VILNIUS UNIVERSITY

LINA DREIŽIENĖ

CLASSIFICATION RISK IN LINEAR DISCRIMINATION OF SPATIAL GAUSSIAN DATA

Doctoral dissertation

Physical sciences, mathematics (01P)

Vilnius, 2018

The dissertation was written in 2014–2018 at Vilnius University.

Scientific supervisor: Prof. Dr. Kęstutis Dučinskas (Vilnius University, Physical sciences, Mathematics – 01 P)

Scientific Advisor: Prof. Dr. Marijus Radavičius (Vilnius University, Physical sciences, Mathematics – 01 P)

VILNIAUS UNIVERSITETAS

LINA DREIŽIENĖ

ERDVINIŲ GAUSO DUOMENŲ KLASIFIKAVIMO RIZIKA NAUDOJANT TIESINES DISKRIMINANTINES FUNKCIJAS

Daktaro disertacija

Fiziniai mokslai, matematika (01P)

Vilnius, 2018

Disertacija rengta 2014 – 2018 metais Vilniaus universitete.

Mokslinis vadovas: prof. dr. Kęstutis Dučinskas (Fiziniai mokslai, matematika - 01 P)

Mokslinis konsultantas: prof. dr. Marijus Radavičius (Fiziniai mokslai, matematika - 01 P)

Acknowledgements

Contents

Notations

$\mathbf{A}'$	transposed matrix
$\mathbf{I}_n$	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 $\mathbf{A}$ ir $\mathbf{B}$
$tr(\mathbf{A})$	trace of quadratic matrix $\mathbf{A}$
$\mathbf{R}$	matrix of spatial correlations
$\mathbf{\Sigma}$	covariance matrix
$\mathbf{X}$	design matrix of training sample
$\Phi(\cdot)$	standard normal distribution function
$\varphi(\cdot)$	standard normal distribution density function
$\mathbf{J}$	Fisher information matrix
$e(x)$	the Heaviside step function
$\delta(t)$	Dirac delta function
N_m	Multivariate normal distribution
AN	Asymptotically normal distribution
$N_{m\times n}$	Matrix-variate normal distribution
$vec(\mathbf{A})$	the vectorization of a matrix $\mathbf{A}$.
$vech(\mathbf{A})$	the half-vectorization of a symmetric matrix $\mathbf{A}$

Matrices and vectors are denoted with bald 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

MT – 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 observable feature. Observations made at different locations may be closer in value than measurements made at locations farther apart, for example, the elevation datasets have similar elevation values close to each other; the majority of minerals has location-dependent distributions; house prices and house value assessments are established by comparisons between a house and similar nearby houses; water (wind) polluters generate negative consequences for those downstream (downwind) of their locations, etc. (Griffith [10]). This phenomenon is called *spatial correlation (autocorrelation)*. The presence of spatial correlation means that a certain amount of information is shared and duplicated among neighbouring 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, Sherman [41]) and classification procedures.

The thesis focuses on discriminant analysis (sometimes called supervised statistical classification) for spatially correlated data. The main purpose is to use the Bayesian classification rule, taking into account the spatial correlation between the data values and assuming that the observation to be classified and the training sample are correlated; to apply the proposed classification procedure to assign Gaussian random field observation to one of several prescribed classes and to assess the risk (probability of misclassification in particular case) of classification. The risk of classification is an indicator of the effectiveness of a discriminant function and could be affected by rejecting hypothesis about existence of spatial correlation. The risk of classification could also be affected by anisotropy, the situation when spatial correlation is stronger in one direction than into another. In situation of geometric anisotropy two additional parameters, anisotropy ratio and anisotropy angle should be included.

Spatial statistics is a relatively young science; it emerged in early 1980s as a hybrid discipline of mining engineering, geology, mathematics, and statistics (Cressie [6]). Therefore there are no many studies in the field of discriminant analysis of spatially correlated data. Many authors (i.e. see Lawoko and McLachlan [35]; Kharin [33]) have investigated the problems concerned with classification of dependent observations (equicorrelated structures, autoregressive models) but Switzer [59] was the first to treat classification of spatial data. Mardia [43] extended this research by including spatial discrimination methods in forming the classification maps. The application of spatial contextual (or supervised) classification methods in geospatial data mining is considered by Shekhar et al. [56]. However these authors did not analyse classification risk. The comprehensive analysis of classification risk associated with discriminant analysis of uncorrelated data is presented in Dučinskas [12]. Later Šaltytė and Dučinskas [54] proposed the approximation of expected risk (in particular expected error rate) for classification of scalar Gaussian random field observation and generalised these results to the multivariate spatiotemporal model (Šaltytė-Benth, Dučinskas [55]). But the all above mentioned papers hold the assumption of independence between observation to be classified and training sample. In practice, such an assumption is often not reasonable, especially if observation of interest is close to the observations from training sample. This assumption was rejected in Dučinskas [13], [14]. It should be noted that here only the mean parameters and scale parameter of covariance function are assumed unknown for the geostatistical Gaussian random field models with continuous spatial support.

In this thesis the extension to the complete parametric uncertainty case in classification of univariate and multivariate geostatistical Gaussian random field observation is analysed. Also the extension to the multiclass case is performed. Finally the extensions to the problem of classification of Gaussian spatial data observed on lattice are made. Conditionally autoregressive models (CAR) are also comprehensively explored.

Aim and objectives (problems)

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

To accomplish the aim of the dissertation, the following tasks are raised:

- To derive formulas for classification risk and analytic expressions for its estimators for geostatistical GRF 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 GRF: the case of two classes.
- To derive the asymptotic approximation formula for the expected error rate of geostatistical GRF for 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 analyse the influence of different parameters to 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 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.

Actuality and novelty

The novelty of the results in the thesis:

- Closed-form expression of asymptotic covariance matrix for geometrically anisotropic exponential covariance model.
- Non-parametric test for spatial geometric anisotropy.
- Actual classification risk and asymptotic approximation of expected risk for complete parametric uncertainty case for univariate and multivariate two-class classification problem.
- Extension to classification problem of GMRF observed over lattice and specified by CAR model for univariate and multivariate two-class case.
- Multiclass classification problem of univariate GGRF observation.
- Closed-form expression of approximation of expected risk for geometrically anisotropic exponential covariance model.

Structure of the dissertation

Dissertation consists of 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 this dissertation concerned with discriminant analysis of spatial data. The last chapter introduces the numerical experiments and applications.

Dissemination of the results

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

Publications:

[A1] Dučinskas K., **Dreižienė L**. (2018). *Risks of Classification of the Gaussian Markov Random Field Observations*. Journal of Classification, 35. Accepted (DOI: 10.1007/s00357-)

[A2] **Dreižienė L**., Dučinskas K., Šaltytė-Vaisiauskė, L. (2018). *Statistical classification of multivariate conditionally autoregressive Gaussian random field observations.* Spatial Statistics, https://doi.org/10.1016/j.spasta.2018.03.006

[A3] Dučinskas K., **Dreižienė L**. (2016). *Expected error rates in classification of Gaussian CAR observations*. Computer Data Analysis and Modeling: Theoretical and applied stochastics: Proceedings of the 11th International Conference: Minsk, September 6-10, 2016. ISBN 978-985-553-366-6 p. 127-130.

[A4] **Dreižienė L**., Dučinskas K., Šaltytė-Vaisiauskė, L. (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. Vol. 26, p. 78–81.

[A5] **Dreižienė L**., Dučinskas K., Paulionienė L. (2015). *Correct Classification Rates in Multi-Category Discriminant Analysis of Spatial Gaussian Data*. Open Journal of Statistics. Vol. 5, no.1. p. 21-26.

[A6] Dučinskas K., **Dreižienė L**., Zikarienė E. (2015). *Multiclass classification of the scalar Gaussian random field observation with known spatial correlation function*. Statistics & Probability Letters. Vol. 98, p.107–114.

[A7] Dučinskas K., Zikarienė E., **Dreižienė 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-Science and Technology Publications, 161-166.

[A8] Dučinskas K., **Dreižienė L**. (2013). *Optimal classification of multivariate GRF observations*. Multivariate Statistics: Theory and Applications, World Scientific, 61-72. ISBN 978-981-4449-39-7 (chapter of Book).

[A9] **Dreižienė L.**, Karaliutė M. (2012). *The influence of training sample size on the expected error rate in spatial classification*. LMD darbai. Vilnius: MII; t 53, p. 24–29, ISSN 0132-2818.

[A10] **Dreižienė L**. (2011). *Linear discriminant analysis of spatial Gaussian data with estimated anisotropy ratio.* Lietuvos matematikos rinkinys. LMD darbai. Vilnius: MII; t. 52, p. 315–320, ISSN 0132-2818

[A11] Dučinskas K., **Dreižienė 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; Vol. 7, p. 212-217, ISSN 1878-0296.

[A12] Dučinskas K., **Dreižienė 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; Vol. 40, No. 1, 2, p. 25-36, ISSN 1026-597X

[A13] Dučinskas K., **Dreižienė 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; Vol. 1, p. 33-36, ISBN 978-9955-476-847-2.

[A14] Dučinskas K., **Dreižienė** L. (2010). *Nonparametric test for spatial geometric anisotropy*. Lietuvos matematikos rinkinys. LMD darbai. Vilnius: MII; t. 51, p. 397-401, ISSN 0132-2818.

[A15] **Dreižienė 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, p. 101-105.

[A16] Dučinskas K., **Dreižienė L**. (2007). *Effect of anisotropy coeficient on error rates of linear discriminant functions.* Lietuvos matematikos rinkinys. Vilnius: MII, 47 (spec. Nr.), p. 359-363.

[A17] **Budrikaitė L**. (2005). *Modeling of zonal anisotropic semivariogramos.* Lietuvos Matematikos rinkinys. 45 (spec. nr.), p. 339-342.

[A18] **Budrikaitė L**., Dučinskas K. (2005). *Modelling of geometric anisotropic spatial variation.* Mathematical Modelling and Analysis. Vilnius: Technika, p. 361-366.

[A19] **Budrikaitė L.**, Dučinskas K. (2004). *Forms of anisotropy for spatial variograms*. Lietuvos matematikos rinkinys, 44 (spec. nr.), p. 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, the linear model and its components to be used in the thesis are described. This section also contains description of anisotropic data and commonly found forms of anisotropy, briefly discusses the methods for determining 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 [42] are presented. In the subsection 1.2.1 the ML estimators for geometrically anisotropic covariance are obtained. Finally, a non-parametric test for 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 [6]; Diggle and Ribeiro [8]; Chiles and Delfiner [5]). Traditionally (e.g. Cressie [6]), 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**), subclass of lattice model for Gaussian data. 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 spatial functions of first two moments determine the complete distribution (Haskard 2007). Gaussian Markov random fields are discrete domain GRFs equipped with a Markov property.

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 location $\mathbf{s}$. Assume that $Z(\mathbf{s})$ is a Gaussian random field observation, then the model of observation $Z(\mathbf{s})$ is given by a general linear model

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

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 [25]) which is generated by a zero-mean GRF $\{\varepsilon(\mathbf{s}) : \mathbf{s} \in D\}$ with covariance function defined by the model for all $\mathbf{s}_1, \mathbf{s}_2 \in D$

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

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 ant let $\{\mathbf{s}_i,\ i = 1..n\}$ be a finite set of spatial locations. Then for real numbers $\{a_i, i = 1..n\}$ 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 $\mathbf{x}'(\mathbf{s}) = (1, x_1(\mathbf{s}), \dots, x_q(\mathbf{s}))$ is a vector of non-random covariates, and $\boldsymbol{\beta} = \left(\beta_0, \dots, \beta_q\right)'$ is a vector of parameters.

In the dissertation the following parametric linear mean models are used:

- *Constant mean model*, if $q = 0$ and $\beta_0 \equiv \mu = const$.
- *Trend surface mean model*, if $x_j(\mathbf{s})$, $j = 1..q$, are polynomials of spatial coordinates.
- *Regression mean model*, if $x_j(\mathbf{s}), j = 1..q$, are regressors (independent variables).

An essential concept related to the analysis of spatial processes is their stationarity or homogeneity (Yaglom [66]). 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. Spatial process $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$ is called *stationary* or *homogeneous* if it satisfies the following properties: $E(Z(\mathbf{s})) \equiv \mu = const$ for all $\mathbf{s} \in D$; $C(\mathbf{s}_1, \mathbf{s}_2) = C(\mathbf{s}_1 - \mathbf{s}_2) = cov\left(Z(\mathbf{s}_1), Z(\mathbf{s}_2)\right)$ for all $\mathbf{s}_1, \mathbf{s}_2 \in D$ is a function of distance only; $E\{|Z(\mathbf{s})|^2\} < \infty$ for all $\mathbf{s} \in D$. These fields are also known as *second-order*, *wide-sense* or *weakly homogeneous* fields (Yaglom [66]; Haining [24]; Cressie [6]).

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 $R(\mathbf{s}) = C(\mathbf{s})/C(\mathbf{0})$.

Strong stationarity implies *stationarity* in a wide sense. The vice versa is in general false but it holds for the case of Gaussian processes.

A more general assumption than *stationarity* leads to modelling spatial variation using the *semivariogram* and is called an *intrinsic stationarity* (Cressie [6], Haining [25]).

Definition 1.3. Suppose $var\big(Z(\mathbf{s}_1) - Z(\mathbf{s}_2)\big) = 2\gamma(\mathbf{s}_1 - \mathbf{s}_2)$ for all $\mathbf{s}_1, \mathbf{s}_2 \in D$. The quantity $2\gamma(\cdot)$ which is a function of increment $\mathbf{s}_1 - \mathbf{s}_2$ is called a *variogram* and $\gamma(\cdot)$ a *semivariogram* (Cressie [6]).

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

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

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

Definition 1.5. The process $\{Z(\mathbf{s}) \colon \mathbf{s} \in D\}$ 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

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

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

- The quantity $\tau^2 = C(\mathbf{0}) - \lim_{|\mathbf{h}|\to +0} C(\mathbf{h})$ 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 this dissertation by α.

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 [73] anisotropy can take three forms: *sill anisotropy*, *range anisotropy* and *nugget anisotropy*. Range anisotropy is usually specified as either *geometric* (elliptical) *range anisotropy* or *non-geometric range anisotropy*. Chiles and Delfiner [5], 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 [32], Goovaerts [21], Cressie [6], Wackernagel [61]). Following these authors *geometric anisotropy* occurs when the range, but not the sill, of the covariance changes in different directions. *Zonal anisotropy* exists when the sill of covariance/semivariogram function changes with direction (Wackernagel [61]). Directional semivariograms corresponding different types of anisotropy are depicted in Figure 1. Forms of anisotropy are also discussed in Budrikaitė and Dučinskas [A19].

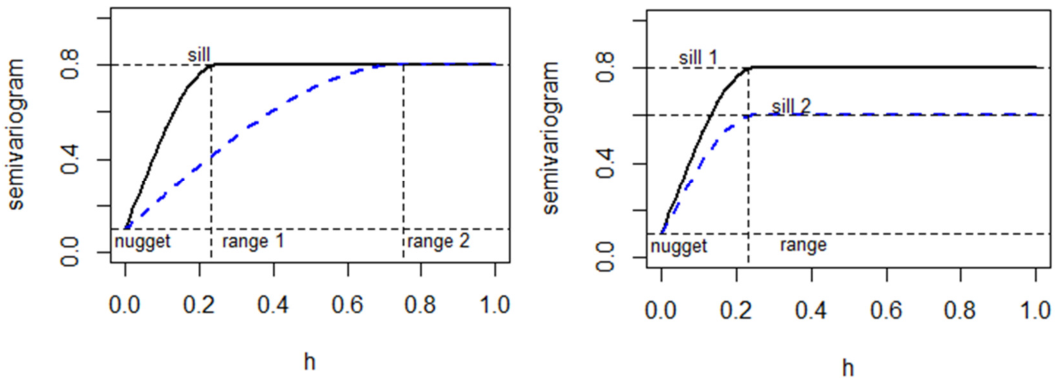

Figure 1. Directional semivariograms in the case of geometric and zonal anisotropy

In this thesis we will focus on the *geometric anisotropy*. 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 major and minor axes of ellipse correspond to the largest and shortest ranges of directional semivariograms. 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 Haskard [27], Sherman [57]. The issues of modelling geometric and zonal spatial variation are also analysed in Budrikaitė and Dučinskas [A18] and Budrikaitė [A17].

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

- *The anisotropy ratio*, denoted by λ, is equal to the ratio of the lengths of the two principal ellipse axes. This parameter is always positive, it could be greater or less than 1, with isotropy corresponding to $\lambda = 1$.
- *The anisotropy angle, φ,* is the angle of rotation which is made by major axis of the ellipse (which is known as 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 due to ease of computation and simpler interpretation. But in many applications spatial isotropy is not reasonable assumption, thus it is very important to verify the existence of anisotropy. There are formal and non-formal methods (graphical techniques), to determine anisotropy. The examples of non-formal methods are: assessing *directional semivariograms*, drawing *rose diagrams* (Isaak, Srivastava [31]; Ecker, Gelfand [19]), *semivariogram* or *contour maps* (Isaak, Srivastava [31]). 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 [37], [38], Guan et al. [22], Maity and Sherman [41]. An original, simple non-parametric test statistic based on directional empirical semivariograms was proposed by Dučinskas and Dreižienė [A14] and is presented in section 1.3.

In Weller and Hoeting [63] a comprehensive review of the different non-parametric methods for testing isotropy is presented. Several of the aforementioned tests were recently implemented in the R package *spTest*, available on CRAN (Weller [62]).

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* (MT) 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 [6], Gupta and Robinson [23], Johansson [30]), *Coding method* (Besag [4], Johansson [30]), *Bayesian method* (Lu, Zhang [39], Dučinskas, Šaltytė [11]), which could be used for estimation of unknown parameters. In this thesis the ML method for theoretical results will be used.

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

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

$$\mathbf{X} = \begin{pmatrix} 1 & x_1(\mathbf{s}_1) & \cdots & x_q(\mathbf{s}_1) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1(\mathbf{s}_n) & \cdots & x_q(\mathbf{s}_n) \end{pmatrix}. \tag{1.2}$$

The only full-rank cases, i.e. $rank(\mathbf{X}) = q + 1$, will be analysed in this thesis. The term $\boldsymbol{\varepsilon}_n$ has a multivariate Gaussian distribution $N_n(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$, where $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ denotes a covariance matrix which for a random vector $\mathbf{Z}_n$ is specified as

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \begin{pmatrix} C(\mathbf{s}_1,\mathbf{s}_1;\boldsymbol{\theta}) & C(\mathbf{s}_1,\mathbf{s}_2;\boldsymbol{\theta}) & C(\mathbf{s}_1,\mathbf{s}_3;\boldsymbol{\theta}) & \dots & C(\mathbf{s}_1,\mathbf{s}_n;\boldsymbol{\theta}) \\ C(\mathbf{s}_2,\mathbf{s}_1;\boldsymbol{\theta}) & C(\mathbf{s}_2,\mathbf{s}_2;\boldsymbol{\theta}) & C(\mathbf{s}_2,\mathbf{s}_3;\boldsymbol{\theta}) & \dots & C(\mathbf{s}_2,\mathbf{s}_n;\boldsymbol{\theta}) \\ C(\mathbf{s}_3,\mathbf{s}_1;\boldsymbol{\theta}) & C(\mathbf{s}_3,\mathbf{s}_2;\boldsymbol{\theta}) & C(\mathbf{s}_3,\mathbf{s}_3;\boldsymbol{\theta}) & \dots & C(\mathbf{s}_3,\mathbf{s}_n;\boldsymbol{\theta}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C(\mathbf{s}_n,\mathbf{s}_1;\boldsymbol{\theta}) & C(\mathbf{s}_n,\mathbf{s}_2;\boldsymbol{\theta}) & C(\mathbf{s}_n,\mathbf{s}_3;\boldsymbol{\theta}) & \dots & C(\mathbf{s}_n,\mathbf{s}_n;\boldsymbol{\theta}) \end{pmatrix}. \tag{1.3}$$

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

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

The covariance matrix for geometric anisotropy case could be factorised as $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{\Gamma}$, where σ^2 is a scale parameter and $\boldsymbol{\Gamma} = \rho\mathbf{I}_n + \mathbf{R}$ is a standardized covariance matrix. Here $\rho = \tau^2/\sigma^2$ is a *relative nugget* or *noise to signal variance* (Diggle, Ribeiro [8]) 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 scale parameter σ^2 are unknown and $\mathbf{\Gamma}$ 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 $\mathbf{\Psi} = (\boldsymbol{\beta}', \boldsymbol{\theta}')'$.

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

$$\ln L = (2\pi)^{-n/2} - \frac{1}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}(\mathbf{Z}_n - \mathbf{X}\boldsymbol{\beta})'\mathbf{\Sigma}^{-1}(\mathbf{Z}_n - \mathbf{X}\boldsymbol{\beta}). \tag{1.4}$$

If the vector of mean parameters β is unknown then ML estimator is

$$\widehat{\boldsymbol{\beta}}_{ML} = (\mathbf{X}'\mathbf{\Sigma}^{-\mathbf{1}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{\Sigma}^{-\mathbf{1}}\mathbf{Z}_n \sim \boldsymbol{N}_{q+1}(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{\Sigma}^{-\mathbf{1}}\mathbf{X})^{-\mathbf{1}}). \tag{1.5}$$

It is obvious that the estimator (1.5) is unbiased and efficient since $cov(\widehat{\boldsymbol{\beta}}_{ML}) = \mathbf{J}_\beta^{-1}$, where $\mathbf{J}_\beta = -E\left(\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right) = \mathbf{X}'\mathbf{\Sigma}^{-\mathbf{1}}\mathbf{X}$ defines the information matrix.

