EXPECTED ERROR RATES IN CLASSIFICATION OF GAUSSIAN CAR OBSERVATIONS

K. Ducinskas[1], L. Dreiziene[2]
[1,2]*Klaipeda University*
[2]*Vilnius University*
[1,2]*Klaipeda and* [2]*Vilnius, LITHUANIA*
e-mail: [1]`kestutis.ducinskas@ku.lt`, [2]`l.dreiziene@gmail.com`

Abstract

Given the neighbourhood structure, the problem of classifying a scalar Gaussian CAR observation into one of two populations specified by different parametric drifts is considered. This paper concerns with classification procedures associated with a parametric plug-in conditional Bayes rule (PBR) obtained by substituting the unknown parameters by their maximum likelihood (ML) estimators in the Bayes rule. For the particular prior distributions of unknown parameters the Bayesian estimators are used. The closed-form expression for the actual error rate associated with aforementioned classification rule and the approximation of the expected error rate (AER) associated with aforementioned PBR is derived. This is the extension of the previous one to the case of complete parametric uncertainty, i.e. when all drift and covariance function parameters are unknown. CAR observations sampled on regular 2-dimensional lattice with respect to the neighbourhood structure based on Euclidean distance between sites is used for simulation experiment.

1 Introduction

Suppose that model of observation $Z(s)$ in population Ω_j is

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

where $x(s)$ is a $q \times 1$ vector of non random regressors and β_j is the $q \times 1$ vector of parameters, $j = 1, 2$. The error term $\varepsilon(s)$ is generated by zero-mean CAR $\{\varepsilon(s) : s \in D\}$ with respect to the undirected graph (nodes with neighbourhood system) that will be described later. For given training sample, consider the problem of classification of the $Z_0 = Z(s_0)$ into one of two populations when $x'(s_0)\beta_1 \neq x'(s_0)\beta_2, s_0 \in D$.

Suppose that the set of spatial locations $\{s_i \in D;\ i = 1, ..., n\}$ forming regular or irregular lattice where training sample $T' = (Z(s_1), ..., Z(s_n))$ is taken, and call it the set of training locations. Indexing spatial locations by integers $0, 1, .., n$, denote lattices by $S_n = 1, ..., n$ and $S_n^0 = S_n \cup \{0\}$. Let $Z(s_i) = Z_i, i = 0, ..., n$ then training sample is defined by $T = (Z_1, ..., Z_n)'$ and $T_0 = (Z_0, Z_1, ..., Z_n)'$. Assume that S_n is partitioned into union of two disjoint subsets, i.e. $S_n = S^{(1)} \cup S^{(2)}$, where $S^{(j)}$ is the subset of S_n that contains n_j locations of feature observations from $\Omega_j, j = 1, 2$.

Assume that lattice S_n^0 is endowed with a neighbourhood system $N_0 = \{N_k : k = 0, 1, ..., n\}$ and lattice S_n is endowed with a neighbourhood system $N = \{N_k : k =$

$1, ..., n\}$ where N_k denotes the collection of sites that are neighbours of site, s_k. So we formed undirected graph G^0 specified by lattice S_n^0 and neigbourhood system N^0 and its subgraph G specified by lattice S_n and neigbourhood system N. Define spatial weight $w_{kl} > 0$ as a measure of similarity between sites k and l, and put $w_{kl} = w_{lk}$ and $h_k = \sum_{l \in N_k} w_{kl}$.

The $n \times 2q$ design matrix of training sample T is denoted by X. Then training sample T would be modeled by the joint distribution (Oliveira and Ferreira, 2011)

$$T \sim N_n(X\beta, \sigma^2 V(\alpha)) \tag{2}$$

where

$$V(\alpha) = (I_n + \alpha H)^{-1} \tag{3}$$

and $\sigma > 0$ is a scale parameter and $\alpha \geq 0$ is a spatial dependence parameter and $n \times n$ matrix $H = (h_{kl} : k, l = 1, ..., n)$ is given by

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

In the following set $\Sigma = \sigma^2 V(\alpha)$. Then the variance-covariance matrix of vector T_0 is $var(T_0) = \sigma^2(I_{n+1} + \alpha H^0)^{-1}$ where $H^0 = (h_{kl}, k, l = 0, 1, ..., n)$.

