Liet. matem. rink., T. 41, spec. nr., 478–484
© 2001 Matematikos ir informatikos institutas

Discriminant analysis of multivariate spatial regressions

Jūratė ŠALTYTĖ, Kęstutis DUČINSKAS (KU)
e-mail: jsaltyte@gmf.ku.lt, duce@gmf.ku.lt

1. Introduction

Let $\{Z(s) : s \in D \subset R^2\}$ be a p-variate Gaussian random field having different means and factorised covariance matrices in populations Ω_1 and Ω_2. Assume that the model of $Z(s)$ in population Ω_l is

$$Z(s) = B_l^T x(s) + \varepsilon_l(s),$$

where $x(s) = \left(x_1(s), \dots, x_q(s)\right)^T$ is a $q \times 1$ vector of nonrandom regressors and B_l is unknown parameter matrix of order $q \times p$, $l = 1, 2$. Assume, that $\{\varepsilon_l(s) : s \in D \subset R^2\}$ is a zero-mean stationary random Gaussian field with spatial covariance defined by a parametric model cov $\left\{\varepsilon_l(t), \varepsilon_l(s)\right\} = \sigma(t - s; \theta_l)$ for all $t, s \in D$, where $\theta_l \in \Theta$ is a $m \times 1$ parameter vector, Θ being an open subset of R^m, $l = 1, 2$. We restrict our attention to the homoscedastic models, i.e., $\sigma(0; \theta) = \Sigma$ for each $\theta \in \Theta$.

Then, in Ω_l the mean function at location s is $\mu_l(s) = B_l^T x(s)$ and the spatial covariance function is cov $\left\{\varepsilon_l(t), \varepsilon_l(s)\right\} = \Sigma\rho(t - s; \theta_l)$, where $\rho(t - s; \theta_l)$ is the spatial correlation function, $l = 1, 2$. It is assumed that the function $\rho(t-s; \theta_l)$ is positive definite (Mardia and Marshall [5]).

Suppose for $t, s \in D$, that

$$\text{cov}\left\{\varepsilon_1(t), \varepsilon_2(s)\right\} = 0. \tag{1}$$

Consider the problem of classification of an observation $Z(r)$, with $r \in D_0 \subset D$, into one of two populations specified above.

Then the probability density function (p.d.f.) of $Z(r)$ in Ω_l is

$$p_l\left(z(r); \mu_l(r), \Sigma\right) = \frac{1}{\sqrt[p]{2\pi}|\Sigma|^{-\frac{1}{2}}} \exp\left\{-\frac{1}{2}\left(z(r) - B_l^T x(s)\right)^T \Sigma^{-1} \left(z(r) - B_l^T x(s)\right)\right\}.$$

Under the assumption that the populations are completely specified and for known prior probabilities of populations $\pi_1(r)$, $\pi_2(r)$ $(\pi_1(r) + \pi_2(r) = 1)$, the Bayes classification rule (BCR) $d_B(\cdot)$ minimizing the probability of misclassification (PMC) is equivalent to assigning $Z(r) = z(r)$ to Ω_l if

$$\pi_l(r) p_l\left(z(r)\right) = \max_{k=1,2} \pi_k(r) p_k\left(z(r)\right),$$

$l = 1, 2$. Then BCR $d_B(z(r))$ is defined as

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

Denote by P_B the PMC of BCR. Usually P_B is called Bayes error rate (see, e.g., Hand (1997)).

In practical applications the parameters of p.d.f. are usually unknown and must be estimated from training samples T_1 and T_2 taken separately from Ω_1 and Ω_2, respectively.

The performance of the plug-in version of the BCR when parameters are estimated from training samples with independent observations are 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 LDF when the parameters are estimated from training samples being a realisations of Gaussian random field. Here the maximum likelihood (ML) estimators of unknown means assuming covariance matrices to be known are considered.

Suppose in a region $D_1 \subset D$, $D_1 \cap D_0 = \varnothing$, we observe the training sample $T = \{T_1, T_2\}$ with $T_l = \{Z_{l1}, \ldots, Z_{lN_l}\}$, where $Z_{l\alpha} = Z(s^l_\alpha)$ denotes the αth observation from Ω_l, $l = 1, 2$, $\alpha = 1, \ldots, N_l$. Assume, that all points in D_0 are beyond the range of spatial correlation function (Cressie, 1993, ch.2) defined for points in D_1. Then $Z(r)$ is independent on T.

