Supervised classification of the observation of spatial Gaussian process with known covariance function

K. Ducinskas[1)]
1) Klaipeda University, Department of Statistics, 92294 H. Manto 84, Klaipeda, Lithuania,
kestutis.ducinskas@ku.lt

Abstract: *The problem of classification of spatial Gaussian process observation into one of two populations specified by different regression mean models and common known covariance function is considered. ML estimators of regression parameters are plugged in the Bayes discriminant function. The asymptotic expansion of the expected error rate associated with Bayes plug – in discriminant function is derived. Numerical analysis of the accuracy of the approximation of the expected error rate based on derived asymptotic expansion in the small training sample case is carried out. This approximation is proposed as optimality criterion function for spatial sampling design.*

Keywords: Bayes discriminant function, spatial correlation, expected error rate.

1. INTRODUCTION

In classical discriminant analysis (DA) sometimes called supervised classification, the observations to be classified and observations in training sample are assumed to be independent. However, in practical situations with temporally and spatially distributed data this is usually not the case. Data that are close together in time or space are likely to be correlated. Thus, to include temporal or spatial dependencies in the classification problem is very important.

When populations are completely specified an optimal classification rule in the sense of minimum misclassification probability is the Bayesian classification rule. In practice, however, the complete statistical description of populations is usually not possible. Training sample is required for the estimation of the probabilistic characteristics of both populations. When estimators of unknown parameters are used, the expressions for the expected error rate are very cumbersome even for the simplest procedures of DA. This makes it difficult to build some qualitative conclusions. Therefore, asymptotic expansions of the expected error rate are especially important.

Many authors have investigated the performance of the plug-in version of the BCR when parameters are estimated from training samples with independent observations, or training samples where observations are temporally dependent (see e.g., [1, 2]). Switzer [3] was the first to treat classification of spatial data, a work that was extended in [4]. However, neither of these authors analyze the error rate of classification. Šaltytė and Dučinskas [5] derived the asymptotic expansion of the expected error rate when classifying the observation of a univariate Gaussian random field into one of two classes with different mean models and common variance. This result was generalized to multivariate spatial-temporal regression model in Šaltytė-Benth and Dučinskas [6]. The influence of the statistical dependence among training sample observations (stationary time series, Markov dependence, autoregressive dependence) the performance of the Bayes Discriminant function is presented in the monograph written by Yu. Kharin [7]. However, in these papers the observation to be classified was assumed independent from training samples in all publications listed above.

In this paper, both restrictions are deleted, i.e. interclass spatial correlations and spatial correlations between observation to be classified and training sample assumed are not equal zero. Performance of the plug-in linear discriminant function when the parameters are estimated from training sample formed by classified observations of Gaussian random field is analyzed. We use the maximum likelihood (ML) estimators of unknown parameters of means and common variance assuming that the spatial correlation is known. Similar problems for group spatial classification are considered in [8].

2. THE MAIN CONCEPTS AND DEFINITIONS

The main objective of this paper is to classify the observations of spatial Gaussian process

$$\left\{Z(s): s \in D \subset R^m\right\}.$$

The model of observation $Z(s)$ in population Ω_l is

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

where $x(s)$ is a $q \times 1$ vector of non random regressors and β_l is a $q \times 1$ vector of parameters, $l = 1,2$. The error term is generated by zero – mean stationary spatial Gaussian process $\left\{\varepsilon(s): s \in D \subset R^m\right\}$ with covariance function defined by nuggetless model for all $s, u \in D$

$$\operatorname{cov}\{\varepsilon(s), \varepsilon(u)\} = r(s-u)\sigma^2,$$

where $r(s-u)$ is the spatial correlation function and σ^2 is variance as a scale parameter.

Consider the problem of classification of the observation $Z_0 = Z(s_0)$ into one of two populations specified above with given training sample T. Training sample T is specified by $T' = (T_1', T_2')$, where T_l is the

$n_l \times 1$ vector of n_l observations of $Z(s)$ from Ω_l, $l = 1,2$, $n = n_1 + n_2$.

Then the model of T is

$$T = X\beta + E$$

where X is the $n \times 2q$ design matrix, $\beta' = (\beta_1', \beta_2')$ and E is the *n*-vector of random errors that has multivariate Gaussian distribution $N_n(0, \sigma^2 R)$.

The design matrix X in (3) is specified by

$$X = X_1 \oplus X_2,$$

where symbol $\oplus$ denotes the direct sum of matrices and X is the $n_l \times q$ matrix of regressors for T_l, $l = 1,2$.

