Liet. matem. rink., T. 43, spec. nr., 554–559

Linear discriminant analysis of interclass correlated spatio-temporal data

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

1. Model

Spatial-temporal data can be considered as a realization of a stochastic process (random field) $\{Z(\mathbf{s};t):\ \mathbf{s}\in D,\ t\in[0,\infty)\}$, where $\mathbf{s}$ and t define spatial and temporal coordinates, respectively.

Suppose that the model of $Z(\mathbf{s};t)$ in population Ω_l is

$$Z(\mathbf{s};t)=B_l^T x(\mathbf{s};t)+\varepsilon_l(\mathbf{s};t),$$

where $x(\mathbf{s};t)=(x_1(\mathbf{s};t),\ldots,x_q(\mathbf{s};t))^T$ is a $q\times 1$ vector of nonrandom regressors and B_l is the unknown parameter matrix of order $q\times p$, $l=1,2$. Assume that $\{\varepsilon_l(\mathbf{s};t):\ \mathbf{s}\in D\subset R^2,\ t\in[0,\infty)\}$ is a p-variate zero-mean intrinsically stationary spatial-temporal Gaussian random field with stationary (in space and time) spatial-temporal covariance function defined by model

$$cov\{\varepsilon_l(\mathbf{s};t),\varepsilon_l(\mathbf{u};v)\}=\sigma(\mathbf{s}-\mathbf{u},t-v)$$

for all $\mathbf{s},\mathbf{u}\in D, t,v>0,\ l=1,2$. We restrict our attention to the homoscedastic models, i.e., $\sigma(\mathbf{0},0)=\Sigma$. Then, in Ω_l the mean function at location $\mathbf{s}$ and time moment t is

$$\mu_l(\mathbf{s};t)=B_l^T x(\mathbf{s};t)$$

and the spatial-temporal covariance function is

$$cov\{\varepsilon_l(\mathbf{s};t),\varepsilon_l(\mathbf{u};v)\}=c(\mathbf{s}-\mathbf{u},t-v)\,\Sigma,$$

where $c(\mathbf{s}-\mathbf{u},t-v)$ is the spatial-temporal correlation function, $l=1,2$. It is assumed that the function $c(\mathbf{s}-\mathbf{u},t-v)$ is positive definite [2].

Assume that, for all $\mathbf{s},\mathbf{u}\in D, t,v>0, \mathbf{s}\neq\mathbf{u}, t\neq v$,

$$cov\{\varepsilon_1(\mathbf{s};t),\varepsilon_2(\mathbf{u};v)\}=c_{12}(\mathbf{s}-\mathbf{u},t-v)\,\Sigma, \tag{1}$$

where $c_{12}(\cdot,\cdot)$ is the interclass spatial correliation function. The case when there is no interclass spatial correlation was considered by Šaltytė and Dučinskas [4].

Consider the problem of classification of the observation $Z^0 = Z(\mathbf{s}_0, t_0)$, with $\mathbf{s}_0 \in D_0 \subset D$, $t_0 > 0$, into one of two populations specified above. 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 BCR $d_B(\cdot)$ minimizing the probability of misclassification (PMC) is

$$d_B\left(z^0\right) = \arg \max_{\{l=1,2\}} \pi_l p_l\left(z^0\right), \tag{2}$$

where z^0 is the realisation of Z^0 and

$$p_l\left(z^0\right) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(z^0 - \mu_l^0\right)^T \Sigma^{-1}\left(z^0 - \mu_l^0\right)\right)$$

is a probability density function (p.d.f.) of Z^0 in Ω_l, $l = 1, 2$. Here $\mu_l^0 = \mu(\mathbf{s}_0; t_0) = B_l^T x^0$ with $x^0 = x(\mathbf{s}_0; t_0)$, $l = 1, 2$.

Denote by P_B the PMC of BCR, usually called Bayes error rate.

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 T_1 and T_2 taken separately from Ω_1 and Ω_2, respectively. When estimators of unknown parameters are used, the plug-in version of BCR is obtained.

Suppose that the spatial-temporal random field is observed at N_l spatial-temporal coordinates in region $D_1 \subset D$, i.e., we observe the training sample $T^T = (T_1^T, T_2^T)$, where T_l is the $N_l \times p$ matrix of N_l observations of p-variate $Z(\mathbf{s}, t)$ from Ω_l, $l = 1, 2$. Then T is the $N \times p$ matrix, where $N = N_1 + N_2$.

Assume that D_1 is beyond the zone of influence of D_0. Then Z^0 is independent on T.

