbf{A}\mathbf{B}'\mathbf{I}'_- \mathbf{x}_0 - \mathbf{x}'_0 \mathbf{I}'_- \mathbf{B}\mathbf{A}\mathbf{B}'\mathbf{I}'_- \mathbf{x}_0 \mathbf{x}'_0 \mathbf{I}'_- \mathbf{A}\mathbf{B}'\mathbf{I}'_- \mathbf{x}_0 / d^2 k_0 \sigma^4) / dk_0, \quad (2.110)$$

where  $\mathbf{J}^{ij}$  is the matrix of zeroes except element  $(i, j)$  that is equal to 1.

The Lagrange remainder is the third-order polynomial with respect to the components of  $\Delta\widehat{\mathbf{B}}$  and  $\Delta\widehat{\sigma}^2$ . The coefficients of this polynomial are the third-order partial derivatives of  $P_0^B(\widehat{\Psi})$  with respect to components of  $\widehat{\mathbf{B}}$  and

$\hat{\sigma}^2$  estimated in the neighborhood of their true values. It is obvious from (2.98) and (2.99) that all third order moments of normally distributed components of  $\Delta\hat{\mathbf{B}}$  are equal to 0 and all third order moments of  $\Delta\hat{\sigma}^2$  components are of order  $O(1/n^2)$ . Third order partial derivatives of  $\Phi(\hat{Q}_l)$  with respect to the elements of  $\hat{\mathbf{B}}$  and  $\hat{\sigma}^2$  are bounded by the uniformly bounded functions in the neighborhood of point  $\hat{\mathbf{B}} = \mathbf{B}, \hat{\sigma}^2 = \sigma^2$ . So we can scrap  $R_3$  in (2.103). Finally, putting (2.104-2.110) into (2.103) and using (2.101), (2.102) we complete the proof of the theorem.

**Remark 2.6.** The formula for the approximation of *expected risk* for univariate GMRF could be found in [A1], and it has the following form

$$AER = R_0^B(\Psi) + g_1\varphi(-d/2 - \gamma^*/d)d(K_\beta + K_\alpha + (\gamma^*)^2K_\theta/d_2)/2,$$

where

$$K_\beta = \mathbf{L}'(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{L}\rho_0, \mathbf{L}' = \alpha\mathbf{w}'_0\mathbf{X}/(1 + \alpha\mathbf{h}_0) - \mathbf{x}'_0(\mathbf{I}_+/2 + \gamma\mathbf{I}_-/d^2), K_\alpha = 2n\mathbf{w}'_0\mathbf{V}\mathbf{w}_0/(\rho_0^3(n \operatorname{tr}\mathbf{V}^2 - (\operatorname{tr}\mathbf{V})^2)),$$

$$K_\theta = \mathbf{v}'\mathbf{I}_\theta\mathbf{v}/\rho_0^2, \text{ and } \mathbf{v}' = (h_0, -1/\sigma_\epsilon^2).$$



# Chapter 3

## Numerical experiments and applications

This chapter demonstrates the results of numerical experiments with simulated data and the application to the real data. In Section 3.1 the empirical power for proposed non-parametric test is calculated. Section 3.2 examines the influence of different covariance parameters and Mahalanobis distance to the proposed AER. The univariate two-class and multiclass cases are analyzed. For the two-class case the comparison of AER values using a symmetric and asymmetric training sample plan is done. For the multiclass case the calculations are performed for grouped and mixed sets of training locations. Actual error rate and its approximation are also accomplished for GMRF. In Section 3.3 the classification problem for real data is solved. The calculations were performed by *geoR*, *gstat* and *INLA*: free and open-source packages for geostatistical analysis included in statistical computing software *R* (<http://www.r-project.org/>). *R* is a language and environment which provides a wide variety of statistical and graphical techniques, and is highly extensible. It allows users to add additional functionality by defining new functions.

The results presented in this section are published in [A6], [A9], [A12], [A14].

### 3.1. The efficacy of non-parametric test

In this section the efficacy analysis of the proposed non-parametric test for isotropy (see Section 1.3) is presented. The numerical experiment is

performed with simulated data, where the geometrically anisotropic GRFs were simulated using package *geoR*. The empirical power of the test is examined for a different number of simulations. Also the comparison for different  $K$ , the largest number of lags in one direction, is presented.

Consider the case with  $D$  being the integer regular 2-dimensional lattice. Set  $\mathbf{h}' = (h_x, h_y)$ , for each  $\mathbf{h} \in D$ . Here  $h_x = x_i - x_j, h_y = y_i - y_j, i, j = 1, \dots, n$ . Simulations are done on  $10 \times 10$  square grid, thus the sample size is  $n = 121$ . We generate the realizations from zero-mean, stationary GRF. The case of nuggetless, factorized covariance model is analyzed

$$C(\mathbf{h}) = \sigma^2 r(\mathbf{h}),$$

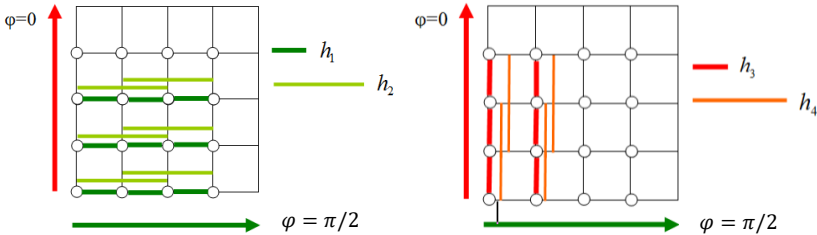
where  $r(\mathbf{h}) = \exp\{-(h_x^2 + \lambda^2 h_y^2)/\alpha^2\}$  is the geometric anisotropic spatial Gaussian correlation function. Suppose that the direction of anisotropy is known, that is, the anisotropy angle  $\varphi$  is fixed and is equal to  $\pi/2$ .

We start with two spatial lags ( $K = 2$ ), that is,  $\mathbf{L} = (h_1, h_2, h_3, h_4)$  and  $|h_1| = |h_3| = 1, |h_2| = |h_4| = 2$ . Then we build the contrast matrix and calculate the test statistic specified in (1.17)

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix},$$

$$\hat{T} = \frac{1}{2} \sum_{i=1}^2 (\hat{\gamma}(h_i) - \hat{\gamma}(h_{i+2}))^2 / \left( \frac{\hat{\gamma}^2(h_i)}{|N(h_i)|} + \frac{\hat{\gamma}^2(h_{i+2})}{|N(h_{i+2})|} \right).$$

Its approximate distribution is the  $\chi_2^2$  distribution.  $N(h_i)$  and  $N(h_{i+2})$  represent the pairs  $(\mathbf{s}_i, \mathbf{s}_j)$  for a certain lag, and  $|N(h_i)|, |N(h_{i+2})|$  are the numbers of such pairs in the two orthogonal directions (see Figure 2).



**Figure 2.** Sampling pairs for directional semivariograms: the main direction of anisotropy,  $\varphi = \pi/2$ , and the orthogonal direction,  $\varphi = 0$ .

As a performance measure of the proposed test statistic we consider the empirical power of test (a frequency of rejecting  $H_0$  for the simulated geometric anisotropic Gaussian data) with significance level,  $p = 0.05$ . For the various values of anisotropy ratio  $\lambda$  and range parameter  $\alpha$  three simulation procedures of size  $M = (150, 300, 600)$  are performed. Table 1 shows that the empirical power of the test increases with increasing of range parameter, but the empirical power is not influenced by the anisotropy ratio. So, we propose to use our test statistic for the particular cases of geometrically anisotropic spatial Gaussian data.

