INFORMATICA, 2002, Vol. 13, No. 2, 227–238

Comparison of ML and OLS Estimators in Discriminant Analysis of Spatially Correlated Observations

Jūratė ŠALTYTĖ, Kęstutis DUČINSKAS
Klaipėda University
H.Manto 84, 5808 Klaipėda, Lithuania
e-mail: jsaltyte@gmf.ku.lt, duce@gmf.ku.lt

Received: November 2001

Abstract. The problem of supervised classification of the realisation of the stationary univariate Gaussian random field into one of two populations with different means and factorised covariance matrices is considered. Unknown means and the common covariance matrix of the feature vector components are estimated from spatially correlated training samples assuming spatial correlation to be known. For the estimation of unknown parameters two methods, namely, maximum likelihood and ordinary least squares are used. The performance of the plug-in discriminant functions is evaluated by the asymptotic expansion of the misclassification error. A set of numerical calculations is done for the spherical spatial correlation function.

Key words: Bayesian classification rule, linear discriminant function, training samples, misclassification probability, estimators, asymptotic expansion.

1. Introduction

The problem of supervised classification (discriminant analysis (DA)) is usually the primary goal of pattern recognition (see, e.g., Raudys, 2000). For example, in the weather prediction the weather may be divided into three classes: fair, rain and possible rain; and the problem is to classify tomorrow's weather into one of these three classes on the basis of data from satellite, when weather masses are observed. In such area like pattern recognition or geostatistics the data of interest are often spatially correlated. And DA of spatially correlated data is of great importance.

When classes are completely specified, an optimal classification rule in the sense of minimum classification error is the Bayesian classification rule. In practice, however, the complete description of classes usually is not possible and for the estimation of probabilistic characteristics of each class the training samples are required. When estimators of unknown parameters are used, the expressions for the expected error rate are very cumbersome even for the simplest procedures of DA. Therefore, asymptotic expansions of the expected error rate are especially important.

In the experimental-design literature traditionally two estimators have been discussed when the observations are correlated, namely, the ordinary least square (OLS) estima-

tor and maximum likelihood (ML) estimator. The OLS estimator is usually chosen because of practical considerations; it is easier to compute than ML estimator since it does not involve the variance, which is frequently unknown. The OLS estimators are always available and yield unbiased estimators; however they may not have minimum variances (among unbiased linear estimators) (Anderson, 1971, p. 560). When those two kinds of estimators are used in context of DA, it is expedient to compare the OLS estimators with the ML estimators in a sense of the asymptotic expansions for the expected error rate.

2. Model and Problem

Let $\{Z(s)\colon\ s \in D \subset \Re^2\}$ be a univariate Gaussian random field having different means and factorised covariance matrices in populations Ω_1 and Ω_2. Then the model of $Z(s)$ in population Ω_l is

$$Z(s) = \mu_l(s) + \varepsilon_l(s), \tag{1}$$

where $\mu_l(s)$ is a mean vector and $\{\varepsilon_l(s)\colon\ s \in D \subset \Re^2\}$ is a zero-mean stationary Gaussian random field with covariance defined by a parametric model $\operatorname{cov}\{\varepsilon_l(t), \varepsilon_l(s)\} = \sigma(h;\theta_l)$, where $h = t - s$, $t, s \in D$, and $\theta_l \in \Theta$ is a $m \times 1$ parameter vector, Θ being an open subset of $\Re^m$, $l = 1, 2$. We restrict the attention to the homoscedastic models, i.e., $\sigma(0;\theta) = \sigma^2$, for each $\theta \in \Theta$. Then the spatial covariance function in Ω_l is $\operatorname{cov}\{\varepsilon_l(t), \varepsilon_l(s)\} = c(h;\theta_l)\sigma^2$, where $c(h;\theta_l)$ is the spatial correlation function, $l = 1, 2$. It is assumed that the function $c(h;\theta_l)$ is positive definite (Mardia and Marshall, 1984). Assume that, for all $t, s \in D$, $t \neq s$,

$$\operatorname{cov}\{\varepsilon_1(t), \varepsilon_2(s)\} = 0.$$

There are several possible mean models: constant mean, regression model and trend surface model. Constant mean model in population Ω_l is

$$E\{Z(s)\} = \mu_l \equiv \text{const},$$

for all $s \in D$, $l = 1, 2$. The regression model in population Ω_l for all $s \in D$ is defined by

$$\mu_l(s) = x^T(s)\beta_l,$$

where $x(s) = \big(x_1(s), \ldots, x_q(s)\big)^T$ is a $q \times 1$ vector of non-random regressors and $\beta_l = (\beta_l^1, \ldots, \beta_l^q)^T \in B$, $l = 1, 2$, is a vector of large-scale variation (trend) parameters, B being an open subset of $\Re^q$. Let X_l be an $N_l \times q$ regressor matrix with j-th column $(x_{1j}, \ldots, x_{N_l j})^T$, where $x_{\alpha j} = x_j(s_\alpha^l)$, $j = 1, \ldots, q$, $\alpha = 1, \ldots, N_l$, $l = 1, 2$, and $|x(s)| \leqslant M < \infty$, for all $s \in D$.