In the case $\mathbf{\Sigma} = \sigma^2\mathbf{\Gamma}$, ML estimators for $\boldsymbol{\beta}$ and σ^2 are

$$\widehat{\boldsymbol{\beta}}_{ML} = (\mathbf{X}'\mathbf{\Gamma}^{-\mathbf{1}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Gamma}^{-\mathbf{1}}\mathbf{Z}_n, \tag{1.6}$$

$$\hat{\sigma}_{ML}^2 = \frac{1}{n}\left(\mathbf{Z}_n - \mathbf{X}\widehat{\boldsymbol{\beta}}_{ML}\right)'\mathbf{\Gamma}^{-1}\left(\mathbf{Z}_n - \mathbf{X}\widehat{\boldsymbol{\beta}}_{ML}\right). \tag{1.7}$$

It is easy to show that

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

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 σ^2 is biased since $E(\hat{\sigma}^2_{ML}) = \sigma^2(n-q-1)/n$ and $\hat{\sigma}^2_{ML} \sim \sigma^2\chi^2_{n-q-1}/n$.

In this dissertation the unbiased estimator of σ^2 will be used

$$\tilde{\sigma}^2_{ML} = \hat{\sigma}^2_{ML} n/(n-q-1). \qquad (1.8)$$

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 [43].

Theorem 1.1. (Mardia and Marshall [43]). Suppose $\mathbf{Z}_n \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$. Let $\lambda_1 \leq \cdots \leq \lambda_n$ be the eigenvalues of the covariance matrix $\boldsymbol{\Sigma}$ and let those of $\boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_{ij}$ be λ^i_k and $\lambda^{ij}_k, k = 1..n$, respectively with $|\lambda^i_1| \leq \cdots \leq |\lambda^i_n|$ and $|\lambda^{ij}_1| \leq \cdots \leq |\lambda^{ij}_n|$ for $i,j = 1..p$. Here $\boldsymbol{\Sigma}_i = \partial\boldsymbol{\Sigma}/\partial\boldsymbol{\theta}_i, \boldsymbol{\Sigma}_{ij} = \partial^2\boldsymbol{\Sigma}/\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j$. Moreover, suppose that as $n \to \infty$:

(a) $\lim \lambda_n = C < \infty, \lim|\lambda^{ij}_n| = C_{ij} < \infty \ \forall\, i,j = 1..p$;

(b) $\|\boldsymbol{\Sigma}_i\|^{-2} = O\left(n^{-\frac{1}{2}-\delta}\right)$ for some $\delta > 0$, for $i = 1..p$;

(c) $\forall\, i,j = 1..p,\ a_{ij} = lim\left\{\frac{t_{ij}}{(t_{ii}t_{jj})^{1/2}}\right\}$ exists, where

$t_{ij} = tr\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_j\right)$ and $\mathbf{A} = (a_{ij})$ is a non-singular matrix;

(d) $\lim(\mathbf{X}'\mathbf{X})^{-1} = 0$.

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

$$\widehat{\boldsymbol{\Psi}} \sim AN(\boldsymbol{\Psi}, \mathbf{J}^{-1}). \qquad (1.9)$$

Here $\mathbf{J} = \mathbf{J}_\beta \oplus \mathbf{J}_\theta$ is an information matrix with components

$$\mathbf{J}_\beta = \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}, \qquad (1.10)$$

$$(\mathbf{J}_\theta)_{ij} = tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_j)/2. \qquad (1.11)$$

It should be noted that the estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$ are asymptotically independent and the asymptotic covariance matrix is simply the inverse of Fisher information matrix. This theorem is proved under the *increasing domain* asymptotic framework in which the minimum distance between sampling points is bounded away from zero and thus the spatial domain of observation is unbounded (Zhang, Zimmerman [69]). 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 [34]). If the *infill asymptotics* framework is considered, the spatial domain is fixed (bounded) and locations of observations get denser, as the number of observations increases (Cressie [6]).

The *mixed domain* asymptotics or *hybrid asymptotics* (Zheng, Zhu 2012) is a combination of *increasing domain* and *infill asymptotics*. Here the sampling region grows to infinity and at the same time the distance between neighbouring sampling sites goes to zero (Lahiri [34]).

The asymptotic behaviour of spatial covariance parameter estimators can be different under the different asymptotic spatial frameworks. For example, Zheng and Zhu [70] 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.

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 [19])

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \tau^2 \mathbf{I}_n + \sigma^2 \mathbf{R}, \tag{1.12}$$

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

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

Lemma 1.1. Consider $\mathbf{Z}_n \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ and assume that conditions formulated in **Theorem 1.1** hold. Then the ML estimator $\widehat{\boldsymbol{\theta}}$ satisfies the following properties (as $n \to \infty$):

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

Similarly

$$\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta} \text{ and } \widehat{\boldsymbol{\beta}} \sim AN_{q+1}(\boldsymbol{\beta}, \mathbf{J}_\beta^{-1}).$$

The asymptotic covariance matrix $\mathbf{J}_\theta$ is a symmetric $p \times p$ matrix (in the case of geometric anisotropy we have 5 covariance parameters, $p = 5$) with elements defined in (1.11). The matrix $\mathbf{J}_\beta$ is defined in (1.10).

In order to build the asymptotic covariance matrix $\mathbf{J}_\theta$ for (1.12), we need to find first order partial derivatives of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ with respect to the $i - th$ covariance parameter, $i = 1..5$. To get the analytic expressions of $\boldsymbol{\Sigma}_i$ we choose anisotropic exponential parametric model of covariance function

$$\mathrm{C}(h, \boldsymbol{\theta}) = \begin{cases} \tau^2 + \sigma^2, \ \ h = 0 \\ \sigma^2 \, exp\left\{-\sqrt{(h_u)^2 + \lambda^2 (h_v)^2}/\alpha\right\}, h \neq 0' \end{cases} \tag{1.13}$$

where

$h_u = \left(h_x \cos\varphi + h_y \sin\varphi\right)/\alpha_{max}$ and $h_v = \left(-h_x \sin\varphi + h_y \cos\varphi\right)/\alpha_{max}$, $h_x = x_i - x_j, h_y = y_i - y_j, i, j = 1..n$. x_i and y_i denotes the X and Y coordinates of location $\mathbf{s}_i$. To get (1.13) the linear coordinate transformation procedure was applied. More about coordinate transformation procedure can be found in Sherman [57], Chiles and Delfiner [5].

Lemma 1.2. Consider the exponential covariance model (1.13). Then the first order partial derivatives, $\mathbf{\Sigma}_i = \partial\mathbf{\Sigma}/\partial\boldsymbol{\theta}_i, i = 1..5$, are

$$\mathbf{\Sigma}_1 = \partial\mathbf{\Sigma}/\partial\tau^2 = \mathbf{I}_n, \tag{1.14}$$

$$\mathbf{\Sigma}_2 = \partial\mathbf{\Sigma}/\partial\sigma^2 = \mathbf{R}, \tag{1.15}$$

$$\mathbf{\Sigma}_3 = \partial\mathbf{\Sigma}/\partial\alpha = \frac{\sigma^2}{\alpha^2}\mathbf{R} \circ \mathbf{H}_\alpha, \tag{1.16}$$

$$\mathbf{\Sigma}_4 = \partial\mathbf{\Sigma}/\partial\lambda = -\frac{\lambda\sigma^2}{\alpha}\mathbf{R} \circ \mathbf{H}_\lambda, \tag{1.17}$$

$$\mathbf{\Sigma}_5 = \partial\mathbf{\Sigma}/\partial\varphi = \frac{\sigma^2(\lambda^2-1)}{\alpha}\mathbf{R} \circ \mathbf{H}_\varphi. \tag{1.18}$$

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

$$(H_\alpha)_{ij} = \sqrt{\left(h_u^{ij}\right)^2 + \lambda^2\left(h_v^{ij}\right)^2}, \tag{1.19}$$

$$(H_\lambda)_{ij} = \left(h_v^{ij}\right)^2 / \sqrt{\left(h_u^{ij}\right)^2 + \lambda^2\left(h_v^{ij}\right)^2}, \tag{1.20}$$

$$\left(H_\varphi\right)_{ij} = h_u^{ij} h_v^{ij} / \sqrt{\left(h_u^{ij}\right)^2 + \lambda^2\left(h_v^{ij}\right)^2}, \tag{1.21}$$

$$h_u^{ij} = \left(h_x^{ij} \cos\varphi + h_y^{ij} \sin\varphi\right)/\alpha, h_v^{ij} = \left(-h_x^{ij} \sin\varphi + h_y^{ij} \cos\varphi\right)/\alpha,$$
$$h_x^{ij} = x_i - x_j,\ h_y^{ij} = y_i - y_j, i,j = 1..n.$$

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 [74]). They will also be applied in the thesis for solving classification problems of spatial data.

1.3. Non-parametric test for spatial geometric anisotropy

A conventional practice when checking for isotropy is to assess plots of empirical semivariograms. However these graphical techniques are open to interpretation. Guan et al. [22] have proposed a formal approach to test isotropy which is based on the asymptotic joint normality of empirical semivariograms for multiple directions. An L_2-consistent subsampling estimator for 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 the simpler test statistic in Gaussian case under the assumption of independence of the classical semivariogram estimators. These results are published in Dučinskas, Dreižienė [A14].

Suppose that spatial data are observations of a 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 $\frac{1}{2}var\big(Z(\mathbf{s}_1) - Z(\mathbf{s}_2)\big) = \gamma(\mathbf{s}_1 - \mathbf{s}_2)$ for all $\mathbf{s}_1, \mathbf{s}_2 \in D$ is called a semivariogram. We consider the geometric anisotropy case which means that semivariograms have the same nugget, same sill but different ranges in to perpendicular directions (see Wackernagel, 2003).

Denote by $\mathbf{S}_n = \{\mathbf{s}_i \in D, i = 1..n\}$ the set of locations where GRF is observed and use the classical semivariogram estimator proposed by Matheron (1962), based on the method of moments (see Cressie [6])

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{|N(\mathbf{h})|}\sum_{(\mathbf{s}_i,\mathbf{s}_j)\in N(\mathbf{h})}\Big(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\Big)^2. \qquad (1.22)$$

Here $\mathrm{N}(\mathbf{h})$ denotes all pairs $(\mathbf{s}_i, \mathbf{s}_j)$ for which $\mathbf{s}_\mathrm{i}, \mathbf{s}_\mathrm{j} \in \mathbf{S}_n$, $\mathbf{s}_\mathrm{i} - \mathbf{s}_\mathrm{j} = \mathbf{h}$ 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 $|h_i| = |h_{i+K}|, i = 1..K$.

The hypothesis of isotropy is expressed as

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

Rejecting this hypothesis means accepting 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 (1.22) obtained 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. [22]. Under some regularity conditions, it was proved that

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

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

Under the hypothesis of isotropy, there exists a full row rank matrix $\mathbf{A}$ such that $\mathbf{A}\boldsymbol{\Gamma} = \mathbf{0}$ (Lu, Zimmerman [37]). $\mathbf{A}$ is called a contrast matrix. Then under the hypothesis of isotropy it follows from continuous mapping theorem that

$$n\left(\mathbf{A}\hat{\boldsymbol{\Gamma}}\right)'\left(\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}}\mathbf{A}'\right)^{-1}\left(\mathbf{A}\hat{\boldsymbol{\Gamma}}\right) \overset{D}{\rightarrow} \chi_r^2 \text{ as } n \rightarrow \infty, \tag{1.23}$$

where r denotes the row rank of $\mathbf{A}$.

Following Cressie [6] we have

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

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

We propose the following estimator of $\mathbf{\Sigma}_{\Gamma}$

$$\mathbf{\Sigma}_{\Gamma} = ndiag\left(\frac{2\gamma^2(h_1)}{|N(h_1)|}, \dots, \frac{2\gamma^2(h_{2K})}{|N(h_{2K})|}\right)$$

and use it to form the test statistic

$$\widehat{\mathbf{T}} = n\left(\mathbf{A}\widehat{\mathbf{\Gamma}}\right)'(\mathbf{A}\mathbf{\Sigma}_{\mathbf{\Gamma}}\mathbf{A}')^{-1}\left(\mathbf{A}\widehat{\mathbf{\Gamma}}\right). \tag{1.24}$$

The proposed test statistic according the Slutsky theorem (Ferguson [20]) follows asymptotic chi-square distribution, $\widehat{\mathbf{T}} \sim \chi_r^2$.

So the hypothesis H_0 is to be rejected if $\widehat{\mathbf{T}} > \chi_{r,p}^2$, where $\chi_{r,p}^2$ is the p-critical value of a chi-squared distribution with r degrees of freedom. If $\widehat{\mathbf{T}} \leq \chi_{r,p}^2$, the hypothesis H_0 is to be accepted.

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

Chapter 2

Classification of spatial GRF observations

This chapter contains the main results of the dissertation. 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 different number of populations are presented. In Section 2.2 the problem of classification of GGRF observation is analysed. 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 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 univariate and multivariate cases are considered. Certain parts of this chapter are published in [A1]-[A4], [A6]-[A8], [A10]-[A13].

2.1. Elements of linear discriminant analysis (main concepts and definitions)

Consider the problem of classification of a single GRF $\{Z(\mathbf{s}): \mathbf{s} \in D \subset \mathbb{R}^d\}$ observation $Z_0 = Z(\mathbf{s}_0)$, $\mathbf{s}_0 \in D$, to one of m populations $\Omega_l, l = 1..m$ (also known as classes or categories). Following Beret and Calder (2016) the spatial location ($\mathbf{s}_0$) of the observation to be classified will be called the *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 Ω_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 terms $\varepsilon(\mathbf{s})$ is generated by the same zero-mean GRF $\{\varepsilon(\mathbf{s}): \mathbf{s} \in \mathrm{D} \subset \mathbb{R}^d\}$ with covariance function defined by the model for all $\mathbf{s}, \mathbf{s} + \mathbf{h} \in D$, $C(\mathbf{h}) = cov(\varepsilon(\mathbf{s}), \varepsilon(\mathbf{s} + \mathbf{h}))$, $l = 1..m$.

Denote by $\mathbf{S}_n = \{\mathbf{s}_i \in D, i = 1..n\}$ the set of locations where 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 training sample (see Shekhar et al. [56]). 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 Ω_l, $l = 1..m$.

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

$$\mathbf{T} = \mathbf{M} + \mathbf{E}, \tag{2.1}$$

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

Since the assumption of independence of training sample and the FO Z_0 was rejected, the conditional $Z(\mathbf{s}_0)$ distributions will be used.

The risk of classification: general definitions

Let $p_l(Z_0|\mathbf{t}), l = 1..m$ denote the class-conditional probability density function of Z_0, given $\mathbf{T} = \mathbf{t}$ in the population Ω_l, describing the distribution of the feature vector in each population (Theodoridis [60]). The *losses of classification* or the *loss function* when an object from the population l is allocated to the population k, is denoted by $L(l, k), l, k = 1..m$. The loss $L(l, l)$ correspond to 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 $\pi_l, l = 1..m$, $\sum_{l=1}^{m} \pi_l = 1$, depend on $\mathbf{s}_0$, but do not depend on training sample $\mathbf{T}$.

(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 $D_t(\bullet): \mathcal{Z} \to \{1..m\}$. Then the *expected loss* or *conditional risk* (McLachlan [46]) of random observation Z_0 given $\mathbf{t}$ from population Ω_l, by prescribed classification rule is given by

$$R(l, D_t(\bullet)) = E_{Z_0|t,l}\{L(l, D_t(Z_0))\}$$

and the *total risk* is the total excepted losses

$$R(D_t(\bullet)) = \sum_{l=1}^{m} \pi_l R(l, D_t(\bullet)). \tag{2.2}$$

The rule minimizing the total risk (2.2) is said to be *Bayes classification rule* (McLachlan [46], Anderson [3]) and will be denoted as $D_t^B(\bullet)$ and for the observation Z_0 it could be expressed as

$$D_t^B(Z_0) = arg\, min_{\{k=1..m\}}\{\sum_{l=1}^{m} \pi_l p_l(Z_0|\mathbf{t}, \mathbf{\Psi}) L(l, k)\}. \tag{2.3}$$

Recall that $\mathbf{\Psi}$ denotes the combined vector of unknown population parameters. Then *Bayes risk*, associated with Bayes classification rule (2.3) is

$$R^B = R(D_t^B(\bullet)) = \sum_{l=1}^{m} \pi_l E_{Z_0|\mathbf{t},l} L(l, D_t^B(Z_0)). \tag{2.4}$$

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

$$G_{lk}^B(Z_0) = \sum_{j=1}^{m} \pi_j\, p_j(Z_0|\mathbf{t}, \mathbf{\Psi}) d(j, l, k), \tag{2.5}$$

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

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

Then the *Bayes risk*, defined in (2.4) and associated with (2.5) could be expressed as (see Dučinskas [12])

$$R^B(\mathbf{\Psi}) = \sum_{l,k=1}^{m} \pi_l E_{Z_0|t,l} L(l,k) \prod_{j=1,j\neq k} e(G_{kj}^B(Z_0)). \tag{2.6}$$

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

Let $\widehat{\mathbf{\Psi}}$ denote the vector of parameters estimates. Thus replacing the vector of the unknown parameters of G_{lk}^B by the vector of estimates we get pairwise PBDF

$$\hat{G}_{lk}^B(Z_0) = G_{lk}^B\big(Z_0, \widehat{\mathbf{\Psi}}\big). \tag{2.7}$$

Definition 2.1. The *actual risk*, given $\mathbf{T} = \mathbf{t}$, for PBDF (2.7) is defined as

$$R^B(\widehat{\mathbf{\Psi}}) = \sum_{l,k=1}^{m} \pi_l E_{Z_0|t,l} L(l,k) \prod_{j=1,j\neq k} e(\hat{G}_{kj}^B(Z_0)) \,. \tag{2.8}$$

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

$$ER = E_T\big(R^B(\widehat{\mathbf{\Psi}})\big).$$

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 e al. [9]).

The risk of classification for $m = 2$

The two-class case is a special case. The most of the author's publications ([A1], [A2], [A3], [A7], [A8], [A9], [A10], [A11], [A12], [A13], [A15], [A16]) deal with two-class case, that is the reason this case is analysed separately.

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

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

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

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 $W_{lk}^B(Z_0, \mathbf{\Psi})$.

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, $W_{12}^B(Z_0)$ is replaced by $W^B(Z_0)$. Then the expression of *Bayes discriminant function* has the form

$$W^B(Z_0, \mathbf{\Psi}) = ln\left(\frac{p_1(Z_0|\mathbf{t}, \mathbf{\Psi})}{p_2(Z_0|\mathbf{t}, \mathbf{\Psi})}\right) + \gamma^*\,, \tag{2.9}$$

here $\gamma^* = \ln\left(\frac{g_1}{g_2}\right)$. According to (2.9) the observation Z_0, given $\mathbf{T} = \mathbf{t}$ is allocated to the population Ω_1, if $W^B(Z_0, \mathbf{\Psi}) \geq 0$ and to the population Ω_2 otherwise.

Then the *Bayes risk* could be evaluated by

$$R^B(\mathbf{\Psi}) = \sum_{l=1}^2 \sum_{k=1}^2 \pi_l L(l,k) P_{lk}, \tag{2.10}$$

where for $l, k = 1,2$, $P_{lk} = P_l((-1)^k W^B(Z_0, \mathbf{\Psi}) < 0)$. 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 (2.10) to a single-sum function

$$R^B(\mathbf{\Psi}) = \sum_{l=1}^2 (\pi_l L(l,l) + g_l PM_l), \tag{2.10a}$$

where $PM_l = P_l((-1)^l W^B(Z_0, \mathbf{\Psi}) > 0)$.

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

$$R^B(\widehat{\mathbf{\Psi}}) = \sum_{l=1}^2 \sum_{k=1}^2 \pi_l L(l,k) \hat{P}_{lk}, \tag{2.10b}$$

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

The probability of misclassification

We analyse 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 δ_{lk} is the Kronecker delta, the risk becomes the *probability of misclassification* or *error rate* (Dučinskas [12]). Such a loss function will be called a *zero-one loss function* and it 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

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

In the following we will use the equivalent discriminant functions

$$W_{lk}^B(Z_0, \boldsymbol{\Psi}) = ln\left(\frac{p_l(Z_0 | \mathbf{t}, \boldsymbol{\Psi})}{p_k(Z_0 | \mathbf{t}, \boldsymbol{\Psi})}\right) + \gamma_{lk}, \qquad (2.11)$$

here $\gamma_{lk} = \ln\left(\frac{\pi_l}{\pi_k}\right)$, $l, k = 1..m, k \neq l$. According to the (2.11) the observation Z_0, given $\mathbf{T} = \mathbf{t}$, is allocated to the population Ω_l if $W_{lk}^B(Z_0, \boldsymbol{\Psi}) \geq 0$, for all $l, k = 1..m, k \neq l$.

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

$$P^B(\boldsymbol{\Psi}) = 1 - \sum_{l=1}^{m} \pi_l PC_l, \qquad (2.12)$$

where $PC_l = P_l(W_{lk}^B(Z_0, \boldsymbol{\Psi}) \geq 0, l = 1..m, l \neq k)$ 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 $p_l(Z_0 | \mathbf{t}, \boldsymbol{\Psi})$.

Plugging in the parameters estimators into (2.12) we get PBDF (2.7). 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

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

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

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 designated as $EER = E_T\left(R^B(\widehat{\mathbf{\Psi}})\right)$.

The probability of misclassification for $m = 2$

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

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

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

$$P^B(\mathbf{\Psi}) = \sum_{l=1}^{2} \pi_l P_l, \tag{2.15}$$

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

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

(A3) The mean models in the populations Ω_l, $l = 1..m$, are different parametric models $\boldsymbol{\mu}_l(\mathbf{s}) = \mathbf{x}'(\mathbf{s})\boldsymbol{\beta}_l$, defined in the section 2.1.