**Table 1.** Empirical powers of test for simulated data

		Range ( $\alpha$ )					
		$M = 150$					
$\lambda$		1	4	7	10	13	16
2		0.03	0.57	0.69	0.8	0.81	0.83
4		0.04	0.53	0.69	0.83	0.81	0.85
6		0.05	0.53	0.75	0.79	0.79	0.78
8		0.02	0.59	0.76	0.70	0.75	0.87
10		0.07	0.53	0.69	0.73	0.80	0.84
		$M = 300$					
2		0.06	0.56	0.68	0.76	0.81	0.82
4		0.04	0.48	0.72	0.77	0.77	0.84
6		0.06	0.58	0.71	0.76	0.77	0.84
8		0.05	0.53	0.67	0.76	0.78	0.82
10		0.05	0.51	0.69	0.78	0.80	0.78
		$M = 600$					
2		0.05	0.57	0.71	0.76	0.80	0.83
4		0.03	0.54	0.71	0.75	0.78	0.81
6		0.04	0.53	0.73	0.76	0.82	0.83
8		0.05	0.54	0.72	0.76	0.80	0.81
10		0.05	0.52	0.67	0.79	0.80	0.84

Increasing the number of lags ( $K = 5$  and  $K = 10$ ) yields the significant increase of empirical power, especially for the low range values. Analyzing the contents of Table 2 we notice that for  $K = 10$  even for small range values the empirical power is approximately 50%.

**Table 2.** Empirical powers of test with different number of lags and various values of range and anisotropy ratio

$\lambda$	Range ( $\alpha$ )					
	$K = 2$					
	1	4	7	10	13	16
2	0.03	0.57	0.69	0.80	0.81	0.83
4	0.04	0.53	0.69	0.83	0.81	0.85
6	0.05	0.53	0.75	0.79	0.79	0.78
8	0.02	0.59	0.76	0.70	0.75	0.87
10	0.07	0.53	0.69	0.73	0.80	0.84
	$K = 5$					
2	0.17	0.78	0.84	0.87	0.94	0.94
4	0.18	0.76	0.91	0.82	0.91	0.87
6	0.19	0.75	0.87	0.87	0.91	0.87
8	0.12	0.76	0.87	0.91	0.88	0.90
10	0.13	0.75	0.83	0.85	0.92	0.89
	$K = 10$					
2	0.52	0.94	0.91	0.97	0.89	0.92
4	0.49	0.96	0.93	0.89	0.95	0.94
6	0.47	0.92	0.89	0.93	0.93	0.89
8	0.44	0.91	0.91	0.91	0.89	0.89
10	0.59	0.96	0.91	0.92	0.88	0.93

## 3.2. The analysis of AER accuracy and influence of statistical parameters to AER

### GGRF two-class case

In order to investigate the performance of the proposed plug-in Bayes discriminant function and to analyze the influence of covariance parameters, also to evaluate the accuracy of the derived AER, a simulation study was carried out. Consider the case of classification of scalar observation  $Z_0$  to the one of two populations,  $\Omega_l, l = 1, 2$ . With an insignificant loss of generality the case with *zero-one loss* function, i.e.  $L(l, k) = 1 - \delta_{lk}, l, k = 1, 2$ , is

analyzed. Also the equal-sized training samples with equal prior probabilities are assumed, that is,  $n_1 = n_2 = 4$  and  $\pi_1 = \pi_2 = 1/2$ .

The observations are assumed to arise from a stationary Gaussian random field with constant mean and nuggetless covariance function given by  $C(\mathbf{h}, \boldsymbol{\theta}) = \sigma^2 r(\mathbf{h})$ , where  $\sigma^2$  is the unknown scale parameter (or partial sill), and  $r(\mathbf{h})$  is the spatial correlation function. The exponential geometrically anisotropic correlation function with the unknown range  $\alpha$  and anisotropy ratio  $\lambda$  is chosen

$$r(\mathbf{h}) = \exp\left\{-\sqrt{h_x^2 + \lambda^2 h_y^2}/\alpha\right\}.$$

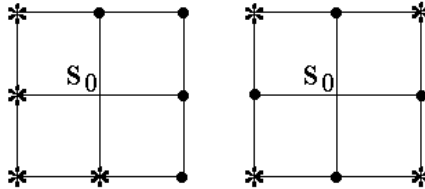
Here  $h_x = x_i - x_j$ ,  $h_y = y_i - y_j$ ,  $i, j = 1, \dots, n$ . The anisotropy angle is assumed known,  $\varphi = \pi/2$ . Hence, the vector of unknown covariance parameters has three components, i.e.  $\boldsymbol{\theta} = (\sigma^2, \lambda, \alpha)'$ .

The training sample  $\mathbf{T} = (\mathbf{T}'_1, \mathbf{T}'_2)' = (Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))'$  is observed on a regular 2-dimensional lattice with unit spacing. Consider the case of  $\mathbf{s}_0 = (1,1)$  and fixed set of training locations (STL)  $\mathbf{S}_n$  which is partitioned into the union of 2 disjoint subsets, i.e.  $\mathbf{S}_n = \mathbf{S}^{(1)} \cup \mathbf{S}^{(2)}$ . Two different STL,  $\xi_1$  and  $\xi_2$ , are analyzed:

$$\xi_1 = \left\{ \mathbf{S}^{(1)} = \{(1,2), (2,2), (2,1), (2,0)\}, \mathbf{S}^{(2)} = \{(1,0), (0,0), (0,1), (0,2)\} \right\},$$

$$\xi_2 = \left\{ \mathbf{S}^{(1)} = \{(1,2), (2,1), (1,0), (0,1)\}, \mathbf{S}^{(2)} = \{(0,0), (0,2), (2,2), (2,0)\} \right\}.$$

The distributions of  $\mathbf{S}_n$  are presented in Figure 3.



**Figure 3.** Two different STL, where  $\bullet$  represents the elements of  $\mathbf{S}^{(1)}$  and  $*$  represents the elements of  $\mathbf{S}^{(2)}$

In order to implement the AER formula some functions included into *geoR* were used, e.g. function *grf* was used for the simulation of Gaussian random field; *variog* was applied for calculating the values of empirical semivariogram, etc. However, it should be noted that most of the calculations were done using the *R* programming language. For example, the *geoR* package contains a *varcov.spatial* function that calculates an isotropic covariance matrix, and there is no function which creates the anisotropic covariance matrix, so it is necessary to build a procedure that allows the anisotropy parameters to be included in the covariance function.

The values of AER specified in (2.33)-(2.36) are calculated for various values of parameters  $\lambda$  and  $\alpha$ . The results of the calculations with Mahalanobis distance  $\Delta = 1$  are presented in Table 3 and Table 4. Analyzing the contents of these tables we can conclude that for both STL the AER values are decreasing with the increase of range parameter  $\alpha$ . It means, the stronger spatial correlation yields the lower AER values.

For low range values ( $\alpha < 2.4$  or  $\alpha < 2$ ) the AER values are also decreasing with the increase of anisotropy ratio  $\lambda$ , which means that higher level of anisotropy gives lower values of AER. However, for high range values the influence of anisotropy ratio becomes insignificant.  $\lambda = 1$  stands for isotropic case, and we can notice that for isotropic data we get better accuracy of proposed classifier.

