Statistical Classification of the Observation of Nuggetless Spatial Gaussian Process with Unknown Sill Parameter

K. Dučinskas

Department of Statistics, Klaipėda University
H. Manto str. 84, 92294 Klaipėda, Lithuania
duce@gmf.ku.lt

Received: 2008-10-02 **Revised:** 2009-04-28 **Published online:** 2009-05-26

Abstract. The problem of classification of spatial Gaussian process observation into one of two populations specified by different regression mean models and common stationary covariance with unknown sill parameter is considered. Unknown parameters are estimated from training sample and these estimators 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 approximation based on derived asymptotic expansion in the small training sample case is carried out. Comparison of two spatial sampling designs based on values of this approximation is done.

Keywords: Gaussian random field, Bayes discriminant function, spatial correlation, actual error rate, expected error rate.

1 Introduction

In classical discriminant analysis 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 (BCR). 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 analyse 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 regression mean models and common variance. This result was generalized to multivariate spatial-temporal regression model in Šaltytė-Benth and Dučinskas [6]. However, in these papers the interclass spatial correlation was assumed equal zero. Also, the observation to be classified were 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 is considered in [7].

2 The main concepts and definitions

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

$$\bigl\{Z(s)\colon s \in D \subset R^m\bigr\}.$$

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

$$Z(s) = x'(s)\beta_l + \varepsilon(s), \tag{1}$$

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 $\{\varepsilon(s)\colon s \in D \subset R^m\}$ with covariance function defined by nuggetless model for all $s, u \in D$

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

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

$$T = X\beta + E, \tag{3}$$

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_l 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

$$\mu_{lt}^0 = E(Z_0|T; \Omega_l) = x_0'\beta_l + \alpha_0(T - X\beta), \quad l = 1, 2, \tag{4}$$

$$\sigma_{0t}^2 = V(Z_0|T; \Omega_l) = \sigma^2 k, \tag{5}$$

where

$$x_0' = x'(s_0), \quad \alpha_0 = r_0'R^{-1}, \quad k = 1 - r_0'R^{-1}r_0. \tag{6}$$

Under the assumption that the populations are completely specified and for known prior probabilities of populations π_1 and π_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

$$W_t(Z_0) = \left(Z_0 - \frac{1}{2}(\mu_{1t}^0 + \mu_{2t}^0)\right)(\mu_{1t}^0 - \mu_{2t}^0)/\sigma_{0t}^2 + \gamma, \tag{7}$$

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

In practical applications the parameters of the PDF 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 $\hat{\mu}_{1T}^0$, $\hat{\mu}_{2T}^0$ and $\hat{\sigma}_{0T}^2$ be the estimators of μ_{1T}^0, μ_{2T}^0 and σ_{0T}^2, respectively, obtained by replacing β and σ^2 in equations (4) and (5) with their estimators $\hat{\beta}$ and $\hat{\sigma}^2$ based on T. Put $\Psi' = (\beta', \sigma^2)$ and $\hat{\Psi}' = (\hat{\beta}', \hat{\sigma}^2)$.

The BPDF is obtained by replacing the parameters β, σ^2 in (7) with their estimators. Then the BPDF for random T is

$$W_T(Z_0; \hat{\Psi}) = \left(Z_0 - \alpha_0(T - X\hat{\beta}) - \frac{1}{2}x_0'H\hat{\beta}\right)(x_0'G\hat{\beta})(k\hat{\sigma}^2) + \gamma, \tag{8}$$

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

$$P(\hat{\Psi}) = \sum_{l=1}^{2} \pi_l \hat{P}_{0l}, \tag{9}$$

where, for $l = 1, 2$,

$$\hat{P}_{0l} = P_{0T}\big((-1)^l W_T(Z_0; \hat{\Psi}) > 0 | \Omega_l\big), \tag{10}$$