2.2. Classification of GGRF observation

In this section we solve 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 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 ML method. These results are the extension to the ones published in Dučinskas [14], [15] where the factorised nuggetless covariance function was considered and the only one covariance parameter σ^2 was assumed to be unknown.

In this section we also present 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 Ω_1 or Ω_2. Under the assumption **(A3)** the model of observation $Z(\mathbf{s})$ in the population Ω_l can be written as

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

Then the model of training sample has the form

$$\mathbf{T} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}, \tag{2.17}$$

where

$$\mathbf{X} = \oplus_{l=1}^{2} \mathbf{X}_l \tag{2.18}$$

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.

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

$$\mathbf{T} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}). \tag{2.19}$$

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

$$p_l(Z_0|\mathbf{t}) = f(Z_0|\mu_{lt}, \sigma_t^2) \tag{2.20}$$

with conditional mean and variance

$$\mu_{lt} = E(Z_0|\mathbf{T} = \mathbf{t}; \Omega_l) = \mathbf{x}_0'\boldsymbol{\beta}_l + \boldsymbol{\alpha}_0'(\mathbf{t} - \mathbf{X}\boldsymbol{\beta}), l = 1,2, \tag{2.21}$$

$$\sigma_t^2 = var(Z_0|\mathbf{T} = \mathbf{t}; \Omega_l) = C(\mathbf{0}) - \mathbf{c}_0'\boldsymbol{\Sigma}^{-1}\mathbf{c}_0, \tag{2.22}$$

where $\mathbf{x}_0' = \left(x_1(\mathbf{s}_0), \dots, x_q(\mathbf{s}_0)\right)$ 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 $\boldsymbol{\Sigma}^{-1} = \sigma^{-2}\boldsymbol{\Gamma}^{-\mathbf{1}}$ the conditional variance could be rewritten as

$$\sigma_t^2 = var(Z_0|\mathbf{T} = \mathbf{t}; \Omega_l) = \sigma^2(1 - \mathbf{r}_0'\boldsymbol{\Gamma}^{-1}\mathbf{r}_0). \tag{2.23}$$

Then BDF specified in (2.9) is a linear Z_0 function

$$W^B(Z_0, \boldsymbol{\Psi}) = (Z_0 - (\mu_{1t} + \mu_{2t})/2)' \, (\mu_{1t} - \mu_{2t})/\sigma_t^2 + \gamma^*. \tag{2.24}$$

Replacing the conditional mean and variance into (2.24) by the expressions given in (2.21) and (2.23) we get the following formula

$$W^B(Z_0, \boldsymbol{\Psi}) = (Z_0 - \boldsymbol{\alpha}_0'(\mathbf{t} - \mathbf{X}\boldsymbol{\beta}) - \mathbf{x}_0'\mathbf{I}_+\boldsymbol{\beta}/2)'(\,\mathbf{x}_0'\mathbf{I}_-\boldsymbol{\beta})/K\sigma^2 + \gamma^*, \tag{2.25}$$

where $\mathbf{I}_+ = (\mathbf{I}_q, \mathbf{I}_q)$, $\mathbf{I}_- = (\mathbf{I}_q, -\mathbf{I}_q)$ and $K = 1 - \mathbf{r}_0'\boldsymbol{\Gamma}^{-\mathbf{1}}\mathbf{r}_0$.

Now, the Bayes risk, associated with this linear BDF (2.25), will be derived, but before that it is essential to introduce 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

$$d^2 = \frac{(\mu_{1t} - \mu_{2t})^2}{\sigma_t^2} = \frac{(\mu_{1t} - \mu_{2t})^2}{\sigma^2 K} \qquad (2.26)$$

be the squared *conditional Mahalanobis distance* between Ω_1 and Ω_2 at the location $\mathbf{s}_0$. Then the squared *marginal Mahalanobis distance* is specified by formula

$$\Delta^2 = (\mu_1 - \mu_2)^2/\sigma^2. \qquad (2.27)$$

These distances will be considered as class separation measures. Using (2.21) and (2.23) it is easy to show that 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)**, **(A3)** hold. Then *Bayes risk* associated with BDF (2.25) for two-class case is

$$R^B(\mathbf{\Psi}) = \sum_{l=1}^{2} \{\pi_l L(l,l) + g_l \Phi(-d/2 + (-1)^l \gamma^*/d)\}. \qquad (2.28)$$

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

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

and variance

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

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

By using the properties of normal distribution we obtain

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

Here $\Phi(\cdot)$ is the standard normal distribution function.

In order to get plug-in BDF we have to obtain the estimators of conditional mean (2.21) and variance (2.23). For the case of complete parametric uncertainty the estimators are the following

$$\hat{\mu}_{lt} = \mathbf{x}_0'\widehat{\boldsymbol{\beta}}_l + \widehat{\boldsymbol{\alpha}}_0'\left(\mathbf{t} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right), \tag{2.29}$$

$$\hat{\sigma}_t^2 = \widehat{\sigma}^2 \widehat{K}. \tag{2.30}$$

Here $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$ denote the estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, thus $\widehat{\boldsymbol{\Psi}} = \left(\widehat{\boldsymbol{\beta}}', \widehat{\boldsymbol{\theta}}'\right)$ denotes the vector of population parameters estimators. Replacing the obtained estimators into (2.25) we get the plug-in BDF

$$W^B\left(Z_0, \widehat{\boldsymbol{\Psi}}\right) = \left(Z_0 - \widehat{\boldsymbol{\alpha}}_0'\left(\mathbf{t} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) - \mathbf{x}_0'\mathbf{I}_+\widehat{\boldsymbol{\beta}}/2\right)'\left(\mathbf{x}_0'\mathbf{I}_-\widehat{\boldsymbol{\beta}}\right)/\widehat{K}\widehat{\sigma}^2 + \gamma^*. \tag{2.31}$$

Lemma 2.2. The *actual risk* for PBDF $W^B\left(Z_0, \widehat{\boldsymbol{\Psi}}\right)$ is defined as

$$R^B(\widehat{\boldsymbol{\Psi}}) = \sum_{l=1}^{2}\left\{\pi_l L(l,l) + g_l \Phi(\hat{Q}_l)\right\}. \tag{2.32}$$

Here

$$\hat{Q}_l = (-1)^l\left((a_l - \hat{b})\mathrm{sgn}(\mathbf{x}_0'\mathbf{I}_-\widehat{\boldsymbol{\beta}})/\sigma_t + \gamma^*\hat{\sigma}_t^2/\left|\mathbf{x}_0'\mathbf{I}_-\widehat{\boldsymbol{\beta}}\right|\sigma_t\right), \tag{2.33}$$

$$a_l = \mathbf{x}_0'\boldsymbol{\beta}_l + \boldsymbol{\alpha}_0'(\mathbf{t} - \mathbf{X}\boldsymbol{\beta}), l = 1,2, \tag{2.34}$$

$$\hat{b} = \widehat{\boldsymbol{\alpha}}_0'\left(\mathbf{t} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) + \mathbf{x}_0'\mathbf{I}_+\widehat{\boldsymbol{\beta}}/2. \tag{2.35}$$

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

$$W^B\left(Z_0, \widehat{\boldsymbol{\Psi}}\right)|\Omega_l \sim N(\mu_l^W, \sigma_W^2),$$

where mean and variance have the following expressions

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

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

Then the probabilities of misclassification are

$$PM_1 = P_1\left(\widehat{W}^B(Z_0) < 0\right) = \Phi\left(-\frac{\mu_1^W}{\sigma_W}\right) =$$

$$= \Phi\left(-\frac{(a_1 - \hat{b})sgn(\mathbf{x}_0'\mathbf{I}_-\widehat{\boldsymbol{\beta}})}{\sigma_t} - \frac{\gamma^* \hat{\sigma}_t^2}{\left|\mathbf{x}_0'\mathbf{I}_-\widehat{\boldsymbol{\beta}}\right|\sigma_t}\right),$$

$$PM_2 = P_2\left(\widehat{W}^B(Z_0) \geq 0\right) = \Phi\left(\frac{\mu_2^W}{\sigma_W}\right) =$$

$$= \Phi\left(\frac{(a_2 - \hat{b})sgn(\mathbf{x}_0'\mathbf{I}_-\widehat{\boldsymbol{\beta}})}{\sigma_t} + \frac{\gamma^* \hat{\sigma}_t^2}{\left|\mathbf{x}_0'\mathbf{I}_-\widehat{\boldsymbol{\beta}}\right|\sigma_t}\right).$$

Using (2.10a) we complete the proof of lemma.

Then the *expected risk* could be evaluated by

$$ER = E_T\left(R^B\left(\widehat{\boldsymbol{\Psi}}\right)\right) = E_T\left\{\sum_{l=1}^2\left(\pi_l L(l,l) + g_l \Phi(\hat{Q}_l)\right)\right\}. \tag{2.36}$$

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, asymptotic approximations of the *expected risk* are especially important.

In this dissertation the *approximation of expected risk* (AER) based on asymptotic expansion is proposed. We focus on the maximum likelihood estimators, since the inverse of the information matrix associated with likelihood function of training sample well approximates the covariance matrix of these estimators. The asymptotic properties of ML estimators showed by Mardia and Marshall [42] 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 $R^B(\widehat{\boldsymbol{\Psi}})$ with respect to $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$ evaluated at the point $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}, \widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ and let $\mathbf{R}_{\beta\theta}^{(2)}$ denote the matrix of the second order derivatives of $R^B(\widehat{\boldsymbol{\Psi}})$ with respect to to $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$ evaluated at the point $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}$. Also the following assumption is made:

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

Restrictive assumption **(A4)** is exploited intensively by many authors (see Zimmerman [74]; Zhu and Stein [71]), since Abt [1] showed that finer approximations of MSPE considering the correlation between $\mathbf{T}$ and $\widehat{\boldsymbol{\theta}}$ do not give better results.

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

Theorem 2.1. Suppose that observation Z_0 to be classified by BPDF (2.31) and let Mardia and Marshall conditions (Theorem 1.1) and assumption **(A4)** hold. Then the approximation of ER is

$$AER = \sum_{l=1}^{2} g_l \Phi(Q_l) + g_1 \varphi(Q_1) d(K_\beta + K_\theta)/2\sigma_t^2, \tag{2.37}$$

$$K_\beta = \boldsymbol{\Lambda}'\mathbf{J}_\beta^{-1}\boldsymbol{\Lambda}, \mathbf{J}_\beta = \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}, \tag{2.38}$$

$$\boldsymbol{\Lambda}' = \boldsymbol{\alpha}_0'\mathbf{X} - \mathbf{x}_0'(\mathbf{I}_+/2 + \gamma^*\mathbf{I}_-/d^2), \tag{2.39}$$

$$K_\theta = tr\big(\boldsymbol{\Sigma}\mathbf{A}_\theta\mathbf{J}_\theta^{-1}\mathbf{A}_\theta'\big) + (\gamma^*)^2\mathbf{s}_\theta'\mathbf{J}_\theta^{-1}\mathbf{s}_\theta/d^2\sigma_t^2. \tag{2.40}$$

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

$$R^B(\widehat{\mathbf{\Psi}}) = R^B(\mathbf{\Psi}) + R_\beta^{(1)}\Delta\widehat{\boldsymbol{\beta}} + \mathbf{R}_\theta^{(1)}\Delta\widehat{\boldsymbol{\theta}} +$$

$$+\frac{1}{2}\left(\left(\Delta\widehat{\boldsymbol{\beta}}\right)'\mathbf{R}_\beta^{(2)}\Delta\widehat{\boldsymbol{\beta}} + 2\left(\Delta\widehat{\boldsymbol{\beta}}\right)'\mathbf{R}_{\beta\theta}^{(2)}\Delta\widehat{\boldsymbol{\theta}} + \left(\Delta\widehat{\boldsymbol{\theta}}\right)'\mathbf{R}_\theta^{(2)}\Delta\widehat{\boldsymbol{\theta}}\right) + R_3, \qquad (2.41)$$

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

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

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

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

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

Partial derivatives of standard normal distribution density function with respect to $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$ are

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

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

Replacing the estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$ by the true values into $\varphi(\hat{Q}_l)$ 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 $R^B(\widehat{\mathbf{\Psi}})$ evaluated at the point $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ are

$$R_\beta^{(1)} = g_1\varphi(Q_1)\sum_{l=1}^{2}Q_{l\beta}^{(1)}, \qquad (2.42)$$

$$R_\theta^{(1)} = g_1\varphi(Q_1)\sum_{l=1}^{2}Q_{l\theta}^{(1)}, \qquad (2.43)$$

$$R_\beta^{(2)} = g_1\varphi(Q_1)\sum_{l=1}^{2}\left(Q_{l\beta}^{(2)} - Q_l Q_{l\beta}^{(1)}\left(Q_{l\beta}^{(1)}\right)'\right), \qquad (2.44)$$

$$R_\theta^{(2)} = g_1\varphi(Q_1)\sum_{l=1}^{2}\left(Q_{l\theta}^{(2)} - Q_l Q_{l\theta}^{(1)}\left(Q_{l\theta}^{(1)}\right)'\right). \qquad (2.45)$$

$$R_{\beta\theta}^{(2)} = g_1\varphi(Q_1)\sum_{l=1}^{2}\left(Q_{l\beta\theta}^{(2)} - Q_l Q_{l\theta}^{(1)}\left(Q_{l\beta}^{(1)}\right)'\right). \qquad (2.45a)$$

Here $Q_{l\beta}^{(k)}$ represents the $k-th$ order partial derivatives of $\hat{Q}_l$ with respect to $\widehat{\boldsymbol{\beta}}$ at the point $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}$, $Q_{l\theta}^{(k)}$ represents the $k-th$ order partial derivatives of $\hat{Q}_l$ with respect to $\widehat{\boldsymbol{\theta}}$ at the point $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}, \widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ and $Q_{l\beta\theta}^{(2)}$ represents the second order partial derivative of $\hat{Q}_l$ with respect to $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$ at the point $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}$. To obtain these derivatives at first we differentiate (2.33)-(2.35) with respect to $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$

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

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

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

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

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

Then we evaluate these derivatives at the point $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ and get the following expressions

$$Q_{l\beta}^{(1)} = (-1)^l\boldsymbol{\Lambda}/\sigma_t,$$

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

It is easy to notice that

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

The application of the above formulas to (2.43)-(2.45a) yields

$$R_\beta^{(1)} = 0, R_\theta^{(1)} = 0, \qquad (2.46)$$

$$R_{\beta}^{(2)} = g_1 d\varphi\left(-\frac{d}{2}-\frac{\gamma^*}{d}\right)\mathbf{\Lambda}\mathbf{\Lambda}'/\sigma_t^2, \tag{2.47}$$

$$R_{\theta}^{(2)} = \frac{g_1 d\varphi(Q_1)}{\sigma_t^2}\left(-\mathbf{A}'_{\theta}(\mathbf{t}-\mathbf{X}\boldsymbol{\beta})+\frac{\gamma^*\mathbf{s}_{\theta}}{d\sigma_t}\right)\left(-\mathbf{A}'_{\theta}(\mathbf{t}-\mathbf{X}\boldsymbol{\beta})+\frac{\gamma^*\mathbf{s}_{\theta}}{d\sigma_t}\right)'. \tag{2.48}$$

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

$$E_T\left(\left(\Delta\widehat{\boldsymbol{\beta}}\right)'\mathbf{R}_{\beta}^{(2)}\left(\Delta\widehat{\boldsymbol{\beta}}\right)\right) \cong \frac{g_1 d\varphi\left(-\frac{d}{2}-\gamma^*/d\right)}{\sigma_t^2}(\mathbf{\Lambda}'\mathbf{J}_{\beta}^{-1}\mathbf{\Lambda}). \tag{2.49a}$$

$$E_T\left(\left(\Delta\widehat{\boldsymbol{\theta}}\right)'\mathbf{R}_{\theta}^{(2)}\left(\Delta\widehat{\boldsymbol{\theta}}\right)\right) \cong \frac{g_1 d\varphi(-d/2-\gamma^*/d)}{\sigma_t^2}\left(tr(\mathbf{\Sigma}\mathbf{A}_{\theta}\mathbf{J}_{\theta}^{-1}\mathbf{A}'_{\theta})+\frac{(\gamma^*)^2\mathbf{s}'_{\theta}\mathbf{J}_{\theta}^{-1}\mathbf{s}_{\theta}}{d^2\sigma_t^2}\right). \tag{2.49b}$$

Then taking the expectation term by term of the right-hand side of (2.41), using (2.28), (2.46)-(2.49ab) and replacing moments of estimators by the corresponding moments of asymptotic distributions specified in (1.9)-(1.11) we complete the proof of theorem.

Remark 2.1. For a nuggetless factorised covariance matrix $\mathbf{\Sigma}=\sigma^2\mathbf{R}$ and for $\boldsymbol{\theta}=\sigma^2$, which means that the only one covariance parameter (*partial sill*) is unknown, the approximation of ER specified in (2.37)-(2.40) coincides with one derived in Dučinskas [14]

$$AER = \sum_{l=1}^{2} g_l\Phi(Q_l) + g_1\varphi(Q_1)(\mathbf{\Lambda}(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})\mathbf{\Lambda}'d/2K + (\gamma^*)^2/(n-2q)d),$$

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 Dreižienė [A10] and is of the following form

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

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

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

$$K_\lambda = tr\left(\boldsymbol{\Sigma}\mathbf{A}_\lambda\mathbf{J}_\lambda^{-\mathbf{1}}\mathbf{A}'_\lambda\right) + (\gamma^*)^2\mathbf{s}'_\lambda\mathbf{J}_\lambda^{-\mathbf{1}}\mathbf{s}_\lambda/d^2\sigma_t^2, \mathbf{A}_\lambda = \partial\widehat{\boldsymbol{\alpha}}_0/\partial\hat{\lambda}.$$

The closed-form expression of AER for exponential covariance model

In order to apply the AER formula (2.37)-(2.40) in practice there might arise difficulties evaluating the term K_θ. This term includes 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 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 $\widehat{\boldsymbol{\theta}} = (\hat{\tau}^2, \hat{\sigma}^2, \hat{\alpha}, \hat{\lambda}, \hat{\varphi})'$ represents the vector of parameters estimators.

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

$$\mathbf{A}_\theta = \left(\frac{\partial\widehat{\boldsymbol{\alpha}}_0}{\partial\hat{\tau}^2}, \frac{\partial\widehat{\boldsymbol{\alpha}}_0}{\partial\hat{\sigma}^2}, \frac{\partial\widehat{\boldsymbol{\alpha}}_\mathbf{0}}{\partial\hat{\alpha}}, \frac{\partial\widehat{\boldsymbol{\alpha}}_\mathbf{0}}{\partial\hat{\lambda}}, \frac{\partial\widehat{\boldsymbol{\alpha}}_0}{\partial\hat{\varphi}}\right), \tag{2.50}$$

where

$$\partial\widehat{\boldsymbol{\alpha}}_0/\partial\hat{\theta}_i = -\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\theta_i}\boldsymbol{\alpha}_0 + \boldsymbol{\Sigma}^{-1}(\mathbf{c}_0)_{\theta_i}, i = 1..5. \tag{2.51}$$

For geometrically anisotropic covariance function, defined in (1.13) we get the following first order partial derivatives:

$$(\mathbf{c}_0)_{\tau^2} = \mathbf{0}_n, \tag{2.52}$$

$$(\mathbf{c}_0)_{\sigma^2} = \mathbf{r}_0, \tag{2.53}$$

$$(\mathbf{c}_0)_\alpha = \frac{\sigma^2}{\alpha^2}\mathbf{r}_0 \circ \mathbf{H}_{0\alpha}, \tag{2.54}$$

$$(\mathbf{c}_0)_\lambda = -\frac{\sigma^2\lambda}{\alpha}\mathbf{r}_0 \circ \mathbf{H}_{0\lambda}, \tag{2.55}$$

$$(\mathbf{c}_0)_{\varphi} = \frac{\sigma^2(\lambda^2-1)}{\alpha}\mathbf{r}_0 \circ \mathbf{H}_{0\varphi}, \tag{2.56}$$

where

$$(H_{0\alpha})_j = \sqrt{\left(h_u^{0j}\right)^2 + \lambda^2\left(h_v^{0j}\right)^2},$$

$$(H_\lambda)_j = \left(h_v^{0j}\right)^2 / \sqrt{\left(h_u^{0j}\right)^2 + \lambda^2\left(h_v^{0j}\right)^2},$$

$$\left(H_\varphi\right)_j = h_u^{0j} h_v^{0j} / \sqrt{\left(h_u^{0j}\right)^2 + \lambda^2\left(h_v^{0j}\right)^2},$$

$h_u^{0j} = \left(h_x^{0j} \, cos\, \varphi + h_y^{0j} \, sin\, \varphi\right)/\alpha$, $h_v^{0j} = \left(-h_x^{i0} \, sin\, \varphi + h_y^{0j} \, cos\, \varphi\right)/\alpha$, $h_x^{0j} = x_0 - x_j$, $h_y^{0j} = y_0 - y_j$, x_0 and y_0 are the coordinates of s_0 and x_j and $y_j, j = 1..n$, represent the coordinates of the $j - th$ $\mathbf{T}$ component.

