Nonlinear Analysis: Modelling and Control, 2002, **v. 7**, No. 1, 31-42

Comparison of Nonlinear Spatial Correlation Models by the Influence of the Data Augmentation to the Classification Risk

J.Šaltytė, K.Dučinskas
Klaipėda University
H.Manto 84, 5808 Klaipėda, Lithuania
jsaltyte@gmf.ku.lt, duce@gmf.ku.lt

Received: 14.03.2002
Accepted: 24.04.2002

Abstract

The Bayesian classification rule used for the classification of the observations of the (second-order) stationary Gaussian random fields with different means and common factorised covariance matrices is investigated. The influence of the observed data augmentation to the Bayesian risk is examined for three different nonlinear widely applicable spatial correlation models. The explicit expression of the Bayesian risk for the classification of augmented data is derived. Numerical comparison of these models by the variability of Bayesian risk in case of the first-order neighbourhood scheme is performed.

Keywords: spatial correlation, nugget effect, sill, Bayesian classification rule, data augmentation.

1 Introduction

In remote sensing and image analysis discriminant analysis (DA) of spatially correlated Gaussian data are of great importance. When classes

are completely specified, an optimal classification rule in sense of minimum classification risk is the Bayesian classification rule (BCR). Evidently the risk of BCR, i.e. Bayesian risk (BR), is decreasing when number of observations to be classified is increasing.

In this paper we examined numerically the BR reduction when the observation to be classified is augmented by observations at the four first-order neighbours on the 2-dimensional lattice. Three different spatial correlation models are compared in terms of values of BR. These models are spherical, exponential and Ornstein-Uhlenbeck correlation functions. The linear spatial correlation was ignored because it cannot correspond to the second-order stationary process (Christensen, 1991, ch.6.).

2 The Problem and Model

We consider the situation when the object with observed feature $\mathbf{Z}(\mathbf{r})$, distributed in some spatial domain $D \subset \Re^2$, may belong to one of two classes Ω_1 or Ω_2 with known prior probabilities $\pi_1(\mathbf{r})$, $\pi_2(\mathbf{r})$, respectively, $\sum_{l=1}^{2} \pi_l(\mathbf{r}) = 1$. The mathematical model of the feature $\mathbf{Z}(\mathbf{r})$ is a p-variate random field $\{\mathbf{Z}(\mathbf{r}) : \mathbf{r} \in D \subset \Re^2\}$, having different means and factorised covariance matrices in classes Ω_1 and Ω_2. This model in Ω_l is of the form

$$\mathbf{Z}(\mathbf{r}) = \boldsymbol{\mu}_l(\mathbf{r}) + \boldsymbol{\varepsilon}_l(\mathbf{r}),$$

where $\boldsymbol{\mu}_l(\mathbf{r}) \in R^p$ is a mean function and $\{\boldsymbol{\varepsilon}_l(\mathbf{r}) : \mathbf{r} \in D\}$ is a p-variate zero-mean second-order stationary spatially correlated random error field, l=1,2.

Assume that considered field is Gaussian with spatially factorised covariance function. Hence, the class-conditional covariance between any two observations $\mathbf{Z}(\mathbf{s})$ and $\mathbf{Z}(\mathbf{t})$ from class Ω_l is

$$\operatorname{cov}\{\mathbf{Z}(\mathbf{s}),\mathbf{Z}(\mathbf{t})\}=\operatorname{cov}\{\boldsymbol{\varepsilon}_l(\mathbf{s}),\boldsymbol{\varepsilon}_l(\mathbf{t})\}=c_l(\mathbf{h})\boldsymbol{\Sigma},$$

where $c_l(\mathbf{h})$ is a spatial correlation function, $\mathbf{h}=\mathbf{s}-\mathbf{t}$, $c_l(\mathbf{0})=1$ and $\boldsymbol{\Sigma}=\operatorname{cov}\{\mathbf{Z}(\mathbf{s}),\mathbf{Z}(\mathbf{s})\}$ is a covariance matrix for the feature vector components, l=1,2. Also let

$$\operatorname{cov}\{\boldsymbol{\varepsilon}_1(\mathbf{s}),\boldsymbol{\varepsilon}_2(\mathbf{t})\}=\mathbf{0},$$

for any $\mathbf{s},\mathbf{t}\in D$, where $\mathbf{0}$ is a $p\times p$ matrix of zeroes.