Haining (1990) suggests represent mean as a polynomial function of coordinates of a specified order, that is

$$\mu_l(s) = a^T(s)\lambda_l,$$

where $a(s)$ is a vector of location coordinates of the point s and their products, and λ_l is a vector of trend surface parameters so that

$$a^T(s) = (1, s_x, s_y, (s_x)^2, (s_y)^2, s_x s_y, \ldots, (s_x)^p (s_y)^q),$$

where s_x and s_y define the coordinates of point s in $\Re^2$, and for, $l = 1, 2$,

$$\lambda_l = (\lambda_l^{10}, \lambda_l^{01}, \lambda_l^{20}, \lambda_l^{02}, \lambda_l^{11}, \ldots, \lambda_l^{pq})^T.$$

Geographical coordinates such as longitude and latitude could also be considered instead of s_x and s_y. The sum $p + q = k$ represents the order of the trend surface. This is so-called trend surface model. Let A_l be an $N_l \times k$ matrix with α-th raw being $a^T(s_\alpha^l)$, $\alpha = 1, \ldots, N_l$, $l = 1, 2$.

It is easy to see, that constant mean model and trend surface model could be considered as special cases of regression model, when $q = 1$, $x(s) = 1$ (constant mean) and $k = 1$, X_l is replaced by A_l (trend surface model). So, further we restrict our attention to the regression model.

Model (1) is generally used in geostatistics, which is usually concerned with optimal linear spatial prediction called kriging. As Cressie (1993) designates, in kriging often the parameters of regression model (called "large-scale-variation" parameters) are of greatest interest, assuming that the "small-scale-variation" parameters associated with error process $\varepsilon_l(s)$ are known. This is why we concentrate bigger attention on the mean model. However, we are solving the problem of DA, and it is useful to find an estimator of the variance as well, because the unknown variance is often the case in practice. Here factorised model of covariance (Mardia, 1984) will be used. More information on the structure of covariance functions can be found in, e.g., Raudys (2000).

Consider the problem of supervised classification (Jain *et al*, 2000) of the observation $Z(r)$ with $r \in D_0 \subset D$ into one of two populations specified above. Under the assumption, that the classes are completely specified and for known prior probabilities of populations $\pi_1(r)$ and $\pi_2(r)$ $(\pi_1(r) + \pi_2(r) = 1)$, the Bayesian classification rule (BCR) $d_B(\cdot)$ minimising the probability of misclassification (PMC) is

$$d_B\big(z(r)\big) = \arg \max_{\{l=1,2\}} \pi_l(r) p_l\big(z(r)\big), \tag{2}$$

where $\pi_l(r)$ is a prior probability of Ω_l, $l = 1, 2$.

Denote by P_B^r the PMC of BCR, usually called the Bayesian error rate.

As it was already mentioned in the introduction, in practical applications the parameters of density function are usually not known and must be estimated. Then the estimators of unknown parameters are found from the training samples T_1 and T_2 taken separately from Ω_1 and Ω_2, respectively. When estimators of unknown parameters are used, the plug-in version of BCB is obtained. The performance of the plug-in version of the BCR when parameters are estimated from training samples with independent observations is widely investigated (see, e.g., Okamoto, 1963). However, it has been founded that the

assumption of independence is frequently violated. Lawoko and McLachlan (1985) investigated the performance of sample linear discriminant function (LDF) when training samples follow a stationary autoregressive process. In this paper we shall consider the performance of the plug-in linear DF when the parameters are estimated from training sample following a Gaussian random field model described above. The ML and OLS procedures for the estimation of unknown means and variance, assuming the spatial dependence parameter is known, are used.