The first order partial derivatives of covariance matrix with respect to parameter $\hat{\theta}_i$ evaluated at $\hat{\theta}_i = \theta_i$ coincide with ones defined in Lemma 1.2. $\mathbf{\Sigma}_i = \mathbf{\Sigma}_{\theta_i}, i = 1..5$. Then replacing $(\mathbf{c}_0)_{\theta_i}$ by (2.52)-(2.56) and $\mathbf{\Sigma}_{\theta_i}$ by (1.14)-(1.18) into (2.51) we get the elements of matrix $\mathbf{A}_\theta$

$$\partial\widehat{\boldsymbol{\alpha}}_0/\partial\hat{\tau}^2 = -\mathbf{\Sigma}^{-\mathbf{1}}\boldsymbol{\alpha}_0,$$

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

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

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

$$\partial\widehat{\boldsymbol{\alpha}}_0/\partial\widehat{\varphi} = \frac{\sigma^2(\lambda^2-1)}{\alpha}\mathbf{\Sigma}^{-1}(\mathbf{R} \circ \mathbf{H}_\varphi\alpha_0 - \mathbf{r}_0 \circ \mathbf{H}_{0\varphi}).$$

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

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

Using (2.52)-(2.56) and (1.14)-(1.18) we obtain the elements of $\mathbf{s}_\theta$:

$$\mathrm{s}_{\tau^2} = \boldsymbol{\alpha}_0'\boldsymbol{\alpha}_0,$$

$$s_{\sigma^2} = 1 + \boldsymbol{\alpha}_0'\mathbf{R}\boldsymbol{\alpha}_0 - 2\boldsymbol{\alpha}_0'\mathbf{r}_0,$$

$$s_\alpha = -\frac{\sigma^2}{\alpha^2}\left(\boldsymbol{\alpha}_0'(\mathbf{R}\circ\mathbf{H}_\alpha)\boldsymbol{\alpha}_0 - 2\boldsymbol{\alpha}_0'(\mathbf{r}_0\circ\mathbf{H}_{0\alpha})\right),$$

$$s_\lambda = -\frac{\lambda\sigma^2}{\alpha}\left(\boldsymbol{\alpha}_0'(\mathbf{R}\circ\mathbf{H}_\lambda)\boldsymbol{\alpha}_0 - 2\boldsymbol{\alpha}_0'(\mathbf{r}_0\circ\mathbf{H}_{0\lambda})\right),$$

$$s_\varphi = \frac{\sigma^2(\lambda^2-1)}{\alpha}\left(\boldsymbol{\alpha}_0'(\mathbf{R}\circ\mathbf{H}_\varphi)\boldsymbol{\alpha}_0 - 2\boldsymbol{\alpha}_0'(\mathbf{r}_0\circ\mathbf{H}_{0\varphi})\right).$$

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 Dreižienė [A10].

Multiclass case

Now consider a *multiclass case* $(m > 2)$ with *zero-one loss* function. So the main goal is to solve the problem of classification of the 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 σ^2 and known spatial correlation function.

The pairwise BDF specified in (2.11) in this case has the expression

$$W_{lk}^B(Z_0, \boldsymbol{\Psi}) = (Z_0 - (\mu_{lt} + \mu_{kt})/2)'\,(\mu_{lt} - \mu_{kt})/\sigma_t^2 + \gamma_{lk}, \qquad (2.57)$$

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

$$\text{Classify } Z_0 \text{ to population } \Omega_l \text{ if } W_{lk}^B(Z_0, \boldsymbol{\Psi}) \geq 0. \qquad (2.57a)$$

Replacing the conditional mean and variance into (2.57) by the expressions given in (2.21) and (2.23) we get the following formula for the pairwise BDF

$$W_{lk}^B(Z_0, \boldsymbol{\Psi}) = (Z_0 - \boldsymbol{\alpha}_0'(\mathbf{t} - \mathbf{X}\boldsymbol{\beta}) - \mu_{lk})'d_{lk}/\sigma\sqrt{K} + \gamma_{lk}, \qquad (2.58)$$

where $\mu_{lk} = \mathbf{x}_0'(\boldsymbol{\beta}_l + \boldsymbol{\beta}_k)/2$ and d_{lk} stands for conditional Mahalanobis distance

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

d_{lk} could be expressed in terms of marginal Mahalanobis distance, Δ_{lk}

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

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

Now, the probability of misclassification (Bayes misclassification probability) associated with 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 μ and variance σ^2 and set $\varphi(x; 0, 1) = \varphi(x)$.

Lemma 2.3. The probability of misclassification for $m > 2$ due to Bayes rule specified in (2.57a) is

$$P^B(\boldsymbol{\Psi}) = 1 - \sum_{l=1}^{m} \pi_l \int_{B_l} \varphi(u) du, \tag{2.59}$$

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

Proof. The probability of misclassification due to Bayes classification rule is

$$P^B(\boldsymbol{\Psi}) = 1 - \sum_{l=1}^{m} \pi_l PC_l, \tag{2.60}$$

where for $l, k = 1..m, l \neq k$, $PC_l = P_l(W_{lk}^B(Z_0, \boldsymbol{\Psi}) \geq 0)$ is the probability of correct classification of Z_0 when it comes from Ω_l with mean μ_{lt} and variance σ_t^2. In the above conditions it follows that

$$Z_0 = \sigma_t u + \mu_{lt}, \tag{2.61}$$

where $u \sim N(0,1)$. After making the change of variables $u \to Z_0$ in (2.60) and putting (2.61), (2.21) and (2.23) into (2.58) we complete 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, $\mathbf{\Psi} = (\boldsymbol{\beta}', \sigma^2)'$. Moreover, $\boldsymbol{\alpha}_0$ for a nuggetless covariance does not depend on σ^2, that is, $\boldsymbol{\alpha}_0' = \mathbf{r}_0'\mathbf{R}^{-1}$. Based on that the estimators of conditional mean and variance are

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

$$\hat{\sigma}_t^2 = \widehat{\sigma}^2 K.$$

Replacing the conditional mean and variance into BDF (2.58) by their estimators specified above we obtain the plug-in BDF,

$$W_{lk}^B\big(Z_0, \widehat{\mathbf{\Psi}}\big) = \big(Z_0 - \boldsymbol{\alpha}_0'\big(\mathbf{t} - \mathbf{X}\widehat{\boldsymbol{\beta}}\big) - \hat{\mu}_{lk}\big)'\hat{d}_{lk}/\hat{\sigma}\sqrt{K} + \gamma_{lk}, \qquad (2.62)$$

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

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

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 (2.62) for $m > 2$ has the following form

$$P^B\big(\widehat{\mathbf{\Psi}}\big) = 1 - \sum_{l=1}^m \pi_l \int_{A_l} \varphi(u)du, \qquad (2.63)$$

where

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

Proof. The proof is completed by making the transformation of random variables (2.61) in the formulas presented in Definition 2.5. and (2.62).

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 spatial correlation parameters are unknown, the likelihood function is intractable. The ML estimators have no closed-form especially in multiclass case and it is impossible to obtain 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) $B(x)\delta\big(B(x)\big) \equiv 0$,

(d2) $B(x)\delta'\big(B(x)\big) + \delta(B(x)) \equiv 0$,

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

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}^{-1}\mathbf{T}\,,$$

$$\widehat{\boldsymbol{\beta}} \sim N_{2q}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}).$$

We use the bias adjusted ML estimator of variance

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

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

Since EER approximation is based on 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 $P^B(\widehat{\boldsymbol{\Psi}})$ with respect to $\widehat{\boldsymbol{\Psi}}$ evaluated at $\boldsymbol{\Psi}$, respectively. Similarly, let $\nabla_\beta P_t$ ant $\nabla_\beta^2 P_t$ denote the vector of the first order partial derivatives and the matrix of second order partial derivatives of $P^B(\widehat{\boldsymbol{\Psi}})$ with respect to $\widehat{\boldsymbol{\beta}}$ evaluated at $\boldsymbol{\beta}$, respectively. By $P_t^{(k)}$ we denote the $k-th$ $(k = 1, 2)$ order partial derivatives of $P^B(\widehat{\boldsymbol{\Psi}})$ with respect to $\hat{\sigma}^2$ evaluated at σ^2. Finally we denote by $\nabla_\beta P_t^{(1)}$ the vector of second order partial derivatives of $P^B(\widehat{\boldsymbol{\Psi}})$ with respect to $\widehat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ evaluated at values $\boldsymbol{\beta}$ and σ^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 $W_{lk}^B(Z, \boldsymbol{\Psi}) = 0$, 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

$$\begin{aligned} G_{lk}^B(z) &= \pi_l\varphi(z; \mu_{lt}, \sigma_t^2) - \pi_k\varphi(z; \mu_{kt}, \sigma_t^2) = \\ &= \pi_l\varphi(z; \mu_{lt}, \sigma_t^2)(1 - exp\{-W_{lk}^B(z)\}). \end{aligned} \qquad (2.64a)$$

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

$$\nabla_\Psi P_t = 0, \qquad (2.65)$$

$$\nabla_\Psi^2 P_t =$$

$$\sum_{l=1}^m \sum_{k>l} \pi_l p_l(z_{lk}) \nabla_\Psi W_{lk}(z_{lk}) \nabla_\Psi' W_{lk}(z_{lk}) \prod_{j \neq l,k} e(W_{lj}(z_{lk}))/W_{lk}', \qquad (2.66)$$

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

Proof. It is easy to see that

$$P^B(\widehat{\mathbf{\Psi}}) = 1 - \sum_l \pi_l \int \prod_{k\neq l} e\left(\widehat{W}^B_{lk}(z)\right) p_l(z)dz =$$

$$1 - \sum_l \pi_l \int \prod_{k\neq l} e\left(\hat{G}_{lk}(z)\right) p_l(z)dz,$$

where

$$\hat{G}_{lk}(z) = \pi_l \varphi(z; \hat{\mu}_{lt}, \hat{\sigma}_t^2) - \pi_k \varphi(z; \hat{\mu}_{kt}, \hat{\sigma}_t^2) =$$

$$= \pi_l \varphi(z; \hat{\mu}_{lt}, \hat{\sigma}_t^2)\left(1 - exp\{-\widehat{W}^B_{lk}(z)\}\right).$$

By using (d4) we obtain

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

By using (d1) in the above equation we obtain (2.65). The Hessian $\nabla^2_\Psi P_t$ evaluated at $\widehat{\mathbf{\Psi}} = \mathbf{\Psi}$ is equal to

$$\nabla^2_\Psi P_t =$$

$$= \sum_{l=1}^m \sum_{k>l} \int \left[G_{kl}(z)\nabla_\Psi G_{lk}(z)\delta'\left(G_{lk}(z)\right)\nabla'_\Psi G_{lk}(z) \prod_j e\left(G_{lj}(z)\right)\right.$$

$$+ G_{kl}(z)\,\nabla^2_\Psi G_{kl}(z)\delta\left(G_{kl}(z)\right)\prod_j e\left(G_{lj}(z)\right)$$

$$\left. + G_{kl}(z)\delta\left(G_{lk}(z)\right)\nabla_\Psi G_{lk}(z)\left(\sum_{j\neq l,k} \delta\left(G_{lj}(z)\right)\nabla'_\Psi G_{lj}(z) \prod_{v\neq j,l,k} e(G_{lv}(z))\right)\right] dz$$

It follows from (d1), (d2) that integral of the second and third terms in the above square brackets are equal to 0.

According to (d3) we have $\delta(G^B_{lk}) = \delta(z - z_{lk})\frac{\pi_l p_l(z_{lk})}{|W'(z_{lk})|}$. Then the proof of (2.66) is completed by using (d2) and (d3) to the integral of the first term.

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

(B1) $n(\mathbf{X}'\mathbf{X})^{-1} \to \mathbf{V}$, as $n \to \infty$, where $\mathbf{V}$ is a positively definite $mq \times mq$ matrix with finite determinant.

(B2) $\lambda_{max}(\mathbf{R}) < \mathrm{v} < +\infty$, as $n \to \infty$,

(B3) $\frac{n_l}{n_k} \to \nu_{lk}$, as $n_l, n_k \to \infty$, $0 < \nu_{lk} < \infty$.

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

$$\begin{array}{cccccccccccc} \boldsymbol{F}_{lk} = & (\mathbf{0}_q & \dots & \mathbf{0}_q & f_l\boldsymbol{I}_q & \mathbf{0}_q & \dots & \mathbf{0}_q & f_k\boldsymbol{I}_q & \mathbf{0}_q & \dots & \mathbf{0}_q)' \\ q \times mq & 1 & \dots & l-1 & l & l+1 & \dots & k-1 & k & k+1 & \dots & m \end{array}$$

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.

Set $\boldsymbol{\Lambda}_{lk} = \mathbf{X}'\boldsymbol{\alpha}_0 - \mathbf{F}_{lk}\mathbf{x}_0$ 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 PBDF and let assumptions (B1)-(B3) hold. Then the asymptotic expansion of EER is

$$EER = P^B(\boldsymbol{\Psi}) + C/2 + D + O(n^{-2}), \tag{2.64}$$

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 (2.64) by ignoring the remainder i.e.

$$AEER = P^B(\boldsymbol{\Psi}) + C/2 + D. \tag{2.64a}$$

Proof. The proof is based on Taylor series expansion of $P^B(\widehat{\mathbf{\Psi}})$ around $\mathbf{\Psi}$ and taking the expectation with respect to 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

$$EER = P^B(\mathbf{\Psi}) + tr\left(\nabla^2_\Psi P_t Var(\widehat{\mathbf{\Psi}})\right)/2 =$$

$$P^B(\mathbf{\Psi})\left(tr\left(\nabla^2_\beta P_t Var(\widehat{\boldsymbol{\beta}})\right) + P_t^{(2)} Var(\hat{\sigma}^2)\right)/2 + E(R_3). \tag{2.67}$$

Then by applying (2.66) we can write

$$\nabla^2_\beta P_t = \sigma\sqrt{K}\sum_{l=1}^m \sum_{k>l} \pi_l p_l(z_{lt}) \nabla_\beta W_{lk}(z_{lk}) \nabla'_\beta W_{lk}(z_{lk}) \times$$

$$\prod_{j\neq l,k} e\left(W_{lj}(z_{lk})\right)/d_{lk}, \tag{2.68}$$

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

Note that using $\nabla_\beta W_{lk} = d_{lk}\Lambda_{lk}/\sigma\sqrt{K}$, $(W_{lk})'_{\sigma^2} = \gamma_{lk}/\sigma^2$ in (2.68) and (2.69) and inserting them into (2.67) we obtain the main term of expansion (2.64). 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 analysed in [A4] and [A5].

2.2.2. Multivariate case

The case of multivariate GGRF with known spatial correlation function is analysed in Dučinskas [15]. Here the error rates are investigated and the factorised 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 Ω_l is

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

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 $\{\varepsilon(\mathbf{s}) : \mathbf{s} \in D\}$ with factorized nuggetless covariance function defined by the following model for all $\mathbf{s}, \mathbf{u} \in D\ cov\{\varepsilon(\mathbf{s}), \varepsilon(\mathbf{u})\} = r(\mathbf{s} - \mathbf{u})\mathbf{S}$, 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

$$\mathbf{T} = \mathbf{X}\mathbf{B} + \mathbf{E} \tag{2.70}$$

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 $\mathbf{E}$ represent 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 $\mathbf{R} = \mathbf{R}(\boldsymbol{\vartheta})$ 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 population Ω_l, the conditional distribution of $\mathbf{Z}_0$ given $\mathbf{T} = \mathbf{t}$ is Gaussian with conditional mean and variance

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

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

Then BDF specified in (2.24) becomes

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

And Bayes risk for the BDF (2.71) has the same form specified in (2.28)

$$R^B = \sum_{l=1}^{2}\{\pi_l L(l,l) + g_l\Phi(-d/2 + (-1)^l\gamma^*/d)\}, \qquad (2.72)$$

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

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

Assume that true values of parameters $\mathbf{B}$, $\mathbf{S}$ and $\boldsymbol{\vartheta}$ are unknown (complete parametric uncertainty).

Replacing the parameters by their estimators in (2.71) and using the expressions of conditional mean and variance we get the following PBDF

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

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

$$R^B\big(\widehat{\mathbf{\Psi}}\big) = \sum_{l=1}^{2}\big\{\pi_l L(l,l) + g_l\Phi(\widehat{Q}_l)\big\}, \qquad (2.74)$$

where

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

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

$$W^B\big(Z_0, \widehat{\mathbf{\Psi}}\big)|\Omega_l \sim N(\boldsymbol{\mu}_l^W, \sigma_W^2), \qquad (2.76)$$

where

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

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

$$\mathbf{b} = \widehat{\boldsymbol{\alpha}}_0'\big(\mathbf{t} - \mathbf{X}\widehat{\mathbf{B}}\big) + \mathbf{x}_0'\mathbf{I}_+\widehat{\mathbf{B}}/2,$$

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

The proof is completed and formulas (2.74), (2.75) are obtained by using the equations (2.73), (2.76)-(2.78) and (2.10b).

The asymptotic expansion of ER for $m = 2$

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

$$\boldsymbol{\beta}_{\mathbf{v}} = vec(\mathbf{B}), \boldsymbol{\eta} = vech(\mathbf{S}), \mathbf{R}_{\vartheta} = \partial vec\mathbf{R}/\partial \boldsymbol{\vartheta}',$$

$$\dim \boldsymbol{\beta}_{\mathbf{v}} = q_0 = 2qn, \dim \boldsymbol{\eta} = m = p(p+1)/2\,, \dim \boldsymbol{\vartheta} = r.$$

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

$$\Lambda(\boldsymbol{\Psi}) = const -$$

$$-1/2\big(p\, ln|\mathbf{R}|\ + n\, ln|\mathbf{S}| + tr(\mathbf{R}^{-1}(\mathbf{T}-\mathbf{XB})\mathbf{S}^{-1}(\mathbf{T}-\mathbf{XB})')\big).$$

Then the information matrices for the corresponding parameters are

$$\mathbf{J}_{\beta} = (\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}) \otimes \mathbf{S}^{-1},$$

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

$$\mathbf{J}_{\theta} = p\mathbf{R}_{\vartheta}'(\mathbf{R}^{-1} \otimes\ \mathbf{R}^{-1})\mathbf{R}_{\vartheta}/2,$$

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

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

It is easy to obtain that

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

Denote by $\mathbf{J} = \begin{pmatrix} \mathbf{J}_{\eta} & \mathbf{J}_{\eta\vartheta} \\ \mathbf{J}_{\vartheta\eta} & \mathbf{J}_{\vartheta} \end{pmatrix}$ and $\mathbf{V} = \mathbf{J}^{-1} = \begin{pmatrix} \mathbf{V}_{\eta} & \mathbf{V}_{\eta\vartheta} \\ \mathbf{V}_{\vartheta\eta} & \mathbf{V}_{\vartheta} \end{pmatrix}$ 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

$$\widehat{\boldsymbol{\beta}} \sim AN_{q_0}(\boldsymbol{\beta}, \mathbf{V}_{\mathrm{B}}\ ), \widehat{\boldsymbol{\eta}} \sim \mathrm{AN}_m(\boldsymbol{\eta}, \mathbf{V}_\eta), \widehat{\boldsymbol{\vartheta}} \sim AN_r(\boldsymbol{\vartheta}, \mathbf{V}_\vartheta). \qquad (2.79)$$

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

Let the assumption **(A4)** which claims about the independence of $\mathbf{T}$ and estimator of covariance parameters $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\eta}}, \widehat{\boldsymbol{\vartheta}})$ hold. Also let $\mathbf{A}_\vartheta = \partial\widehat{\boldsymbol{\alpha}}_0/\partial\boldsymbol{\vartheta}'$ be the $n \times k$ matrix of partial derivatives evaluated at the point $\widehat{\boldsymbol{\vartheta}} = \boldsymbol{\vartheta}$, $\mathbf{K}_\vartheta = \partial K/\partial\boldsymbol{\vartheta}'$ be the $k \times 1$ vector of partial derivatives evaluated at the point $\widehat{\boldsymbol{\vartheta}} = \boldsymbol{\vartheta}$ and let $\varphi(\cdot)$ be the standard normal distribution density function.
Set $= \mathbf{X}'\boldsymbol{\alpha}_0 - (\mathbf{I}_+'/2 + \gamma^*\mathbf{I}_-'/d^2)\mathbf{x}_0$, $\mathbf{R}_0 = (\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}$, $\Delta\boldsymbol{\mu}\ = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$,
$\varphi_1 = \varphi(-d/2 - \gamma^*/d)$.