This is the case, when spatial classified training data are collected at fixed locations. Let t denote the realization of T. Set $k = 1 + \alpha h_0$.

Since Z_0 follows model specified in (1)-(3), the conditional distribution of Z_0 given $T = t, \Omega_j$ is Gaussian with mean

$$\mu_{jt}^0 = E(Z_0 | T = t; \Omega_j) = x_0' \beta_j + \alpha_0'(t - X\beta), j = 1, 2 \tag{4}$$

and variance

$$\sigma_0^2 = var(Z_0 | T = t; \Omega_j) = \sigma^2 / k, \tag{5}$$

where $\alpha_0' = \alpha w_0' / k$ and $w_0' = (w_{01}, ..., w_{0n})$.

Under the assumption of complete parametric certainty of populations, the Bayes discriminant fuction (BDF) minimizing the overall misclassification probability (OMP) is specified by (McLachlan, 2004)

$$W_t(Z_0, \Psi) = \Big(Z_0 - \frac{1}{2}(\mu_{1t}^0 + \mu_{2t}^0)\Big)(\mu_{1t}^0 - \mu_{2t}^0)/\sigma_0^2 + \gamma, \tag{6}$$

where $\gamma = ln(\pi_1/\pi_2)$ and $\Psi = (\beta', \theta')'$, $\theta' = (\alpha, \sigma^2)$. Here $\pi_1, \pi_2 (\pi_1 + \pi_2 = 1)$ are respectively prior probabilities of the populations Ω_1 and Ω_2, for observation at location s_0.

The squared Mahalanobis distance between conditional distributions of Z_0 given $T = t$ is specified by

$$\Delta_0^2 = (\mu_{1t}^0 - \mu_{2t}^0)^2 / \sigma_0^2 = x_0'(\beta_1 - \beta_2) k / \sigma^2.$$

Denote by $P(\Psi)$ the OMP for BDF defined in (6). Then using the properties of normal distribution we obtain

$$P(\Psi) = \sum_{j=1}^{2} \Big(\pi_j \Phi(-\Delta_0/2 + (-1)^j \gamma/\Delta_0) \Big), \tag{7}$$

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

2 Error rates for plug-in BDF

As it follows we shall write hat above parameters for their estimators based on realization of training sample $T = t$. Put $\hat{\Psi} = (\hat{\beta}', \hat{\theta}')'$ and $\hat{\theta} = (\hat{\alpha}, \hat{\sigma}^2)$, $\hat{\alpha}_0 = \hat{\alpha} w_0/(1 + \hat{\alpha} h_0)$. Then by using (4), (5) we get the estimators of conditional mean and conditional variance

$$\hat{\mu}_{jt}^0 = x_0' \hat{\beta}_j + \hat{\alpha}_0'(t - X\beta), j = 1, 2$$
$$\hat{\sigma}_0^2 = \hat{\sigma}^2/(1 + \hat{\alpha} h_0).$$

Then replacing parameters with their estimators in (6) we form the plug-in BDF

$$W_t(Z_0; \hat{\Psi}) = \Big(Z_0 - \hat{\alpha}_0'(t - X\hat{\beta}) - \frac{1}{2} x_0' I^+ \hat{\beta} \Big) (x_0' I^- \hat{\beta})/\hat{\sigma}_2 + \gamma \tag{8}$$

with $I^+ = (I_q, I_q)$ and $I^- = (I_q, -I_q)$, where I_q denotes the identity matrix of order q.