Suppose in region $D_1 \subset D$, $D_1 \cap D_0 = \emptyset$, we observe the training sample $T = \{T_1, T_2\}$ with $T_l = \{Z_{l1}, \ldots, Z_{lN_l}\}$, where $Z_{l\alpha} = Z(s_\alpha^l)$ denotes the α-th observation from Ω_l, $l = 1, 2$, $\alpha = 1, \ldots, N_l$. Assume that D_1 is beyond the range (or the zone of influence) of D_0. Then $Z(r)$ is independent on T.

Let $\hat{\mu}_l(r)$ and $\hat{\sigma}^2$ be the estimators of $\mu_l(r)$ and σ^2, respectively, based on T. The plug-in rule $d_B\big(z(r); \hat{\mu}_1(r), \hat{\mu}_2(r), \hat{\sigma}^2\big)$ is obtained by replacing the parameters in (2) with their estimators. Then the corresponding plug-in LDF $\hat{W}^r$ (McLachlan, 1974a), for $g(r) = \ln\left(\frac{\pi_1(r)}{\pi_2(r)}\right)$, is

$$\hat{W}^r = \left(z(r) - \frac{1}{2}\big(\hat{\mu}_1(r) + \hat{\mu}_2(r)\big)\right)\big(\hat{\mu}_1(r) - \hat{\mu}_2(r)\big)\frac{1}{\hat{\sigma}^2} + g(r).$$

DEFINITION 1. The actual error rate for $d_B\big(z(r); \hat{\mu}_1(r), \hat{\mu}_2(r), \hat{\sigma}^2\big)$ is defined as

$$P^r\big(\hat{\mu}_1(r), \hat{\mu}_2(r), \hat{\sigma}^2\big) = \sum_{l=1}^{2} \pi_l(r) \int\limits_Z L\Big(l, d_B\big(z(r); \hat{\mu}_1(r), \hat{\mu}_2(r), \hat{\sigma}^2\big)\Big) p_l\big(z(r); \mu_l(r), \sigma^2\big)\, \mathrm{d}\, z(r).$$

In our case the actual error rate for $d_B\big(z(r); \hat{\mu}_1(r), \hat{\mu}_2(r), \hat{\sigma}^2\big)$ is defined as

$$P^r\big(\hat{\mu}_1(r), \hat{\mu}_2(r), \hat{\sigma}^2\big) = \sum_{l=1}^{2} \pi_l(r) \Phi\left((-1)^l \frac{\Big(\mu_l(r) - \frac{1}{2}\big(\hat{\mu}_1(r) + \hat{\mu}_2(r)\big)\Big)\big(\hat{\mu}_1(r) - \hat{\mu}_2(r)\big) + \hat{\sigma}^2 g(r)}{\sigma\sqrt{\big(\hat{\mu}_1(r) - \hat{\mu}_2(r)\big)^2}}\right),$$

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

DEFINITION 2. The expectation of the actual error rate with respect to distribution of T designated as $E_T\big\{P^r\big(\hat{\mu}_1(r), \hat{\mu}_2(r), \hat{\sigma}^2\big)\big\}$ is called the expected error rate (EER) for the $d_B\big(z(r);\ \hat{\mu}_1(r),\ \hat{\mu}_2(r),\ \hat{\sigma}^2\big)$.

The goal of this paper is to find asymptotic expansions of EER associated with plug-in LDF for different estimators. The case of normally distributed observations in training sample from the one of two classes with equal feature vector covariances was firstly considered in Okamoto (1963). Dučinskas (1997) has been made the generalization for

the case of arbitrary number of classes ($l \geqslant 2$) and regular class-conditional densities. McLachlan (1974b) presented EER for the case of equicorrelated Gaussian observations. Mardia (1984) considered similar problem of classifying the spatially distributed Gaussian observations with constant means, but he did not analyse the EER of PMC. In this paper we present the asymptotic expansion up to the order $O(N^{-2})$, where $N = N_1+N_2$, for the EER of classifying spatially distributed Gaussian observation with different means and common spatially factorised covariance. Terms of higher order are omitted from the asymptotic expansion since their contribution is in generally negligible (Schervish, 1981). The ML and OLS estimators of means and the bias-adjusted ML and bias-adjusted OLS estimators of the covariance are used in the plug-in version of the BCR. A set of calculations for a certain neighbourhood structure and spherical spatial correlation model is performed in order to estimate the plug-in BCR.