Let $\widehat{B}_1$ and $\widehat{B}_2$ be the ML estimators of B_1 and B_2, respectively, based on T, and let $\widehat{\mu}_l(r) = \widehat{B}_l^T x(r)$. The plug-in rule $d_B\big(z(r), \widehat{\mu}_1(r), \widehat{\mu}_2(r)\big)$ is obtained by replacing the parameters in (2) with their estimators, i.e.,

$$d_B\big(z(r), \widehat{\mu}_1(r), \widehat{\mu}_2(r)\big) = \arg\max_{\{l=1,2\}} \pi_l(r)p_l\big(z(r); \widehat{\mu}_l(r)\big). \tag{3}$$

Then the corresponding discriminant function $L(r)$, also known simply as the sample LDF (see, McLachlan (1974)), is defined as

$$L(r) = \left(z(r) - \frac{1}{2}(\widehat{\mu}_1(r) + \widehat{\mu}_2(r))\right)^T \Sigma^{-1}(\widehat{\mu}_1(r) - \widehat{\mu}_2(r)) + \gamma(r),$$

where $\gamma(r) = \ln \frac{\pi_1(r)}{\pi_2(r)}$.

DEFINITION 1. The actual error rate of $d_B\big(z(r); \widehat{\mu}_1(r), \widehat{\mu}_2(r)\big)$ is defined as

$$\begin{aligned}&P^r\big(\widehat{\mu}_1(r), \widehat{\mu}_2(r)\big)\\ &\triangleq \sum_{l=1}^{2} \pi_l(r) \int \Big(1 - \delta\Big(l, d_B\big(z(r); \widehat{\mu}_1(r), \widehat{\mu}_2(r)\big)\Big)p_l\big(z(r); \mu_l(r), \Sigma\big)\Big)\, \mathrm{d}z(r).\end{aligned}$$

In our case the actual error rate for $d_B\big(z(r);\widehat{\mu}_1(r),\widehat{\mu}_2(r)\big)$ is then given by

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

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

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

The goal of this paper is to find asymptotic expansions of ER for the plug-in LDF. The case of independent normally distributed observations in training sample from one of two classes with equal feature covariance matrices, was considered in [3]. Dučinskas [7] has been made the generalization for the case of arbitrary number of classes ($l \geqslant 2$) and regular class-conditional densities. McLachlan [10] presented ER for the case of equi-correlated Gaussian observations. Mardia [9] considered similar problem of classifying spatially distributed Gaussian observations with constant means, but he did not analyse ER or probabilities of misclassification.

In this paper we obtain the asymptotic expansion up to the order $O(N^{-1})$, where $N = N_1 + N_2$, for the ER of classifying spatially distributed Gaussian observation with different means and spatially factorised covariance matrices. Terms of higher order are omitted from the asymptotic expansion since the contribution of higher order terms is generally negligible [10]. ML estimators of means are used in plug-in version of Bayes classification rule. We also make a comparison for the accuracy of our asymptotic approximation with Monte Carlo simulations when training sample sizes are small.

2. Main results

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

$$\mu_l^+ = \big(\mu_l^T(s_1),\ldots,\mu_l^T(s_{N_l})\big)^T,$$

and

$$\Sigma_l^+ = C_l \otimes \Sigma,$$

where C_l is the spatial correlation matrix of order $N_l \times N_l$, whose (α,β)th element is $\rho_l(s_\alpha - s_\beta)$, $\alpha,\beta = 1,\ldots,N_l$, $l = 1,2$. Suppose, that Σ_l^+ is known, $l = 1,2$. Without loss of generality, we assume, that $\Sigma = I$, where I is the unity matrix.

Let X_l be an $N_l \times q$ regressor matrix with i-th column $(x_{1i},\dots,x_{N_l i})^T$, where $x_{ki} = x_i(s_k)$, $i = 1,\dots,q$, $k = 1,\dots,N_l$, $l = 1,2$. Denote by $\mathbf{T}_l^*$ the matrix of order $N_l \times p$, whose j'th column is $\big(Z_{lj}(s_1),\dots,Z_{lj}(s_{N_l})\big)^T$.