Let $\widehat{B}_1$, $\widehat{B}_2$ and $\widehat{\Sigma}$ be the ML estimators of B_1, B_2 and Σ, respectively, based on T, and let $\hat{\mu}_l(\mathbf{s}; t) = \widehat{B}_l^T x(\mathbf{s}; t)$, $l = 1, 2$. Put $\Psi = (\mu_1^0, \mu_2^0, \Sigma)$ and $\widehat{\Psi} = (\hat{\mu}_1^0, \hat{\mu}_2^0, \widehat{\Sigma})$.

The plug-in rule $d_B(z^0; \widehat{\Psi})$ is obtained by replacing the parameters in (2) with their estimators. Then the corresponding sample LDF is defined as

$$W\left(z^0; \widehat{\Psi}\right) = \left(z^0 - \frac{1}{2}\left(\hat{\mu}_1^0 + \hat{\mu}_2^0\right)\right)^T \widehat{\Sigma}^{-1}\left(\hat{\mu}_1^0 - \hat{\mu}_2^0\right) + \gamma,$$

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

DEFINITION 1. The actual error rate for $d_B(z^0; \widehat{\Psi})$ is defined as

$$P\left(\widehat{\Psi}\right) = \sum_{l=1}^{2} \pi_l \int \left(1 - \delta\left(l, d_B(z^0; \widehat{\Psi})\right) p_l(z^0; \Psi)\right) \mathrm{d}z^0,$$

where $\delta(\cdot, \cdot)$ is the Kronecker delta.

In the considered case the actual error rate for $d_B(z^0; \widehat{\Psi})$ can be rewritten as

$$P(\widehat{\Psi}) = \sum_{l=1}^{2} \pi_l^0 \Phi\left((-1)^l \frac{\left(\mu_l^0 - \frac{1}{2}\left(\hat{\mu}_1^0 + \hat{\mu}_2^0\right)\right)^T \widehat{\Sigma}^{-1}\left(\hat{\mu}_1^0 - \hat{\mu}_2^0\right) + \gamma}{\sqrt{\left(\hat{\mu}_1^0 - \hat{\mu}_2^0\right)^T \widehat{\Sigma}^{-1} \Sigma \widehat{\Sigma}^{-1}\left(\hat{\mu}_1^0 - \hat{\mu}_2^0\right)}}\right),$$

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, designated as $E_T\{P(\widehat{\Psi})\}$, is called the expected error rate (EER) for the $d_B(z^0;\widehat{\Psi})$.

2. Results

Let X be an $N\times q$ regressor matrix of T. Denote by $C=\begin{pmatrix} C_1 & C_{12} \\ C_{21} & C_2 \end{pmatrix}$ the $N\times N$ spatial-temporal correlation matrix of the joint training sample T. Assume that the mathematical model of T is

$$T = XB + E,$$

where $X = X_1 \oplus X_2$, $B = \left(B_1^T, B_2^T\right)^T$ and $E \sim N_{N\times p}(O, C\otimes\Sigma)$.

Suppose that $2q \times 2q$ matrix $D = \left(X^T C^{-1} X\right)^{-1}$ is partitioned as $D = \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix}$, where D_{11}, D_{12}, D_{21} and D_{22} are $q\times q$ submatrices. Set $D_l = (D_{l1}, D_{l2})$, $l = 1, 2$.

Lemma. *For $l = 1, 2$, ML estimators of B_l and Σ based on T are*

$$\widehat{B}_l = D_l X^T C^{-1} T \tag{3}$$

and

$$\widehat{\Sigma} = \frac{1}{N}(T - X\widehat{B})^T C^{-1}(T - X\widehat{B}), \tag{4}$$

where

$$\widehat{B} = \left(X^T C^{-1} X\right)^{-1} X^T C^{-1} T. \tag{5}$$

Proof. Since $T \sim N_{N\times p}(XB, C\otimes\Sigma)$, the log-likelihood of T is

$$\ln L = const - \frac{1}{2}\left(p\ln|C| + N\ln|\Sigma|\right) - \frac{1}{2} tr\left(C^{-1}(T - XB)\,\Sigma^{-1}(T - XB)^T\right).$$

If $\widehat{B}$ satisfies

$$tr\left(C^{-1}(T - X\widehat{B})\Sigma^{-1}(T - X\widehat{B})^T\right) = \min_B\left(tr\left(C^{-1}(T - XB)\,\Sigma^{-1}(T - XB)^T\right)\right), \tag{6}$$

then $\widehat{B}$ is said to be the generalised least square estimator of B (Christensen, 2001). In normal case, the generalised least square estimator is ML estimate as well. By minimising

(6), we obtain the ML estimator for B. After simple algebra we derive (3) from (5). Solving the equation $\frac{\partial \ln L}{\partial \Sigma} = 0$, we complete the proof of Lemma.