**Table 3.** Values of AER for STL  $\xi_1$  with  $\Delta = 1$  and various  $\lambda$  and  $\alpha$ .

	$\alpha$							
$\lambda$	0.6	0.8	1.2	1.6	2	2.4	2.8	3.2
1	0.32844	0.30726	0.27095	0.24097	0.21575	0.19421	0.17555	0.15922
2	0.33414	0.31400	0.28144	0.25502	0.23240	0.21267	0.19525	0.17974
3	0.33440	0.31362	0.28081	0.25490	0.23292	0.21372	0.19672	0.18153
4	0.33394	0.31314	0.28047	0.25472	0.23291	0.21387	0.19699	0.18190
5	0.33324	0.31255	0.28020	0.25460	0.23288	0.21390	0.19708	0.18202
6	0.33254	0.31200	0.27994	0.25450	0.23285	0.21391	0.19711	0.18207
7	0.33193	0.31154	0.27972	0.25440	0.23282	0.21391	0.19713	0.18210
8	0.33140	0.31116	0.27953	0.25431	0.23278	0.21390	0.19714	0.18212
9	0.33095	0.31084	0.27938	0.25424	0.23275	0.21389	0.19714	0.18212
10	0.33056	0.31058	0.27925	0.25417	0.23272	0.21388	0.19713	0.18213

**Table 4.** Values of AER for STL  $\xi_2$  with  $\Delta = 1$  and various  $\lambda$  and  $\alpha$ 

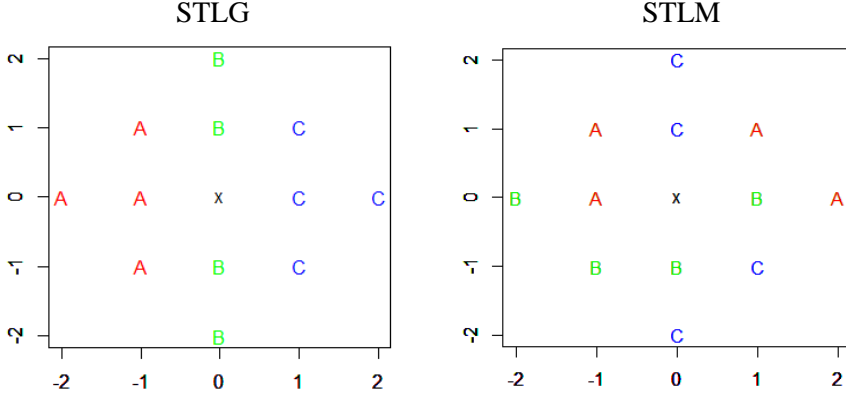
$\lambda$	$\alpha$							
	0.6	0.8	1.2	1.6	2	2.4	2.8	3.2
1	0.33011	0.31038	0.27675	0.24879	0.22499	0.20442	0.18641	0.17048
2	0.33529	0.31632	0.28604	0.26146	0.24026	0.22159	0.20494	0.18999
3	0.33563	0.31612	0.28566	0.26159	0.24100	0.22284	0.20660	0.19196
4	0.33520	0.31575	0.28561	0.26179	0.24140	0.22340	0.20729	0.19275
5	0.33451	0.31520	0.28550	0.26193	0.24168	0.22378	0.20773	0.19324
6	0.33381	0.31467	0.28533	0.26198	0.24186	0.22403	0.20803	0.19356
7	0.33320	0.31421	0.28514	0.26197	0.24196	0.22419	0.20823	0.19379
8	0.33267	0.31383	0.28497	0.26194	0.24201	0.22430	0.20837	0.19395
9	0.33222	0.31351	0.28482	0.26189	0.24203	0.22436	0.20846	0.19406
10	0.33183	0.31325	0.28470	0.26184	0.24203	0.22439	0.20852	0.19413

Comparing the values of two STL it was noticed that the symmetric STL ( $\xi_1$ ) gives lower AER values than asymmetric STL ( $\xi_2$ ) since the values of the ratio  $AER_{\xi_2}/AER_{\xi_1}$  for all parametric structures is greater than 1 (for more details see [A12]).

The influence of the anisotropy ratio on the approximation of expected error rates in classification of GGRF observation with nuggetless and known correlation function is studied in [A15]. Here AER values for different Mahalanobis distance and various anisotropy ratios are calculated.

## GGRF multiclass case

Now an example of classifying a scalar observation  $Z_0$  for the three class case will be carried out. In this example observations are assumed to arise from stationary GGRF with constant mean and isotropic exponential covariance function given by  $C(\mathbf{h}, \boldsymbol{\theta}) = \sigma^2 \exp\{-\mathbf{h}/\alpha\}$ . The set of training locations  $\mathbf{S}_{12}$  that forms the third order neighborhood for  $\mathbf{s}_0 = (0,0)$  is considered. We will analyze two different STL: with grouped labels (STLG) and mixed labels (STLM). The distributions of STLG and STLM are depicted in Figure 4.



**Figure 4.** STL  $\mathbf{S}_{12}$  with different labels distributions. The points indicated by A, B and C belong to  $\mathbf{S}^{(1)}$ ,  $\mathbf{S}^{(2)}$  and  $\mathbf{S}^{(3)}$ , respectively. Sign  $\times$  denotes  $\mathbf{s}_0$

All of the simulations have considered small training sample sizes, i.e.  $n_l = 4$  and equal prior probabilities  $\pi_l = 1/3, l = 1, 2, 3$ . 1000 simulations (runs) were performed for each STL. For each simulated training sample the actual error rate  $P_0^B(\hat{\Psi})$  specified in Lemma 2.4 (2.61) was evaluated. The expected error rate, obtained by averaging actual error rates over runs, is denoted by  $\overline{\text{EER}}$ . AEER, derived in Theorem 2.2, is also calculated and its accuracy is evaluated by  $\eta = |\text{AEER} - \overline{\text{EER}}|$ .

We considered the parametric structure with  $\mu_1 = b, \mu_2 = 0, \mu_3 = -3b$  and  $\sigma^2 = 1$ , then  $\Delta_{12} = b, \Delta_{13} = 4b, \Delta_{23} = 3b$ . It follows that  $b$  represents the level of separation between classes and is called the *separation step*.

Table 5 contains the values of AEER and  $\overline{\text{EER}}$  calculated for various levels of spatial correlation and class separation specified by parameters  $\alpha$  and  $b$ . They show that EER and its approximation decreases as values of these parameters increases for both labels distributions. That is quite logical since the Mahalanobis distances  $|d_{lk}|$  between the classes are proportional to  $b/\sqrt{K}$ ,  $K = 1 - \mathbf{r}_0^T \mathbf{R}^{-1} \mathbf{r}_0$  and  $K$  decreases as  $\alpha$  increases. So the separation between classes increases with increasing of  $\alpha$  and  $b$ .

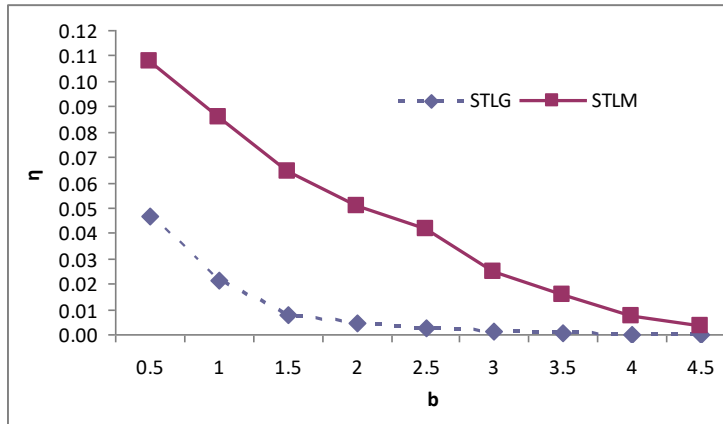
Slight difference in AEER and  $\overline{\text{EER}}$  decreasing rates could be caused by increasing of the asymptotic expansion remainder values for the strongly correlated cases.