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

$$AER = R^B + g_1\varphi_1\{\boldsymbol{\Lambda}'R^B\boldsymbol{\Lambda}d/K + (p-1)\mathbf{x}_0'\mathbf{I}_-R^B\mathbf{I}_-'\mathbf{x}_0/Kd + tr\big(\mathbf{F}_1\mathbf{V}_\eta\big) +$$
$$tr(\mathbf{F}_2\mathbf{V}_\vartheta) + 2tr\big(\mathbf{F}_3\mathbf{V}_{\eta\vartheta}\big)\}/2, \qquad (2.80)$$

where $R^B\ = R^B(\boldsymbol{\Psi})$ is Bayes risk, specified in (2.72).

$$F_1 = \mathbf{D}_p'\big((\mathbf{S}^{-1}\Delta\boldsymbol{\mu}\Delta\boldsymbol{\mu}'\mathbf{S}^{-1} \otimes \mathbf{S}^{-1}\Delta\boldsymbol{\mu}\Delta\boldsymbol{\mu}'\mathbf{S}^{-1})(\gamma^*)^2K/\Delta^4 + \mathbf{S}^{-1}\Delta\boldsymbol{\mu}\Delta\boldsymbol{\mu}'\mathbf{S}^{-1} \otimes$$
$$(\mathbf{S}^{-1} - \mathbf{S}^{-1}\Delta\boldsymbol{\mu}\Delta\boldsymbol{\mu}'\mathbf{S}^{-1}/d^2)\big)\mathbf{D}_p/(d\sqrt{K}), \qquad (2.81)$$

$$F_2 = (tr(\mathbf{A}_\vartheta'\mathbf{R}\mathbf{A}_\vartheta\mathbf{V}_\vartheta)\Delta^2 + (\gamma^*)^2\mathbf{K}_\vartheta'\mathbf{V}_\vartheta\mathbf{K}_\vartheta)/\Delta^3\sqrt{K}, \qquad (2.82)$$

$$F_3 = \mathbf{D}_p'(\mathbf{S}^{-1}\Delta\boldsymbol{\mu} \otimes \mathbf{S}^{-1}\Delta\boldsymbol{\mu})(\gamma^*)^2\mathbf{K}_\vartheta/\Delta^4\sqrt{K}. \qquad (2.83)$$

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

$$E_T(R^B) = R^B + E_T\Big(\big(\Delta\widehat{\boldsymbol{\beta}}\big)'\mathbf{R}^{(2)}_{\beta}\Delta\widehat{\boldsymbol{\beta}} + 2\big(\Delta\widehat{\boldsymbol{\vartheta}}\big)'\mathbf{R}^{(2)}_{\eta\theta}\,\Delta\widehat{\boldsymbol{\eta}} + \big(\Delta\widehat{\boldsymbol{\vartheta}}'\big)\mathbf{R}^{(2)}_{\theta}\big(\Delta\widehat{\boldsymbol{\vartheta}}\big) + (\Delta\widehat{\boldsymbol{\eta}})'\mathbf{R}^{(2)}_{\eta}(\Delta\boldsymbol{\eta})\Big)/2 + R_3, \tag{2.84}$$

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

$$R^{(2)}_{\beta} = g_1\varphi_1\Big((\mathbf{S}^{-1}\Delta\boldsymbol{\mu}\Delta\boldsymbol{\mu}'\mathbf{S}^{-1} \otimes \boldsymbol{\Lambda}\boldsymbol{\Lambda}')/K + (\mathbf{S^{-1}} - \mathbf{S^{-1}}\Delta\boldsymbol{\mu}\Delta\boldsymbol{\mu}'\mathbf{S}^{-1}/\Delta^2) \otimes (\mathbf{I}'_{-}\mathbf{x}_0\mathbf{x}'_0\mathbf{I}_{-})/\Delta\sqrt{K}\Big), \tag{2.85}$$

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

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

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

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

$$\boldsymbol{E}\Big(\big(\Delta\widehat{\boldsymbol{\vartheta}}\big)'\mathbf{R}^{(2)}_{\vartheta}\big(\Delta\widehat{\boldsymbol{\vartheta}}\big)\Big) \cong g_1\varphi_1(tr(\mathbf{A}'_{\vartheta}\mathbf{R}\mathbf{A}_{\vartheta}\mathbf{V}_{\vartheta})\Delta^2 + (\gamma^*)^2\mathbf{K}'_{\vartheta}\mathbf{V}_{\vartheta}\mathbf{K}_{\vartheta})/\Delta^3\sqrt{K}, \tag{2.89}$$

$$E\Big((\Delta\widehat{\boldsymbol{\eta}})'\mathbf{R}^{(2)}_{\eta\vartheta}\big(\Delta\widehat{\boldsymbol{\vartheta}}\big)\Big) \cong g_1\varphi_1(\gamma^*)^2tr\big(\mathbf{D}'_m(\mathbf{S^{-1}}\Delta\boldsymbol{\mu} \otimes \mathbf{S^{-1}}\Delta\boldsymbol{\mu})\mathbf{K}_{\vartheta}\mathbf{V}_{\eta\vartheta}\big)/\Delta^4\sqrt{K}. \tag{2.90}$$

Then using (2.85), (2.86), (2.89), (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 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 classifying a multivariate GGRF observation into one of the several populations specified by different parametric mean models is investigated in Dreižienė et al. [A5]. In this 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 at al. [36]). In contrast to 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 an undirected graph. Thus 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 [52] 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 part 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.

2.3.1 Univariate case

In this section we focus on classification of scalar GMRF $\{Z(\mathbf{s}): \mathbf{s} \in D \subset R^2\}$ observation into one of two populations Ω_1 or Ω_2, when training sample is given. The model of observation $Z(\mathbf{s})$ in populations $\Omega_l, l = 1{,}2$ has the same form as for GGRF, defined in (2.16)

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

The main difference comparing with GGRF is that the error term $\varepsilon(\mathbf{s})$ is generated on lattice by zero-mean GMRF $\{\varepsilon(\mathbf{s}): \mathbf{s} \in D\}$ with respect to the neighbourhood 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 $Z(\mathbf{s})$ 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 $Z(\mathbf{s})$ from Ω_l, $l = 1{,}2$, $n = n_1 + n_2$. The focal location, $\mathbf{s}_0$, is indexed by $\{0\}$.

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

$$\mathbf{w}_0' = (w_{01}, \dots, w_{0n}), \mathbf{w}_i' = (w_{i1}, \dots, w_{ii-1}, w_{ii+1}, \dots, w_{in}), i = 1..n.$$

Following de Oliveira and Ferreira [52] we construct matrices $\mathbf{H}^0 = (h_{kl}^0: k, l = 0{,}1..n)$ and $\mathbf{H} = (h_{kl}: k, l = 1..n)$ with dimensions $(n+1) \times (n+1)$ and $(n \times n)$, respectively. The elements of these matrices are defined as follows

$$h_{kl}^0 = \begin{cases} h_k^0 & if \quad k = l \\ -w_{kl} & if \ k \in N_l^0, \\ 0 & otherwise \end{cases} \quad h_{kl} = \begin{cases} h_k & if \quad k = l \\ -w_{kl} & if \ k \in N_l, \\ 0 & otherwise \end{cases}$$

where

$$h_k^0 = \sum_{l \in N_k^0} w_{kl}\,, k, l = 0,1..n,$$

$$h_k = \sum_{l \in N_k} w_{kl}\,, k, l = 1..n.$$

These matrices, assumed to be known, allow the modelling of different patterns of spatial correlation by the specification of different neighbourhood systems and weights (w_{lk}) (de Oliveira Fereirra [52]).

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

For simplicity, we use the following notations:

$$Z(\mathbf{s}_i) = Z_i,\ \varepsilon(\mathbf{s}_i) = \varepsilon_i,\ x(\mathbf{s}_i) = x_i, i = 0..n.$$

$$\mathbf{Z} = (Z_0, Z_1, \dots, Z_n)', \mathbf{Z}_{-0} = (Z_1, \dots, Z_n)', \boldsymbol{\varepsilon}_{-0} = (\varepsilon_1,\ \dots,\ \varepsilon_n)',$$

Then let for $i = 1..n$

$$\boldsymbol{\varepsilon}_{-i} = (\varepsilon_0, \dots, \varepsilon_{i-1}, \varepsilon_{i+1}, \dots, \varepsilon_n)', \mathbf{Z}_{-i} = (Z_0, \dots, Z_{i-1},\ Z_{i+1}, \dots,\ Z_n)'.$$

The full conditionals for $i = 0,1..n, l = 1,2$ are specified as

$$\varepsilon_i | \varepsilon_{-i} \sim N(\mu^i,\ \sigma_i^2),$$

where $\mu^i = \boldsymbol{\alpha}_i' \boldsymbol{\varepsilon}_{-i},\ \sigma_i^2 = \sigma^2/(1 + \alpha h_i), \boldsymbol{\alpha}_i' = \alpha \mathbf{w}_i'/(1 + \alpha h_i),$ $\mathbf{w}_i' = (w_{i1}, w_{ii-1},\ w_{ii+1}, \dots, w_{in}),\ i = 1..n.$ $\alpha \geq 0$ is a spatial dependence parameter and $\sigma > 0$ is a scale parameter. Then $\boldsymbol{\varepsilon}$ has the a multivariate Gaussian distribution

$$\boldsymbol{\varepsilon} \sim N_{n+1}\big(\mathbf{0},\ \sigma^2 (\mathbf{I}_{n+1} + \alpha \mathbf{H}^0)^{-1}\big).$$

The precision matrix (i.e. inverse of covariance matrix) of vector $\boldsymbol{\varepsilon}$ is $\big(var(\boldsymbol{\varepsilon})\big)^{-1} = (\mathbf{I}_{n+1} + \alpha \mathbf{H}^0)/\sigma^2$, where $\mathbf{I}_n$ denotes the identity matrix of $n - th$ order.

So, we can supplement the formulation of the classification problem: for a given training sample $\mathbf{T} = \mathbf{Z}_{-0}$ consider the problem of classification of the observation Z_0 into one of two populations. Let $\mathbf{t}$ denote the realization of $\mathbf{T}$.

Put $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$. Then the conditional distribution of Z_0 given $\mathbf{T} = \mathbf{t}$ in the population $\Omega_l, l = 1{,}2$ is Gaussian with mean and variance

$$\mu_{lt} = E(Z_0|\mathbf{T} = \mathbf{t}|\Omega_l) = \mathbf{x}_0'\boldsymbol{\beta}_l + \boldsymbol{\alpha}_0'(\mathbf{t} - \mathbf{X}\boldsymbol{\beta}), l = 1..2 \tag{2.91}$$

$$\sigma_t^2 = var(Z_0|\mathbf{T} = \mathbf{t}; \Omega_l) = \sigma^2/(1 + \alpha h_0)^{\cdot} \tag{2.92}$$

The training sample $\mathbf{T}$ would be modelled by the joint distribution (see de Oliveira and Ferreira [52])

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

where $\mathbf{X}$ denotes the $n \times 2q$ design matrix of training sample $\mathbf{T}$ specified in (2.18). The covariance matrix could be written in a factorized form, i.e. $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \sigma^2\mathbf{V}(\alpha)$, $\mathbf{V}(\alpha) = (\mathbf{I}_n + \alpha\mathbf{H})^{-1}$ and $\boldsymbol{\theta} = (\alpha, \sigma^2)'$. Parameter α controls the strength of correlation between the components of $\mathbf{T}$. When $\alpha = 0$, the components of $\mathbf{T}$ become independent random variables. In the following for brevity we will use $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ and $\mathbf{V} = \mathbf{V}(\alpha)$.

Recall that Bayes discriminant function (BDF) minimizing the risk of classification for two-class case has the following form

$$W^B(Z_0, \boldsymbol{\Psi}) = (Z_0 - (\mu_{1t} + \mu_{2t})/2)'\,(\mu_{1t} - \mu_{2t})/\sigma_t^2 + \gamma^*, \tag{2.93}$$

with $\gamma^* = ln(g_1/g_2)$, $g_l = \pi_l(L(l, 3 - l) - L(l,\ l))$, where $\pi_l, l = 1{,}2$, are prior probabilities. The combined vector of population parameters $\boldsymbol{\Psi}$ includes parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, i.e. $\boldsymbol{\Psi} = (\boldsymbol{\beta}', \boldsymbol{\theta}')$, where $\boldsymbol{\theta}' = (\alpha, \sigma^2)$.

The risk for the BDF $W^B(Z_0, \boldsymbol{\Psi})$ is the same as specified in (2.10), i.e.

$$R^B = \textstyle\sum_{l=1}^{2}\sum_{k=1}^{2}\pi_l L(l,k)P_{lk}.$$

The squared *marginal Mahalanobis distance* is $\Delta^2 = (\mu_1 - \mu_2)^2/\sigma^2$ and the squared *conditional Mahalanobis distance* between conditional distributions of Z_0 given $\mathbf{T} = \mathbf{t}$ is then specified by

$$d^2 = \frac{(\mu_{1t} - \mu_{2t})^2}{\sigma_t^2} = \Delta^2(1 + \alpha h_0) = \Delta^2\rho_0.$$

In the population Ω_l, the conditional distribution of $W^B(Z_0, \boldsymbol{\Psi})$ given

$\mathbf{T} = \mathbf{t}$ is Gaussian distribution with mean and variance

$$E(W^B(Z_0, \boldsymbol{\Psi})|\mathbf{T} = \mathbf{t}; \Omega_l) = (-1)^{l+1} d^2/2 + \gamma^*,$$

$$Var(W^B(Z_0, \boldsymbol{\Psi})|\mathbf{T} = \mathbf{t}; \Omega_l) = d^2, l = 1..2.$$

By using the properties of the Gaussian distribution we obtain the closed-form expression for Bayes risk which has same form as one presented in (2.28). The explicit expression of the overall misclassification probability (special case of Bayes risk) associated with BDF for CAR model is derived in Dučinskas et al. 2013).

As it follows we shall denote the ML estimators of parameters by $\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}$ and put $\widehat{\boldsymbol{\Psi}} = \left(\widehat{\boldsymbol{\beta}}', \widehat{\boldsymbol{\theta}}'\right)$, where $\widehat{\boldsymbol{\theta}}' = (\hat{\alpha}, \hat{\sigma}^2)$. Then using (2.91), (2.92) we get the estimators of conditional mean and conditional variance

$$\hat{\mu}_{lt} = \mathbf{x}_0' \widehat{\boldsymbol{\beta}}_l + \widehat{\boldsymbol{\alpha}}_0' \left(\mathbf{t} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right), l = 1,2, \tag{2.94}$$

$$\hat{\sigma}_t^2 = \hat{\sigma}^2/(1 + \hat{\alpha} h_0). \tag{2.95}$$

Then by replacing the parameters with their estimators in (2.93) we form the PBDF

$$W^B\left(Z_0, \widehat{\boldsymbol{\Psi}}\right) =$$

$$\left(Z_0 - \widehat{\boldsymbol{\alpha}}_0'\left(\mathbf{t} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) - \mathbf{x}_0' \mathbf{I}_+ \widehat{\boldsymbol{\beta}}/2\right)' (\, \mathbf{x}_0' \mathbf{I}_- \widehat{\boldsymbol{\beta}})(1 + \hat{\alpha} h_0)/\hat{\sigma}^2 + \gamma^*. \tag{2.96}$$

Definition 3.1. The actual risk for PBDF $W^B\left(Z_0, \widehat{\boldsymbol{\Psi}}\right)$ specified in (2.96) is defined as

$$R^B\left(\widehat{\boldsymbol{\Psi}}\right) = \textstyle\sum_{l=1}^2 \left\{\pi_l L(l, l) + g_l \Phi(\hat{Q}_l)\right\}, \tag{2.97}$$

where the argument of $\Phi(\cdot)$ is

$$\hat{Q}_l = (-1)^l \left((a_l - \hat{b}) \operatorname{sgn}(\mathbf{x}_0' \mathbf{I}_- \widehat{\boldsymbol{\beta}})/\sigma_t + \gamma^* \hat{\sigma}_t^2 / \left|\mathbf{x}_0' \mathbf{I}_- \widehat{\boldsymbol{\beta}}\right| \sigma_t\right), \tag{2.98}$$

$$a_l = \mathbf{x}_0' \beta_l + \boldsymbol{\alpha}_0' (\mathbf{t} - \mathbf{X}\boldsymbol{\beta}), l = 1,2, \tag{2.99}$$

$$\hat{b} = \widehat{\boldsymbol{\alpha}}_0' \left(\mathbf{t} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) + \mathbf{x}_0' \mathbf{I}_+ \widehat{\boldsymbol{\beta}}/2. \tag{2.100}$$

Applying the asymptotic properties of the ML estimators established by Mardia and Marshall [43] (see Theorem 1.1) we conclude that the ML estimator $\widehat{\boldsymbol{\Psi}}$ is weakly consistent and asymptotically Gaussian, i.e.

$$\widehat{\boldsymbol{\Psi}} \sim AN_{2q+2}(\boldsymbol{\Psi}, \mathbf{J}^{-1}), \tag{2.101}$$

here the expected information matrix is given by

$$\mathbf{J} = \mathbf{J}_\beta \oplus \mathbf{J}_\theta, \tag{2.102}$$

where

$$\mathbf{J}_\beta = \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}, \tag{2.103}$$

$$\mathbf{J}_\theta = \frac{1}{2}\begin{pmatrix} n - 2tr\mathbf{V} + tr\mathbf{V}^2 & (tr\mathbf{V} - n)/\sigma^2 \\ (tr\mathbf{V} - n)/\sigma^2 & n/\sigma^4 \end{pmatrix}. \tag{2.104}$$

or

$$\mathbf{J}_\theta = \frac{1}{2}\begin{pmatrix} \alpha^2 \sum_{i=1}^{n-1}\left(\frac{\lambda_i}{1+\alpha\lambda_i}\right)^2 & -\alpha\sum_{i=1}^{n-1}\frac{\lambda_i}{(1+\alpha\lambda_i)\sigma^2} \\ -\alpha\sum_{i=1}^{n-1}\frac{\lambda_i}{(1+\alpha\lambda_i)\sigma^2} & n/\sigma^4 \end{pmatrix}.$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ are the ordered eigenvalues of $\mathbf{H}$.

Using the results of Theorem 1.1, the following asymptotic conclusions are valid

$$\widehat{\boldsymbol{\beta}} \sim AN_{2q}(\boldsymbol{\beta}, \mathbf{I}_\beta), \widehat{\boldsymbol{\theta}} \sim AN_2(\boldsymbol{\theta}, \mathbf{I}_\theta), E_T(\Delta\boldsymbol{\theta}(\Delta\widehat{\boldsymbol{\beta}})') \cong 0$$

where

$$\mathbf{I}_\beta = \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \tag{2.105}$$

$$\mathbf{I}_\theta = \mathbf{J}_\theta^{-1} = \frac{2\sigma^4}{n\,tr(\mathbf{V}^2) - (tr\mathbf{V})^2}\begin{pmatrix} n/\sigma^4 & (-tr\mathbf{V} + n)/\sigma^2 \\ (-tr\mathbf{V} + n)/\sigma^2 & n - 2tr\mathbf{V} + tr\mathbf{V}^2 \end{pmatrix}. \tag{2.106}$$

It could also be written in terms of eigenvalues

$$\boldsymbol{I}_\theta = \frac{2\sigma^4}{\alpha^2\left(\left(n\sum_i^{n-1}\frac{\lambda_i}{1+\alpha\lambda_i}\right)^2-\left(\sum_i^{n-1}\frac{\lambda_i}{1+\alpha\lambda_i}\right)^2\right)}\begin{pmatrix} n/\sigma^4 & \alpha\sum_{i=1}^{n-1}\frac{\lambda_i}{(1+\alpha\lambda_i)\sigma^2} \\ \alpha\sum_{i=1}^{n-1}\frac{\lambda_i}{(1+\alpha\lambda_i)\sigma^2} & \alpha^2\sum_{i=1}^{n-1}\left(\frac{\lambda_i}{1+\alpha\lambda_i}\right)^2 \end{pmatrix}.$$

Let $R_\beta^{(k)}$, $R_\theta^{(k)}$, $k = 1,2$ denote the $k-th$ order derivatives of $R^B(\widehat{\boldsymbol{\Psi}})$ with respect to $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$ evaluated at the point $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ and let $\mathbf{R}_{\beta\theta}^{(2)}$ denote the matrix of the second order partial derivatives of $R^B(\widehat{\boldsymbol{\Psi}})$ with respect to $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$ evaluated at the point $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}, \widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ and let the assumption **(A4)** hold.

Let $\mathbf{A}_\theta = \partial\widehat{\boldsymbol{\alpha}}_0/\partial\boldsymbol{\theta}'$ be the $n \times 2$ matrix and $\mathbf{s}_\theta = \partial\hat{\sigma}_t^2/\partial\widehat{\boldsymbol{\theta}}'$ two component vector of partial derivatives evaluated at the point $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}$. Hence we have

$$\mathbf{A}_\theta = (\mathbf{w}_0, \mathbf{0}_n)/(1+\alpha h_0)^2,\ \mathbf{s}_\theta = (-\sigma_t^2 h_0, 1)/(1+\alpha h_0). \qquad (2.107)$$

Put $\mathbf{L}' = \alpha\mathbf{w}_0'\mathbf{X}/(1+\alpha\mathbf{h}_0) - \mathbf{x}_0'(\mathbf{I}_+/2 + \gamma\mathbf{I}_-/d^2)$ and $\mathbf{v}' = (h_0, -1/\sigma_t^2)$.