Lemma (Kai-Tai and Yao-Ting, 1997). *For $l = 1,2$ ML of B_1 and B_2 based on T are*

$$\widehat{B}_l = (X_l^T C_l^{-1} X_l)^{-1} X_l^T C_l^{-1} \mathbf{T}_l^*, \quad l = 1,2.$$

Let $\lambda_{N_l}(C_l)$ be the largest eigenvalue of C_l, $l = 1,2$.

Assumption 1. Assume, that $x^T(r)(X_l^T X_l)x(r) = O(\frac{1}{N_l})$, as $N_l \to \infty$, $l = 1,2$.

Assumption 2. Suppose, that $\lambda_{N_l}(C_l) < \nu_l$, $\nu_l < \infty$, as $N_l \to \infty$, $l = 1,2$.

Assumption 3. Assume, that $\frac{N_1}{N_2} \to v$, as $N_1, N_2 \to \infty$, $0 < v < \infty$, $l = 1,2$.

Put $\Delta\widehat{\mu}_l(r) = \widehat{\mu}_l(r) - \mu_l(r) = \big(\widehat{B}_l - B_l\big)^T x(r)$. Let $\varphi(\cdot)$ denote the standard normal p.d.f.

Define

$$a_l(r) = x^T(r)(X_l^T C_l^{-1} X_l)^{-1} x(r), \tag{4}$$

and $\Delta^2(r) = x^T(r)(B_1 - B_2)\Sigma^{-1}(B_1 - B_2)^T x(r)$ (square of the Mahalanobis distance) for any $r \in D$ and $l = 1,2$.

For simplicity we omit the superscript r on $P^r(\cdot)$.

Let $P_l^{(1)}$ be the first-order derivatives of $P\big(\widehat{\mu}_1(r),\widehat{\mu}_2(r)\big)$ with respect to $\widehat{\mu}_l(r)$ evaluated at $\mu_l(r)$, and $P_{k,l}^{(2)}$ denotes the second-order derivatives of $P\big(\widehat{\mu}_1(r),\widehat{\mu}_2(r)\big)$ with respect to $\widehat{\mu}_l(r)$ and $\widehat{\mu}_k(r)$ evaluated at $\mu_l(r)$ and $\mu_k(r)$, respectively, $(l,k = 1,2)$.

Theorem. *Suppose Assumptions 1–3 hold for training samples T_1, T_2. Then the asymptotic expansion of the expected error rate for the $d_B\big(z(r),\widehat{\mu}_1(r),\widehat{\mu}_2(r)\big)$ is*

$$E_T\Big\{P\big(\widehat{\mu}_1(r),\widehat{\mu}_2(r)\big)\Big\} = \sum_{l=1}^{2}\pi_l(r)\Phi\left(-\frac{\Delta(r)}{2} + (-1)^l\frac{\gamma(r)}{\Delta(r)}\right)$$
$$+\frac{\pi_1(r)}{2\Delta(r)}\varphi\left(-\frac{\Delta(r)}{2} - \frac{\gamma(r)}{\Delta(r)}\right)\sum_{l=1}^{2} a_l(r)\left(-\frac{\Delta(r)}{2} + (-1)^l\frac{\gamma(r)}{\Delta(r)}\right)^2 + O(N^{-2}).$$

Proof. Since $P\big(\widehat{\mu}_1(r),\widehat{\mu}_2(r)\big)$ is invariant under linear transformations of data we use the convenient canonical form of $\mu_k^T(r) = \big((-1)^{k+1}\frac{\Delta(r)}{2},0,\dots,0\big)$, $k = 1,2$. Expand $P\big(\widehat{\mu}_1(r),\widehat{\mu}_2(r)\big)$ in Taylor series about the point $\widehat{\mu}_1(r) = \big(\frac{\Delta(r)}{2},0,\dots,0\big)^T$, $\widehat{\mu}_2(r) = \big(-\frac{\Delta(r)}{2},0,\dots,0\big)^T$. Taking the expectation with respect to the distribution of T and dropping the third order terms we have

$$E_T\Big(P\big(\widehat{\mu}_1(r),\widehat{\mu}_2(r)\big)\Big) = P_B + \sum_{l=1}^{2}(P_l^{(1)})^T E_T\{\Delta\widehat{\mu}_l(r)\}$$

$$+\frac{1}{2}\sum_{l,k=1}^{2}\operatorname{tr}\Big(E_T\{\Delta\widehat{\mu}_k^T(r)P_{k,l}^{(2)}\Delta\widehat{\mu}_l(r)\}\Big). \tag{5}$$