If $p_l(\mathbf{z}(\mathbf{r}))$ denotes the p.d.f. of $\mathbf{Z}(\mathbf{r})=\mathbf{z}(\mathbf{r})$, then

$$p_l(\mathbf{z}(\mathbf{r}))=(2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{z}(\mathbf{r})-\boldsymbol{\mu}_l(\mathbf{r}))^T\boldsymbol{\Sigma}^{-1}(\mathbf{z}(\mathbf{r})-\boldsymbol{\mu}_l(\mathbf{r}))\right).$$

We denote by $d(\mathbf{z}(\mathbf{r}))$ a classification rule, where $d(\mathbf{z}(\mathbf{r}))=l$ implies that the object with observation $\mathbf{Z}(\mathbf{r})=\mathbf{z}(\mathbf{r})$ is to be assigned to the class Ω_l, l=1,2. The losses of classification when an object from class l is allocated to class k, is denoted by $L(l,k)$. Then the risk of classification based on rule $d(\cdot)$ can be expressed as

$$R=R(d(\cdot))=\sum_{l=1}^{2}\pi_l(\mathbf{r})\int_{\mathbf{Z}}L(l,d(\mathbf{z}(\mathbf{r})))p_l(\mathbf{z}(\mathbf{r}))d\mathbf{z}(\mathbf{r}).$$

The BCR $d_B(\cdot)$ minimising R is defined as

$$d_B(\cdot)=\arg\max_{\{l=1,2\}}g_l(\mathbf{r})p_l(\mathbf{z}(\mathbf{r}))$$

where, for l=1,2,

$$g_l(\mathbf{r})=\pi_l(\mathbf{r})(L(l,3-l)-L(l,l)).$$

The risk for the BCR in considered case is

$$R_B = \sum_{l=1}^{2}\left(\pi_l(\mathbf{r})L(l,l)+ g_l(\mathbf{r})\Phi\left(-\frac{\Delta(\mathbf{r})}{2}+(-1)^l\frac{g(\mathbf{r})}{\Delta(\mathbf{r})}\right)\right), \qquad (1)$$

here $\Delta(\mathbf{r})$ is the Mahalanobis distance

$$\Delta(\mathbf{r}) = \left((\boldsymbol{\mu}_1(\mathbf{r})-\boldsymbol{\mu}_2(\mathbf{r}))^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1(\mathbf{r})-\boldsymbol{\mu}_2(\mathbf{r}))\right)^{\frac{1}{2}},$$

$\Phi(\cdot)$ is the standard normal distribution function and $g(\mathbf{r}) = \ln\left(\frac{g_1(\mathbf{r})}{g_2(\mathbf{r})}\right)$.

In this paper it will be shown how the Bayesian risk (1) changes when the point is classified on the basis of augmented observation vector.

3 Bayesian Risk for augmented Data

The usual multivariate *DA* is used to classify an object at location **r** on the basis of the feature vector for that object. However, using more observations through the neighbours in the classification procedure, it is expected that the risk of classification will be reduced.

Suppose there are *m* neighbours in the vicinity of **r**. Define the neighbourhood of point **r** as $\mathbf{N}_\mathbf{r} \equiv \{\mathbf{r}_1,...,\mathbf{r}_m\}$. We assume that the points in $\mathbf{N}_\mathbf{r}$ and classifying point **r** belong to the same class. Let $\mathbf{Z}_{\mathbf{N}_\mathbf{r}}$ contains the observations on objects at locations in $\mathbf{N}_\mathbf{r}$, i.e. $\mathbf{Z}_{\mathbf{N}_\mathbf{r}} = (\mathbf{Z}(\mathbf{r}_1),...,\mathbf{Z}(\mathbf{r}_m))^T$. Then the mean vector for augmented observations $\mathbf{Z}^+(\mathbf{r}) = \left(\mathbf{Z}^T(\mathbf{r}), \mathbf{Z}_{\mathbf{N}_\mathbf{r}}^T\right)^T$ is