3. Asymptotic Expansion

The expectation vector and the covariance matrix of the vectorised training sample T_l defined by $T_l^V = (Z_{l1}, \ldots, Z_{lN_l})^T$ are

$$\mu_l^V = \big(\mu_l(s_1^l), \ldots, \mu_l(s_{N_l}^l)\big)^T \quad \text{and} \quad \Sigma_l^V = \sigma^2 C_l,$$

respectively, where C_l is the spatial correlation matrix of order $N_l \times N_l$, whose $\alpha\beta$-th element is $c(s_\alpha^l - s_\beta^l)$, $\alpha, \beta = 1, \ldots, N_l$, $l = 1, 2$. Suppose, that C_l is known, and $\hat{\mu}_l^\upsilon(s)$ and $\hat{\sigma}_\upsilon^2$ are the estimators of $\mu_l(s)$ and σ^2, respectively, based on T; here υ can take the value ML or OLS, $l = 1, 2$.

When the regression model of mean is used, the estimator of mean is of the form $\hat{\mu}_l^\upsilon(s) = x^T(s)\hat{\beta}_l^\upsilon$, where $\hat{\beta}_l^\upsilon$ is the estimator of the corresponding regression parameters, $l = 1, 2$.

Lemma 1. *For $l = 1, 2$, the maximum likelihood estimators of $\mu_l(s)$ and σ^2, based on T are*

$$\hat{\mu}_l^{ML}(s) = x^T(s)(X_l^T C_l^{-1} X_l)^{-1} X_l^T C_l^{-1} T_l^V x(s), \tag{3}$$

$$\hat{\sigma}_{ML}^2 = \frac{1}{N} \sum_{l=1}^{2} \big(T_l^V - \hat{\mu}_l^{ML}(s)\big)^T C_l^{-1} \big(T_l^V - \hat{\mu}_l^{ML}(s)\big). \tag{4}$$

Proof. The log-likelihood of T_l, $l = 1, 2$, is

$$\ln L_l = \text{const} - \frac{1}{2}(N_l \ln \sigma^2 + \ln |C_l|) - \frac{1}{2\sigma^2}(T_l^V - \mu_l^V)^T C_l^{-1}(T_l^V - \mu_l^V).$$

Solving the equations $\dfrac{\partial \ln L_l}{\partial \mu_l(s)} = 0$, $l = 1, 2$, and $\sum\limits_{l=1}^{2} \dfrac{\partial \ln L_l}{\partial \sigma^2} = 0$, we complete the proof of lemma.

Lemma 2. *For $l = 1, 2$, the ordinary least squares estimators of $\mu_l(s)$ and σ^2, based on T are*

$$\hat{\mu}_l^{OLS}(s) = x^T(s)(X_l^T X_l)^{-1} X_l^T T_l^V x(s), \tag{5}$$

$$\hat{\sigma}^2_{OLS} = \frac{1}{N}\sum_{l=1}^{2}\left(T_l^V - \hat{\mu}_l^{OLS}(s)\right)^T\left(T_l^V - \hat{\mu}_l^{OLS}(s)\right). \tag{6}$$

Proof. The proof of lemma is similar to that of Lemma 1. Only the difference is, that the classical (non-spatial) assumptions of iid errors, i.e., $C_l = I$, $l = 1, 2$, is used.

Since $E\{\hat{\sigma}^2_{ML}\} = \dfrac{N-2q}{N}\sigma^2$ and $E\{\hat{\sigma}^2_{OLS}\} = \dfrac{N-\omega}{N}\sigma^2$, where $\omega = \sum\limits_{l=1}^{2}(X_l^T X_l)^{-1} \times X_l^T C_l X_l$, further we will use the bias-adjusted ML and OLS estimators of σ^2:

$$\tilde{\sigma}^2_{ML} = \frac{N}{N-2q}\sigma^2 \quad \text{and} \quad \tilde{\sigma}^2_{OLS} = \frac{N}{N-\omega}\hat{\sigma}^2_{OLS}. \tag{7}$$

It can be easily shown that $\hat{\mu}_l^v(s)$ for finite N have known exact distributions $\hat{\mu}_l^{ML}(s) \sim N\left(x^T(s)\beta_l, \delta_l^{ML}\right)$, where

$$\delta_l^{ML} = \sigma^2 x^T(s)(X_l^T C_l^{-1} X_l)^{-1} x(s), \tag{8}$$

and $\hat{\mu}_l^{OLS}(s) \sim N\left(x^T(s)\beta_l, \delta_l^{OLS}\right)$, where

$$\delta_l^{OLS} = \sigma^2 x^T(s)(X_l^T X_l)^{-1} X_l^T C_l X_l (X_l^T X_l)^{-1} x(s). \tag{9}$$

