Liet. matem. rink., T. 40, spec. nr., 304–310

Discrimination of CAR models

Jūratė ŠALTYTĖ (KU)
e-mail: jsaltyte@gmf.ku.lt

1. Introduction

Let $\{Z(s) : s \in D \subset \mathbf{Z}^2\}$ be a random field on a lattice D. When building models for data on a specific lattice, no possibility is given to a realization occurring on non-lattice locations. Such spatial models on lattices are analogues of time-series models. Here it is more appropriate to think of the (finite) data set of Z's as a part of an increasing lattice whose sites tend out to infinity in at least one direction of Euclidean space. The concept of lattice is closely related to the concepts of the neighbors and neighborhood. According to Cressie (1993), a site t is defined to be a *neighbor* of site s if the conditional distribution of $Z(s)$, given all other site values, depends functionally on $z(t)$, for $t \neq s$. Also define

$$\mathbf{N}_s \equiv \{t : t \text{ is a neighbour of } s\} \tag{1}$$

to be the *neighborhood* set of a site s.

The notion that data close together in space are likely to be correlated is natural. And the most obvious departure from the independence model is to assume that considered spatial process is Markov random field.

DEFINITION 1. Any probability measure whose conditional distributions define a neighborhood structure $\{\mathbf{N}_s : s = 1, \ldots, n\}$ through (1) is defined to be a *Markov random field* [1].

The most useful class of Markov random field models for continuous data is the class of so called *conditional Gaussian models.*

Assume that the model of $Z(s)$ in population Ω_l is

$$Z_l(s) = x_l^T(s)\beta_l + \varepsilon_l(s),$$

where $x_l^T(s) = (x_l^1(s), \ldots, x_l^q(s))$ are $q \times 1$ vectors of nonrandom regressors and $\beta_l = (\beta_l^1, \ldots, \beta_l^q)^T \in B$, $l = 1, 2$, are parameter vectors, B being an open subset of R^q. Suppose, that $\{\varepsilon_l(s) : s \in D \subset \mathbf{Z}^2\}$ is a univariate zero-mean random Gaussian field with spatial covariance defined by a parametric model $\operatorname{cov}\{\varepsilon_l(s), \varepsilon_l(t)\} = \sigma(s - t; \theta_l)$ for all $s, t \in D$, where $\theta_l \in \Theta$ is a $p \times 1$ parameter vector, Θ being an open subset of R^p, $l = 1, 2$; this means that the considered random field is *intrinsically stationary* field. (Intrinsic stationarity of $\{\varepsilon_l(s) : s \in D \subset R^2\}$ means that $E\{(\varepsilon_l(t) - \varepsilon_l(s))^2\}$

depends only on $(s-t)$, for $l=1,2$ and $s,t \in D$.) The attention will be restricted to the homoscedastic models, i.e., $\sigma_s(0;\theta)=\sigma_s^2$, for all $\theta \in \Theta$. Suppose that σ_s^2 are known and let $|x(s)| \leqslant M < \infty$, for $s \in D$.

If $p_l(z(s)/\{z(t),\ t \neq s\})$ denotes the conditional probability density function of $Z(s)$ for Ω_l, then

$$p_l(z(s)/\{z(t), t \neq s\}) = \frac{1}{\sqrt{2\pi\tau_{ls}^2}} \exp\left[-\left\{z(s)-\lambda_{ls}\left(\{z(t): t \neq s\}\right)\right\}^2/2\tau_{ls}^2\right], \quad (2)$$

where $\lambda_{ls}(\cdot)$ and τ_{ls}^2 are its conditional mean and variance at location s, respectively, $l=1,2$, $s=1,\ldots,n$.

Suppose that considered model satisfies a condition of "pairwise-only dependence" between sites, i.e.,

$$\lambda_{ls}(\{z(t): t \neq s\}) = \mu_{ls} + \sum_t q_{st}^l (z(t)-\mu_{lt}),$$

where $q_{st}^l = q_{ts}^l$, $q_{ss}^l = 0$, and $q_{su}^l = 0$ for $u \notin \mathbf{N}_s$ [2]. Here q_{st}^l is a function of the unknown parameter α, i.e., $q_{st}^l(\alpha)$, but for notational convenience suppressed notation $q_{st}^l = q_{st}^l(\alpha)$ will be used.