Theorem 2.2. Suppose that observation Z_0 is to be classified by BPDF and let the conditions from Theorem 1.1 and the assumption (A4) hold. Then the approximation of ER is

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

where

$$K_\beta = \mathbf{L}'(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{L}\rho_0,$$

$$K_\alpha = 2n\mathbf{w}_0'\mathbf{V}\mathbf{w}_0/(\rho_0^3(n\, tr\mathbf{V}^2 - (tr\,\mathbf{V})^2))$$

$$K_\theta = \mathbf{v}'\mathbf{I}_\theta\mathbf{v}/\rho_0^2.$$

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

$$R^B(\widehat{\boldsymbol{\Psi}}) = R^B(\boldsymbol{\Psi}) + \mathbf{R}_\beta^{(1)}\Delta\widehat{\boldsymbol{\beta}} + \mathbf{R}_\theta^{(1)}\Delta\widehat{\boldsymbol{\theta}} + \frac{1}{2}\left((\Delta\widehat{\boldsymbol{\beta}})'\mathbf{R}_\beta^{(2)}\Delta\widehat{\boldsymbol{\beta}} + 2(\Delta\widehat{\boldsymbol{\beta}})'\mathbf{R}_{\beta\theta}^{(2)}\Delta\widehat{\boldsymbol{\theta}} + (\Delta\widehat{\boldsymbol{\theta}})'\mathbf{R}_\theta^{(2)}\Delta\widehat{\boldsymbol{\theta}}\right) + R_3, \qquad (2.108)$$

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

By using similar arguments as in Theorem 2.1 we obtain

$$\mathbf{R}_{\beta}^{(1)} = 0, \mathbf{R}_{\theta}^{(1)} = 0, \tag{2.109}$$

$$\mathbf{R}_{\beta}^{(2)} = g_1 d\varphi(-d/2 - \gamma^*/d)\mathbf{LL}'/\sigma_t^2, \tag{2.110}$$

$$\mathbf{R}_{\theta}^{(2)} = \frac{g_1 d\varphi(Q_1)}{\sigma_t^2}\left(-\mathbf{A}'_{\theta}(\mathbf{T} - \mathbf{X}\boldsymbol{\beta}) + \frac{\gamma^*\mathbf{s}_{\theta}}{d\sigma_t}\right)\left(-\mathbf{A}'_{\theta}(\mathbf{T} - \mathbf{X}\boldsymbol{\beta}) + \frac{\gamma^*\mathbf{s}_{\theta}}{d\sigma_t}\right)'. \tag{2.111}$$

It is easy to show that all elements of matrix $\mathbf{R}_{\beta\theta}^{(2)}$ are finite. From (2.106) we have the following approximation

$$E_T\left(\left(\Delta\widehat{\boldsymbol{\beta}}\right)'\mathbf{R}_{\beta}^{(2)}\left(\Delta\widehat{\boldsymbol{\beta}}\right)\right) \cong tr\left(\mathbf{R}_{\beta}^{(2)}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\right) = \sigma^2 tr\left(\mathbf{R}_{\beta}^{(2)}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\right).$$

Then using the assumption **(A4)** and replacing $E_T\left(\Delta\widehat{\boldsymbol{\theta}}\Delta\widehat{\boldsymbol{\theta}}'\right)$ by its asymptotic approximation $\mathbf{I}_{\theta}$ we get the following approximation

$$E_T\left(\left(\Delta\widehat{\boldsymbol{\theta}}\right)'\mathbf{R}_{\theta}^{(2)}\left(\Delta\widehat{\boldsymbol{\theta}}\right)\right) \cong tr\left(\mathbf{R}_{\theta}^{(2)}\mathbf{I}_{\theta}\right) =$$

$$g_1 d\varphi(-d/2 - \gamma^*/d)(tr(\boldsymbol{\Sigma}\mathbf{A}_{\boldsymbol{\theta}}\mathbf{J}_{\theta}^{-1}\mathbf{A}_{\boldsymbol{\theta}}{}') + (\gamma^*)^2\mathbf{s}'_{\theta}\mathbf{I}_{\theta}\mathbf{s}_{\theta}/d^2\sigma_t^2)/\sigma_t^2. \tag{2.112}$$

Then taking the expectation term by the term of the right–hand side of (2.108) with the dropped residual term, using (2.28), (2.107), (2.109)-(2.112), and replacing the moments of estimators by the corresponding approximations specified in (2.101)-(2.106) we complete the proof of theorem.

2.3.2 Multivariate case

In present section we are concerned with classification problems for a multivariate GMRF. Mardia [44] introduced a multivariate GMRF, and more recently Jin et al. [29], Sain and Cressie [53] explored these models. We focus on a subclass of a multivariate GMRF with parametrical structure proposed by Pettit et al. [49]. 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 [52] 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 two-class case with zero-one loss function. The main objective is to classify a single observation of multivariate GMRF $\{Z(\mathbf{s}): \mathbf{s} \in D \subset R^2\}$ when the training sample is given. The model of observation $\mathbf{Z}(\mathbf{s})$ is

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

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 error term is generated by p-variate zero-mean multivariate stationary GMRF $\{\varepsilon(\mathbf{s}): \mathbf{s} \in D\}$ and is specified with respect to the undirected graph which is described in the Section 2.2.2.

Suppose that $\mathbf{\Lambda}$ 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. Hence putting the matrix element equal to zero, gives conditional independence between the components of $\mathbf{Z}$. The full conditionals for $i = 0..n$ are specified as

$$\varepsilon_i | \varepsilon_{-i} \sim N(\mu_l^i, \mathbf{\Sigma}^i), \tag{2.113}$$

$$\mu_l^i = (\boldsymbol{\alpha}_i' \otimes \mathbf{I}_p)\varepsilon_{-i}, \ l = 1,2, \ \mathbf{\Sigma}^i = \boldsymbol{\rho}_i \mathbf{\Lambda}^{-1}. \tag{2.114}$$

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

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

Denote a training sample in vector form by $\mathbf{T} = \mathbf{Z}_{-0}$ and in matrix form by $\mathbf{T}^* = (\mathbf{Z}_1, \dots, \mathbf{Z}_n)'$. Design matrix $\mathbf{X}$ for training sample $\mathbf{T}$ is specified in (2.18). Then under some regularity conditions (Mardia [44]), the joint distribution for training sample in vector form is

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

and in matrix form follows $n \times p$ 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 $\mathbf{V}(\mathrm{\alpha})=(\mathbf{I}_n+\mathrm{\alpha}\mathbf{H})^{-1}$ denotes the spatial correlation matrix for $\mathbf{T}$.

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

$$(\mathbf{Z}_0|\mathbf{T}=\mathbf{t};\Omega_l)\sim N_p(\mu_{lt},\mathbf{S}_t), \tag{2.115}$$

where for $l=1{,}2$

$$\mu_{lt}=\mathbf{B}_l'\mathbf{x}_0+\left(\boldsymbol{\alpha}_0'\otimes\mathbf{I}_p\right)\left(\mathbf{t}-\mathrm{vec}(\mathbf{XB})\right)=$$

$$=\mathbf{B}_l'\mathbf{x}_0+\left(\boldsymbol{\alpha}_0'\otimes\mathbf{I}_p\right)(\mathbf{t}^{*}-\mathbf{XB})'\boldsymbol{\alpha}_0, \tag{2.116}$$

$$\mathbf{S}_t=\mathrm{\rho}_0\mathbf{\Lambda}^{-1}. \tag{2.117}$$

In the following let P_{0l} denote the conditional distribution specified in (2.115)-(2.117), for $l=1{,}2$. The squared Mahalanobis distance between the populations based on the conditional distribution for the observation taken at location $\mathbf{s}_0$ is $d^2=(\boldsymbol{\mu}_1^0-\boldsymbol{\mu}_2^0)'\mathbf{\Lambda}(\boldsymbol{\mu}_1^0-\boldsymbol{\mu}_2^0)/\rho_0$, where $\boldsymbol{\mu}_l^0=\mathbf{B}_l'\mathbf{x}_0, l=1{,}2$.

Then Bayes discriminant function minimizing the probability of misclassification is formed by log-ratio of conditional likelihood of distribution specified in (2.115)-(2.117), that is

$$W^B(\mathbf{Z}_0,\mathbf{\Psi})=(1+\alpha h_0)\left(\mathbf{Z}_0-\tfrac{1}{2}(\boldsymbol{\mu}_{1t}+\boldsymbol{\mu}_{2t})\right)'\mathbf{\Lambda}(\boldsymbol{\mu}_{1t}-\boldsymbol{\mu}_{2t})/\sigma^2+\gamma, \tag{2.118}$$

where $\gamma=\ln(\pi_1/\pi_2)$. Using (2.116), (2.117) and replacing $\mathbf{t}^{*}$ by $\mathbf{T}^{*}$ in (2.118) we get

$$W^B(\mathbf{Z}_0,\mathbf{\Psi})=$$

$$=(1+\alpha h_0)(\mathbf{Z}_0-\boldsymbol{\alpha}_0'(\mathbf{T}^{*}-\mathbf{XB})-\mathbf{x}_0'\mathbf{I}_{+}\mathbf{B}/2)'\mathbf{\Lambda}\mathbf{x}_0'\mathbf{I}_{-}\mathbf{B}/\sigma^2+\gamma. \tag{2.119}$$

Lemma 2.8. Bayes error rate for $W^B(\mathbf{Z}_0,\mathbf{\Psi})$ specified in (2.119) is

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

Proof. The proof of Lemma 2.8 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 $P^B = \sum_{l=1}^{2} \pi_l P_l((-1)^l W^B(\mathbf{Z}_0) \geq 0)$. The conditional distribution of $W^B(\mathbf{Z}_0, \boldsymbol{\Psi})$ in population Ω_l, given $\mathbf{T} = \mathbf{t}$, is Gaussian with mean and variance

$$E_l\big(W^B(\mathbf{Z}_0, \boldsymbol{\Psi})\big) = (-1)^{l+1} d^2/2 + \gamma,\, l = 1{,}2.$$

$$Var\big(W^B(\mathbf{Z}_0, \boldsymbol{\Psi})\big) = d^2.$$

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 σ^2 are unknown and the ML estimators $\widehat{\mathbf{B}}$ and $\hat{\sigma}^2$ based on $\mathbf{T}$ are used. Then the vector of parameters that are to be estimated and the vector of their estimators are denoted by $\boldsymbol{\Psi} = (\mathbf{B}, \sigma^2)$ and $\widehat{\boldsymbol{\Psi}} = \big(\widehat{\mathbf{B}}, \hat{\sigma}^2\big)$, respectively. After replacing $\boldsymbol{\Psi}$ by $\widehat{\boldsymbol{\Psi}}$ in (2.119), we get the PBDF

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

$$\times \Big(\mathbf{Z}_0 - \big(\mathbf{T}^* - \mathbf{X}\widehat{\mathbf{B}}\big)' \boldsymbol{\alpha}_0 - \widehat{\mathbf{B}}' \mathbf{I}_+' \mathbf{x}_0/2\Big)' \boldsymbol{\Lambda} \widehat{\mathbf{B}}' \mathbf{I}_-'\, \mathbf{x}_0/\hat{\sigma}^2 + \gamma. \qquad (2.120)$$

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

$$P^B\big(\widehat{\boldsymbol{\Psi}}\big) = \sum_{l=1}^{2} \pi_l P_{0l}\big((-1)^l W^B\big(\mathbf{Z}_0, \widehat{\boldsymbol{\Psi}}\big) > 0\big). \qquad (2.121)$$

Lemma 2.9. The *actual error rate* for PBDF specified in (32) is

$$\mathrm{P}^B\big(\widehat{\boldsymbol{\Psi}}\big) = \sum_{l=1}^{2} \pi_l \Phi(\hat{Q}_l), \qquad (2.122)$$

where

$$\hat{Q}_l = (-1)^l \frac{(1+\alpha h_0)\Big(x_0'\big(\mathbf{B}_l - \mathbf{I}_+ \widehat{\mathbf{B}}/2\big) + \boldsymbol{\alpha}_0' \mathbf{X}\big(\Delta\widehat{\mathbf{B}}\big)\Big) \boldsymbol{\Lambda} \widehat{\mathbf{B}}' \mathbf{I}_-' \mathbf{x}_0 + \gamma \hat{\sigma}^2}{\hat{\sigma} \sqrt{\mathbf{x}_0' \mathbf{I}_- \widehat{\mathbf{B}} \boldsymbol{\Lambda} \widehat{\mathbf{B}}' \mathbf{I}_-' \mathbf{x}_0 (1+\alpha h_0)}} \qquad (2.123)$$

with $\Delta\widehat{\mathbf{B}} = \widehat{\mathbf{B}} - \mathbf{B}$.

Proof. Proof of lemma is established by essentially exploiting the proof of Lemma 2.8, formulas (2.120), (2.121) and properties of multivariate Gaussian distribution.

In order to derive the approximation of expected error rate (AEER) we will use the ML estimator of the regression coefficients $2q \times p$ matrix $\widehat{\mathbf{B}} = (\mathbf{X}'\mathbf{V}^{-\mathbf{1}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-\mathbf{1}}\mathbf{T}^{*}$ and the bias adjusted ML estimator of feature covariance matrix

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

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

$$\widehat{\mathbf{B}} \sim N_{2q\times p}(\mathbf{B}, \sigma^2(\mathbf{X}'\mathbf{V}^{-\mathbf{1}}\mathbf{X})^{-1} \otimes \boldsymbol{\Lambda}^{-1}) \qquad (2.124)$$

and

$$\hat{\sigma}^2 \sim \frac{\sigma^2\chi^2(np-2q)}{np-2q}. \qquad (2.125)$$

Theorem 2.3. Suppose that observation $\mathbf{Z}_0$ to be classified by BPDF specified in (2.120). Then the approximation of EER is

$$AEER = \sum_{l=1}^{2} \pi_l \Phi(-d/2 + (-1)^l \gamma/d) + \pi_1 \varphi(-d/2 - \gamma/d) \times$$

$$\{\mathbf{F}_0'\mathbf{R}_B\mathbf{F}_0 d/k_0 + (p-1)\mathbf{x}_0'\mathbf{I}_-\mathbf{R}_B\mathbf{I}_-'\mathbf{x}_0/(k_0 d) + 2\gamma^2/d(np-2q)\}/2,$$

where

$$F_0 = \mathbf{X}'\boldsymbol{\alpha}_0 - (\mathbf{I}_+'/2 + \gamma\mathbf{I}_-/d^2)\mathbf{x}_0,$$

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

Proof. Let $\Delta\widehat{\mathbf{B}} = \widehat{\mathbf{B}} - \mathbf{B}$, $\Delta\hat{\sigma}^2 = \hat{\sigma}^2 - \sigma^2$. Using (2.124)-(2.125) it is easy to show (e.g. Magnus and Neudecker [40]) that

$$E_T(\Delta\widehat{\mathbf{B}}) = 0, E_T(\Delta\hat{\sigma}^2) = 0, E_T(\Delta\hat{\sigma}^2\Delta\widehat{\mathbf{B}}) = 0 \qquad (2.127)$$

and

$$E_T\left(vec(\varDelta\widehat{\mathbf{B}}')\left(vec(\varDelta\widehat{\mathbf{B}}')\right)'\right) = \mathbf{R}_B \otimes \mathbf{\Sigma}, \tag{2.128}$$

$$E_T(\varDelta\hat{\sigma}^2) = \sigma^4/(np - 2q). \tag{2.129}$$

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

$P_B^{(1)}(\alpha,\beta) = \partial P(\widehat{\mathbf{\Psi}})/\partial\hat{b}_{\alpha\beta}$,

$P_B^{(2)}(\alpha\beta,\gamma\delta) = \partial^2 P(\widehat{\mathbf{\Psi}})/\partial\hat{b}_{\alpha\beta}\partial\hat{b}_{\gamma\delta}$,

$P_\sigma^{(1)} = \partial P(\widehat{\mathbf{\Psi}})/\partial\hat{\sigma}_{ij}$,

$P_\sigma^{(2)} = \partial_\sigma^2 P(\widehat{\mathbf{\Psi}})/\partial^2\hat{\sigma}^2$ and

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

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

Make a Taylor expansion of $P^B(\widehat{\mathbf{\Psi}})$ at the points $\widehat{\mathbf{B}} = \mathbf{B}$ and $\hat{\sigma}^2 = \sigma^2$ 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 (2.127) we get

$$E_T\{P(\widehat{\mathbf{\Psi}})\} = P^B(\mathbf{\Psi}) + \left(\sum_{\alpha,\gamma=1}^{2q}\sum_{\beta,\delta=1}^{p} P_B^{(2)}(\alpha\beta,\gamma\delta)E_T\{\Delta\hat{b}_{\alpha\beta}\Delta\hat{b}_{\gamma\delta}\} + P_\sigma^{(2)}E_T\{\Delta\hat{\sigma}^2\}\right)/2 + R_3, \tag{2.13}$$

where R_3 is the expectation of remainder term. Note that

$$\pi_1\varphi(Q_1) = \pi_2\varphi(Q_2). \tag{2.131}$$

By using the chain rule and equation (2.131) we have

$$P_B^{(2)}(\alpha\beta,\gamma\delta) = \pi_1\varphi(Q_1)\sum_{l=1}^{2}(-1)^l\left(Q_l Q_{lB}^{(1)}(\alpha\beta)Q_{lB}^{(1)}(\gamma\delta) - Q_{lB}^{(2)}(\alpha\beta,\gamma\delta)\right), \tag{2.132}$$

$$P_\sigma^{(2)} = \pi_1\varphi(Q_1)\sum_{l=1}^{2}(-1)^l\left(Q_l\left(Q_{l\sigma}^{(1)}\right)^2 - Q_{l\sigma}^{(2)}\right). \tag{2.133}$$

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

$$Q_{lB}^{(1)}(\alpha\beta) =$$

$$= (-1)^l(\mathbf{X}'\boldsymbol{\alpha}_0 - (\mathbf{I}_+'/2 + \gamma\mathbf{I}_-'/d^2)\mathbf{x}_0)\mathbf{J}^{\alpha\beta}\boldsymbol{\Lambda}\mathbf{B}'\mathbf{I}_-\mathbf{x}_0/(dk_0\sigma^2), \quad (2.134)$$

$$\sum_{l=1}^{2} Q_{lB}^{(2)}(\alpha\beta,\gamma\delta) = (\sigma^2)^{-1}\big(\mathbf{x}_0'\mathbf{I}_-\mathbf{J}^{\alpha\gamma}\mathbf{I}_-'\mathbf{x}_0 -$$

$$\mathbf{x}_0'\mathbf{I}_-\mathbf{J}^{\alpha\beta}\boldsymbol{\Lambda}\mathbf{B}'\mathbf{I}_-'\mathbf{x}_0\mathbf{x}_0'\mathbf{I}_-\mathbf{J}^{\gamma\delta}\boldsymbol{\Sigma}^{-1}\mathbf{B}'\mathbf{I}_-'\mathbf{x}_0/d^2k_0\big)/(dk_0), \quad (2.134)$$

$$Q_{l\sigma}^{(1)} = (-1)^l\gamma\mathbf{x}_0'\mathbf{I}_-\mathbf{B}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{B}'\mathbf{I}_-\mathbf{x}_0'/(d^3k_0\sigma^2), \quad (2.135)$$

$$\sum_{l=1}^{2} Q_{l\sigma}^{(2)} =$$

$$(\mathbf{x}_0'\mathbf{I}_-\mathbf{B}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\mathbf{B}'\mathbf{I}_-'\mathbf{x}_0 - \mathbf{x}_0'\mathbf{I}_-\mathbf{B}\boldsymbol{\Lambda}\mathbf{B}'\mathbf{I}_-'\mathbf{x}_0\mathbf{x}_0'\mathbf{I}_-\boldsymbol{\Lambda}\mathbf{B}'\mathbf{I}_-'\mathbf{x}_0/d^2k_0\sigma^4)/(dk_0), \quad (2.136)$$

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

Remember, that the Lagrange remainder is the third-order polynomial with respect to the components of $\Delta\widehat{\mathbf{B}}$ and $\Delta\hat{\sigma}^2$. Coefficients of this polynomial are the third-order partial derivatives of $P^B(\widehat{\boldsymbol{\Psi}})$ with respect to components of $\widehat{\mathbf{B}}$ and $\hat{\sigma}^2$ estimated in the neighbourhood of their true values. It is obvious from (2.124) and (2.125), that all third order moments of normally distributed components of $\Delta\widehat{\mathbf{B}}$ 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 elements of $\widehat{\mathbf{B}}$ and $\hat{\sigma}^2$ are bounded by the uniformly bounded functions in the neighbourhood of point $\widehat{\mathbf{B}} = \mathbf{B}$, $\hat{\sigma}^2 = \sigma^2$. So we can scrap R_3 in (2.130).

Finally, putting (2.131-2.136) into (2.130) and using (2.128), (2.129) we complete the proof of the theorem.

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. Univariate two-class and multiclass cases are analysed. For the two-class case the comparison of AER values using a symmetric and asymmetric training sample plan is done. For the multiclass case calculations are performed for grouped and mixed training sets of locations. Actual error rate and its approximation are also accomplished for GMRF. Section 3.3 illustrates the effects of two different spatial sampling designs on AER. In section 3.4 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 (Section 1.3) is presented. The numerical experiment is performed with simulated data, where geometrically anisotropic GRFs were simulated using package geoR. The empirical power of test is examined for 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 integer regular 2-dimensional lattice. Set $h' = (h_x, h_y)$ for each $h \in D$. Simulations are done on 10×10 square grid, thus the sample size is $n = 121$. We generate the realizations from zero-mean, stationary GRF. The case of nuggetless covariance model, $C(h) = \sigma^2 r(h)$, with geometric anisotropic spatial Gaussian correlation function $r(h) = exp\{-(h_x^2 + \lambda^2 h_y^2)/\alpha^2\}$ is considered. Suppose that the anisotropy angle φ is equal to $\frac{\pi}{2}$.

We will start with $K = 2$, $|h_1| = |h_3| = 1$ and $|h_2| = |h_4| = 2$. It implies that only two lags are used. Then we build the contrast matrix and calculate the test statistic specified in (1.24)

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

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

Its approximate distribution is the χ_2^2 distribution. $N(h_i)$ and $N(h_{i+2})$ represent the pairs (s_i, s_j) for a certain lag and $|N(h_i)|$, $|N(h_{i+2})|$ are the numbers of such pairs in two orthogonal directions (see Figure 2).

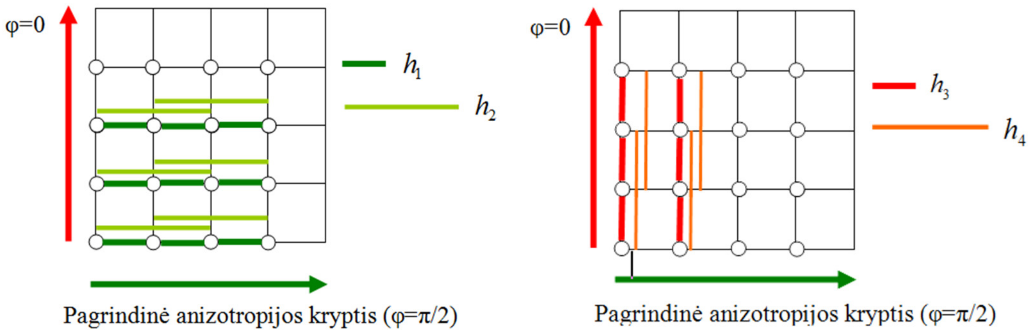
φ=0
h_1
h_2
Pagrindinė anizotropijos kryptis (φ=π/2)
φ=0
h_3
h_4
Pagrindinė anizotropijos kryptis (φ=π/2)

Figure 2. Sampling pairs for directional semivariograms: the main direction of anisotropy, $\varphi = 0^0$, and the orthogonal direction, $\varphi = \frac{\pi}{2}$

As a performance measure of the proposed test statistic we considered the empirical power of test (frequency of rejecting H_0 for simulated geometric anisotropic Gaussian data) with significance level, $p = 0.05$. For various values of anisotropy ratio λ and range parameters α, three simulation procedures of size $M = (150, 300, 600)$ are performed. Table 1 shows that empirical power of test increases with increasing of range parameter, but 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

	α					
	M=150					
λ	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	0.8	0.78
	M=600					
2	0.05	0.57	0.71	0.76	0.8	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	0.8	0.81
10	0.05	0.52	0.67	0.79	0.8	0.84

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

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

	α					
	K=2					
λ	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, to analyse the influence of covariance parameters and to evaluate the accuracy of the derived AER, a simulation study was carried out. Consider the case of classification the scalar observation Z_0 to 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 analysed. Also the equal-sized training samples with equal prior probabilities are assumed, $n_1 = n_2 = 4,\ \pi_1 = \pi_2 = 1/2$.

