QUADRATIC DISCRIMINATION THROUGH ORTHOGONAL TRANSFORMATION AND ITS APPLICATION TO LONG-RANGE FORECASTING OF DROUGHT AND EXCESSIVE RAINFALL

Shi Neng (施 能)

Nanjing Institute of Meteorology, Nanjing

and *Chen Jiyi* (程极益)

Nanjing Agricultural University, Nanjing

Received November 20, 1984

ABSTRACT

The method of quadratic discrimination through orthogonal transformation is introduced to optimize quadratic discriminant function, which has been proved more effective than the method of stepwise multip'e discrimination. It is noteworthy that, in the case of a large number of predictors, it is better to make a preliminary choice by using stepwise multiple discrimination or stepwise regression so that the calculation can be made more stable and the effectiveness of discrimination can be improved.

I. METHOD

Assume that the vector of predictors $\mathbf{X}=(x_1, x_2, \ldots, x_p)^T$ satisfies the p-dimensional multivariate normal distribution and the predictand is divided into two categories ω_1 and ω_2, then the conditional probability density functions can be represented by

$$P(\mathbf{X}/\omega_1)=N_1(\boldsymbol{\mu}(1), \boldsymbol{\Sigma}(1)),\ P(\mathbf{X}/\omega_2)=N_2(\boldsymbol{\mu}(2), \boldsymbol{\Sigma}(2)), \tag{1}$$

where $\boldsymbol{\mu}(1)$, $\boldsymbol{\mu}(2)$, $\boldsymbol{\Sigma}(1)$, $\boldsymbol{\Sigma}(2)$ are column vectors of mathematical expectation and covariance matrices for the two categories, respectively.

The equation of the discriminant function can be written as

$$D(\mathbf{X})=\ln\frac{P(\mathbf{X}/\omega_1)}{P(\mathbf{X}/\omega_2)}=\frac{1}{2}\Big[(\mathbf{X}-\boldsymbol{\mu}(2))^T\boldsymbol{\Sigma}^{-1}(2)(\mathbf{X}-\boldsymbol{\mu}(2))-(\mathbf{X}-\boldsymbol{\mu}(1))^T\boldsymbol{\Sigma}^{-1}(1)$$

$$(\mathbf{X}-\boldsymbol{\mu}(1))\Big]-\frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}(1)|}{|\boldsymbol{\Sigma}(2)|}. \tag{2}$$

$\mathbf{X}$ can be classified into two categories: the first category if $D(\mathbf{X})>0$ and the second one if $D(\mathbf{X})<0$. In Eq. (2) the superscripts T and -1 indicate the transposition and inverse operations, respectively, and the symbol $|\ |$ represents the determinant of a matrix. No assumption that $\boldsymbol{\Sigma}(2)=\boldsymbol{\Sigma}(1)$ is found in Eq. (2), thus $D(\mathbf{X})$, composed of the quadratic functions of $x_1, x_2, \cdots, x_p$ is called the quadratic discriminant function so as to

distinguish from the linear discriminant function on the assumption that $\boldsymbol{\Sigma}(1)=\boldsymbol{\Sigma}(2)$. In Eq. (2) the inverse matrix is involved, which is not easy to calculate. Refs. [1] and [2] present the method of eliminating the inverse matrix while expressing $D(\mathbf{X})$ as

$$D(\mathbf{X})=\sum_{i=1}^{p}\sum_{j=1}^{p}a_{ij}x_ix_j+\sum_{i=1}^{p}b_ix_i+c\ . \tag{3}$$

There are specific formulas to calculate a_{ij}, b_j, c $(i,j=1,2,\cdots,p)$. However, when p is large, $D(\mathbf{X})$ involves too many terms to be calculated, and some of the terms may not be necessary to provide adequate discriminant information. In order to overcome this disadvantage, we should reduce the dimensionalities of $\mathbf{X}$. Thus, it is helpful to extract the principal component of X, that is, the covariance matrix should be diagonalized. This can be realized by the following method[4].

First, move the origin of the coordinate axes to the point in space of the first mathematical expectation. If the first covariance matrix is a positively definite (generally, so is $\boldsymbol{\Sigma}(1)$), turn the coordinate axes so that $\boldsymbol{\Sigma}(1)$ is diagonalized. Then compress the space of the predictors in order to transform the diagonalized matrix into an identity matrix, and finally rotate the second covariance matrix $\boldsymbol{\Sigma}(2)$ into a diagnoal form while $\boldsymbol{\Sigma}(1)$ remains an identity matrix. These three steps can be written as