$$\boldsymbol{\mu}_l^+(\mathbf{r}) = \mathbf{1}_{m+1}\otimes\boldsymbol{\mu}_l(\mathbf{r}), \qquad (2)$$

where $\mathbf{1}_{m+1}$ is $(m+1)\times 1$ vector of ones, and $\otimes$ is Kronecker's product (see e.g. Mardia, 1984). The covariance matrix of $\mathbf{Z}^+(\mathbf{r})$, given that $\mathbf{r}$ belongs to Ω_l, is

$$\boldsymbol{\Sigma}_l^+ = \mathbf{P}_l \otimes \boldsymbol{\Sigma}, \qquad (3)$$

where $\mathbf{P}_l$ is the spatial correlation matrix of order $(m+1)\times(m+1)$, whose $\alpha\beta$'th element is $c_{l;\alpha\beta} = c_l(\mathbf{r}_{\alpha-1} - \mathbf{r}_{\beta-1})$, $\alpha, \beta = 1,...,m$, and $\mathbf{r}_0=\mathbf{r}$.

The classification is accomplished by implementing the assignment of point $\mathbf{r}$ on the basis of value $\mathbf{z}^+(\mathbf{r})$ of augmented data vector $\mathbf{Z}^+(\mathbf{r})$. Under the assumptions above the l'th class conditional distribution of $\mathbf{Z}^+(\mathbf{r})$ is $(m+1)\times p$-variate normal with mean (2) and covariance matrix (3). Let $p_l^+(\mathbf{z}^+(\mathbf{r}))$ denotes the p.d.f. of $\mathbf{z}^+(\mathbf{r})$ in the l'th class.

Denote by $d^+(\mathbf{z}^+(\mathbf{r}))$ a classification rule based on augmented observation $\mathbf{Z}^+(\mathbf{r}) = \mathbf{z}^+(\mathbf{r})$. Then the risk of classification based on rule $d^+(\cdot)$ is

$$R^+ = R^+(d^+(\cdot)) = \sum_{l=1}^{2} \pi_l^+(\mathbf{r}) \int_{\mathbf{Z}} L(l, d^+(\mathbf{z}^+(\mathbf{r}))) p_l^+(\mathbf{z}^+(\mathbf{r})) d\mathbf{z}^+(\mathbf{r}),$$

where $\pi_l^+(\mathbf{r})$ is a prior probability that observations at locations $\mathbf{r}_0,...,\mathbf{r}_m$ belong to the l'th class, l=1,2. The Bayesian classification rule $d_B^+(\cdot)$ minimising R^+ is defined as

$$d_B^+(\cdot) = \arg\max_{\{l=1,2\}} g_l^+(\mathbf{r}) p_l^+(\mathbf{z}^+(\mathbf{r})),$$

where, for l=1,2,

$$g_l^+(\mathbf{r}) = \pi_l^+(\mathbf{r})(L(l,3-l) - L(l,l)). \qquad (4)$$

Since the goal is the evaluation of the performance of the linear *DA*, it is necessary assume, that $c_1(\cdot)=c_2(\cdot)=c(\cdot)$. Put $\rho^{\bullet\bullet}=\mathbf{1}_{m+1}^T\mathbf{P}^{-1}\mathbf{1}_{m+1}$, where $\mathbf{1}_{m+1}$ is (m+1)-dimensional vector of ones and $\mathbf{P}$ is the spatial correlation matrix defined above. Denote $g^+(\mathbf{r})=\ln\dfrac{g_1^+(\mathbf{r})}{g_2^+(\mathbf{r})}$, where $g_l^+(\mathbf{r})$ is defined in (4), l=1,2.

LEMMA. Let $d^+(\mathbf{z}^+(\mathbf{r}))$ is used for classification of $\mathbf{r}\in D$ on the basis of augmented vector $\mathbf{Z}^+(\mathbf{r})$. Then the Bayesian risk of classification is equal

$$R_B^+=\sum_{l=1}^{2}\left(\pi_l^+(\mathbf{r})L(l,l)+g_l^+(\mathbf{r})\Phi\left(-\frac{\Delta^+(\mathbf{r})}{2}+(-1)^l\frac{g^+(\mathbf{r})}{\Delta^+(\mathbf{r})}\right)\right), \qquad (5)$$

where

$$\Delta^+(\mathbf{r})=\sqrt{\rho^{\bullet\bullet}}\,\Delta(\mathbf{r}). \qquad (6)$$