On the base of factorization theorem [2], it is not difficult to show that

$$\mathbf{Z} \sim N(\mu_l, C_l^{-1}M_l) \quad (3)$$

provided $C_l = I - Q_l$ is invertible and $C_l^{-1}M_l$ is symmetric and positive-definite matrix describing spatial dependence between observations at different locations; here $\mathbf{Z} \equiv (Z_{l1},..,Z_{ln})^T$, $\mu_l = (\mu_{l1},\ldots,\mu_{ln})^T$, $\mu_{ls} = x_l^T(s)\cdot\beta_l$ $(s=1,\ldots,n)$, Q_l is an $n \times n$ matrix, whose (s,t)th element is q_{st}^l, and $M \equiv \operatorname{diag}(\tau_{l1}^2,\ldots,\tau_{ln}^2)$ is an $n \times n$ diagonal matrix.

In order the asumption about the homoscedascity of model to be valid, the condition $(C_1^{-1}M_1)_{ss} = (C_2^{-1}M_2)_{ss}$, where $(C_l^{-1}M_l)_{ss}$ means a diagonal element of the matrix $(C_l^{-1}M_l)$, $l=1,2$, for the unconditional variance at location s $\sigma_s^2 = (C_l^{-1}M_l)_{ss}$ must be satisfied. The expression (3) is called a *conditional Gaussian autoregression* (CAR) *model*.

Let $Z(r)$ be an observation at $r \in D$ from one of the two populations Ω_1 and Ω_2. Under the assumption that the populations are completely specified and for known prior probabilities of populations π_{1r}, π_{2r} $(\pi_{1r}+\pi_{2r}=1)$, the Bayesian classification rule (BCR) $d_B(\cdot)$ minimizing the probability of misclassification (PMC) is

$$d_B(z(r)) = \arg\max_{\{l=1,2\}} \pi_l p_l(z(r)), \quad (4)$$

where $z(r)$ is the observed value of $Z(r)$. Denote by P_B^r the PMC of BCR. Usually P_B^r is called *Bayes error rate* (see, e.g., Hand 1997, ch 7.).

In practice, however, complete description of classes most often is not possible. Thus, the probabilistic characteristics of each class must be estimated from training samples. Most widely used statistical approach for the estimation of unknown mean and variance, assuming the spatial dependence parameter to be known, is the maximum likelihood (ML) method. So, the plug-in rule is usually formed with using ML estimates. Such kind of plug-in rules are extensively applied in practice. But the spatial dependence parameter need to be estimated also. In ML function this parameter appears in normalizing constant and analytical expression for its estimator in general case is not available. R.K. Pace and D. Zou (2000) consider special case of correlation matrix and particular neighbourhood structure and present analytical expression of estimation of spatial dependence parameter as the real solution of cubic equation. However, as Cressie (1993) designates, the method of ML for estimation of lattice-model parameters is no longer automatically the method of choice. In the general case for conditionally specified models, a number of modified likelihood-based estimation procedures have been proposed. One of these procedures is pseudo maximum likelihood (PML) estimation method introduced by Besag (see, e.g., [1]). The PML estimator would be consistent as the training sample size is increased. When the PML estimation procedure is used, the efficiency loss can occur because the maximization of the objective function not always yield functions of a minimal sufficient statistic, unlike for the ML estimator. On the other hand, working with exact likelihood's unwieldy normalizing constant is avoided. It is useful technique for the estimation of the unknown spatial dependence parameter.

Let $T_l = \{Z_{l1}, \ldots, Z_{lN_l}\}$ be the training samples, where $Z_{lk} = Z(s_k^l)$ denotes the kth observation from Ω_l, $l = 1, 2$. In this paper it is supposed that α, the parameter of C_l ($l = 1, 2$), and σ_s^2 are known, and ML estimators of β_l, $l = 1, 2$, based on T_l are used.