Denote by r_0 the vector of correlations between Z_0 and T. Since Z_0 is correlated with training sample, we have to deal with conditional distribution of Z_0 given $T = t$ with means μ_{lt}^0 and variance σ_{0t}^2 that are defined by

$$\mu_{lt}^0 = E(Z_0 | T = t; \Omega_l) = x_0'\beta_l + \alpha(t - X\beta), l = 1,2 \quad (1)$$

$$\sigma_{0t}^2 = V(Z_0 | T = t; \Omega_l) = \sigma^2 k, \quad (2)$$

where

$$x_0' = x'(s_0),\ \alpha = r_0' R^{-1},$$

$$k = 1 - r_0' R^{-1} r_0.$$

Under the assumption that the populations are completely specified and for known prior probabilities of populations π_1 and $\pi_2 (\pi_1 + \pi_2 = 1)$, the Bayes discriminant function (BDF) minimizing the probability of misclassification (PMC) is formed by the log-ratio of conditional densities

$$W_t(Z_0) = \left(Z_0 - \frac{1}{2}(\mu_{1t}^0 + \mu_{2t}^0)\right)(\mu_{1t}^0 - \mu_{2t}^0)/\sigma_{0t}^2 + \gamma, \quad (3)$$

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

In practical applications the parameters of the p. d. f. are usually not known. Then the estimators of unknown parameters can be found from training samples taken separately from Ω_1 and Ω_2. When estimators of unknown parameters are used, the plug - in version of BDF (BPDF) is obtained.

Let, $\hat{\mu}_{1T}^0$, $\hat{\mu}_{2T}^0$ be the estimators of μ_{1T}^0, μ_{2T}^0 respectively, obtained by replacing β in equation (1) with their estimator $\hat{\beta}$ based on T.

The BPDF is obtained by replacing the parameters $\hat{\beta}$ in (3) with their estimators. Then the BPDF for random T is

$$W_T(Z_0; \hat{\beta}) = \left(Z_0 - \alpha(T - X\hat{\beta}) - \frac{1}{2}x_0' H\hat{\beta}\right)(x_0' G\hat{\beta})/(k\sigma^2) + \gamma, \quad (4)$$

with $H = (I_q, I_q)$ and $G = (I_q, -I_q)$, where I_q denotes the identity matrix of order *q*.

Definition 1. The actual error rate for BPDF is defined as

$$P(\hat{\beta}) = \sum_{l=1}^{2} \pi_l \hat{P}_{0l}, \quad (5)$$

where, for $l = 1,2$,

$$\hat{P}_{0l} = P_{0T}\left((-1)^l W_T(Z_0; \hat{\beta}) > 0 | \Omega_l\right), \quad (6)$$

is the conditional probability that $W_T(Z_0; \hat{\beta})$ misclassifies Z_0 when it comes from Ω_l (conditional probability is based on conditional distribution of Z_0 with mean μ_{lT}^0 and variance σ_{0T}^2).

In the considered case, the actual error rate specified in (5), (6) for $d_B(z^0; \hat{\beta})$ can be rewritten as

$$P(\hat{\beta}) = \sum_{l=1}^{2} \pi_l \Phi(\hat{Q}_l),$$

where $\Phi(\cdot)$ is the standard normal distribution function, and

$$\hat{Q}_l = (-1)^l \left((a_l + b\hat{\beta})' x_0' G\hat{\beta} + \sigma^2 \gamma\right) \Big/ \left(\sigma \sqrt{\hat{\beta}' G' x_0 x_0' G\hat{\beta} k}\right),$$

where for $l = 1,2$

$$a_l = x_0'\beta_l - \alpha X\beta, b = \alpha X - x_0' H/2.$$

Definition 2. The expectation of the actual error rate with respect to the distribution of T, designated as $E_T\{P(\hat{\beta})\}$, is called the expected error rate (EER).

In this paper we use estimators of β based on T

$$\hat{\beta} = X(X'R^{-1}X)^{-1}X'R^{-1}T$$

It is easy to show that

$$\hat{\beta}_{ML} \sim N_{2q}(\beta, \Sigma_\beta),\ \Sigma_\beta = \sigma^2(X'R^{-1}X)^{-1}, \quad (7)$$

Put

$$\Delta_0^2 = (x_0' G\beta)^2 / (k\sigma^2). \quad (8)$$

Let $\lambda_{\max}(R)$ be the largest eigenvalue of R and let $\varphi(\cdot)$ be the standard normal distribution density function.