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

In the considered case, the actual error rate specified in (9), (10) for $d_B(z^0; \hat{\Psi})$ can be rewritten as

$$P(\hat{\Psi}) = \sum_{l=1}^{2} \pi_l \Phi(\hat{Q}_l), \tag{11}$$

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

$$\hat{Q}_l = (-1)^l \big((a_l + b\hat{\beta})' x_0' G\hat{\beta} + \hat{\sigma}^2 \gamma k\big) \Big/ \Big(\sigma\sqrt{\hat{\beta}' G' x_0}\, x_0' G\hat{\beta} k\Big), \tag{12}$$

where for $l = 1, 2$

$$a_l = x_0' \beta_l - \alpha_0 X\beta, \quad b = \alpha_0 X - x_0' H/2. \tag{13}$$

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

It is known (see [8]), that the ML estimators of β and σ^2 based on T are

$$\hat{\beta}_{ML} = X\big(X' R^{-1} X\big)^{-1} X' R^{-1} T, \tag{14}$$

$$\hat{\sigma}^2_{ML} = (T - X\hat{\beta}_{ML})' R^{-1} (T - X\hat{\beta}_{ML})/n. \tag{15}$$

Using the properties of multivariate Gaussian distribution it is easy to prove that

$$\hat{\beta}_{ML} \sim N_{2q}(\beta, \Sigma_\beta), \quad \Sigma_\beta = \sigma^2\big(X' R^{-1} X\big)^{-1}, \tag{16}$$

$$\hat{\sigma}^2_{ML} \sim \sigma^2 \chi^2_{n-2q}/(n - 2q). \tag{17}$$

ML estimator of β and bias adjusted ML estimator of σ^2 are used in BPDF, i.e. $\hat{\beta} = \hat{\beta}_{ML}$, $\hat{\sigma}^2 = \hat{\sigma}^2_{ML} n/(n - 2q)$.

Then by using (14)–(17) it is easy to show (see e.g., [9]) that

$$E_T(\Delta\hat{\beta}) = 0, \quad E_T(\Delta\hat{\beta}' \Delta\hat{\beta}) = \Sigma_\beta, \quad E_T\big(\Delta\hat{\sigma}^2 \Delta\hat{\beta}\big) = 0, \tag{18}$$

$$E_T\big(\Delta\hat{\sigma}^2\big) = 0, \quad E_T(\Delta\hat{\sigma})^2 = 2\sigma^4/(n - 2q), \tag{19}$$

where

$$\Delta\hat{\beta} = \hat{\beta} - \beta, \quad \Delta\hat{\sigma}^2 = \hat{\sigma}^2 - \sigma^2.$$

Put

$$\Delta_0^2 = (\mu_{1T} - \mu_{2T})^2 / \big(k\sigma^2\big). \tag{20}$$

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) $\operatorname{rank}(X) = 2q;\ \ \lambda_{max}(R) < v < +\infty$, 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 and let assumptions* (A1)–(A3) *hold. Then the asymptotic expansion of EER is*

$$E_T\big(P(\hat\Psi)\big) = \sum_{l=1}^{2} \pi_l \Phi(Q_l) + \pi_1 \varphi(Q_1)\big\{C + 2\gamma^2/(n-2q)\big\}/2\Delta_0 + O\big(1/n^2\big), \tag{21}$$

where for $l = 1, 2$

$$Q_l = -\Delta_0/2 + (-1)^l \gamma/\Delta_0, \tag{22}$$

$$C = \Lambda \Sigma_\beta \Lambda' \Delta_0^2/k, \quad \Lambda = \alpha_0 X - x_0'\big(H/2 + \gamma G/\Delta_0^2\big). \tag{23}$$

Proof. Expanding $P(\hat\Psi)$ in the Taylor series about points $\hat\beta = \beta$ and $\hat\sigma^2 = \sigma^2$, we have

$$P\big(\hat\Psi\big) = P_\beta + P_\beta' \Delta\hat\beta + \hat P_\sigma \Delta\hat\sigma^2 + \frac{1}{2}\big(\Delta'\hat\beta \hat P_\beta'' \Delta\hat\beta + 2\Delta\hat\beta' \hat P_{\beta\sigma^2}'' + \hat P_{\sigma^2}''(\Delta\hat\sigma)^2\big) + R_3, \tag{24}$$

where R_3 is Lagrange remainder.