Put $T = \{T_1, T_2\}$, $N = N_1 + N_2$. Let $\widehat{\beta}_1, \widehat{\beta}_2$ be the estimators of β_1, β_2, respectively, based on T, and let $\widehat{\mu}_l(r) = x_l^T(r)\widehat{\beta}_l$. The plug-in rule $d_B(z(r), \widehat{\mu}_{1r}, \widehat{\mu}_{2r}, \sigma_r^2)$ is obtained by replacing the parameters in (4) with their estimators, i.e.,

$$d_B\big(z(r), \widehat{\mu}_{1r}, \widehat{\mu}_{2r}, \sigma_r^2\big) = \arg \max_{\{l=1,2\}} \pi_l p_l\big(z(r); \widehat{\mu}_{lr}, \sigma_r^2\big).$$

Then the corresponding discriminant function W_r also known simply as the sample linear discriminant function (see McLachlan, 1974) is defined as

$$W_r = \left(z(r) - \frac{1}{2}(\widehat{\mu}_{1r} + \widehat{\mu}_{2r})\right)(\widehat{\mu}_{1r} - \widehat{\mu}_{2r})/\sigma_r^2 + \gamma_r.$$

DEFINITION 2. The *actual error rate* for $d_B\left(z(r), \widehat{\mu}_1, \widehat{\mu}_2, \sigma^2\right)$ is defined as

$$P^r(\widehat{\mu}_1, \widehat{\mu}_2, \sigma^2) \triangleq \sum_{l=1}^{2} \pi_l \int \left(1 - \delta\big(l, d_B\big(z(r), \widehat{\mu}_{1r}, \widehat{\mu}_{2r}, \sigma_r^2\big)\big) p_l\big(z(r); \mu_{lr}, \sigma_r^2\big)\right) \mathrm{d}z(r).$$

DEFINITION 3. The expectation of the actual error rate with respect to the distribution of T designated as $E_T\{P^r(\widehat{\mu}_1, \widehat{\mu}_2, \sigma^2)\}$ is called the *expected error rate* (EER) for the $d_B\big(z(r), \widehat{\mu}_{1r}, \widehat{\mu}_{2r}, \sigma_r^2\big)$.

Asymptotic approximations and asymptotic expansions for EER in case of independent observations were considered by many authors (see, e.g., Okamoto 1963, Dučinskas 1997). Mardia (1984) considered similar problem of classifying spatially distributed Gaussian observations with constant means. But he did not analyze EER and probabilities of misclassification. In this paper the asymptotic approximation for the EER of classifying an observation from CAR model with different means depending on the locations is obtained. ML estimators of means are used in the plug-in version of Bayesian classification rule. A comparison for the accuracy of obtained asymptotic approximation with Monte Carlo simulations when training sample sizes are small is made.

2. Approximation for the EER

The attention will be restricted to the case when the effect of cross-correlations between observations from different populations is negligible. In this paper it is supposed, that if $Z(s)$ is from Ω_1 and $Z(t)$ is from Ω_2, then $\operatorname{cov}\big(Z(s), Z(t)\big) = 0$.

The expectation vector and the covariance matrix of $T_l^V = (Z_{l1}, \ldots, Z_{lN_l})^T$ are $\mu_l = (\mu_{l1}, \ldots, \mu_{lN_l})^T$ and $\Sigma_l = C_l^{-1} M_l$, respectively, where C_l is the matrix of order $N_l \times N_l$, whose (s,t)-th element is $c_{st}^l(h_l) = c^l(s_l - t_l)$, $s, t = 1, \ldots, N_l$, and $M \equiv \operatorname{diag}(\tau_{l1}^2, \ldots, \tau_{lN_l}^2)$, $l = 1, 2$. Let X^l be an $N_l \times q$ regressor matrix with ith column $(x_{1i}^l, \ldots, x_{N_l i}^l)'$, where $x_{ki}^l = x_i(s_k^l)$, $i = 1, \ldots, q$, $k = 1, \ldots, N_l$, $l = 1, 2$.

Lemma. *For $l = 1, 2$, ML estimators of β_1, β_2 based on T are*

$$\widehat{\beta}_l = \left(X_l^T M_l^{-1} C_l X_l\right)^{-1} X_l^T M_l^{-1} C_l T_l^V, \quad l = 1, 2.$$

Proof. The log-likelihood of T_l is

$$\ln L_l = \text{const} + \frac{1}{2} \ln |C_l| - \frac{1}{2} \ln |M_l| - \frac{1}{2} (T_l^V - X_l \beta_l)^T M_l^{-1} C_l (T_l^V - X_l \beta_l),$$

$l = 1, 2$. Solving the equations $\frac{\partial \ln L_l}{\partial \beta_l} = 0$, $l = 1, 2$, we complete the proof of Lemma.