Proof. The square of Mahalanobis distance between classes Ω_1 and Ω_2 based on augmented observation $\mathbf{Z}^+(\mathbf{r})$ is

$$\begin{aligned}(\Delta^+(\mathbf{r}))^2&=(\boldsymbol{\mu}_1^+(\mathbf{r})-\boldsymbol{\mu}_2^+(\mathbf{r}))^T(\boldsymbol{\Sigma}^+)^{-1}(\boldsymbol{\mu}_1^+(\mathbf{r})-\boldsymbol{\mu}_2^+(\mathbf{r}))=\\&=(\mathbf{1}_{m+1}\otimes\boldsymbol{\mu}_1(\mathbf{r})-\mathbf{1}_{m+1}\otimes\boldsymbol{\mu}_2(\mathbf{r}))^T(\mathbf{P}\otimes\boldsymbol{\Sigma})^{-1}(\mathbf{1}_{m+1}\otimes\boldsymbol{\mu}_1(\mathbf{r})-\mathbf{1}_{m+1}\otimes\boldsymbol{\mu}_2(\mathbf{r})).\end{aligned}$$

Using the property $(\mathbf{P}\otimes\boldsymbol{\Sigma})^{-1}=\mathbf{P}^{-1}\otimes\boldsymbol{\Sigma}^{-1}$ and taking an inverse of $\mathbf{P}$ we obtain that $(\Delta^+(\mathbf{r}))^2=\rho^{\bullet\bullet}\Delta^2(\mathbf{r})$. This completes the proof of the lemma.

REMARK. In the case of independent observations the value of $\rho^{\bullet\bullet}$ is equal to $m+1$, since $\mathbf{P}$ becomes the identity matrix $\mathbf{I}$ in such a case.

Switzer (1980) proposed another way to augment the *p*-variate observation $\mathbf{Z}(\mathbf{r})$. It is so-called simple augmentation, when data are augmented with the mean of observations in neighbouring locations. In such case we have $\mathbf{Z}_M^+(\mathbf{r}) = \left(\mathbf{Z}^T(\mathbf{r}), \mathbf{Z}_{\mathbf{N}_\mathbf{r}^M}^T\right)^T$, where

$$\mathbf{Z}_{\mathbf{N}_\mathbf{r}^M} = \frac{1}{m}\mathbf{Z}(\mathbf{r}_i),$$

i=1,…,m. Then the *DA* is performed on the basis of augmented observation vector $\mathbf{Z}_M^+(\mathbf{r})$. The spatial covariance matrix (3) in such case is

$$\mathbf{\Sigma}_M^+ = \mathbf{P} \otimes \mathbf{\Sigma}_M,$$

where $\mathbf{\Sigma}_M = \mathrm{cov}\left(\mathbf{Z}(\mathbf{r}), \mathbf{Z}_{\mathbf{N}_\mathbf{r}^M}(\mathbf{r})\right)$.

The results of the lemma stated above can be easily adapted to the situation just described.

The influence of the data augmentation to the BR is evaluated by the risk reduction rate defined by

$$\mathrm{QR} = \frac{R_B^+}{R_B}.$$

4 Example

As an example we consider the integer regular 2-dimensional lattice and assume that the point **r** to be classified and its neighbours are inside the lattice. We deal with the first-order neighbourhood (see, e.g. Besag (1974)). The considered situation is presented in Figure 1.

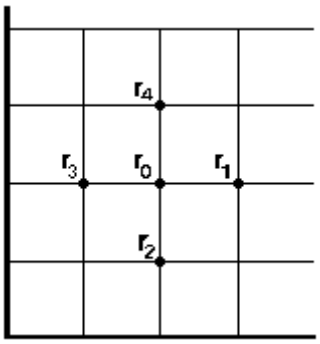

Figure 1. The positions of classifying point and its first-order neighbours (signed by "•") on the lattice

To estimate an influence of data augmentation on the classification risk, we consider three spatial correlation models.