Taking the expectation of the right side of (24) and using (18), (19) we get

$$E_T\big(P(\hat\Psi)\big) = P_\beta + \frac{1}{2}\operatorname{tr}(\hat P_\beta'' \Sigma_\beta) + \hat P_{\sigma^2}'' \frac{\sigma^4}{n-2q} + E_T(R_3). \tag{25}$$

Note that

$$\hat P_\beta'' = \pi_1 \varphi(Q_1)\big(\Lambda' x_0' G \beta \beta' G' x_0 \Lambda / k^2\big) \tag{26}$$

and

$$\hat P_{\sigma^2}'' = \pi_1 \varphi(Q_1) \gamma^2 / \big(\sigma^4 \Delta_0\big). \tag{27}$$

Remember, that Lagrange remainder R_3 is the third order polynomial with respect to the components of $\Delta\hat\beta$ and $\Delta\hat\sigma^2$. Coefficients of this polynomial are the third order partial derivatives of $P(\hat\Psi)$ with respect to $\hat\beta$ and $\hat\sigma^2$ estimated in the neighbourhood of their true values.

It is obvious that all third order moments of components of normally distributed vector $\Delta\hat{\beta}$ are equal 0 and

$$E_T\big(\Delta\hat{\sigma}^2\big)^3 = 8/(n-2q)^2 = O\big(1/n^2\big).$$

Third order partial derivatives of $\Phi(\hat{\theta}_l)$ with respect to $\hat{\beta}$ and $\hat{\sigma}^2$ are bounded by the uniformly integrable functions in the same neighbourhood.

Then we can conclude that

$$E_T(R_3) = O\big(1/n^2\big). \tag{28}$$

Notice that

$$\Delta_0^2 = (x_0'G\beta)^2/(k\sigma)^2. \tag{29}$$

Putting (26)–(29) into (25) 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]).

4 Example and discussions

The first 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 $r(h) = \exp\{-|h|/\alpha\}$. Then semivariogram has the form $\gamma(h) = \sigma^2(1-\exp\{h/\alpha\})$.

With an insignificant loss of generality the cases with $m = 1, \quad n_1 = n_2 = n_0$ 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 (5), (6) and (20) it follows that $k = 1 - r_0'R^{-1}r_0, \ \ \Delta_0 = \Delta/\sqrt{k}, \ \ \gamma = 0$.

Denote theoretical values of EER by TER.

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 spatial sampling design (SSD) (see [10]) for $\rho \in [0.25;1)$ and $n_1 = n_2 = 2$ we have

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

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 and the values of index of relative accuracy of proposed asymptotic expansion specified by

$$\eta = |\text{AER} - \text{TER}|/\text{TER}$$

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

Independent observations case ($\rho = 0$) is included in Table 1 in order to estimate the effect of the spatial correlation to the expected error rate.

Table 1 shows that AER values increases with spatial correlation.

Table 1. Values of AER, η for the K-optimal SSD $n_1 = n_2 = 2$ and $\pi_1 = \pi_2 = 0.5$

Δ	AER	η	AER	η
	$\rho = 0$		$\rho = 0.25$	
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

Analysing the content of the Table 1 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 ($\Delta < 1$).

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.

The second example numerically illustrates the comparison of two SSD based on the minimum of AER criterion.

Assume that D is a 2×2 square grid points on R_+^2 with unit spacing. Let the observation to be classified is taken at point $s_0 = (1, 1)$ and T is taken in the second order neighbourhood of s_0 i.e. $n = 8$.

Consider two SSD ξ_1 and ξ_2 specified by

$$\xi_1 = \{s_0,\ D_1 = \{(1,2),(2,2),(2,1),(2,0)\},\ D_2 = \{(1,0),(0,0),(0,1),(0,2)\}\},$$
$$\xi_2 = \{s_0,\ D_1 = \{(1,2),(2,1),(0,1),(1,0)\},\ D_2 = \{(0,0),(0,2),(2,0),(2,2)\}\}.$$

They are illustrated in Fig. 1.