Define

$$\gamma_{ML} = \frac{2(\sigma^2)^2}{N-2q}, \tag{10}$$

$$\gamma_{OLS} = \frac{2(\sigma^2)^2}{(N-\omega)^2}\sum_{l=1}^{2}\Big(\mathrm{tr}\mathrm{C}_\mathrm{l}^2 - 2\mathrm{tr}\big((\mathrm{X}_\mathrm{l}^\mathrm{T}\mathrm{X}_\mathrm{l})^{-1}\mathrm{X}_\mathrm{l}^\mathrm{T}\mathrm{C}_\mathrm{l}^2\mathrm{X}_\mathrm{l}\big) + \mathrm{tr}\big((\mathrm{X}_\mathrm{l}^\mathrm{T}\mathrm{X}_\mathrm{l})^{-1}\mathrm{X}_\mathrm{l}^\mathrm{T}\mathrm{C}_\mathrm{l}^2\mathrm{X}_\mathrm{l}\big)^2\Big). \tag{11}$$

For simplicity we omit the superscript "r" in $P^r(\cdot)$. Put $\Delta\hat{\mu}_l^v(r) = \hat{\mu}_l^v(r) - \mu_l(r)$, $\Delta\tilde{\sigma}^2_v = \tilde{\sigma}^2_v - \sigma^2$. Let $\varphi(\cdot)$ denotes the standard normal density function.

Denote by

$$P_l^{(1)} = \frac{\partial P(\cdot)}{\partial\hat{\mu}_l^v(r)}, \quad P_{k,l}^{(2)} = \frac{\partial^2 P(\cdot)}{\partial\hat{\mu}_k^v(r)\partial\hat{\mu}_l^v(r)}, \quad P_{\tilde{\sigma}^2_v}^{(1)} = \frac{\partial P(\cdot)}{\partial\tilde{\sigma}^2_v},$$

$$P_{(\tilde{\sigma}^2_v)^2}^{(2)} = \frac{\partial^2 P(\cdot)}{\partial(\tilde{\sigma}^2_v)^2}, \quad P_{l,\tilde{\sigma}^2_v}^{(2)} = \frac{\partial^2 P(\cdot)}{\partial\hat{\mu}_l^v(r)\partial\tilde{\sigma}^2_v}$$

the partial derivatives of $P\left(\hat{\mu}_1^v(r), \hat{\mu}_2^v(r), \tilde{\sigma}^2_v\right)$ up to the second order with respect to the corresponding parameters evaluated at $\hat{\mu}_l^v(r) = \mu_l(r)$ and $\tilde{\sigma}^2_v = \sigma^2$, $l = 1, 2$.

Let $\lambda(C_l)$ be the largest eigenvalue of C_l, $l = 1, 2$.

ASSUMPTION 1. Assume, that $\operatorname{rank}(X_l) = q$, for $l = 1, 2$.

ASSUMPTION 2. Suppose, that $\lambda(C_l) < \kappa_l$, $0 < \kappa_l < \infty$, $l = 1, 2$.

ASSUMPTION 3. Assume, that $\dfrac{N_1}{N_2} \to \tau$, as $N_1, N_2 \to \infty$, $0 < \tau < \infty$.

Theorem 1. *Suppose that assumptions* 1–3 *hold for training samples* T_1, T_2. *Then the asymptotic expansion of the expected risk for the* $d_B\big(z(r); \hat{\mu}_1^{\upsilon}(r), \hat{\mu}_2^{\upsilon}(r), \tilde{\sigma}_{\upsilon}^2\big)$, *where* υ *can take the value* ML *or* OLS, *is*

$$E_T\Big\{P\big(\hat{\mu}_1^{\upsilon}(r), \hat{\mu}_2^{\upsilon}(r), \tilde{\sigma}_{\upsilon}^2\big)\Big\} = \sum_{l=1}^{2} \pi_l(r)\Phi\bigg(-\frac{\Delta(r)}{2} + (-1)^l \frac{g(r)}{\Delta(r)}\bigg)$$
$$+\frac{\pi_1(r)}{2\Delta(r)}\varphi\bigg(-\frac{\Delta(r)}{2} - \frac{g(r)}{\Delta(r)}\bigg) \sum_{l=1}^{2} \bigg(\delta_l^{\upsilon}\bigg(-\frac{\Delta(r)}{2} + (-1)^l \frac{g(r)}{\Delta(r)}\bigg)^2 + g^2(r)\gamma_{\upsilon}\bigg)$$
$$+O(N^{-2}),$$

with δ_l^{υ} *defined in* (8), (9) *and* γ_{υ} *defined in* (10), (11).