The isotropic spherical correlation function is given by expression

$$c_S(\mathbf{h}) = \begin{cases} \dfrac{\kappa_1}{\kappa_0 + \kappa_1}\left(1 - \dfrac{3}{2}\dfrac{|\mathbf{h}|}{\eta} + \dfrac{1}{2}\dfrac{|\mathbf{h}|^3}{\eta^3}\right), & 0 \le |\mathbf{h}| \le \eta, \\ 1, & |\mathbf{h}| = 0, \\ 0, & |\mathbf{h}| > \eta, \end{cases}$$

for nonnegative κ_0, κ_1, η. The nugget effect is κ_0 and the sill is $\kappa_0 + \kappa_1$. For this model, observations more than η units apart are uncorrelated, so the range is η.

The exponential correlation function is

$$c_E(\mathbf{h}) = \begin{cases} \dfrac{\kappa_1}{\kappa_0 + \kappa_1}\exp\left(-\eta\sqrt{t^2 h_1^2 + h_2^2}\right), & |\mathbf{h}| > 0, \\ 1, & |\mathbf{h}| = 0, \end{cases}$$

for nonnegative κ_0, κ_1, η. Here t is the parameter of anisotropy. When t=1, the exponential correlation function becomes isotropic one; otherwise it is anisotropic. The nugget effect is κ_0, the sill is $\kappa_0 + \kappa_1$, and the range is infinite. While the range is infinite, correlations decrease very rapidly as **h** increases. Of course, this phenomenon depends on the value of η.

The Ornstein – Uhlenbeck correlation function is defined as follows

$$c_{OU}(\mathbf{h}) = \begin{cases} \dfrac{\kappa_1}{\kappa_0 + \kappa_1} \exp\left(-\eta\left(t^2 h_1^2 + h_2^2\right)\right), & |\mathbf{h}| > 0, \\ 1, & |\mathbf{h}| = 0, \end{cases}$$

for nonnegative κ_0, κ_1, η. It is anisotropic correlation function, when $t \neq 1$. In the case of t=1 it becomes a well known isotropic correlation function often called Gaussian correlation function. The behaviour of the Ornstein – Uhlenbeck model is similar to that of the exponential model. However, the correlations at distances greater than one approach zero much more rapidly than in the exponential model. Also, for small distances, the correlation approaches the value $\dfrac{\kappa_1}{\kappa_0 + \kappa_1}$ much more rapidly then the exponential does.

From the comparison of expressions for BR (1) and (5) it is seen that they differ only in term $\sqrt{\rho^{\bullet\bullet}}$ in the argument of function $\Phi(\cdot)$. Therefore we need define the spatial correlation matrices for considered situation of position of classifying point and its neighbours, compute the sum of elements of inverses of these matrices, and then find the analytical expressions for classification risk.

Denote by $R^{+}_{B,S}$, $R^{+}_{B,E}$ and $R^{+}_{B,OU}$ the BR of classification, when spherical, exponential and Ornstein-Uhlenbeck correlation are used, respectively; then QR_S, QR_E and QR_{OU} are values of the QR for the same three models.

To illustrate the influence of data augmentation on the classification risk a set of numerical calculations in tables below is presented. Consider for simplicity that $L(l,k)=1-\delta_{lk}$, where δ_{lk} is the Kronecker's delta, and $\pi_l=\frac{1}{2}$, l,k=1,2.

Assume, primarily, that there is no nugget effect, i.e. $\kappa_0=0$, thus the ratio $\frac{\kappa_1}{\kappa_0+\kappa_1}=1$. From the Figure 1 it is seen, that an appropriate value for the range for the spherical correlation function is $\eta=2.2$. Assume, that the parameter of anisotropy in the exponential and Ornstein-Uhlenbeck correlation functions is t=1.5. Whereas t>1, it is clear that the behaviour of process described by these functions in the east-west direction is more intensive than that in the south-north direction. The quantity η is the spatial dependence parameter for the exponential and Ornstein-Uhlenbeck functions. Let $\eta=0.4$ for both models. Values of the BR and QR for considered set of parameters are presented in Table 1.

For all described cases the values of BR approach zero when distance $\Delta(\mathbf{r})$ increases. Also it is obvious that bigger number of observations determines smaller risk, which is reasonable thing.