**Table 5.** AEER and  $\overline{\text{EER}}$  values for different  $\alpha$  and  $b$

		$b$								
$\alpha$		0.5	1	1.5	2	2.5	3	3.5	4	
STLG	AEER	1	0.37768	0.20695	0.12516	0.07692	0.04460	0.02397	0.01190	0.00545
		2	0.30566	0.14607	0.07632	0.03612	0.01489	0.00533	0.00163	0.00044
		3	0.25606	0.10991	0.04781	0.01700	0.00489	0.00114	0.00021	0.00003
	$\overline{\text{EER}}$	1	0.41488	0.22851	0.13270	0.07926	0.04604	0.02497	0.01263	0.00591
		2	0.35213	0.16729	0.08409	0.04084	0.01731	0.00669	0.00228	0.00065
		3	0.30367	0.13187	0.05636	0.02115	0.00676	0.00188	0.00040	0.00007
STLM	AEER	1	0.37474	0.20388	0.12249	0.07452	0.04265	0.02258	0.01102	0.00495
		2	0.30188	0.14293	0.07363	0.03415	0.01374	0.00478	0.00142	0.00037
		3	0.25275	0.10737	0.04584	0.01590	0.00444	0.00100	0.00017	0.00002
	$\overline{\text{EER}}$	1	0.42585	0.24952	0.15246	0.10027	0.06332	0.04002	0.02274	0.01283
		2	0.40950	0.22844	0.13768	0.08506	0.05517	0.02940	0.01727	0.00768
		3	0.39864	0.22117	0.13910	0.08201	0.05133	0.02975	0.01607	0.00879

The values of accuracy  $\eta$ , are depicted in Figure 5. It shows the advantage of STLG against STLM. It means that the proposed approximation of EER is more precise when the classes are not mixed over the region.



**Figure 5.** Comparison of  $\eta$  for different labels distributions and  $b$ , for  $\alpha = 2$

The results presented in this chapter are published in Dučinskas and Drežienė [A12] and Dučinskas et al. [A6].

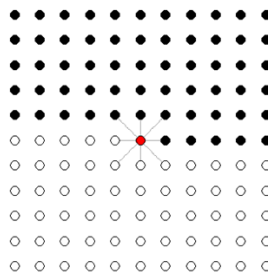
Summing up, the results of numerical analysis give us strong arguments to expect that the proposed approximation of the expected risk (or expected error rate) could be effectively used for the performance evaluation of the plug-in Bayes rule applied to the classification of spatial Gaussian process observation in particular parametric structure cases and even small training samples.

## GMRF two-class case

---

The similar calculations are done for GMRF observation. The values of actual risk and the approximation of expected risk (Section 2.3) in the finite training sample case are calculated, and the influence of statistical parameters is demonstrated. The case of  $n_1 = n_2 = 60$ ,  $\pi_1 = \pi_2 = 1/2$  and  $L(l, k) = 1 - \delta_{lk}$ ,  $l, k = 1, 2$ , is considered. The simulations of GMRF were performed by INLA, the package included in *R*. The parameters to be varied in the simulation experiment are the spatial dependence parameter  $\alpha$  and the marginal Mahalanobis distance  $\Delta$ .

Assume that GMRFs are sampled on the  $11 \times 11$  regular unit spacing lattice  $\mathbf{S}_{120}$  with the focal location in the center of lattice (see Figure 6). Here we use the power distance weights of the form  $w_{ij} = d_{ij}^{-m}$ , where  $d_{ij}$  refers to the Euclidean distance between sites  $i$  and  $j$ , and  $m$  is any positive integer.



**Figure 6.** Set of training locations with focal location.

The points indicated by ● and ○ belong to  $\mathbf{S}^{(1)}$  and  $\mathbf{S}^{(2)}$ , respectively

For the simulations the true values of the parameters are fixed at  $\beta_1 = 0$ ,  $\beta_2 = 1$ , and  $\sigma^2 = 1$ . The numerical illustration of parameters estimates is

presented in Table 6. This table shows that for all selected values of spatial dependence parameter  $\alpha$  there are no significant biases of the considered estimators.

**Table 6.** ML estimates of  $\beta, \sigma^2$  and Mahalanobis distance  $d$  with  $\hat{d}$  for various values of  $\alpha$ .

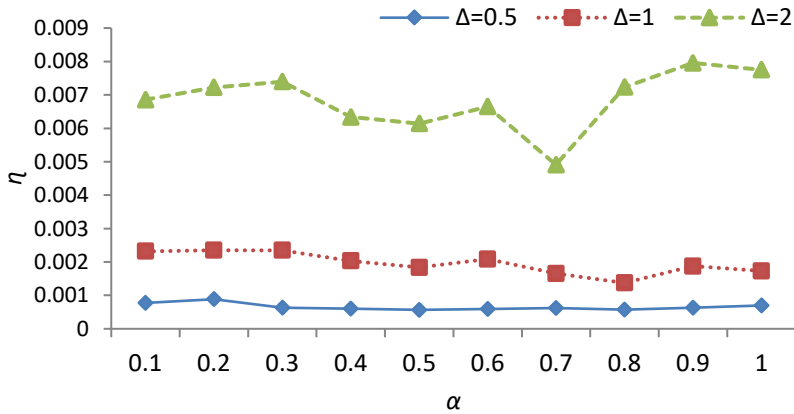
$\alpha$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\sigma}^2$	$\hat{d}$	$d$
0.1	-0.00093	1.00875	1.00861	1.34530	1.29724
0.2	-0.00047	1.00824	1.00861	1.59157	1.53808
0.3	-0.00005	1.00774	1.00861	1.80306	1.74600
0.4	-0.00071	0.99931	0.98609	1.99365	1.93168
0.5	-0.00019	0.99941	0.98609	2.16557	2.10100
0.6	0.00840	0.99145	0.99907	2.26333	2.25767
0.7	0.00739	0.99129	0.99907	2.41081	2.40414
0.8	0.00649	0.99116	0.99907	2.54989	2.54219
0.9	-0.00300	1.01759	1.04422	2.71571	2.67312
1	-0.00284	1.01755	1.04422	2.84076	2.79793

The approximation of expected risk (AER) (Theorem 2.2.) is also calculated and the accuracy of AER is evaluated by relative error  $\eta = |\text{AER} - \overline{\text{ER}}|/\overline{\text{ER}}$ . Table 7 provides the values of AER and  $\overline{\text{ER}}$  calculated with respect to  $\alpha$  and  $\Delta$ . The results show that all AER and  $\overline{\text{ER}}$  values are decreasing while  $\alpha$  and  $\Delta$  are increasing. That means, the greater separation between classes and the greater spatial dependence parameter give better accuracy of the proposed AER.

**Table 7.**  $\overline{\text{ER}}$  and AER values for different  $\alpha$  values and different class separation ( $\Delta$ )

$\alpha$	$\overline{\text{ER}}$			AER		
	$\Delta=0.5$	$\Delta=1$	$\Delta=2$	$\Delta=0.5$	$\Delta=1$	$\Delta=2$
0.1	0.37319	0.25877	0.09775	0.37348	0.25937	0.09842
0.2	0.35053	0.22129	0.06228	0.35084	0.22181	0.06273
0.3	0.33151	0.19160	0.04056	0.33172	0.19205	0.04086
0.4	0.31482	0.16733	0.02683	0.31501	0.16767	0.02700
0.5	0.29993	0.14699	0.01791	0.30010	0.14726	0.01802
0.6	0.28642	0.12966	0.01203	0.28659	0.12993	0.01211
0.7	0.27407	0.11486	0.00815	0.27424	0.11505	0.00819
0.8	0.26269	0.10204	0.00553	0.26284	0.10218	0.00557
0.9	0.25211	0.09081	0.00377	0.25227	0.09098	0.00380
1	0.24223	0.08103	0.00258	0.24240	0.08117	0.00260

Figure 7 shows that the accuracy of AER is sufficiently stable with respect to the increase in  $\alpha$ . However, it was noticed that the general trend for the relative error of the AER is an increase in the distance between the populations.