Since $P\big(\widehat{\mu}_1(r),\widehat{\mu}_2(r)\big)$ is minimized at $\mu_l^T(r)=\big((-1)^{l+1}\frac{\Delta(r)}{2},0,\ldots,0\big)$, then

$$P_l^{(1)}=\mathbf{0}_p,\quad l=1,2, \tag{6}$$

where $\mathbf{0}_p$ is p-dimensional vector of zeroes.

Using Lemma and (4) we get, that

$$E_T\{\Delta\widehat{\mu}_l(r)\Delta\widehat{\mu}_l^T(r)\}=a_l(r)\mathbf{I}_p,\quad l=1,2, \tag{7}$$

where $\mathbf{I}_p$ is $p\times p$ unity matrix.

From (1) it follows, that

$$E_T\{\Delta\widehat{\mu}_1(r)\Delta\widehat{\mu}_2^T(r)\}=\mathbf{0}_{p\times p}, \tag{8}$$

where $\mathbf{0}_{p\times p}$ is $p\times p$ matrix of zeroes.

Note that

$$P_{l,l}^{(2)}=\frac{\pi_1(r)}{\Delta(r)}\varphi\Big(-\frac{\gamma(r)}{\Delta(r)}-\frac{\Delta(r)}{2}\Big)\begin{pmatrix}\Big(-\frac{\Delta(r)}{2}+(-1)^l\frac{\gamma(r)}{\Delta(r)}\Big)^2 & \mathbf{0}_{p-1}^T\\ \mathbf{0}_{p-1} & \mathbf{I}_{p-1}\end{pmatrix}, \tag{9}$$

here $\mathbf{0}_{p-1}$ is a $(p-1)$-dimensional vector of zeroes and $\mathbf{I}_{p-1}$ is $(p-1)\times(p-1)$ unity matrix.

Using Assumptions 1, 2 and inequality

$$x^T(r)(X_l^TC_l^{-1}X_l)^{-1}x(r)<\lambda_{N_l}x^T(r)(X_l^TX_l)^{-1}x(r),$$

we can conclude that

$$a_l(r)=\mathrm{O}\Big(\frac{1}{N_l}\Big),\quad\text{as}\quad N_l\to\infty,\ l=1,2.$$

Omitted terms of expression (5) are of order $a_1^k(r)$ and $a_2^k(r)$, $k\geqslant 2$, and so under Assumption 3 they are of order $\mathrm{O}(N^{-2})$.

Then putting (6)–(9) into (5) we complete the proof of the stated theorem.

The asymptotic ER, AER, defined by

$$AER=\sum_{l=1}^{2}\pi_l(r)\Phi\Big(-\frac{\Delta(r)}{2}+(-1)^l\frac{\gamma(r)}{\Delta(r)}\Big)$$
$$+\frac{\pi_1(r)}{2\Delta(r)}\varphi\Big(-\frac{\Delta(r)}{2}-\frac{\gamma(r)}{\Delta(r)}\Big)\sum_{l=1}^{2}a_l(r)\Big(\Big(-\frac{\Delta(r)}{2}+(-1)^l\frac{\gamma(r)}{\Delta(r)}\Big)^2+p-1\Big),$$

can be used for the performance evaluation and comparison of the considered LDF. The minimum of the AER for fixed training sample sizes can be considered as criterion for a optimality of spatial sampling design.

3. Example

Here we make the comparison of AER with ER estimated from Monte Carlo simulations (denoted by P_{MC}) for one special case. In the example we assume D to be the integer regular 2-dimensional lattice and use the second-order neighbourhood scheme for training samples. There are four spatially symmetric observations in training sample for each class: empty circles for T_1 and filled once – for T_2 (Fig. 1).