Since $E\left\{\widehat{\Sigma}\right\} = \frac{N-2q}{N}\Sigma$, further we will use bias adjusted estimator

$$\widetilde{\Sigma} = \frac{N}{N-2q}\widehat{\Sigma}.$$

Put $\Delta\hat{\mu}_l^0 = \hat{\mu}_l^0 - \mu_l^0 = \left(\widehat{B}_l - B_l\right)^T x^0, l = 1, 2, \Delta\widehat{\Sigma} = \widehat{\Sigma} - \Sigma$. Let $\varphi(\cdot)$ be the standard normal p.d.f. Denote by $P_l^{(1)} = \partial P\left(\widehat{\Psi}\right)/\partial\hat{\mu}_l^0$, $P_{l,k}^{(2)} = \partial^2 P\left(\widehat{\Psi}\right)/\partial\hat{\mu}_l^0\partial\left(\hat{\mu}_k^0\right)^T$, $P_{\tilde{\sigma}_{ij}}^{(1)} = \partial P\left(\widehat{\Psi}\right)/\partial\tilde{\sigma}_{ij}$, $P_{\tilde{\sigma}_{ij},\tilde{\sigma}_{mq}}^{(2)} = \partial^2 P\left(\widehat{\Psi}\right)/\partial\tilde{\sigma}_{ij}\partial\tilde{\sigma}_{mq}$, $P_{lm,\tilde{\sigma}_{ij}}^{(2)} = \partial^2 P\left(\widehat{\Psi}\right)/\partial\hat{\mu}_{lm}^0\partial\tilde{\sigma}_{ij}$ the partial derivatives up to second order of $P\left(\widehat{\Psi}\right)$ with respect to the corresponding parameters evaluated at $\hat{\mu}_1^0 = \mu_1^0$, $\hat{\mu}_2^0 = \mu_2^0$ and $\widetilde{\Sigma} = \Sigma$, where $\hat{\mu}_{lm}^0$ denotes the m'*th* component of vector $\hat{\mu}_l^0$ and $\tilde{\sigma}_{ij}$ is ij'*th* element of matrix $\widetilde{\Sigma}$, $l, k = 1, 2$, $i, j, m, q = 1, \ldots, p$.

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

Assumption 2. Assume that $(x^0)^T D_{kl} x^0 = \mathrm{O}\left(\frac{1}{N}\right), k.l = 1, 2$ as $N \to \infty$.

Theorem. *Suppose assumptions* 1–2 *hold for training sample* T. *Then the asymptotic expansion of the expected error rate for the* $d_B(z^0; \widehat{\Psi})$ *is*

$$E_T\left\{P\left(\widehat{\Psi}\right)\right\} = P_B + \frac{1}{2}\pi_1\varphi\Big(-\frac{\Delta}{2} - \frac{\gamma}{\Delta}\Big)\Big(\sum_{l=1}^{2}\frac{a_l}{\Delta} + b_{lk} + \frac{c}{N-2q}\Big) + \mathrm{O}(N^{-2}),$$

where, for $l, k = 1, 2$,

$$a_l = \left(\frac{\Delta}{2} + (-1)^{l+1}\,\frac{\gamma}{\Delta}\right)^2 D_{ll}^{-1},$$

$$b_{lk} = \left(\frac{1}{2} - \frac{\gamma}{\Delta^2}\right)\left(\frac{1}{2} + \frac{\gamma}{\Delta^2}\right)\Delta\left(x^0\right)^T D_{lk}^{-1} x^0,$$

$$c = \frac{\gamma^2}{\Delta} + (p-1)\,\Delta,$$

and

$$\Delta = \sqrt{(\mu_1^0 - \mu_2^0)^T\,\Sigma{-}1\,(\mu_1^0 - \mu_2^0)}.$$

Proof. Since $P(\widehat{\Psi})$ is invariant under linear transformations of data we use the convenient canonical form of

$$\mu_l^0 = (-1)^{l+1}\,\frac{\Delta}{2}1_0, \quad \Sigma = I, \tag{7}$$

where 1_0 is a p-variate vector of zeroes except first element, which is equal to 1, $l = 1, 2$.