The observations are assumed to arise from a stationary Gaussian random field with constant mean and nuggetless covariance function given by $C(h,\mathbf{\theta}) = \sigma^2 r(h)$, where σ^2 is the unknown variance (partial sill) and $r(h)$ is a spatial correlation function. The exponential geometric anisotropic correlation function with unknown range parameter, α, unknown anisotropy ratio, λ, and known anisotropy angle $\varphi = \pi/2$, specified by

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

is considered. Here $h_x = x_i - x_j, h_y = y_i - y_j, i,j = 1...n.$

Hence, the vector of unknown covariance parameters has three components, i.e. $\mathbf{\theta} = (\sigma^2,\ \lambda, \alpha)'$.

The training sample $\mathbf{T} = (\mathbf{T}_1', \mathbf{T}_2')' = (Z(\mathbf{s}_1), Z(\mathbf{s}_2), \dots, Z(\mathbf{s}_n))'$ is observed on a regular 2-dimensional lattice with unit spacing.

Consider the case $\mathbf{s}_0 = (1,1)$ and fixed set of training locations $\mathbf{S}_n$ 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 ξ_1 and ξ_2 are analysed:

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

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

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

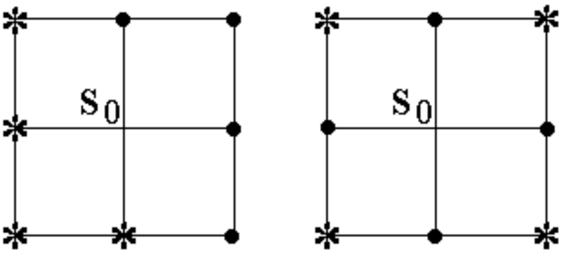

Figure 3. Two different STL, where ● represents the elements of $\mathbf{S}^{(1)}$ and * represents the elements of $\mathbf{S}^{(2)}$.

In order to realise AER formula some functions included into *geoR* were used, i.e. function *grf()* was used for simulation of Gaussian random field; *variog()* was applied for calculating 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 anisotropic covariance matrix, so it is necessary to create a procedure that allows the anisotropy parameters to be included in the covariance function structure.

The values of AER specified in (2.37)-(2.40) are calculated for various values of parameters λ and α. The results of the calculations with $\Delta = 1$ are presented in Table 3 and Table 4. Analysing the contents of the tables we can conclude that for both STL, AER values are decreasing with the increase of anisotropy ratio λ and are decreasing with the increase of range parameter α. It means the higher level of anisotropy, the lower are the values of AER. And the stronger spatial correlation yields the lower AER values.

Table 3. Values of AER for STL ξ_1 with $\Delta = 1$ and various λ and α.

	α							
λ	0.6	0.8	1.2	1.6	2	2.4	2.8	3.2
1	0.328438	0.307256	0.270947	0.240968	0.215754	0.194206	0.175548	0.159223
2	0.334138	0.313996	0.281444	0.255017	0.232404	0.212669	0.195248	0.179737
3	0.334396	0.313624	0.280805	0.254897	0.232919	0.213723	0.196722	0.181530
4	0.333937	0.313140	0.280474	0.254719	0.232910	0.213868	0.196990	0.181895
5	0.333237	0.312551	0.280198	0.254600	0.232881	0.213904	0.197076	0.182018
6	0.332543	0.312004	0.279944	0.254496	0.232850	0.213913	0.197112	0.182072
7	0.331927	0.311540	0.279721	0.254400	0.232816	0.213911	0.197128	0.182101
8	0.331398	0.311157	0.279533	0.254314	0.232781	0.213903	0.197135	0.182116
9	0.330946	0.310840	0.279378	0.254238	0.232747	0.213892	0.197136	0.182124
10	0.330559	0.310576	0.279249	0.254173	0.232716	0.213879	0.197134	0.182128

Table 4. Values of AER for STL ξ_2 with $\Delta = 1$ and various λ and α

	α							
λ	0.6	0.8	1.2	1.6	2	2.4	2.8	3.2
1	0.330108	0.310382	0.276753	0.248792	0.224998	0.204422	0.186405	0.170478
2	0.335293	0.316316	0.286035	0.261460	0.240264	0.221588	0.204943	0.189991
3	0.335625	0.316118	0.285660	0.261593	0.241000	0.222838	0.206599	0.191959
4	0.335198	0.315747	0.285613	0.261787	0.241399	0.223403	0.207291	0.192748
5	0.334506	0.315202	0.285500	0.261927	0.241684	0.223781	0.207734	0.193236
6	0.333814	0.314669	0.285326	0.261978	0.241861	0.224032	0.208032	0.193564
7	0.333199	0.314210	0.285141	0.261971	0.241960	0.224194	0.208232	0.193789
8	0.332669	0.313828	0.284971	0.261935	0.242008	0.224295	0.208368	0.193946
9	0.332217	0.313511	0.284823	0.261886	0.242026	0.224357	0.208458	0.194056
10	0.331831	0.313247	0.284698	0.261836	0.242026	0.224393	0.208519	0.194134

Comparing the values of two STL it was noticed that the symmetric STL (ξ_1) was more optimal than asymmetric STL (ξ_2) by the AER minimum criterion since the values of the ratio $AER_{\xi_2}/AER_{\xi_1}$for all parametric structures is greater than 1 [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 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 $C(h) = \sigma^2 \exp\{-h/\alpha\}$. The set of training locations $\mathbf{S}_{12}$ that forms the third order neighbourhood for $s_0 = (0,0)$ is considered. We will analyse two different STL: with grouped labels (STLG) and mixed labels (STLM). The distributions of STLG and STLM are shown in Figure 4.

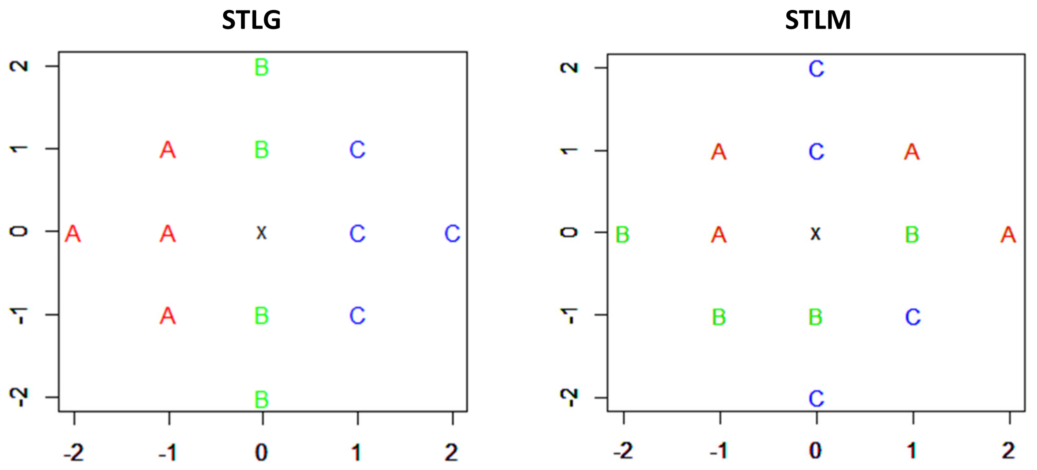

Figure 4. STL $\mathbf{S}_{12}$ with different labels distributions. The points indicated by A, B and C belong to $\mathbf{S}^{(1)}$, $\mathbf{S}^{(2)}$ and $\mathbf{S}^{(3)}$, respectively. Sign × denotes s_0.

All of the simulations have considered small training sample sizes, i.e. $n_l = 4$ and equal prior probabilities $\pi_l = \frac{1}{3}, l = 1..3$.

1000 simulations (runs) were performed for each STL. For each simulated training sample the actual error rate $P^B(\widehat{\boldsymbol{\Psi}})$ specified in Lemma 2.4 (2.63) was calculated. Expected error rate obtained by averaging actual error rates over runs is denoted by $\overline{EER}$. AEER derived in Theorem 2.2 is also calculated and its accuracy is evaluated by $\eta = |AEER - \overline{EER}|$.

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

Table 5 contains the values of AEER, $\overline{EER}$ calculated for various levels of spatial correlation and class separation specified by parameters α 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 Mahalanobis distances $|d_{lk}|$ between classes are proportional to $b/\sqrt{K}$, $K = 1 - \mathbf{r}_0'\mathbf{R}^{-1}\mathbf{r}_0$ and K decreases as α increases. So the separation between classes increases with increasing of α and b.

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

Table 5. AEER and $\overline{EER}$ values for different α and b

			b							
		α	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
	$\overline{EER}$	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
	$\overline{EER}$	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, η, are depicted in Figure 5. It shows the advantage of STLG against STLM. It means that proposed approximation of EER is more precise when classes are not mixed over the region.

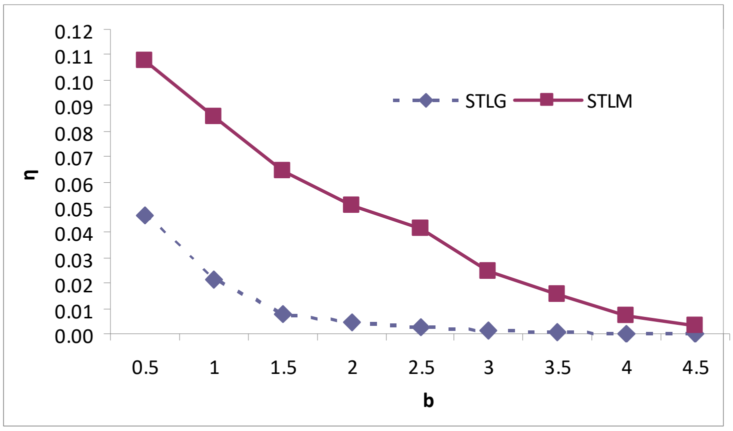

Figure 5. Comparison of η for different labels distributions and $\alpha = 2$

The described results are published in Dučinskas and Dreižienė [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 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 the actual risk and the approximation of the expected risk (section 2.3) in the finite training sample case are calculated and the influence of statistical parameters is demonstrated. The case $n_1 = n_2 = 60$, $\pi_1 = \pi_2 = 1/2$ 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, α, and the marginal Mahalanobis distance, Δ.

Assume that GMRFs are sampled on the 11×11 regular unit spacing lattice $\mathbf{S}_{120}$ with the focal location in the centre of the 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.

For the simulations, the true values of the parameters are fixed at $\beta_1 = 0, \beta_2 = 1, \sigma^2 = 1$. Numerical illustration of the estimates of parameters is presented in Table 6. This table shows that for all selected values of the spatial dependence parameter α, there are no significant biases of the considered estimators.

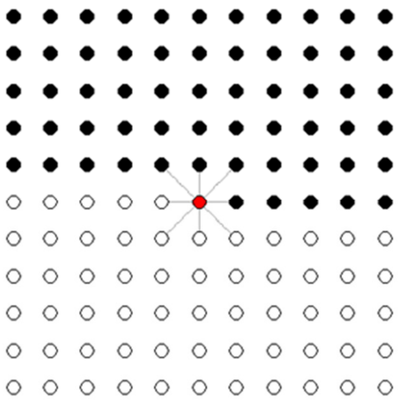

Figure 6. Set of training locations with focal location. The points indicated by ● and ○ belong to $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$, respectively

Table 6. ML estimates of β, σ^2 and Mahalanobis distance d with $\hat{d}$ for various values of α.

α	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\sigma}^2$	$\hat{d}$	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 the AER is evaluated by relative error $\eta = |AER - \overline{ER}|/\overline{ER}$. Table 7 shows the values of the AER and $\overline{ER}$ calculated with respect to α and Δ. The results show that all AER and $\overline{ER}$ values are decreasing while α and Δ are increasing. That means the greater separation between classes and the greater spatial dependence parameter give better accuracy of the proposed AER.

Figure 6 shows that the accuracy of the AER is sufficiently stable with respect to the increase in α. However, it was noticed that the general trend for the relative error of the AER is an increase in the distance between populations.

Table 7. $\overline{\mathrm{ER}}$ and AER values for different values of α and different class separation (Δ)

		$\overline{ER}$			AER	
α	Δ=0.5	Δ=1	Δ=2	Δ=0.5	Δ=1	Δ=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

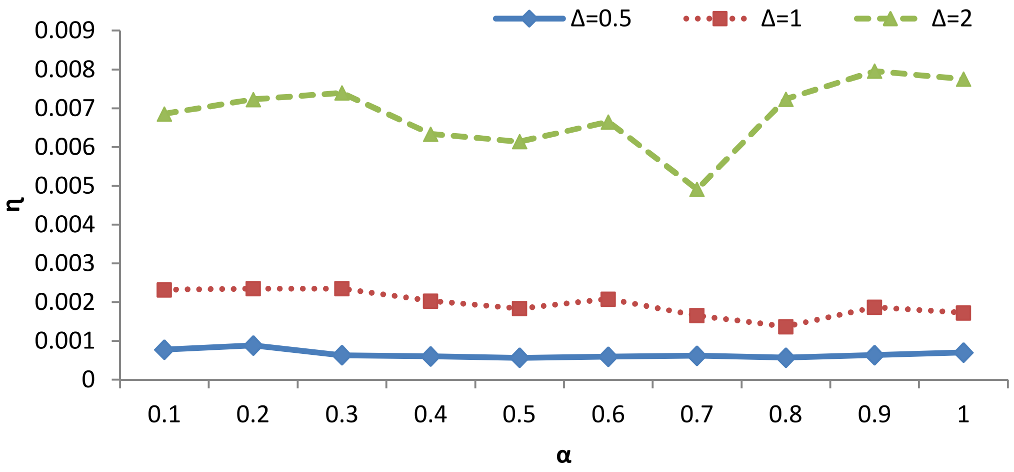

Figure 6. Relative $^{\text{error}}$ of the AER for various values of α and three values of *Δ*.

3.3. The influence of sample size to the AER

The results presented in section 3.2 are based on small training samples, e.g., $n = 8$. It is obvious that increasing training sample the better classification accuracy could be achieved, since the greater training sample gives more information about population parameters. As it was mentioned in the section 1.1 there are three ways to increase the training sample: increasing domain asymptotics framework, infill asymptotics and mixed domain asymptotics.

This numerical experiment is considered to investigate the influence of training sample increase on AER using infill asymptotics and increasing domain frameworks.

Assume that D is a regular 2-dimensional lattice with unit spacing. Consider the case $s_0 = (4, 4)$ and eight fixed STL S_{mn}, $m = 1, 2$, where 1 denotes infill asymptotic sampling framework, 2 denotes increasing domain asymptotic sampling framework and n represents the size of training sample, $n = 8, 16, 32, 98$. For example S_{m8} contains 8 neighbors of s_0, S_{m16} contains 16 neighbors of s_0, and so on. S_{mn} is partitioned into a union of two disjoint subsets, i.e., $S_{mn} = S^{(1)} \cup S^{(2)}$, where $S^{(l)}, l = 1{,}2$ is the subset of S_{mn} that contains n_j locations of feature observations from Ω_l and let $n_1 = n_2$.

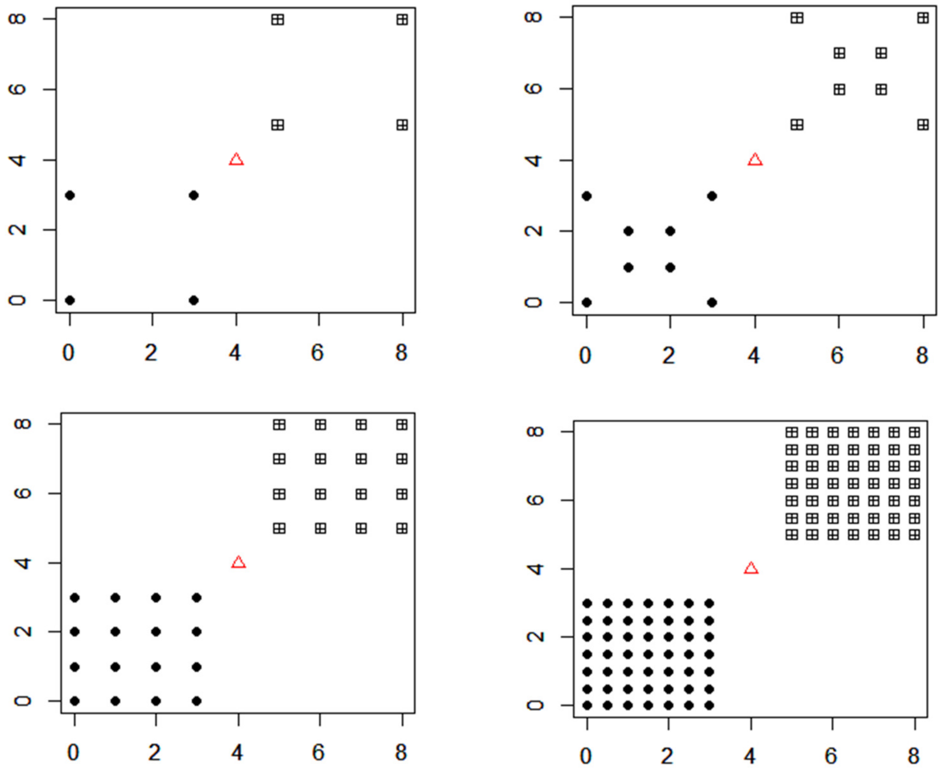

Figure 7. Infill asymptotic sampling framework with different training sample sizes $n = \{8, 16, 32, 98\}$; the symbols ●, ⊞ and △ represent $S^{(1)}, S^{(2)}$ and $\mathbf{s}_0$, respectively.

Figure 7 represents STL for infill asymptotic sampling framework (S_{1n}), here extra locations are taken from between observed locations. The STL for increasing domain asymptotic sampling framework (S_{2n}), are shown in figure

8. In this case the extra locations are taken by increasing the domain of observations.

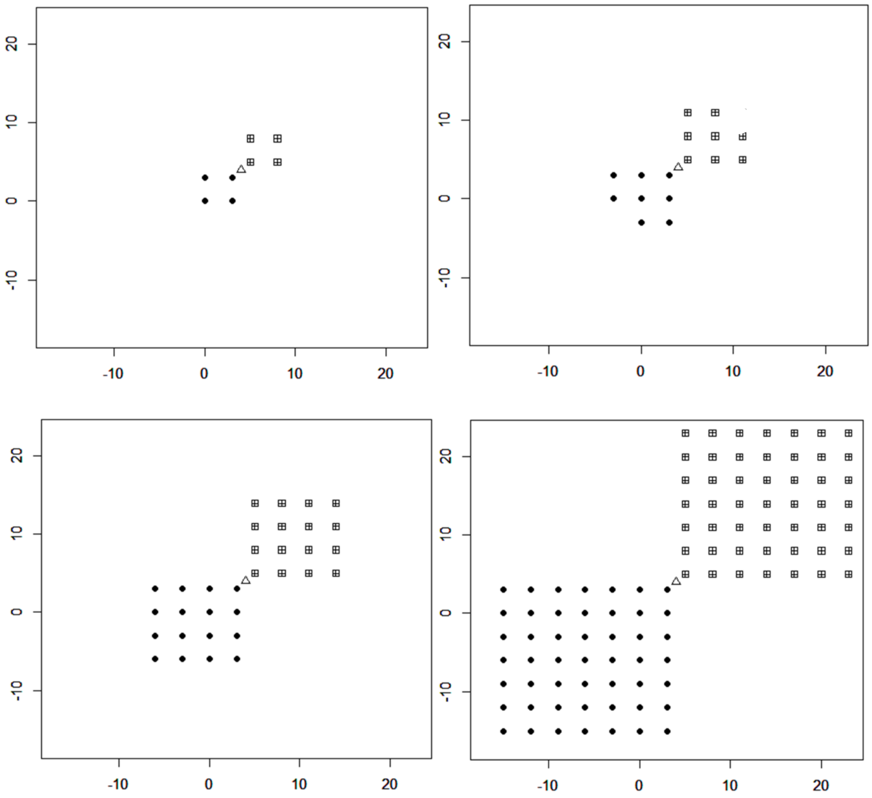

Figure 8. Increasing domain asymptotics sampling framework with different training sample sizes $n\ = \{8,16,32,98\}$; the symbols ●, ⊞ and △ represent $S^{(1)}, S^{(2)}$ and $\mathbf{s}_0$, respectively.

With an insignificant loss of generality the case with $\pi_j = 0{,}5$ and $L(i,j) = 1 - \delta_{ij}, i,j = 1,2$ is considered. Observations are assumed to arise from stationary GRF with different constant mean and common nuggetless covariance function given by $C(h) = \sigma^2 r(h)$, where σ^2 is variance (partial sill) and $r(h) = exp\{-\sqrt{h_x^2 + \lambda^2 h_y^2}/\alpha\}$ is the exponential geometric anisotropic correlation function with anisotropy ratio λ and anisotropy angle $\varphi = \pi/2$. We consider the case with unknown mean and anisotropy ratio parameters.