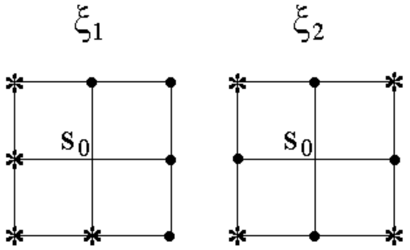

Fig. 1. Two different SSD with D_1 and D_2 points signed as • and $*$, respectively.

Let $d_i^{(l)}$ be the sum of distances from s_0 to pionts in D_i, for SSD ξ_l, $i = 1, 2$, $l = 1, 2$. Then $d_{12}^{(l)} = |d_1^{(l)} - d_2^{(l)}|$ represents the degree the population labels assymetry in training sample. In the considered situation we have $d_{12}^{(1)} = 0$, $d_{12}^{(2)} = 4(\sqrt{2} - 1)$.

Two levels of populations seperability i.e. $\Delta = 0.2$ and $\Delta = 2.0$ are considered. Optimality of the SSD for supervised classification ξ_l is evaluated by AER$_l$, $l = 1, 2$.

The values of AER$_l$ for $l = 1, 2$ are given in Table 2 for various values of ρ and α, that represent the range of spatial correlation between observations of spatial Gaussian process.

Analyzing the figures in Table 2 we can conclude that optimality of SSD depends an degree of population labels assymmetry in training sample, i.e. the minimum of proposed criterion is attained for symmetric SSD ξ_1 $(d_{12} = 0)$. The larger value of AER is obtained for ξ_2 with larger d_{12}.

The conclusions described above are valid for both levels of populations serepability ($\Delta = 0.2$ and $\Delta = 2.0$) and for various values of the range for spatial correlations.

Table 2. Values of AER_l, $l = 1, 2$ for $\Delta = 0.2$ and $\Delta = 2.0$ and $\pi_1 = \pi_2$

ρ	α	AER_1	AER_2	AER_1	AER_2
		$\Delta = 0.2$		$\Delta = 2.0$	
0.14	0.5	0.45954	0.45977	0.15497	0.15613
0.37	1.0	0.45111	0.45220	0.10962	0.11477
0.51	1.5	0.44275	0.44464	0.07493	0.08171
0.62	2.0	0.43514	0.43769	0.05123	0.05801
0.67	2.5	0.42822	0.43130	0.03521	0.04127
0.72	3.0	0.42186	0.42540	0.02434	0.02946

References

1. C. R. O. Lawoko, G. L. McLachlan, Discrimination with autocorrelated observations, *Pattern Recogn.*, **18**(2), pp. 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), pp. 367–376, 1980.

4. K. V. Mardia, Spatial discrimination and classification maps, *Commun. Stat.-Theor. M.*, **13**(18), pp. 2181–2197, 1974.

5. J. Šaltytė, K. Dučinskas, Comparison of ML and OLS estimators in discriminant analysis of spatially correlated observations, *Informatica*, **13**(2), pp. 297–238, 2002.

6. J. Šaltytė-Benth, K. Dučinskas, Linear discriminant analysis of multivariate spatial-temporal regressions, *Scand. J. Stat.*, **32**, pp. 281–294, 2005.

7. K. Dučinskas, Approximation of the expected error rate in classification of the Gaussian random field observations, *Statistics and Probability Letters*, **79**, pp. 138–144, 2009.

8. R. Christensen, *Advanced Linear Modelling*, 2nd ed., Springer-Verlag, New York, 2001.

9. J. R. Magnus, H. Neudecker, *Matrix Differential Calculus and Applications in Statistics and Econometrics*, Wiley, New York, 2002.

10. D. L. Zimmerman, Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction, *Environmetrics*, **17**, pp. 635–652, 2006.