$$\mathbf{X}-\boldsymbol{\mu}(1); \tag{4}$$

$$\mathbf{A}\boldsymbol{\Sigma}(1)\mathbf{A}^T=\mathbf{I};\ \mathbf{A}\boldsymbol{\Sigma}(2)\mathbf{A}^T=\boldsymbol{\Lambda}; \tag{5}$$

where $\mathbf{I}$ is an identity matrix, $\boldsymbol{\Lambda}$ is the diagonal matrix with elements being $\lambda_1,\ \lambda_2,\cdots,\lambda_p$. Thus, it is easy to prove that $\lambda_1,\lambda_2,\cdots,\lambda_p$ are given by p roots of the following equation system[1]

$$|\boldsymbol{\Sigma}(2)-\lambda\boldsymbol{\Sigma}(1)|=0. \tag{6}$$

Transformation of matrix A may also be obtained from (4) and (5) (see section III). When A is determined, the linear transformation

$$\mathbf{Y}=(y_i)=\mathbf{A}(\mathbf{X}-\boldsymbol{\mu}(1)) \tag{7}$$

$$(m_i)=\mathbf{A}(\boldsymbol{\mu}(2)-\boldsymbol{\mu}(1)) \tag{8}$$

can be accomplished. From (5) we obtain

$$-\frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}(1)|}{|\boldsymbol{\Sigma}(2)|}=\frac{1}{2}\sum_{i=1}^{p}\ln\lambda_i\ . \tag{9}$$

From Eqs. (5), (7) and (8), we have

$$(\mathbf{X}-\boldsymbol{\mu}(1))^T\boldsymbol{\Sigma}^{-1}(1)(\mathbf{X}-\boldsymbol{\mu}(1))=\sum_{i=1}^{p}y_i^2\ , \tag{10}$$

$$(\mathbf{X}-\boldsymbol{\mu}(2))^T\boldsymbol{\Sigma}^{-1}(2)(\mathbf{X}-\boldsymbol{\mu}(2))=\sum_{i=1}^{p}\frac{1}{\lambda_i}(y_i^2+m_i^2-2y_im_i). \tag{11}$$

Thus, $D(\mathbf{X})$ may become

$$D(\mathbf{Y})=\frac{1}{2}\sum_{i=1}^{p}\left(\ln\lambda_i-y_i^2+\frac{1}{\lambda_i}(y_i^2+m_i^2-2m_iy_i)\right)\ , \tag{12}$$

1) In Ref. [4], the diagonal elements of diagonal matrix which satisfies conditions (4) and (5) are the root of the $|\boldsymbol{\Sigma}(1)-\lambda\boldsymbol{\Sigma}(2)|=0$. However, this result is not correct.

or

$$D(\mathbf{Y})=\sum_{i=1}^{p}D(y_i)\,, \tag{13}$$

$$D(y_i)=\frac{1}{2}\left(\ln\lambda_i-y_i^2+\frac{1}{\lambda_i}(y_i^2+m_i^2-2m_iy_i)\right). \tag{14}$$

We may classify X as the first category if $D(\mathbf{Y})>0$; and as the second category if $D(\mathbf{Y})<0$.

II. ESTIMATION OF INFORMATION OF VARIABLES

Eq. (14) is the quadratic function of the components $y_1, y_2, \cdots, y_p$ of Y after the linear transformation of X, and is also a quadratic discriminant function. If transformation matrix A is determined; Y can be obtained. The merit of Eqs. (13) and (14) is that the contributions of y_i to the discrimination of velues of the function are independent of each other. Thus, if the classification errors of y_i can be predetermined, the number of predictors can be reduced. It is adequate that the summation $\sum_{i=1}^{p}$ is performed only for those y_i which are more effective on the discrimination. Thus, if y_i can be arranged according to their discriminant efficiency, then the discriminant functions may be optimized. By using the Kullback divergence[4] of y_i, the discriminant errors may be roughly estimated,

$$[\mathrm{div}]_p=\int_{\mathbf{X}}P(\mathbf{X}/\omega_1)\ln\frac{P(\mathbf{X}/\omega_1)}{P(\mathbf{X}/\omega_2)}d\mathbf{X}+\int_{\mathbf{X}}P(\mathbf{X}/\omega_2)\ln\frac{P(\mathbf{X}/\omega_2)}{P(\mathbf{X}/\omega_1)}d\mathbf{X}\,, \tag{15}$$

where the subscript p denotes the number of predictors. The discriminant errors are decreased with the increase of $[\mathrm{div}]_p$.