Figure 9 shows the values of AER using infill asymptotics and increasing domain asymptotics sampling frameworks. AER are calculated assuming Mahalanobis distance between marginal distributions $\Delta = 1$ and $\alpha = 0.6$, $\sigma^2 = 1$. The results show that less values of AER are obtained using increasing domain sampling framework. AER values are decreasing while training sample size increases for both sampling frameworks and for both isotropic and anisotropic cases ($\lambda = 1$ and $\lambda = 2$).

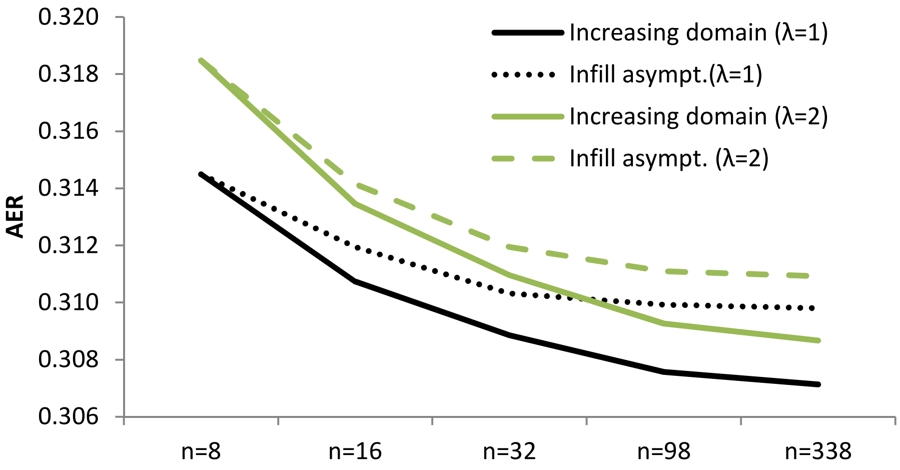

Figure 9. AER values using different asymptotic frameworks: isotropic case ($\lambda = 1$), anisotropic case ($\lambda = 2$)

Table 8 shows the ratio of AER calculated using an increasing domain asymptotic sampling framework (AER_{inc}) to infill asymptotic sampling framework (AER_{inf}). It is obvious that AER_{inc} /AER_{inf} increases while α is increasing for all training sample sizes. This leads to the conclusion that for greater α values and greater training sample size the infill asymptotics sampling framework gives lower values of AER in comparison with the increasing domain asymptotics sampling framework.

Table 8. AER_{inc}/AER_{inf} with different α values and fixed $\Delta = 1$ and $\lambda = 1$.

	AER_{Inc}/AER_{Inf}			
α	N=16	N=32	N=98	N=338
0.8	0.9954	0.9947	0.9925	0.9920
1.2	0.9960	0.9962	0.9955	0.9960
1.6	0.9972	0.9983	0.9988	1.0001
2.0	0.9981	1.0000	1.0015	1.0034
2.4	0.9987	1.0014	1.0037	1.0061
2.8	0.9992	1.0025	1.0055	1.0084
3.2	0.9996	1.0035	1.0070	1.0104

3.4. 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 10). 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 litre 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 ($1.584km^2$), shallow (average depth 3.8m) 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 3–4m depth and occurring on both hard substrates and soft bottoms (Zaiko, Daunys [68]).

Figure 10. 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 analyse the influence of spatial correlation to the misclassification errors.

The training sample consists of $n = 39$ observations (see Fig. 11). The red dots, $n_1 = 22$, represent the locations (stations in the Curonian Lagoon) were 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 and the remaining ones could be included into design matrix as covariates.

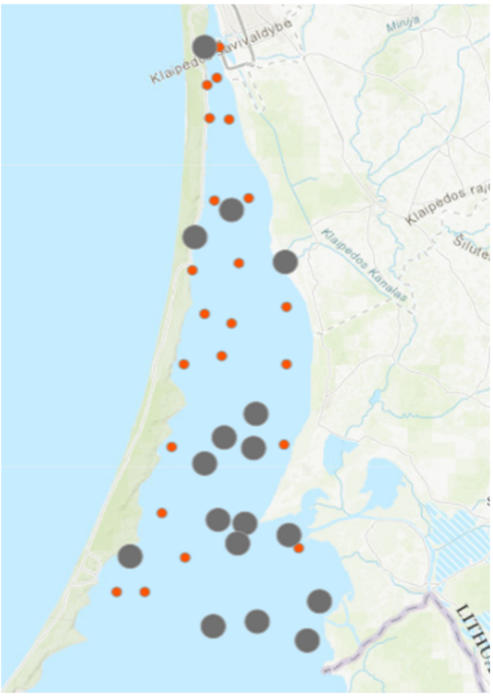

Figure 11. The set of training locations. Red ● and grey ⬤ dots represent absence and presence of zebra mussel, respectively

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. The existence of spatial correlation could also be confirmed by semivariograms (see Figure 12).

Table 9. 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

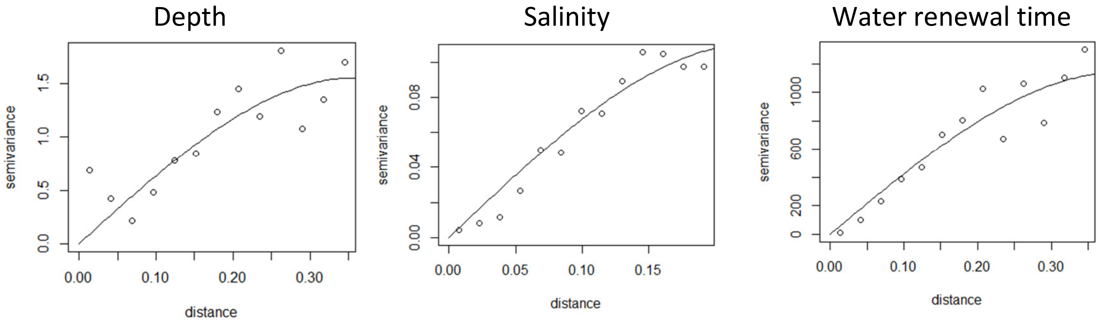

Figure 12. Semivariograms for the observed variables

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

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & y_1 & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & y_{n1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & x_{n1+1} & y_{n1+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 1 & x_n & y_n \end{pmatrix}.$$

Lastly, spatial information could be used for the estimation of prior probabilities. The simplest way is to assume that the populations are equiprobable, that is, $\pi_1 = \pi_2$, 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 acount. 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 $\pi_l = \frac{n_l}{n}, l = 1..2$ (Theodoritis 2009). Another way is to include only the nearest neighbours of $\mathbf{s}_0$. Then

$\pi_l = \frac{n_{0l}}{n_0}$, where n_{0l} is the number of nearest neighbours of $\mathbf{s}_0$ in population $\mathbf{T}_l, l = 1..2$ and $n_0 = n_{01} + n_{02}$.

For every observed variable different means models were used: 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 analysed.

Let us start with univariate case of GMRF where the covariance is defined as $\mathbf{\Sigma} = \sigma^2(\mathbf{I}_n + \alpha\mathbf{H})^{-1}$. Let σ^2 be the unknown parameter and α is a known and equal to 0.5. To construct the matrix $\mathbf{H}$ a different number of neighbours 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 realised and according to the sign of $W^B(Z_0, \widehat{\mathbf{\Psi}})$ the decision was made. To evaluate the performance of the discriminant function the cross validation procedure was applied.

Table 10 shows the results for the variable *Depth*. There the probabilities of correct classification are presented. These probabilities were calculated for different mean models and for different number of neighbours. The last column corresponds to the situation with no spatial correlation. The remaining columns correspond to the situations with different number of neighbours. The best model which gives greatest (73.7%) correct classification probability is

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

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

Table 10. Correct classification probabilities for variable Depth

		Neighborhood structure				
	Distance	$d_{\max} = 0.1$	$d_{\max} = 0.2$	$d_{\max} = 0.3$	$d_{\max} = 0.4$	*No spatial correlation*
	Number of neighbours	[4-16]	[15-35]	[24-38]	[36-38]	0
Mean model	~1	0.718	0.641	0.641	0.641	0.385
	~1+XY	**0.737**	0.658	0.684	0.711	0.632
	~1+XY+Restime	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+Restime+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 11), but this time the best model which gives the greatest (74.4%) correct classification probability is

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

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

Table 11. Correct classification probabilities for variable Salinity

		Neighbourhood structure				
	Distance	$d_{\max} = 0.1$	$d_{\max} = 0.2$	$d_{\max} = 0.3$	$d_{\max} = 0.4$	$d_{\max} = 0$
	Number of neighbours	[4-16]	[15-35]	[24-38]	[36-38]	0
Mean model	~1	**0.744**	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

Using *Water renewal time* for classification we get the results which are presented in the Table 12. The best model here, which gives the greatest (76.9%) correct classification probability, is

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

where $\mathbf{X\beta}$ is the constant mean model.

Table 12. Correct classification probabilities for variable *Water renewal time*

		Neighborhood structure				
	Distance	$d_{\max} = 0.1$	$d_{\max} = 0.2$	$d_{\max} = 0.3$	$d_{\max} = 0.4$	$d_{\max} = 0$
	Number of neighbors	[4-16]	[15-35]	[24-38]	[36-38]	0
Mean model	~1	**0.769**	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

Analysing Figure 13 it is ease to notice that including spatial correlation into covariance gives higher correct classification probabilities. The influence of prior probabilities is depicted in Figure 14. The upper line corresponds to the case of equiprobable populations. The bottom line shows the correct classification probabilities when priors are estimated icluding only nearest neighbours but not the whole training sample.

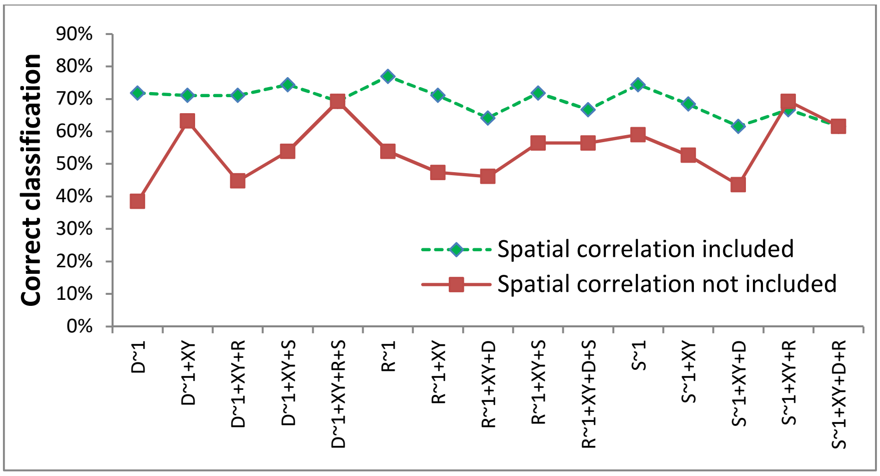

Figure 13. The influence of spatial correlation

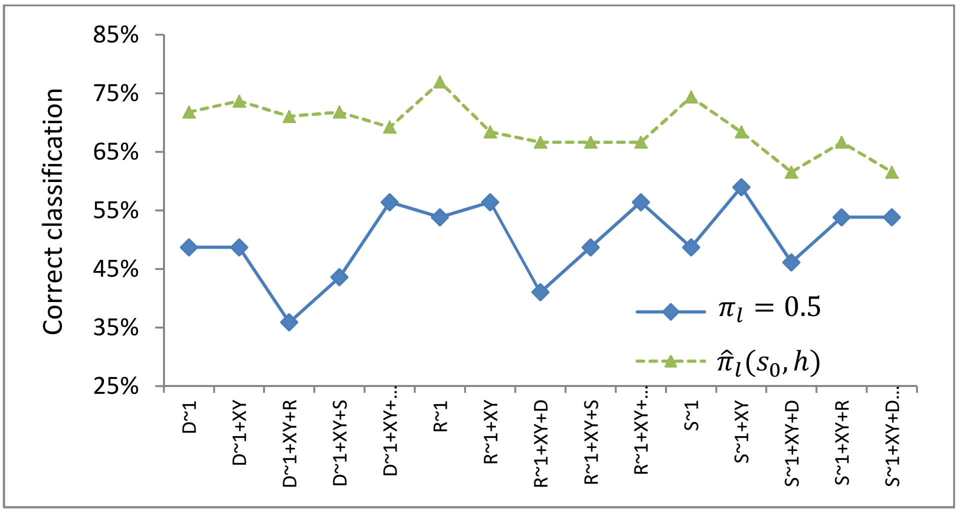

Figure 14. The influence of prior probabilities

The analogous calculations are done for GGRF (see Table 13). Here spherical covariance model was used.

Table 13. Correct classification probabilities for geostatistical model

		Neighborhood structure				
	Distance	$d_{\max} = 0.1$	$d_{\max} = 0.2$	$d_{\max} = 0.3$	$d_{\max} = 0.4$	$d_{\max} = 0$
Variable	No of neighbours	[4-16]	[15-35]	[24-38]	[36-38]	0
Depth	~1	**0.684**	0.658	0.658	0.658	0.447
	~1+XY	0.632	0.579	0.605	0.605	0.421
	~1+XY+Restime	**0.684**	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	**0.763**	0.526
Salinity	~1	**0.737**	0.684	0.711	**0.737**	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

- Closed-form expression of asymptotic covariance matrix for 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 the range parameter which determines the level of spatial correlation.
- The derived AER and AEER formulas could be applied as a target function constructing the optimality criterion for the spatial sampling design.
- A simulation study to examine accuracy of the proposed classifiers and to investigate the influence of population parameters to AER was included. According to the results the conclusions could be made:

  …

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. https://doi.org/10.1007/s11004-015-9594-x
3. Anderson, T.W. (2003). An Introduction to Multivariate Statistical Analysis. Wiley, New York.
4. Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). J Roy Stat Soc, 36:192-236.
5. Chiles, J.P., Delfiner, P. (2012). Geostatistics: Modeling Spatial Uncertainty. Second Edition. John Wiley & Sons, New York.
6. Cressie, N. (1993). Statistics for Spatial Data. Wiley & Sons, New York.
7. Cressie, N., Hawkins, D. H. (1980). Robust estimation of the variogram: I. Mathematical Geology, 12(2):115-125.
8. Diggle, P. J., Ribeiro, P. J. (2007). Model-based Geostatistics. Springer.
9. 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.
10. Griffith, 2009
11. Dučinskas, K., Šaltytė-Benth, J. (2003). Erdvinė statistika. KU, Klaipėda.
12. Dučinskas, K. (1997). An asymptotic analysis of the regret risk in discriminant analysis under various training schemes. Lithuanian Mathematical Journal, 37(4):337-351.

13. Dučinskas, K. (2009). Approximation of the expected error rate in classification of the Gaussian random field observations. Statistics and Probability Letters, 79:138-144.
14. 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.
15. Dučinskas, K. (2011). Error rates in classification of multivariate Gaussian random field observation. Lithuanian Mathematical Journal, 51:477-485.
16. 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.
17. Dučinskas, K., Dreižienė, L. (2011). Supervised classification of the scalar Gaussian random field observations under a deterministic spatial sampling design. Austrian Journal of Statistics, 40:25-36.
18. Ecker, M. D., Gelfand, A. E. (2003). Spatial modelling and prediction under stationary non-geometric range anisotropy. Environmental and Ecological Statistics, 10:165-178.
19. Ecker, M. D., Gelfand, A. E. (1999). Spatial Modelling and Prediction under Range anisotropy. Environmental and Ecological Statistics, 10:165-178.
20. Ferguson, T. A., (1996). Course in large sample theory. Chapman and Hall, London, UK.
21. Goovaerts, P. (1997). Geostatistics for natural resources evaluation. Oxford Univ. Press, New York.
22. 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.
23. Gupta, A., Robinson, P.M. (2018). Pseudo maximum likelihood estimation of spatial autoregressive models with increasing dimension. Journal of Econometrics, 202:92–107.

24. Haining, R. (1990). Spatial data analysis in the social and environmental sciences. Cambridge University Press, Cambridge, UK.
25. Haining, R. (2004). Spatial data analysis: Theory and practice. Cambridge University Press, Cambridge, UK.
26. Handbook of spatial statistics / [edited by] Alan E. Gelfand … [et al.]. p. cm. (Chapman & Hall/CRC handbooks of modern statistical methods), 2010.
27. Haskard, K.A., An anisotropic Matern spatial covariance model: RELM estimation properties. Doctoral thesis.
28. Hirst, D. (1996). Error-rate estimation in multiply-group linear discriminant analysis. Technometrics, 38:389-399.
29. Jin, X., Carlin, B. P., Banerjee, S. (2005). Generalized hierarchical multivariate CAR models for areal data. Biometrics, 61:950–961.
30. 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.
31. Isaaks, E. H., Srivastava, R. M. (1989). An Introduction to applied geostatistics. Oxford University Press.
32. Journel, A., Huijbregts, C. (1978). Mining geostatistics. Academic Press, New York.
33. Kharin, Y. (1996). Robustness in statistical pattern recognition. Dordrecht: Kluwer Academic Publishers.
34. Lahiri, 2003
35. Lawoko, C. R. O., McLachlan, G. L. (1985). Discrimination with autocorrelated observations. Pattern Recognition, 18(2):145-149.
36. Lindgren, F., Rue, H., Lindström, J. (2011). Journal of the Royal Statistical Society. Series B: Statistical Methodology, 73(4):423-498

37. Lu, H., Zimmerman D. L. (2001). Testing for isotropy and other directional symmetry properties of spatial correlation. Preprint.

38. 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.

39. Lu, J., Zhang, L., (2010). Evaluation of Parameter Estimation Methods for Fitting Spatial Regression. Forest Science, 56(5):505–514

40. Magnus, J. R., Neudecker, H. (2002). Matrix differential calculus and applications in statistics and econometrics. Wiley, New York.

41. Maity, A., Sherman, M. (2012). Testing for spatial isotropy under general designs. Journal of Statistical Planning and Inference, 142:1081–1091.

42. Mardia, K. V., Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. Biometrika, 71:135-146.

43. Mardia, K. V., (1984). Spatial discrimination and classification maps. Comm. Statist. Theory Methods, 13:2181-2197.

44. Mardia, K. V., (1988). Multi-dimensional multivariate Gaussian Markov random fields with application to image processing. J. Multivariate Anal., 24:265-284.

45. Matheron, G. (1963). Principles of geostatistics. Economic Geology, 58:1246-1266.

46. Mclachlan, G. J. (2004). Discriminant analysis and statistical pattern recognition. Wiley, New York.

47. Myers, D. E, Journel, A. (1990). Variograms with zonal anisotropies and noninvertible kriging systems. Mathematical Geology, 22(7): 779–785.

48. Petersen, K. B., Pedersen, M. S. (2006). The matrix cookbook. http://www.math.uwaterloo.ca/~hwolkowi//matrixcookbook.pdf (2018-07-29)

49. Pettit, 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. Stat. Comput., 12:353–367.
50. Rue, H., Held, L. (2005). Gaussian Markov random fields: theory and applications. Chapman & Hall, Boca Raton.
51. Okamoto, M. (1963). An asymptotic expansion for the distribution of linear discriminant function. Ann. Math. Statist., 34:1286-1301.
52. 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.
53. Sain, S. R., Cressie, N. (2007). A spatial model for multivariate lattice data. J. Econom., 140:226–259.
54. Šaltytė, J., Dučinskas, K. (2002). Comparison of ML and OLS estimators in discriminant analysis of spatially correlated observations. Informatica, 13(2):297-238.
55. Šaltytė-Benth, J., Dučinskas, K. (2005). Linear discriminant analysis of multivariate spatial-temporal regressions. Scand. J. Statist., 32:281-294.
56. 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.
57. Sherman, M. (2011). Spatial statistics and spatio-temporal data: covariance functions and directional properties. Wiley & Sons.
58. Schervish, M. J. (1984). Linear discrimination for three known normal populations. Journal of Statistical Planning and Inference, 10:167-175.
59. Switzer, P. (1980). Extensions of linear discriminant analysis for statistical classification of remotely sensed satellite imagery. Math. Geol., 12(4):367-376.
60. Theodoridis 2009

61. Wackernagel, H. (2003). Multivariate geostatistics: an introduction with applications. Springer-Verlag, Berlin Heidelberg.

62. 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

63. Weller, Z. D., Hoeting, J. A. (2016). A review of nonparametric hypothesis tests of isotropy properties in spatial data. Statist. Sci., 31(3):305-324.

64. 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.

65. Xu, T., Wang, J. (2013). An efficient model-free estimation of multiclass conditional probability. Journal of Statistical Planning and Inference, 143:2079-2088.

66. Yaglom, A. N. (1987). Correlation theory of stationary and related random functions I. Springer.

67. Ying, Z. (2001). Specification of variogram structures with geometric anisotropy. Stanford Center for Reservoir Forecasting Department of Petroleum Engineering Stanford University.

68. Zaiko, A., Daunys, D. (2015). Invasive ecosystem engineers and biotic indices: Giving a wrong impression of water quality improvement? Ecol. Indic., 52:292–299.

69. Zhang, H., Zimmerman, D. L. (2005). Towards reconciling two asymptotic frameworks in spatial statistics. Biometrika, 92(4):921–936.

70. 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

71. Zhu, Z., Stein, M. L. (2006). Spatial sampling design for prediction with estimated parameters. Journal of Agricultural, Biological, and Environmental Statistics, 11:24-44.

72. Zhu, Z., Zhang, H. (2006). Spatial design under the infill asymptotic framework. Environmetrics, 17:323–337.

73. Zimmerman, D. L. (1993). Another look at anisotropy in geostatistics. Mathematical Geology, 25(4):453-70.

74. Zimmerman, D. L. (2006). Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. Environmetrics, 17:635-652.