Haining (1990) suggested represent mean as a polynomial function of coordinates of a specified order. This is so-called trend surface model. The trend surface model could be considered as a special case of polynomial regression model. Here we use the first-order trend surface model, which corresponds to the case, when regressors are simply the coordinates of considered locations ($q = 2$).

Suppose 2-dimensional observation is taken in the each of considered locations. Let

$$B_1 = \begin{pmatrix} 0.1 & 0.4 \\ 0.3 & 0.2 \end{pmatrix} \quad \text{and} \quad B_2 = \begin{pmatrix} 0.1 & 0.4+\delta \\ 0.3 & 0.2+\delta \end{pmatrix},$$

i.e., the classes differ in the parameter values for the second component of the observation.

Assume the spherical correlation function

$$c(|h|, \kappa_l) = \begin{cases} 1 - \frac{3}{2}\frac{|h|}{\kappa_l} + \frac{1}{2}\frac{|h|^3}{\kappa_l^3}, & 0 \leqslant |h| \leqslant \kappa_l, \\ 0, & h| > \kappa_l, \end{cases}$$

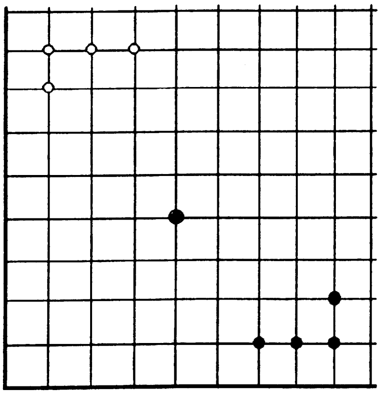

Fig. 1. Training sample design (locations of observations in T_1 and T_2, and the location of observation to be classified)

Table 1

Comparison of approximation with simulation ($\pi_1 = 0.3$)

$\Delta(r)$	AER	P_{MC}	AER/P_{MC}
1,0	0.523	0.261	2.002
1,4	0.296	0.234	1.267
1,8	0.203	0.187	1.086
2,2	0.145	0.167	0.872
2,6	0.104	0.126	0.822
3,0	0.073	0.112	0.650

with range κ_l, $l = 1, 2$. It is easy to see from the Fig. 1, that $\kappa_1 = \sqrt{6}$ and $\kappa_2 = \sqrt{7}$ are appropriate ranges for two training samples.

In Table 1 the values of AER and values obtained by Monte Carlo simulation taking 100 replications at each location are presented. Column with ratio AER/P_{MC} allow us to estimate the accuracy of the proposed expansion. We can conclude that this expansion is sometimes appropriate even for small training sample sizes.

References

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

[2] D.J. Hand, *Construction and Assessment of Classification Rules*, John Wiley & Sons, New York (1997).

[3] M. Okamoto, An asymptotic expansion for the distribution of the linear discriminant function, *Ann. Math. Statist.*, **34**, 1286–1301 (1963).

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

[5] N.A.C. Cressie, *Statistics for Spatial Data*, Wiley Sons, New York (1993).

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

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

[8] G.J. McLachlan, Further results on the effect of intraclass correlation among training samples in discriminant analysis, *Pattern Recognition*, **8**, 273–275 (1974).

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

[10] M.J. Schervish, Asymptotic expansions of the means and variances of error rates, *Biometrika*, **68**, 295–299 (1981).

[11] F. Kai-Tai, Z. Yao-Ting, *Generalised Multivariate Analysis*, Springer-Verlag (1997).

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

Diskriminantinė daugiamačių erdvinių regresijų analizė

K.Dučinskas, J. Šaltytė

Straipsnyje nagrinėjamas daugiamačių Gauso stebėjimų, pasiskirsčiusių erdvėje, klasifikavimo uždavinys. Gautas pirmos eilės asimptotinis tikėtinos klasifikavimo klaidos skleidinys atvejui, kai į Bajeso klasifikavimo taisyklę įstatome maksimalaus tikėtinumo vidurkių įverčius pagal erdvėje koreliuotas mokymo imtis. Atliktas skaitinis asimptotinės klasifikavimo klaidos palyginimas su klaida, sumodeliuota Monte Karlo metodu.