**Figure 7.** Relative error of AER for various values of  $\alpha$  and three values of  $\Delta$ .

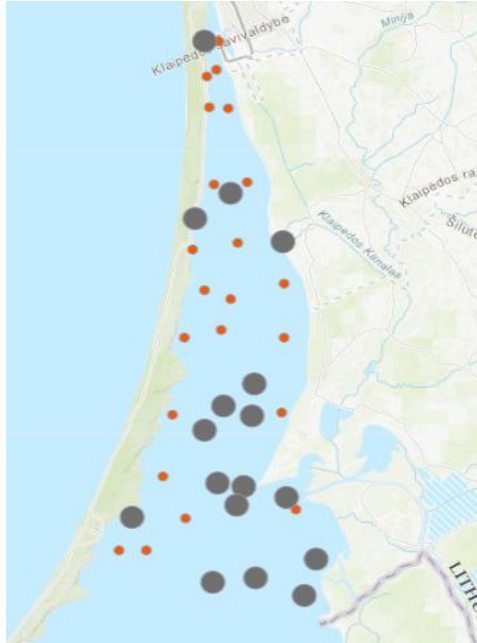
### 3.3. Application of PBDf to the mapping of presence and absence of zebra mussels in the Curonian Lagoon

The *zebra mussel* (*Dreissena polymorpha*) is a small freshwater mussel (see Figure 8). They are commonly found on the bottom of ships and eat the algae that are food for fish. Nevertheless, the zebra mussels process up to one liter of water per day, per mussel and could be used to improve water clarity. These are the reasons why scientists are interested in mussels. They are trying to control the mussels and to be able to cultivate them in the certain areas. The Curonian Lagoon is a large (1.584 km<sup>2</sup>), shallow (average depth 3.8 m) coastal waterbody connected to the Baltic Sea by the narrow Klaipeda Strait. Currently, zebra mussels are highly abundant in the Curonian Lagoon, occupying the littoral zone down to 3–4 m depth and occurring on both hard substrates and soft bottoms (Zaiko, Daunys [76]).



**Figure 8.** Zebra mussels

The main purpose of this section is to apply the proposed discriminant functions to the mapping of presence and absence of zebra mussels in the Curonian Lagoon; to build the model with minimal misclassification error; to analyze the influence of spatial correlation to the misclassification errors.



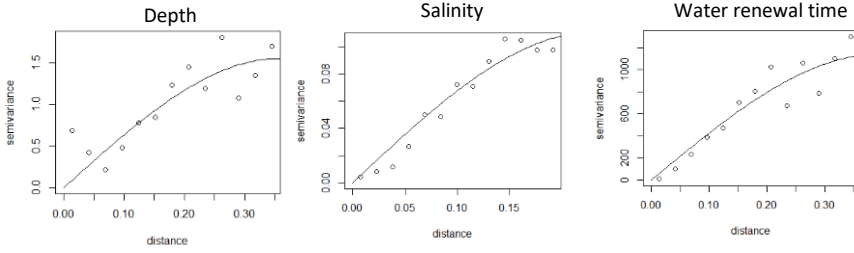
**Figure 9.** Set of training locations. Red ● and grey ● dots represent absence and presence of zebra mussel, respectively

The training sample consists of  $n = 39$  locations (stations in the Curonian Lagoon) (see Figure 9). The red dots,  $n_1 = 22$ , represent the locations where zebra mussels were not found, and the grey dots,  $n_2 = 17$ , represent the presence of zebra mussels. In addition, three variables were observed at those locations: *salinity*, *water renewal time* and *depth*. All these variables could be used for classification as a feature vector  $\mathbf{Z}$ , and they could also be included into the design matrix as covariates.

At first it is necessary to verify if the data are spatially correlated. *Moran I index*, one of the oldest statistics used to examine spatial autocorrelation, shows significant spatial correlation for all variables (Table 8). The existence of spatial correlation could also be confirmed by semivariograms (see Figure 10).

**Table 8.** The values of Moran I index

	Depth	Salinity	Water renewal time
Moran I	0.586	0.954	0.886
p-value	4.69E-06	2.62E-12	3.13E-11



**Figure 10.** Semivariograms for the observed variables

Spatial information could be included into the model through different ways. Firstly it could be included through the covariance matrix. Then, it could be done through the mean model involving coordinates of the locations into the design matrix. For example, the first order trend surface model includes coordinates of spatial locations. Then the design matrix has the following form

$$\mathbf{X} = \begin{pmatrix} 1 & x_l & y_l & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & x_{nl} & y_{nl} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & x_{nl+1} & y_{nl+1} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 1 & x_n & y_n \end{pmatrix}.$$

Lastly, the spatial information could be used for the estimation of prior probabilities. The simplest way is to assume that the populations are equiprobable, i.e.  $\pi_1 = \pi_2$ , but in this situation this does not seem reasonable because the training samples are of a different size and the area of interest (spatial domain  $D$ ) is large enough. The better decision is to take the sizes of  $\mathbf{T}_l$  into account. Since the total number of elements in the training sample  $\mathbf{T}$  is  $n$ , and  $n_1$  and  $n_2$  of them belong to  $\mathbf{T}_1$  and  $\mathbf{T}_2$ , respectively, then the prior probabilities could be evaluated by  $\hat{\pi}_l = n_l/n, l = 1,2$  (Theodoritis [67]). Another way is to include only the nearest neighbors of focal location  $\mathbf{s}_0$ .

Then  $\hat{\pi}_l = n_{0l}/n_0$ , where  $n_{0l}$  is the number of nearest neighbours of  $\mathbf{s}_0$  in population  $\Omega_l$ ,  $l = 1, 2$ , and  $n_0 = n_{01} + n_{02}$ .

For every observed variable different mean models were used: the constant mean model, first order trend surface model and other, more complicated models which include coordinates and other covariates. In total, 15 different models were analyzed.

Let us focus on the univariate case of GMRF where the covariance is defined by  $\Sigma = \sigma^2(\mathbf{I}_n + \alpha\mathbf{H})^{-1}$ . Let  $\sigma^2$  be the unknown parameter and  $\alpha$  is known and equal to 0.5. To construct the matrix  $\mathbf{H}$  a different number of neighbors was included according to the maximum allowed distance ( $d_{max}$ ), and the spatial weights of the form  $w_{ij} = d_{ij}^{-1}$ , where  $d_{ij}$  refers to the Euclidean distance, were used. The formula (2.96) was realized and, according to the sign of  $W^B(Z_0, \hat{\Psi})$ , the decision was made. To evaluate the performance of the discriminant function the cross validation procedure was applied.