3. THE ASYMPTOTIC EXPANSION OF EER

Make the following assumptions:

(A1) $n(X'X)^{-1} \to V$, as $n \to \infty$, where V is positively definite $2q \times 2q$ matrix with finite determinant;

(A2) $rank(X) = 2q$; $\lambda_{\max}(R) < v < +\infty$, as $n \to \infty$;

(A3) $n_1/n_2 \to u$,as $n_1, n_2 \to \infty$, $0 < u < \infty$.

Theorem 1. Suppose that observation Z_0 to be classified by BPDF given in (4) and let assumptions (A1) - (A3) hold. Then the asymptotic expansion of EER is

$$E_T(P(\psi')) = \sum_{l=1}^{2} \pi_l \Phi(Q_l) + \pi_1 \varphi(Q_1) C/2 + O(1/n^2), \quad (9)$$

where for $l = 1,2$

$Q_l = -\Delta_0/2 + (-1)^l \gamma/\Delta_0$,

$C = \Lambda \Sigma_\beta \Lambda' \Delta_0 / k$,

$\Lambda = \alpha X - x_0'(H/2 + \gamma G/\Delta_0^2)$.

Proof. Expanding $P(\hat{\beta})$ in the Taylor series about points $\hat{\beta} = \beta$ we have

$$P(\hat{\beta}) = P_\beta + P_\beta' \Delta\hat{\beta} + \frac{1}{2} \Delta'\hat{\beta} P_\beta'' \Delta\hat{\beta} + R_3 \quad (10)$$

where R_3 is Lagrange remainder.

Taking the expectation of the right side of (10) and using (7) we get

$$E_T(P(\psi')) = P_\beta + \frac{1}{2} tr(P_\beta'' \Sigma_\beta) + E_T(R_3). \quad (11)$$

Note that

$$P_\beta'' = \pi_1 \varphi(Q_1)(\Lambda' x_0' G \beta \beta' G' x_0 \Lambda / k^2) \quad (12)$$

Remember, that Lagrange remainder R_3 is the third order polynomial with respect to the component of $\Delta\hat{\beta}$.

Third order partial derivatives of $\Phi(\hat{\theta}_l)$ with respect to $\hat{\beta}$ is bounded by the uniformly integrable functions in the same neighborhood.

So we can conclude that

$$E_T(R_3) = O(1/n^2). \quad (13)$$

Putting (12) - (13) into (11) we complete the proof of the theorem.

It is easy to notice that this formula agrees with the formulas derived before by other authors (see e.g., [2]).

The approximation of EER is formed by dropping the remainder from the right-hand side of (9). Denoting it by AER we have

$$AER = \sum_{l=1}^{2} \pi_l \Phi(Q_l) + \pi_1 \varphi(Q_1) C/2 \quad (14)$$

The values of AER can be considered as natural measure of the performance or the optimality criterion function of the spatial sampling design for supervised classification.

4. EXAMPLE AND DISCUSSIONS

Numerical example is considered to confirm the accuracy of the approximation based on proposed asymptotic expansion of the expected error rate in the finite (even small) training sample case.

In this example, observations are assumed to arise from univariate spatial Gaussian process on D with unknown constant mean and an isotropic exponential correlation function given by $r(h) = \exp\{-|h|/\theta\}$.

With an insignificant loss of generality the cases with $m = 1$, $n_1 = n_2 = n_0$ and $\pi_1 = \pi_2 = 0.5$ are considered. The Machalanobis distance between marginal distributions of Z_0 is specified by $\Delta = |(\beta_1 - \beta_2)/\sigma|$ Then from (8) it follows that $\Delta_0 = \Delta/\sqrt{k}$, $\gamma = 0$.

Denote theoretical values of EER by TER. They are obtained by numerical integration with MAPLE 9.5.

Assume that D is a 5×5 square grid points on R_+^2 with unit spacing.

For greater interpretability, correlation $r(h)$ function is reparametrized as $r(h) = \rho^{|h|}$ where ρ represents the correlation between adjacent points in D . Using K-optimal sampling design (see [9]) for $\rho \in [0.25;1]$ and $n_1 = n_2 = 2$ we have

$$D_1 = \{(0,3),(3,4)\},\ D_2 = \{(1,0),(4,3)\}$$

where D_i is the set of points in D , where training sample T_i is taken, $i = 1,2$.

Let the observation to be classified is taken at point $s_0 = (2,2)$.