It is obvious, that for any $r \in D$ $\widehat{\mu}_{lr} = x_l^T(r) \widehat{\beta}_l$ for finite N have known exact distribution of the form

$$\widehat{\mu}_{lr} \sim N(\mu_{lr}, a_l^r), \tag{5}$$

where $a_l^r = x_l^T(r)(X_l^T M_l^{-1} C_l X_l)^{-1} x_l(r)$.

To give an approximation the assumption $\operatorname{rank}(X^l) = q$, $l = 1, 2$, is needed.

Put $\gamma_r = \ln \frac{\pi_{1r}}{\pi_{2r}}$, $\Delta \widehat{\mu}_{lr} = \widehat{\mu}_{lr} - \mu_{lr}$, $\Delta_r^2 = (\mu_{1r} - \mu_{2r})^2 / \sigma^2$. Let $\Phi(\cdot)$ and $\varphi(\cdot)$ denote standard normal distribution and density functions, respectively.

Then the actual risk for $d_B(x_r, \widehat{\mu}_{1r}, \widehat{\mu}_{2r}, \sigma^2)$ (see McLaclan (1974)) is

$$P^r(\widehat{\mu}_{1r}, \widehat{\mu}_{2r}, \sigma^2) = \pi_{1r}\Phi\left(-\frac{\left(\mu_{1r} - \frac{1}{2}\left(\widehat{\mu}_{1r} + \widehat{\mu}_{2r}\right)\right)\left(\widehat{\mu}_{1r} - \widehat{\mu}_{2r}\right)/\sigma^2 + \gamma_r}{\sqrt{\left(\widehat{\mu}_{1r} - \widehat{\mu}_{2r}\right)^2/\sigma^2}}\right)$$
$$+ \pi_{2r}\Phi\left(\frac{\left(\mu_{2r} - \frac{1}{2}\left(\widehat{\mu}_{1r} + \widehat{\mu}_{2r}\right)\right)\left(\widehat{\mu}_{1r} - \widehat{\mu}_{2r}\right)/\sigma^2 + \gamma_r}{\sqrt{\left(\widehat{\mu}_{1r} - \widehat{\mu}_{2r}\right)^2/\sigma^2}}\right).$$

For notational convenience, we shall henceforth omit the subscript or superscript r on $\mu_{lr}, P^r(\ ), a_l^r \ldots$.

Let $P_i^{(1)} = \partial P(\ )/\partial\widehat{\mu}_i$, $P_{ij}^{(2)} = \partial^2 P(\ )/\partial\widehat{\mu}_i\partial\widehat{\mu}_j$, $P_{ijk}^{(3)} = \partial^3 P(\ )/\partial\widehat{\mu}_i\partial\widehat{\mu}_j\partial\widehat{\mu}_k$, $P_{ijkl}^{(4)} = \partial^4 P(\ )/\partial\widehat{\mu}_i\partial\widehat{\mu}_j\partial\widehat{\mu}_k\partial\widehat{\mu}_l$ $(i,j,k,l = 1,2)$ be the partial derivatives up to fourth order of $P\left(\widehat{\mu}_1, \widehat{\mu}_2, \sigma^2\right)$ evaluated at $\widehat{\mu}_1 = \mu_1, \widehat{\mu}_2 = \mu_2$.

Theorem. *Suppose that stated assumption holds. Then the approximation of EER for the* $d_B\left(z(r), \widehat{\mu}_{1r}, \widehat{\mu}_{2r}, \sigma^2\right)$ *is*

$$E_T\left\{P\left(\widehat{\mu}_1, \widehat{\mu}_2, \sigma^2\right)\right\} \simeq P_B + \frac{\Delta}{8}\varphi\left(\frac{\Delta}{2}\right)$$
$$\times\left(a_1^2 + a_2^2 + \frac{\Delta - 12}{128}\left(a_1^2 + a_2^2\right) + \frac{\Delta - 12}{16}a_1a_2\right).$$