Table 9 provides the results for the variable *Depth*. Here the probabilities of correct classification are presented. These probabilities were calculated for different mean models and for different number of neighbors. The last column corresponds to the situation with no spatial correlation. The remaining columns correspond to the situations with different number of neighbors. The best model which gives the greatest (73.7%) correct classification probability is

$$\mathbf{Depth} = \mathbf{X}\boldsymbol{\beta} + \varepsilon(\mathbf{s}),$$

where  $\mathbf{X}\boldsymbol{\beta}$  is the first order trend surface model. Spatial information in this model is included through the covariance, where only the closest neighbors are used, through the prior probabilities which are obtained using the same set of neighbors, and through the mean model.



**Table 9.** Correct classification probabilities for variable Depth. Water renewal time is abbreviated as WRT

Neighborhood structure						
Distance	$d_{\max} = 0.1$	$d_{\max} = 0.2$	$d_{\max} = 0.3$	$d_{\max} = 0.4$	No spatial correlation	
Number of neighbors	[4-16]	[15-35]	[24-38]	[36-38]	0	
<b>Mean model</b>	~1	0.718	0.641	0.641	0.641	0.385
	~1+XY	<b>0.737</b>	0.658	0.684	0.711	0.632
	~1+XY+WRT	0.711	0.605	0.632	0.632	0.447
	~1+XY+Salinity	0.718	0.667	0.641	0.641	0.538
	~1+XY+WRT+Salinity	0.692	0.692	0.692	0.692	0.692

Performing the classification by the variable *Salinity* and including *Depth* and *Water renewal time* as covariates we get very similar results (see Table 10), but this time the best model which gives the greatest (74.4%) correct classification probability is

$$\mathbf{Salinity} = \mathbf{X}\boldsymbol{\beta} + \varepsilon(\mathbf{s}),$$

where  $\mathbf{X}\boldsymbol{\beta}$  is the constant mean model. It means that spatial information is not included into the mean model but it is the component of covariance, and it is also used to evaluate the prior probabilities.

**Table 10.** Correct classification probabilities for variable Salinity

Neighborhood structure						
Distance	$d_{\max} = 0.1$	$d_{\max} = 0.2$	$d_{\max} = 0.3$	$d_{\max} = 0.4$	No spatial correlation	
Number of neighbors	[4-16]	[15-35]	[24-38]	[36-38]	0	
<b>Mean model</b>	~1	<b>0.744</b>	0.692	0.692	0.692	0.590
	~1+XY	0.684	0.658	0.658	0.658	0.526
	~1+XY+Depth	0.615	0.590	0.615	0.615	0.436
	~1+XY+WRT	0.667	0.667	0.667	0.667	0.692
	~1+XY+Depth+WRT	0.615	0.590	0.641	0.641	0.615

**Table 11.** Correct classification probabilities for variable Water renewal time

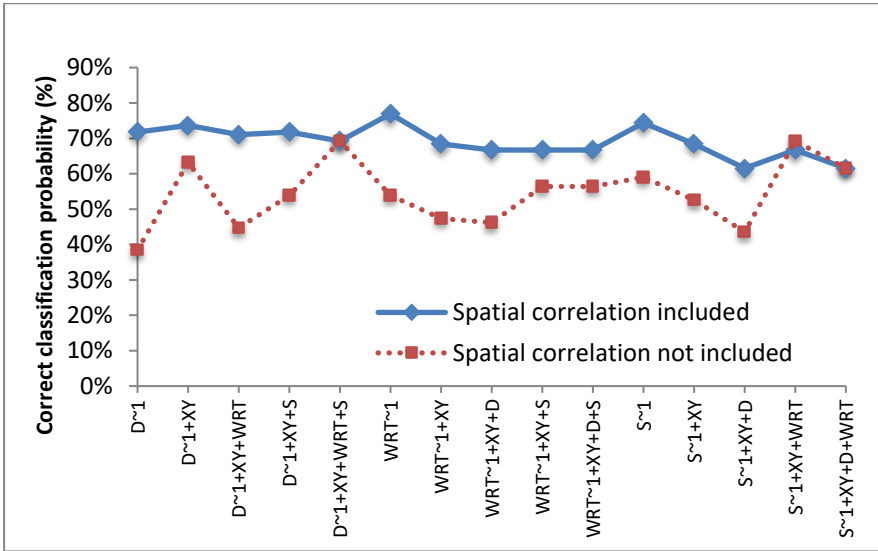
Neighborhood structure						
Distance	$d_{\max} = 0.1$	$d_{\max} = 0.2$	$d_{\max} = 0.3$	$d_{\max} = 0.4$	No spatial correlation	
Number of neighbors	[4-16]	[15-35]	[24-38]	[36-38]	0	
<b>Mean model</b>	~1	<b>0.769</b>	0.744	0.718	0.692	0.538
	~1+XY	0.684	0.684	0.632	0.605	0.474
	~1+XY+Depth	0.667	0.667	0.615	0.590	0.462
	~1+XY+Salinity	0.667	0.667	0.667	0.641	0.564
	~1+XY+Depth+Salinity	0.667	0.641	0.615	0.615	0.564

Using *Water renewal time* for classification we get the results which are presented in Table 11. The best model here, which gives the greatest (76.9%) correct classification probability, is

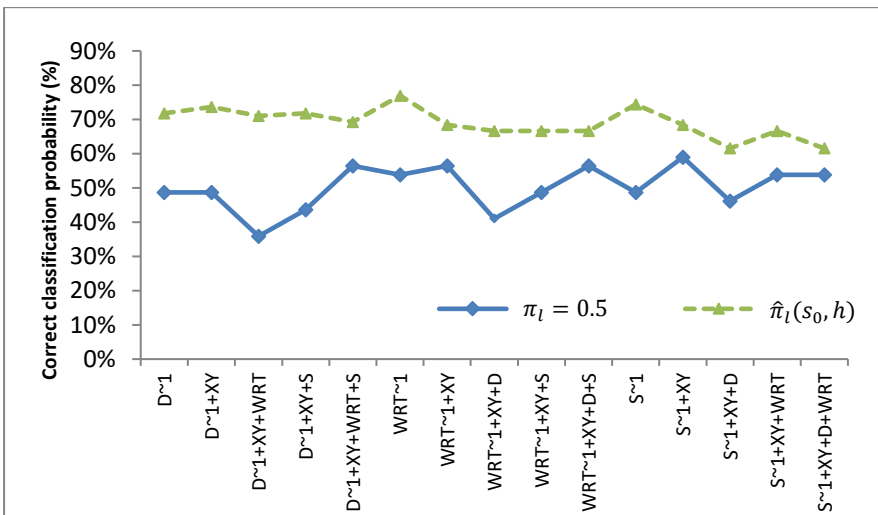
$$\mathbf{Water\ renewal\ time} = \mathbf{X}\boldsymbol{\beta} + \varepsilon(\mathbf{s}),$$

where  $\mathbf{X}\boldsymbol{\beta}$  is the constant mean model. In this model the spatial information is included into the covariance, and it is also used to evaluate the prior probabilities.

Analyzing Figure 11 it is easy to notice that including spatial correlation into the covariance gives higher correct classification probabilities. The influence of prior probabilities is depicted in Figure 12. The upper line corresponds to the case of equiprobable populations. The bottom line shows the correct classification probabilities when priors are estimated including only the nearest neighbors but not the whole training sample.



**Figure 11.** The influence of spatial correlation.  
D-Depth, S-Salinity, WRT-Water renewal time



**Figure 12.** The influence of prior probabilities  
D-Depth, S-Salinity, WRT-Water renewal time

The analogous calculations are done for GGRF (see Table 12). Here spherical covariance model was used. The best model was obtained by using *Water renewal time* as dependent variable. *Depth*, *Salinity* and geographical information were included in the structure of mean model. Moreover, the

whole training sample was included constructing the covariance matrix and estimating prior probabilities.