Proof. . Without loss of generality we use the convenient canonical form of $\sigma^2 = 1$ and $\mu_1(r) = \dfrac{\Delta(r)}{2}$, $\mu_2(r) = -\dfrac{\Delta(r)}{2}$ (see, e.g., McLachlan, 1992). By a Taylor expansion of the $P\big(\hat{\mu}_1^{\upsilon}(r), \hat{\mu}_2^{\upsilon}(r), \tilde{\sigma}_{\upsilon}^2\big)$, for $\upsilon = ML$ or OLS, about the true values of parameters we have

$$\begin{aligned} P\big(\hat{\mu}_1^{\upsilon}(r), \hat{\mu}_2^{\upsilon}(r), \tilde{\sigma}_{\upsilon}^2\big) \;=\; & P_B + \sum_{l=1}^{2} P_l^{(1)} \Delta\hat{\mu}_l^{\upsilon}(r) + P_{\tilde{\sigma}_{\upsilon}^2}^{(1)} \Delta\tilde{\sigma}_{\upsilon}^2 \\ & + \frac{1}{2}\bigg(\sum_{k,l=1}^{2} P_{k,l}^{(2)} \Delta\hat{\mu}_k^{\upsilon}(r)\Delta\hat{\mu}_l^{\upsilon}(r) + P_{(\tilde{\sigma}_{\upsilon}^2)^2}^{(2)} (\Delta\tilde{\sigma}_{\upsilon}^2)^2 \\ & + \sum_{l=1}^{2} P_{l,\tilde{\sigma}_{\upsilon}^2}^{(2)} \Delta\hat{\mu}_l^{\upsilon}(r)\Delta\tilde{\sigma}_{\upsilon}^2\bigg) + O_3, \end{aligned} \tag{12}$$

where

$$P_B = \sum_{l=1}^{2} \pi_l(r)\Phi\bigg(-\frac{\Delta(r)}{2} + (-1)^l \frac{g(r)}{\Delta(r)}\bigg),$$

and O_3 is the third and higher order terms of $\Delta\hat{\mu}_l^{\upsilon}(r)$ and $\Delta\tilde{\sigma}_{\upsilon}^2$ and their products. Since $P\big(\hat{\mu}_1^{\upsilon}(r), \hat{\mu}_2^{\upsilon}(r), \tilde{\sigma}_{\upsilon}^2\big)$ is minimised at $\hat{\mu}_l^{\upsilon}(r) = (-1)^{l+1}\dfrac{\Delta(r)}{2}$ $(l = 1, 2)$ and $\tilde{\sigma}_{\upsilon}^2 = 1$, then

$$P_l^{(1)} = 0 \quad \text{and} \quad P_{\tilde{\sigma}_{\upsilon}^2}^{(1)} = 0. \tag{13}$$

Using (1)–(11), for $l = 1, 2$, under the independence of estimators $\hat{\mu}_l^{\upsilon}$ and $\tilde{\sigma}_{\upsilon}^2$, for $\upsilon = ML$ or OLS, we have

$$E\{\Delta\hat{\mu}_l^{ML}\} = E\{\Delta\hat{\mu}_l^{OLS}\} = E\{\Delta\hat{\mu}_1^{\upsilon}\Delta\hat{\mu}_2^{\upsilon}\} = E\{\Delta\tilde{\sigma}_{\upsilon}^2\} = E\{\Delta\hat{\mu}_l^{\upsilon}\Delta\tilde{\sigma}_{\upsilon}^2\} = 0, \quad (14)$$

$$E\{(\Delta\hat{\mu}_l^{ML})^2\} = x^T(r)(X_l^T C_l^{-1} X_l)^{-1} x(r), \quad (15)$$

$$E\{(\Delta\hat{\mu}_l^{OLS})^2\} = x^T(r)(X_l^T X_l)^{-1} X_l^T C_l X_l (X_l^T X_l)^{-1} x(r). \quad (16)$$

$$E\{(\Delta\tilde{\sigma}_{ML}^2)^2\} = \frac{2}{N-2q}, \quad (17)$$

$$E\{(\Delta\tilde{\sigma}_{OLS}^2)^2\} = \frac{2}{(N-\kappa)^2}\sum_{l=1}^{2}\Big(\operatorname{tr} C_l^2 - 2\operatorname{tr}\left((X_l^T X_l)^{-1} X_l^T C_l^2 X_l\right) + \operatorname{tr}\left((X_l^T X_l)^{-1} X_l^T C_l^2 X_l\right)^2\Big) \quad (18)$$