Lemma 1. *The actual error rate for $W_t(Z_0; \hat{\Psi})$ specified in (8) is*

$$P(\hat{\Psi}) = \sum_{j=1}^{2} \Big(\pi_j \Phi(\hat{Q}_j) \Big). \tag{9}$$

Here

$$\hat{Q}_j = (-1)^j \Big((a_j - \hat{b}) sgn(x_0' I^- \hat{\beta})/\sigma_0 + \hat{\sigma}_0^2 \gamma/(\sigma_0 |x_0' I^- \hat{\beta}|) \Big),$$

where for $j = 1, 2$

$$a_j = x_0' \beta_j + \alpha_0'(t - X\beta) \quad and \quad \hat{b} = \hat{\alpha}_0'(t - X\hat{\beta}) + x_0' I^+ \hat{\beta}/2.$$

Definition 1. The expectation of the actual risk with respect to the distribution of T is called the expected error rate (EER) and is designated as $E_T\Big(P(\hat{\Psi})\Big)$.

We will use the ML estimators of parameters based on training sample. The asymptotic properties of ML estimators established by Mardia and Marshall (1984) under increasing domain asymptotic framework and subject to some regularity conditions are essentially exploited. Hence, the ML estimator $\hat{\Psi}$ is weakly consistent and asymptotically Gaussian, i.e.

$$\hat{\Psi} \sim AN(\Psi, J^{-1}),$$

here the expected information matrix is given by $J = J_\beta \oplus J_\theta$, where $J_\beta = X'\Sigma^{-1}X$ and $(i,j)-th$ element of J_θ is $tr(\Sigma^{-1}\Sigma_i\Sigma^{-1}\Sigma_j)/2$.

Henceforth, denote by (MM) conditions the regularity conditions of Theorem 1 from Mardia and Marshall (1984) and make the following assumption:
(A1) training sample T and estimator $\hat{\theta}$ are statistically independent.

Theorem 1. *Suppose that observation Z_0 to be classified by plug-in PDF and let conditions (MM) and assumption (A1) hold. Then the approximation of EER is*

$$AER = R(\Psi) + \pi_1^*\varphi(-\Delta_0/2 - \gamma/\Delta_0)\Delta_0(K_\beta + K_\alpha + \gamma^2 K_\theta/\Delta_0^2)/2. \quad (10)$$

Here

$$K_\beta = \Lambda' V_\beta \Lambda k,$$
$$\Lambda' = \alpha w_0' X/k - x_0'(I^+/2 + \gamma I^-/\Delta_0^2),$$
$$V_\beta = (X'(I + \alpha H)X)^{-1},$$
$$K_\alpha = w_0'(I + \alpha H)^{-1} w_0 J_{11}^{-1}/k^3,$$
$$K_\theta = \nu' J_\theta^{-1}\nu/k2 \quad where \quad \nu' = (h_0, -1/\sigma_0^2).$$

3 Numerical experiment

In order to demonstrate the results of Theorem 1 simulation experiment was carried out. CAR observations were sampled on regular 2-dimensional lattice with respect to the neighbourhood structure based on Euclidean distance between sites. AER and $P(\hat{\Psi})$ were calculated for different parametric structures. The results of the numerical analysis show that proposed error rates and its approximation formulas could be used as performance evaluation of classification procedures.

References

[1] Ducinskas K., Borisenko I., Simkiene I. (2013). Statistical classification of Gaussian spatial data generated by conditional autoregressive model. *Computational Science and Techniques*. Vol. **1(2)**, pp. 69-79.

[2] Mardia K.V., Marshall R.J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*. Vol. **71**, pp. 135-146.

[3] McLachlan G.J. (2004). *Discriminant analysis and statistical pattern recognition*. Wiley, New York.

[4] de Oliveira V., Ferreira M.A.R. (2011). Maximum likelihood and restricted maximum likelihood estimation for class of Gaussian Markov random fields. *Metrika*. Vol. **74(2)**, pp. 167-183.