**Table 12.** Correct classification probabilities for geostatistical model

Distance	Neighborhood structure				No spatial correlation	
	$d_{\max} = 0.1$	$d_{\max} = 0.2$	$d_{\max} = 0.3$	$d_{\max} = 0.4$		
Number of neighbors	[4-16]	[15-35]	[24-38]	[36-38]	0	
Depth	~1	<b>0.684</b>	0.658	0.658	0.658	0.447
	~1+XY	0.632	0.579	0.605	0.605	0.421
	~1+XY+WRT	<b>0.684</b>	0.579	0.579	0.553	0.526
	~1+XY+Salinity	0.658	0.632	0.658	0.632	0.579
	~1+XY+WRT+Salinity	0.658	0.632	0.632	0.605	0.500
Water renewal time	~1	0.711	0.711	0.684	0.684	0.553
	~1+XY	0.658	0.658	0.684	0.632	0.579
	~1+XY+Depth	0.605	0.579	0.579	0.579	0.553
	~1+XY+Salinity	0.658	0.684	0.658	0.658	0.526
	~1+XY+Depth+Salinity	0.684	0.711	0.737	<b>0.763</b>	0.526
Salinity	~1	<b>0.737</b>	0.684	0.711	<b>0.737</b>	0.632
	~1+XY	0.711	0.658	0.658	0.632	0.526
	~1+XY+Depth	0.684	0.632	0.632	0.632	0.368
	~1+XY+WRT	0.711	0.684	0.658	0.684	0.526
	~1+XY+Depth+WRT	0.632	0.553	0.605	0.605	0.553

# Conclusions

The closed-form expression of asymptotic covariance matrix for the geometrically anisotropic exponential covariance model is obtained.

The proposed non-parametric test for detecting geometric anisotropy is easy to implement and could be used as an alternative to the ones proposed by other authors. The simulation study has shown that the empirical power of the test is increasing with the increase of range parameter which determines the level of spatial correlation.

The derived AER and AEER formulas for complete parametric uncertainty could be applied as a target function constructing the optimality criterion for the spatial sampling design.

The closed-form expression of AER for the geometrically anisotropic exponential covariance model is obtained and could be used implementing the proposed formulas.

A simulation study to examine the accuracy of the proposed classifiers and to investigate the influence of population parameters to the AER was included. According to the results the conclusions could be made:

- Including spatial correlation into the model improves the efficiency of the proposed classifier;
- The approximation of expected risk is mostly affected by the Mahalanobis distance and the range parameter, which means that the stronger spatial correlation and the greater separation between classes yield the lower AER values;
- AER values are decreasing with the increase of anisotropy ratio  $\lambda$  only for low range values, and the influence is not significant;

- The symmetric set of training locations gives lower AER values than asymmetric one; the proposed AER is more precise when the classes are not mixed over the region.

The application of the proposed classification procedures to the mapping of presence and absence of zebra mussels in the Curonian Lagoon leads to the conclusions:

- Including spatial information into the model significantly increases the accuracy of classification;
- Involving only the nearest neighbors in estimation of prior probabilities gives better accuracy of classification;
- GMRF gives better results in comparison with and GGRF.

# Bibliography

1. Abt, M. (1999). Estimating the prediction mean squared error in Gaussian stochastic processes with correlation structure. *Scandinavian Journal of Statistics*, 26:563-578.
2. Allard, D., Senoussi, R., and Porcu, E. (2016). Anisotropy models for spatial data. *Mathematical Geosciences*, 48(3):305-328.
3. Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.
4. Atkinson, P. M., Lewis, P. (2000). Geostatistical classification for remote sensing: an introduction. *Computers & Geosciences*, 26(4):361-371.
5. Batsidis, A., Zografos, K. (2006). Discrimination of observations into one of two elliptic populations based on monotone training samples. *Metrika*, 64:221-241.
6. Batsidis, A., Zografos, K. (2011). Errors of misclassification in discrimination of dimensional coherent elliptic random field observations. *Statistica Neerlandica*, 65:446-461.
7. Berret, C., Calder, C. A. (2016). Bayesian spatial binary classification. *Spatial Statistics*, 16:72–102.
8. Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, 36:192-236.
9. Chiles, J. P., Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty*. Second Edition. John Wiley & Sons, New York.
10. Cressie, N. (1993). *Statistics for Spatial Data*. Wiley & Sons, New York.
11. Cressie, N., Hawkins, D. H. (1980). Robust estimation of the variogram: I. *Mathematical Geology*, 12(2):115-125.
12. Cressie, N. Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Wiley, Hoboken, NJ.
13. Diggle, P. J., Ribeiro, P. J. (2007). *Model-based Geostatistics*. Springer.

14. Diggle, P. J., Ribeiro, P. J., and Christensen, O. F. (2003). *An introduction to model-based geostatistics*. In: Møller J. (eds) *Spatial Statistics and Computational Methods*. Lecture Notes in Statistics, vol 173. Springer, New York, NY.
15. Dučinskas, K., Šaltytė-Benth, J. (2003). *Erdvinė statistika*. KU, Klaipėda.
16. Dučinskas, K. (1997). An asymptotic analysis of the regret risk in discriminant analysis under various training schemes. *Lithuanian Mathematical Journal*, 37(4):337-351.
17. Dučinskas, K. (2009). Approximation of the expected error rate in classification of the Gaussian random field observations. *Statistics and Probability Letters*, 79:138-144.
18. Dučinskas, K. (2009). Statistical classification of the observation of nuggetless spatial Gaussian process with unknown sill parameter. *Nonlinear Analysis: Modelling and Control*, 14(2):155-163.
19. Dučinskas, K. (2011). Error rates in classification of multivariate Gaussian random field observation. *Lithuanian Mathematical Journal*, 51:477-485.
20. Dučinskas, K., Borisenko, I., Šimkienė, I. (2013). Statistical classification of Gaussian spatial data generated by conditional autoregressive model. *Computational Science and Techniques*, 1(2):69-79.
21. Ecker, M. D., Gelfand, A. E. (2003). Spatial modelling and prediction under stationary non-geometric range anisotropy. *Environmental and Ecological Statistics*, 10:165-178.
22. Ecker, M. D., Gelfand, A. E. (1999). Spatial modelling and prediction under range anisotropy. *Environmental and Ecological Statistics*, 10:165-178.
23. Ferguson, T. A., (1996). *Course in large sample theory*. Chapman and Hall, London, UK.
24. Goovaerts, P. (1997). *Geostatistics for natural resources evaluation*. Oxford Univ. Press, New York.



25. Guan, Y., Sherman, M., Calvin, J. A. (2004). A nonparametric test for spatial isotropy using subsampling. *Journal of the American Statistical Association*, 99(467):810-821.
26. Gupta, A., Robinson, P. M. (2018). Pseudo maximum likelihood estimation of spatial autoregressive models with increasing dimension. *Journal of Econometrics*, 202:92-107.
27. Haining, R. (1990). *Spatial data analysis in the social and environmental sciences*. Cambridge University Press, Cambridge, UK.
28. Haining, R. (2004). *Spatial data analysis: Theory and practice*. Cambridge University Press, Cambridge, UK.
29. *Handbook of spatial statistics* / [edited by] Alan E. Gelfand ... [et al.]. p. cm. (Chapman & Hall/CRC handbooks of modern statistical methods), 2010.
30. Haskard, K. A., (2007). An anisotropic Matern spatial covariance model: RELM estimation properties. Doctoral thesis.
31. Hirst, D. (1996). Error-rate estimation in multiply-group linear discriminant analysis. *Technometrics*, 38:389-399.
32. Jin, X., Carlin, B. P., Banerjee, S. (2005). Generalized hierarchical multivariate CAR models for areal data. *Biometrics*, 61:950-961.
33. Johansson, J. O. (2001). Parameter estimation in the auto-binomial model using the coding and pseudo likelihood method approached with simulated annealing and numerical optimization. *Pattern recognition letters*, 22:1233-1246.
34. Isaaks, E. H., Srivastava, R. M. (1989). *An Introduction to applied geostatistics*. Oxford University Press.
35. Journel, A., Huijbregts, C. (1978). *Mining geostatistics*. Academic Press, New York.
36. Kharin, Y. (1996). Robustness in statistical pattern recognition. Dordrecht: Kluwer Academic Publishers.
37. Klein, R., Press, S. J. (1992). Adaptive Bayesian classification of spatial data. *Journal of the American Statistical Association*, 87:844-851.
38. Lahiri, S. N. (2003). *Resampling methods for dependent data*. Springer Series in Statistics.