Note that

$$P_{l,l}^{(2)} = \frac{\pi_1(r)}{\Delta(r)}\varphi\Big(-\frac{\Delta(r)}{2} - \frac{g(r)}{\Delta(r)}\Big)\Big(-\frac{\Delta(r)}{2} + (-1)^l\frac{g(r)}{\Delta(r)}\Big)^2 \quad (19)$$

and

$$P_{(\tilde{\sigma}_{\upsilon}^2)^2}^{(2)} = \frac{\pi_1(r)}{\Delta(r)} g^2(r)\varphi\Big(-\frac{\Delta(r)}{2} - \frac{g(r)}{\Delta(r)}\Big). \quad (20)$$

By substituting the estimators (3)–(7) into (12), taking the expectation of the right side of (12) and using (13)–(20) we complete the proof of the theorem.

As the contribution of higher order terms in the presented asymptotic expansion is in generally negligible (Scherwish, 1981), for the evaluation of the performance of LDF the asymptotic expected error regret (AEER)

$$AEER_{\upsilon} = \frac{\pi_1(r)}{2\Delta(r)}\varphi\Big(-\frac{\Delta(r)}{2} - \frac{g(r)}{\Delta(r)}\Big)\sum_{l=1}^{2}\Big(\delta_l^{\upsilon}\Big(-\frac{\Delta(r)}{2} + (-1)^l\frac{g(r)}{\Delta(r)}\Big)^2 + g^2(r)\gamma_{\upsilon}\Big)$$

for $\upsilon = ML$ or OLS, is used. Minimum of AEER could also be used as a criterion for optimal training sample design.

The numerical comparison of these two regrets is given in the example below.

4. Numerical Example

Here we compare the AEERs when the ML and OLS estimators of unknown parameters are used. Obtained results of this comparison are presented in Table 1.

Table 1

Comparison of the asymptotic expansions (for $\pi_1 = 0.3$ and $t = 1$)

Δ	$AEER_{ML}$	$AEER_{OLS}$	$\frac{AEER_{ML}}{AEER_{OLS}}$	Δ	$AEER_{ML}$	$AEER_{OLS}$	$\frac{AEER_{ML}}{AEER_{OLS}}$
0.4	0.3638	0.5140	0.7078	2.8	0.0782	0.1171	0.6675
0.6	0.3651	0.5461	0.6686	3.0	0.0727	0.1076	0.6761
0.8	0.2406	0.3850	0.6251	3.2	0.0667	0.0976	0.6835
1.0	0.1636	0.2773	0.5899	3.4	0.0603	0.0875	0.6898
1.2	0.1251	0.2186	0.5721	3.6	0.0538	0.0775	0.6951
1.4	0.1070	0.1871	0.5717	3.8	0.0474	0.0678	0.6998
1.6	0.0987	0.1694	0.5826	4.0	0.0412	0.0586	0.7038
1.8	0.0948	0.1583	0.5986	4.2	0.0354	0.0500	0.7072
2.0	0.0923	0.1499	0.6155	4.4	0.0300	0.0422	0.7103
2.2	0.0898	0.1423	0.6313	4.6	0.0251	0.0352	0.7129
2.4	0.0868	0.1345	0.6453	4.8	0.0208	0.0290	0.7153
2.6	0.0829	0.1262	0.6573	5.0	0.0170	0.0236	0.7174

As an example consider the integer regular 2-dimensional lattice. We use the training samples of size 4 for each class.

Consider for both classes the spherical correlation function for observations $Z(s)$ and $Z(t)$ (Cressie, 1993):

$$c(|h|) = \begin{cases} \dfrac{\kappa_1}{\kappa_0 + \kappa_1}\left(1 - \dfrac{3}{2}\dfrac{|h|}{\eta} + \dfrac{1}{2}\dfrac{|h|^3}{\eta^3}\right), & 0 \leqslant |h| \leqslant \eta, \\ 0, & |h| > \eta, \end{cases}$$

for nonnegative κ_0, κ_1, η and $h = s - t$. The nugget effect is κ_0 and the sill is $\kappa_0 + \kappa_1$.