The values of AER (14) and the values of index of relative accuracy of proposed asymptotic expansion specified by

$$\eta = |AER - EER|/EER$$

are given in Table 1 for various values of and for training sample design described above.

Independent observations case $(\rho = 0)$ is included in Table1 in order to estimate the effect of the spatial correlation to the expected error rate.

Table1 shows that AER values increases with spatial correlation.

Analyzing the content of the Table1 we can conclude the proposed approximation of EER based on derived asymptotic expansion is sufficiently accurate even in small training sample $(n=4)$ case, because the values of the index of relative accuracy is not so large $(\eta \in [0.0241;0.25848])$. It is interesting to notice that η attains its minimal and maximal values (these values are underlined in the Table1) in the same case with strongest dependence among observations (i.e., $\rho = 0.9$) but with different degree of separation between populations (i.e., $\Delta = 0.3$ and $\Delta = 0.6$). It is to be noted that in case of strongly separated populations $(\Delta \geq 1)$ the proposed approximation often is more accurate, than in case of close populations $(\Delta < 1)$.

	AER η	AER η
Δ	$\rho = 0$	$\rho = 0.25$
0.2	0.46513 0.05910	0.46352 0.06198
0.6	0.39639 0.12350	0.39174 0.13057
1.0	0.33054 0.13503	0.32337 0.14497
1.4	0.26929 0.11267	0.26036 0.12446
1.8	0.21400 0.07451	0.20419 0.08703
2.2	0.16562 0.03693	0.15578 0.04898
2.6	0.12465 0.01061	0.11546 0.02105
3.0	0.09109 0.00141	0.08304 0.00632
	$\rho = 0.5$	$\rho = 0.7$
0.2	0.45788 0.07155	0.44693 0.08900
0.6	0.37549 0.15162	0.34448 0.18464
1.0	0.29842 0.17120	0.25234 0.20497
1.4	0.22948 0.15163	0.17516 0.17812
1.8	0.17049 0.11192	0.11491 0.12797
2.2	0.12223 0.06952	0.07109 0.07652
2.6	0.08446 0.03648	0.04141 0.03823
3.0	0.05619 0.01638	0.02268 0.01613
	$\rho = 0.8$	$\rho = 0.9$
0.2	0.43512 0.10673	0.40788 0.14390
0.6	0.31204 0.21332	0.24227 0.25848
1.0	0.20702 0.22758	0.12200 0.24158
1.4	0.12642 0.18748	0.05144 0.16326
1.8	0.07075 0.12474	0.01799 0.08168
2.2	0.03617 0.06730	0.00519 0.03076
2.6	0.01685 0.02970	0.00123 0.00912
3.0	0.00714 0.01091	0.00024 0.00241

Table 1. **Values of AER and η for the K-optimal sampling design with $n_1 = n_2 = 2$ and $\pi_1 = \pi_2 = 0.5$**

So the results of numerical analysis give us strong arguments to hope that proposed asymptotic expansion will yield useful approximations of expected error rate of classification of spatially correlated Gaussian observations in finite training (even small) sample case.

5. REFERENCES

[1] C.R.O. Lawoko and G.L. McLachlan. *Discrimination with autocorrelated observations.* Pattern Recognition, 18, 2:145-149, 1985.

[2] G.L. McLachlan. *Discriminant analysis and statistical patter recognition*, Wiley,New York, 2004.

[3] P. Switzer. *Extensions of linear discriminant analysis for statistical classification of remotely sensed satellite imagery*. Math. Geol., 12, 4, 1980.

[4] K.V. Mardia. *Spatial discrimination and classification maps*. Comm. Statist. Theory Methods, 13, 18:2181-2197, 1974

[5] J. Saltyte and K. Ducinskas. *Comparison of ML and OLS estimators in discriminant analysis of spatially correlated observations.* Informatica, 13, 2:297-238, 2002.

[6] J. Saltyte-Benth and K. Ducinskas. *Linear discriminant analysis of multivariate spatial-temporal regressions.* Scand. J. Statist., 32:281-294, 2005.

[7] Yu. Kharin. *Robustness in statistical pattern recognition.* Dordrecht, Boston, London, Kluwer Academic Publishers, 302, 1996.

[8] K. Ducinskas. *Approximation of the expected error rate in classification of the Gaussian random field observations.*, Statistics and Probalibity Letters, 79:138-144, 2009.

[9] D.L. Zimmerman, *Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction.* Environmetrics, 17:635-652,2006.