The values of risks are smallest for the spherical correlation function. However, using the Ornstein-Uhlenbeck correlation functions gives the risk not much bigger than that obtained with using spherical correlation function. It can be concluded that assuming that there is no nugget effect and using parameters defined above the spherical correlation function is

the best one, whereas the exponential function gives the biggest risk which approaches zero (when the distance increases) slower then does the risk obtained by using other two correlation functions. But the influence of the data augmentation is the strongest one for the exponential spatial correlation model, since the values of the QR are the smallest.

Table 1. Values of the BR and QR, when $\kappa_0 = 0$.

$\Delta(\mathbf{r})$	R_B	$R^+_{B,S}$	$R^+_{B,E}$	$R^+_{B,OU}$	QR_E	QR_S	QR_{OU}
0,25	0,390	0,415	0,435	0,418	0,921	0,966	0,929
0,50	0,288	0,333	0,372	0,340	0,830	0,926	0,847
0,75	0,201	0,259	0,312	0,268	0,732	0,881	0,757
1,00	0,132	0,194	0,256	0,204	0,629	0,831	0,663
1,25	0,081	0,140	0,206	0,151	0,528	0,776	0,568
1,50	0,047	0,098	0,163	0,108	0,432	0,719	0,475
1,75	0,025	0,066	0,126	0,074	0,344	0,660	0,389
2,00	0,013	0,042	0,095	0,049	0,266	0,600	0,311
2,25	0,006	0,026	0,070	0,032	0,201	0,540	0,242
2,50	0,003	0,016	0,051	0,019	0,147	0,481	0,184
2,75	0,001	0,009	0,036	0,012	0,105	0,424	0,137
3,00	0,000	0,005	0,025	0,007	0,072	0,370	0,099

Suppose now that we detect a measurement error, i.e. the nugget effect $\kappa_0 = \frac{3}{4}$, thus the ratio $\frac{\kappa_1}{\kappa_0 + \kappa_1} = \frac{1}{4}$. Let the values of other parameters be the same as it was considered above. The comparison of the values of BR and QR are presented in Table 2.

Table 2. The values of the BR and QR, when $\kappa_0 = \frac{3}{4}$.

$\Delta(\mathbf{r})$	R_B	$R^+_{B,S}$	$R^+_{B,E}$	$R^+_{B,OU}$	QR_E	QR_S	QR_{OU}
0,25	0,390	0,400	0,410	0,404	0,888	0,911	0,898
0,50	0,288	0,306	0,325	0,314	0,762	0,809	0,783
0,75	0,201	0,223	0,247	0,234	0,630	0,699	0,661

1,00	0,132	0,155	0,181	0,167	0,501	0,588	0,540
1,25	0,081	0,102	0,128	0,113	0,383	0,480	0,425
1,50	0,047	0,064	0,086	0,073	0,281	0,380	0,323
1,75	0,025	0,038	0,056	0,045	0,197	0,292	0,237
2,00	0,013	0,021	0,034	0,026	0,132	0,217	0,167
2,25	0,006	0,011	0,020	0,015	0,085	0,156	0,113
2,50	0,003	0,006	0,011	0,008	0,052	0,108	0,073
2,75	0,001	0,003	0,006	0,004	0,031	0,073	0,046
3,00	0,000	0,001	0,003	0,002	0,017	0,047	0,028

From the comparison of Table 1 and Table 2 it is obvious that the risks of classification are smaller in the case when there is nugget effect $\kappa_0 = \frac{3}{4}$ assumed. They approach zero quicker than in the case of no nugget effect. Thus, the detecting of nugget effect may be important in attempting to decrease the values of classification risk. From the Table 2 we can conclude that the influence of data augmentation is also strongest for the exponential model.

5 References

1. Christensen R. *Linear models for multivariate, time series and spatial data*. Springer-Verlag, 1991.
2. Mardia K.V., Marshall R.J. "Maximum Likelihood Estimation of Models for Residual Covariance in Spatial Regression". *Biometrica*, **71**(1), 135-46, 1984.
3. Switzer P. Extensions of linear discriminant analysis for statistical classification of remotely sensed satellite imagery. *Math. Geol.*, **12**(4), 1980.
4. Besag J. Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society B*, **36**,192-225, 1974.