39. Lawoko, C. R. O., McLachlan, G. L. (1985). Discrimination with autocorrelated observations. *Pattern Recognition*, 18(2):145-149.
40. Lindgren, F., Rue, H., Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 73(4):423-498.
41. Lu, H., Zimmerman, D. L. (2001). Testing for isotropy and other directional symmetry properties of spatial correlation. Preprint.
42. Lu, N. and Zimmerman, D. L. (2005). Testing for directional symmetry in spatial dependence using the periodogram. *Journal of Statistical Planning and Inference*, 129(1):369-385.
43. Lu, J., Zhang, L., (2010). Evaluation of Parameter Estimation Methods for Fitting Spatial Regression. *Forest Science*, 56(5):505-514
44. Magnus, J. R., Neudecker, H. (2002). Matrix differential calculus and applications in statistics and econometrics. Wiley, New York.
45. Maity, A., Sherman, M. (2012). Testing for spatial isotropy under general designs. *Journal of Statistical Planning and Inference*, 142:1081-1091.
46. Mardia, K. V., Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71:135-146.
47. Mardia, K. V. (1984). Spatial discrimination and classification maps. *Communications in Statistics - Theory and Methods*, 13:2181-2197.
48. Mardia, K. V. (1988). Multi-dimensional multivariate Gaussian Markov random fields with application to image processing. *Journal of Multivariate Analysis*, 24:265-284.
49. Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58:1246-1266.
50. McLachlan, G. J. (2004). *Discriminant analysis and statistical pattern recognition*. Wiley, New York.
51. Myers, D. E, Journel, A. (1990). Variograms with zonal anisotropies and noninvertible kriging systems. *Mathematical Geology*, 22(7):779-785.

52. Okamoto, M. (1963). An asymptotic expansion for the distribution of linear discriminant function. *Annals of Mathematical Statistics*, 34:1286-1301.
53. de Oliveira, V., Ferreira, M. A. R. (2011). Maximum likelihood and restricted maximum likelihood estimation for class of Gaussian Markov random fields. *Metrika*, 74(2):167-183.
54. Petersen, K. B., Pedersen, M. S. (2006). The matrix cookbook. <http://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf> (2018-07-29)
55. Pettitt, A. N., Weir, I. S., Hart, A. G. (2002). A conditional autoregressive Gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. *Statistics and Computing*, 12:353-367.
56. Rue, H., Held, L. (2005). *Gaussian Markov random fields: theory and applications*. Chapman & Hall, Boca Raton.
57. Sain, S. R., Cressie, N. (2007). A spatial model for multivariate lattice data. *Journal of Econometrics*, 140:226-259.
58. Šaltytė, J. (2001). The asymptotic expansion of the expected risk for the LDA of spatially correlated Gaussian observations (Doctoral dissertation).
59. Šaltytė, J., Dučinskas, K. (2002). Comparison of ML and OLS estimators in discriminant analysis of spatially correlated observations. *Informatica*, 13(2):297-238.
60. Šaltytė-Benth, J., Dučinskas, K. (2005). Linear discriminant analysis of multivariate spatial-temporal regressions. *Scandinavian Journal of Statistics*, 32:281-294.
61. Shekhar, S., Schrater, P. R., Vatsavai, R. R., Wu, W., and Chawla, S. (2002). Spatial contextual classification and prediction models for mining geospatial data. *IEEE Transactions on Multimedia*, 4(2):174-188.
62. Sherman, M. (2011). *Spatial statistics and spatio-temporal data: covariance functions and directional properties*. Wiley & Sons.
63. Schervish, M. J. (1984). Linear discrimination for three known normal populations. *Journal of Statistical Planning and Inference*, 10:167-175.

64. Schervish, M. J. (1981). Asymptotic expansions for the means and variances of error rates. *Biometrika*, 68:295-299.
65. Stabingienė, L., (2012). Vaizdų analizė naudojant Bajeso diskriminantines funkcijas (Doctoral dissertation).
66. Switzer, P. (1980). Extensions of linear discriminant analysis for statistical classification of remotely sensed satellite imagery. *Mathematical Geology*, 12(4):367-376.
67. Theodoritis, S., Kountroumbas K. (2009). *Pattern recognition*. Academic Press.
68. Wackernagel, H. (2003). *Multivariate geostatistics: an introduction with applications*. Springer-Verlag, Berlin Heidelberg.
69. Weller, Z. D. (2018). spTest: An R package implementing nonparametric tests of isotropy. *Journal of statistical software*, 83(4). doi:10.18637/jss.v083.i04
70. Weller, Z. D., Hoeting, J. A. (2016). A review of nonparametric hypothesis tests of isotropy properties in spatial data. *Statistical Science*, 31(3):305-324.
71. Williams, C. K. I., Barber, D. (1998). Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:1342-1351.
72. Wu, T. F., Lin, C. J., and Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975-1005.
73. Xu, T., Wang, J. (2013). An efficient model-free estimation of multiclass conditional probability. *Journal of Statistical Planning and Inference*, 143:2079-2088.
74. Yaglom, A. N. (1987). *Correlation theory of stationary and related random functions I*. Springer.
75. Younso, A. (2017). On the consistency of a new kernel rule for spatially dependent data. *Statistics and Probability Letters*, 131: 64-71.
76. Zaiko, A., Daunys, D. (2015). Invasive ecosystem engineers and biotic indices: Giving a wrong impression of water quality improvement? *Ecological Indicators*, 52:292-299.

77. Zhang, H., Zimmerman, D. L. (2005). Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika*, 92(4):921-936.
78. Zheng, Y., Zhu, J. (2012). On the asymptotics of maximum likelihood estimation for spatial linear models on a lattice. *The Indian Journal of Statistics*, 74(1):29-56.
79. Zhu, Z., Stein, M. L. (2006). Spatial sampling design for prediction with estimated parameters. *Journal of Agricultural, Biological, and Environmental Statistics*, 11:24-44.
80. Zhu, Z., Zhang, H. (2006). Spatial design under the infill asymptotic framework. *Environmetrics*, 17:323-337.
81. Zimmerman, D. L. (1993). Another look at anisotropy in geostatistics. *Mathematical Geology*, 25(4):453-70.
82. Zimmerman, D. L. (2006). Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. *Environmetrics*, 17:635-652.

# NOTES

# NOTES

Vilniaus universiteto leidykla  
Universiteto g. 1, LT-01513 Vilnius  
El. p. [info@leidykla.vu.lt](mailto:info@leidykla.vu.lt),  
[www.leidykla.vu.lt](http://www.leidykla.vu.lt)  
Tiražas 15 egz.