With a normal distribution, the substitution of (1) for $P(\mathbf{X}/\omega_1)$, $P(\mathbf{X}/\omega_2)$ in Eq. (15) and integration of (15) will yield

$$[\mathrm{div}]_p=\frac{1}{2}\mathrm{tr}\left((\boldsymbol{\Sigma}(1)-\boldsymbol{\Sigma}(2))(\boldsymbol{\Sigma}^{-1}(2)-\boldsymbol{\Sigma}^{-1}(1))+(\boldsymbol{\mu}(2)-\boldsymbol{\mu}(1))^T\left(\frac{\boldsymbol{\Sigma}^{-1}(1)+\boldsymbol{\Sigma}^{-1}(2)}{2}\right)\right.$$
$$\left.\cdot(\boldsymbol{\mu}(2)-\boldsymbol{\mu}(1))\right) \tag{16}$$

where "tr" denotes the trace of the matrix in Eq. (16).

Assuming that $\boldsymbol{\Sigma}(1)=\boldsymbol{\Sigma}(2)=\boldsymbol{\Sigma}$ in Eq. (16) is given and $[\mathrm{div}]_p$ becomes the Mahalanobis distance D_p^2 in the linear condition.

By using Eq. (5), Eq. (15) becomes again the following form

$$[\mathrm{div}]_p=\frac{1}{2}\sum_{i=1}^{p}\left(m_i^2\left(1+\frac{1}{\lambda_i}\right)+\left(\lambda_i+\frac{1}{\lambda_i}-2\right)\right)=\sum_{i=1}^{p}(\mathrm{div})_i\,,$$

where

$$(\mathrm{div})_i=\frac{1}{2}\left[m_i^2\left(1+\frac{1}{\lambda_i}\right)+\left(\lambda_i+\frac{1}{\lambda_i}-2\right)\right]. \tag{17}$$

Divergence $[\mathrm{div}]_p$ is also reduced to the sum of the composed predictors. This means that $(\mathrm{div})_p$ is the divergence of the composite predictor y_i. Thus, substituting m_i and λ_i of the corresponding predictor y_i into Eq. (17) can obtain $(\mathrm{div})_i$, $i=1, 2, \cdots, p$. Then, the quadratic discriminant functions can be optimized upon arranging $y_1, y_2, \cdots, y_p$ according

to the magnitude of $(\text{div})_i$ and substituting them into (14). Therefore the best quadratic discriminant function can be found in each $D(y_i)$ or the composite $D(y_i)$.

III. PROCEDURE FOR CALCULATION

The matrix of the initial data is given as follows

$$\mathbf{X}=(x_{ij}),\quad (i=1,2,\cdots,p;j=1,2,\cdots,n)$$

where p is the number of the predictors and n is the sample size. The predictands are divided into two categories: ω_1 and ω_2; n_1 is the sample size for ω_1 and n_2 that for ω_2, $n_1+n_2=n$.

(1) *Computation of the mean vectors and covariance matrices for the two categories*

The mean vectors, which are p-dimensional vectors, are represented by $\boldsymbol{\mu}(1)$ and $\boldsymbol{\mu}(2)$ for the two categories respectively. The covariance matrices, which are symmetric matrices of $p\times p$ dimensions, are represented by $\boldsymbol{\Sigma}(1)$ and $\boldsymbol{\Sigma}(2)$ for the two categories respectively.

(2) *Computation of the characteristic roots of* $\boldsymbol{\Sigma}(1)$ *and the corresponding characteristic vectors*

Let the characteristic roots be 1, 2, $\cdots$, p. The characteristic vectors are used as the rows to compose matrix $\mathbf{A}^{(1)}$. We have

$$\mathbf{A}^{(1)}=(a_{ij}^{(1)}),$$

where $a_{ij}^{(1)}$ is the jth component of the characteristic vector corresponding to the characteristic root ζ_i of $\boldsymbol{\Sigma}(1)$. By using the square root $\sqrt{\zeta_i}$ of the characteristic root ζ_i, a diagonal matrix of $\mathbf{A}^{(2)}$ is composed. The diagonal elements of $\mathbf{A}^{(2)}$ are $\sqrt{\zeta_i}$, $i=1, 2, \cdots, p$, the others being 0. Thus

$$\mathbf{A}^{(2)}\cdot\mathbf{A}^{(1)}=\mathbf{Q}.$$

(3) *Computation of the characteristic roots and vectors of* $\mathbf{Q}\boldsymbol{\Sigma}(2)\mathbf{Q}^T$

Suppose that λ_1, $\lambda_2,\cdots,\lambda_p$ are the characteristic roots and $(a_{i1}^{(3)}\ a_{i2}^{(3)}\cdots a_{ip}^{(3)})$ are the characteristic vectors corresponding to λ_i. By using p characteristic vectors, the matrix $\mathbf{A}^{(3)}$ is composed, i. e.

$$\mathbf{A}^{(3)}=(a_{ij}^{(3)}),\quad (i=1,2,\cdots,p;\ j=1,2,\cdots,p)$$

(4) *Computation of the matrix* $\mathbf{A}=\mathbf{A}^{(3)}\cdot\mathbf{Q}$

Matrix $\mathbf{A}$, which is the transformation matrix, can certainly satisfy Eq. (5), that is, $\boldsymbol{\Sigma}(1)$ can be reduced to an identity matrix and $\boldsymbol{\Sigma}(2)$ to a diagonal matrix.