Expand $P(\widehat{\Psi})$ in Taylor series about the point $\mu_1^0 = \frac{\Delta}{2}1_0$, $\mu_2^0 = -\frac{\Delta}{2}1_0$, $\Sigma = I$. Taking the expectation with respect to the distribution of T and dropping the third order terms we have

$$E_T\left\{P(\widehat{\Psi})\right\} = P_B + \sum_{l=1}^{2}(P_l^{(1)})^T E_T\left\{\Delta\hat{\mu}_l^0\right\} + \sum_{i,j=1}^{p}(P_{\tilde{\sigma}_{ij}}^{(1)})^T E_T\left\{\Delta\tilde{\sigma}_{ij}\right\}$$
$$+\frac{1}{2}\sum_{l,k=1}^{2} tr\left(P_{l,k}^{(2)} E_T\left\{\Delta\hat{\mu}_l^0\left(\Delta\hat{\mu}_k^0\right)^T\right\}\right)$$
$$+\frac{1}{2}\sum_{i,j,q,m=1}^{p} P_{\tilde{\sigma}_{ij},\tilde{\sigma}_{qm}}^{(2)} E_T\left\{\Delta\tilde{\sigma}_{ij}\Delta\tilde{\sigma}_{qm}\right\}$$
$$+\frac{1}{2}\sum_{l,k=1}^{2}\sum_{i,j,m=1}^{p} P_{lm,\tilde{\sigma}_{ij}}^{(2)} E_T\left\{\Delta\tilde{\sigma}_{ij}\Delta\hat{\mu}_{lm}^0\right\}. \tag{8}$$

Since $P(\widehat{\Psi})$ is minimized at (7), then, for $l = 1, 2$,

$$P_l^{(1)} = \mathbf{0}_p, \tag{9}$$

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

$$P_{\tilde{\sigma}_{ij}}^{(1)} = 0. \tag{10}$$

Using Lemma we get, that, for $l, k = 1, 2$,

$$E_T\left\{\Delta\hat{\mu}_l^0\left(\Delta\hat{\mu}_k^0\right)^T\right\} = (x^0)^T D_{lk}^{-1} x^0 \Sigma. \tag{11}$$

Also

$$E\left(\Delta\tilde{\sigma}_{ij}\Delta\tilde{\sigma}_{km}\right) = \frac{1}{N-2q}\left(\sigma_{ik}\sigma_{jm} + \sigma_{im}\sigma_{jk}\right) \tag{12}$$

and

$$E_T\left\{\Delta\tilde{\sigma}_{ij}\Delta\hat{\mu}_{lm}^0\right\} = 0, \tag{13}$$

because of properties of Gaussian variables [3]. Note that

$$P_{l,l}^{(2)} = \frac{\pi_1}{\Delta}\varphi\Big(-\frac{\Delta}{2} - \frac{\gamma}{\Delta}\Big)\left(I + \left(\left(\frac{\Delta}{2} + (-1)^{l+1}\frac{\gamma}{\Delta}\right)^2 - 1\right)1_0 1_0^T\right) \tag{14}$$

and

$$P_{1,2}^{(2)} = \frac{\pi_1}{\Delta}\varphi\Big(-\frac{\Delta}{2} - \frac{\gamma}{\Delta}\Big)\Delta\Big(\frac{1}{2} - \frac{\gamma}{\Delta^2}\Big)\Big(\frac{1}{2} + \frac{\gamma}{\Delta^2}\Big)1_0 1_0^T. \tag{15}$$

Using assumptions 1–2 to (11)–(13), we can conclude that all terms in the right side of (8) are of order $\mathrm{O}\left(\frac{1}{N}\right)$, as $N \to \infty$. Obviously the higher order moments of parameter estimators are of order $\mathrm{O}\left(N^{-k}\right)$, $k > 1$. So the proof of theorem is completed.

References

[1] R. Christensen, *Advanced Linear Modeling. Multivariate, Time Series, and Spatial Data; Nonparametric Regression and Response Surface Maximization*, Springer (2001).

[2] N. Cressie, H.-C. Huang, Classes of nonseparable, spatio-temporal stationary covariance functions, *American Statistical Association. Theory and Methods*, **94**(448), 1330–1340 (1999).

[3] J.R. Magnus, H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Aconometrics*, Wiley and Sons (2002).

[4] J. Šaltytė–Benth, K. Dučinskas, Diskriminant analysis of spatial-temporal data, *Lietuvos Matematikos Rinkinys*, **42**, 493–496 (2002).

Tiesinė diskriminantinė tarp klasių koreliuotų erdvinių-laikinių duomenų analizė

J. Šaltytė-Benth, K. Dučinskas

Straipsnyje nagrinėjamas daugiamačių Gauso laukų su faktorizuota kovariacijų funkcija klasifikavimo uždavinys. Gautas vidutinės klasifikavimo klaidos asimptotinis skleidinys atvejui, kai klasių parametrai vertinami pagal mokymo imtis, pasižyminčias tarpklasine koreliacija.