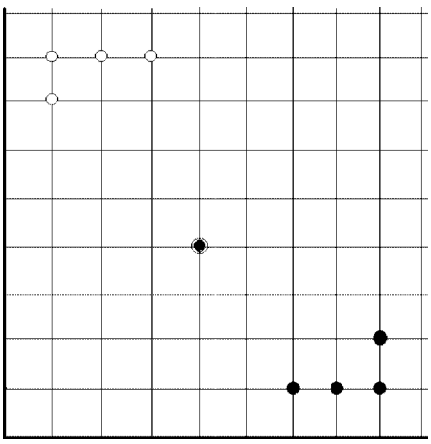

Fig. 1. Training sample design (locations of observations in T_1 and T_2 are signed as $\circ$ and $\bullet$, respectively; $\odot$ denotes the location of r).

For this model, observations more than η units apart are uncorrelated, so the range is η.

Assume, that there is no nugget effect, i.e., $\kappa_0 = 0$, and $\eta = 3$. Let say $q = 1$ and regressors are of the form $x^T(s) = \frac{1}{|s|^2+t}$ and $x^T(r) = \frac{1}{t}$, $l = 1, 2$.

Concluding Remarks

As it was expected, the AEERs are decreasing, when the distance increases. It is seen from the table that the AEER when the ML estimators are used is smaller than that obtained by using the OLS estimators. This difference is higher for small distances. Thus the ML estimators would be especially appropriate for the estimation of parameters, when the distance between classes is insignificant. When classes are more separated we can use OLS estimators, which are easier to calculate.

References

Anderson, T.W. (1971). *The Statistical Analysis of Time Series*. Wiley&Sons.

Cressie, N.A.C. (1993). *Statistics for Spatial Data*. Wiley&Sons.

Dučinskas, K. (1997). An asymptotic analysis of the regret risk in discriminant analysis under various training schemes. *Lith. Math. J.*, **37**(4), 337–351.

Haining, R.P. (1990). *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge University Press.

Jain, A.K., R.P.W. Duin, J. Mao (2000). Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(1), 4–38.

Lawoko, C.R.O., G.J. McLachlan (1985). Discrimination with autocorrelated observations. *Pattern Recognition*, **18**(2), 145–149.

Mardia, K.V. (1984). Spatial discrimination and classification maps. *Commun. Statist. – Theor. Meth.*, **13**(18), 2181–2197.

Mardia, K.V., R.J. Marshall (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrica*, **71**(1), 135–146.

McLachlan, G.J. (1974a). The asymptotic distributions of the conditional error rate and risk in discriminant analysis. *Biometrika*, **61**(1), 131–135.

McLachlan, G.J. (1974b). The asymptotic unbiased technique for estimating the error rates in discriminant analysis. *Biometrics*, **30**(3), 239–249.

McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*, Wiley&Sons.

Okamoto, M. (1963, 1968). An asymptotic expansion for the distribution of the linear discriminant function. *Ann. Math. Statist.*, **34**, 1286–1301. Correction. *Ann. Math. Statist.*, **89**, 1358–1359.

Raudys, Š. (2000). *Statistical and Neural Classifiers*. Springer, 2000.

Scherwish, M.J. (1981). Asymptotic expansions for correct classification rates in discriminant analysis. *The Annals of Statistics*, **9**(5), 1002–1009.

J. Šaltytė graduated from the Klaipėda University in 1997 in system research, where received doctor degree in 2001. Present research interests include discriminant analysis of spatially correlated data, geostatistics.

K. Dučinskas graduated from the Vilnius University in 1976 in applied mathematics, where received doctor degree in 1983. He is a head of System Research Department and an associate professor of Klaipėda University. Present research interests include discriminant analysis of spatially correlated data, geostatistics.

Maksimalaus tikėtinumo ir mažiausių kvadratų įverčių palyginimas diksriminantinėje erdvėje koreliuotų stebėjimų analizėje

Jūratė ŠALTYTĖ, Kęstutis DUČINSKAS

Straipsnyje sprendžiamas stacionaraus vienmačio atsitiktinio Gauso lauko stebėjimų klasifikavimo į vieną iš dviejų klasių su skirtingais vidurkiais ir faktorizuotomis kovariacijų matricomis uždavinys. Nežinomi požymių vektoriaus komponenčių vidurkiai ir bendra kovariacijų matrica vertinami pagal erdvėje koreliuotas mokymo imtis, laikant, kad erdvinės koreliacijos yra žinomos. Nežinomų parametrų vertinimui naudojami du metodai: maksimalaus tikėtinumo ir mažiausių kvadratų metodas. “Plug-in” diskriminantinė funkcija vertinama klaidingos klasifikavimo tikimybės asimptotiniu skleidiniu. Asimptotinis klaidos prieaugis įvertintas ir skaitiškai, naudojant sferinę koreliacijų funkciją.