Proof. Since $P(\widehat{\mu}_1, \widehat{\mu}_2, \sigma^2)$ is invariant under linear transformations of data we use the convenient canonical form of $\sigma^2 = 1$ and $\mu_1 = \Delta$, $\mu_2 = 0$ (see Dunn (1971)). Expand $P\left(\widehat{\mu}_1, \widehat{\mu}_2, \sigma^2\right)$ in Taylor series about the point $\widehat{\mu}_1 = \Delta$, $\widehat{\mu}_2 = 0$. Taking the expectation with respect to the distribution of T and dropping the fifth order terms we have

$$E_T\left(P\left(\widehat{\mu}_1, \widehat{\mu}_2, \sigma^2\right)\right) \simeq P_B + \sum_{l=1}^{2} P_l^{(1)} E_T\left\{\Delta\widehat{\mu}_l\right\} + \frac{1}{2}\sum_{l,k=1}^{2} P_{kl}^{(2)} E_T\left\{\Delta\widehat{\mu}_l\Delta\widehat{\mu}_k\right\}$$
$$+ \frac{1}{3!}\sum_{k,l,m=1}^{2} P_{klm}^{(3)} E_T\left\{\Delta\widehat{\mu}_k\Delta\widehat{\mu}_l\Delta\widehat{\mu}_m\right\}$$
$$+ \frac{1}{4!}\sum_{k,l,m,n=1}^{2} P_{klmn}^{(4)} E_T\left\{\Delta\widehat{\mu}_k\Delta\widehat{\mu}_l\Delta\widehat{\mu}_m\Delta\widehat{\mu}_n\right\}. \qquad (6)$$

Then from (5), the following hold

$$E_T\left\{\left(\Delta\widehat{\mu}_l\right)^2\right\} = a_l, \quad E_T\left\{\Delta\widehat{\mu}_1\Delta\widehat{\mu}_2\right\} = 0,$$
$$E_T\left\{\left(\Delta\widehat{\mu}_l\right)^3\right\} = 0, \quad E_T\left\{\left(\Delta\widehat{\mu}_l\right)^4\right\} = 3\sigma^4 a_l^2, \qquad (7)$$

$l = 1, 2$. Since $P(\widehat{\mu}_1, \widehat{\mu}_2, \sigma^2)$ is minimized at $\widehat{\mu}_l = \mu_l$, then

$$P_l^{(1)} = 0, \quad l = 1, 2. \tag{8}$$

Then putting (7), (8) and higher order derivatives of $P(\widehat{\mu}_1, \widehat{\mu}_2, \sigma^2)$ into (6) we complete the proof of the stated theorem.

3. Example

As an example the integer regular 2-dimensional lattice is considered and the second-order neighborhood scheme (Fig. 1.) for training sample used. It is assumed that there are 4 spatially symmetric observations in training sample and that the spatial structure is of the same form for each class. The comparison of obtained approximation of EER (denoted by P_A) with Monte Carlo simulations (denoted by P_{MC}) for one special case, with $\pi_1 = 0.2$ is presented. Suppose, that regressor is of the form $x(s) = 1/(|s|^2 + 2.5)$ and spatial dependece is described by $q_{st}(\alpha) = \begin{cases} \alpha h_{st}^{-1}, & \text{if } s \neq t \\ 0, & \text{if } s = t \end{cases}$.

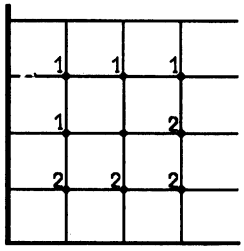

Fig. 1. Second-order neighborhood scheme.

Table 1
Comparison of approximation with simulation ($\alpha = \frac{1}{2}$)

Δ	P_A	P_{MC}	P_A/P_{MC}
1.0	0.32796	0.43852	0.74788
1.4	0.26678	0.38447	0.69389
1.8	0.21201	0.34784	0.60950
2.2	0.16439	0.27031	0.72005
2.6	0.12422	0.22831	0.54407
3.0	0.09133	0.22274	0.41005

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

References

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

[2] J.E. Besag, Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society* B, **36**, 192–225 (1974).

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

[4] R.K. Pace, D. Zou, Closed-form maximum likelihood estimates of nearest neighbor spatial dependence, *Geographical Analysis*, **32**(2), 140–172 (2000).

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

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

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

[9] O.J. Dunn, Some expected values for probabilities of correct classification in discriminant analysis, *Technometrics*, **13**, 345–353 (1971).

Sąlyginių autoregresinių laukų diskriminantinė analizė

J. Šaltytė

Straipsnyje nagrinėjamas uždavinys apie objektų iš srities $D \subset R^2$ klasifikavimą pagal sąlyginių Gauso autoregresinių laukų stebėjimus. Pateiktas asimptotinis klaidos tikimybės skleidinys, kuris lyginamas su Monte Karlo metodu gauta klaidos tikimybe.