(5) *Linear transformation*

$$\mathbf{A}\cdot(\boldsymbol{\mu}(2)-\boldsymbol{\mu}(1))=(m_1,m_2,\cdots,m_p)^T$$

(6) *Computation of divergenece* $(\text{div})_i$, $i=1, 2, \cdots, p$

The computation formula is Eq. (17). Arrange $(\text{div})_i$ according to its magnitude and regulate λ_i, m_i corresponding to $(\text{div})_i$. This step is, in fact, the regulation of the rows of $\mathbf{A}$.

(7) *Computation of p-composed predictors y_i with $i=1, 2, \cdots, p$*

We have

$$(y_1, y_2, \cdots, y_p)^T = \mathbf{A} \cdot (\mathbf{X} - \boldsymbol{\mu}(1)).$$

The discriminant efficiency of p-composed predictors has been regulated from high to low.

(8) *Computation of discriminant functions*

The computation formula is shown in Eqs. (13) and (14).

IV. CASES, COMPARISON OF METHODS AND APPLICATION TO WEATHER PREDICTION

In order to predict the total rainfall during June through August (R_{6-8}) in North China, the mean rainfall at the three stations of Taiyuan, Jinan and Tianjin in the same period is used as the predictand and the following five predictors are selected, i. e. x_1, mean circulation index; x_2, area index of the subtropical high; x_3, intensity index of the subtropical high; x_4, center of the north polar vortex (the positive is taken for east longitude and the negative for west); and x_5, the point to which the subtropical high can extend westward; with $p=5$ and $n=22$. Let the above mean of R_{6-8} be of the first category and the other be of the second.

First, calculate the mean vector $\boldsymbol{\mu}(1)$ and $\boldsymbol{\mu}(2)$, their difference vectors $\boldsymbol{\mu} = \boldsymbol{\mu}(2) - \boldsymbol{\mu}(1)$ and the covariance matrices $\boldsymbol{\Sigma}(1)$ and $\boldsymbol{\Sigma}(2)$, i. e.

$\boldsymbol{\mu} = (-0.114 \quad -3.617 \quad -7.367 \quad 15.683 \quad 103.083)^T$, ($\boldsymbol{\Sigma}(1)$ and $\boldsymbol{\Sigma}(2)$ not shown). Next, calculate transformation matrix A:

$$\mathbf{A} = \begin{pmatrix} 0.539 & 0.208 & -0.066 & 0.084 & 0.025 \\ -1.238 & -0.047 & 0.024 & -0.074 & 0.010 \\ 2.202 & 0.437 & -0.136 & -0.006 & -0.001 \\ 4.204 & -0.254 & 0.108 & 0.004 & 0.001 \\ -0.804 & -0.231 & 0.144 & -0.004 & 0.002 \end{pmatrix}$$

Matrix A satisfies the following relationships:

$$\mathbf{A} \cdot \boldsymbol{\Sigma}(1) \mathbf{A}^T = \mathbf{I}$$

$$\mathbf{A} \cdot \boldsymbol{\Sigma}(2) \mathbf{A}^T = \begin{pmatrix} 10.574 & & & & \\ & 2.182 & & & \\ & & 0.740 & & \\ & & & 0.260 & \\ & & & & 0.160 \end{pmatrix}$$

and

$$(m_i) = \mathbf{A} \cdot \boldsymbol{\mu} = (3.550 \quad 0.088 \quad -1.001 \quad -0.203 \quad -0.050)^T.$$

Then, calculate $(\mathrm{div})_i$, $i=1, 2, 3, 4, 5$; and rearrange m_i, λ_i according to the magnitude of $(\mathrm{div})_i$. The results are listed in Table 1.

The values of m_i and λ_i in Table 1 are, in fact, the mean and variance of the composite predictor y_i in the second category, respectively. Since the orthogonal transformation of (5) and the linear transformation of (7) and (8) cause the mean to be 0 and the variance to be 1 for each of y_i in the first category, the relative importance of each composite predictor can be easily analyzed by the same criterion.

Inserting the values in Table 1 into Eq. (14), we obtain $D(y_i)$ with $i=1, 2, 3, 4, 5$, i. e., the quadratic discriminant function that makes use of composite predictor y_i separately.

The composite quadratic discriminant function can also be obtained. Table 2 shows the accuracy of these discriminant functions for the samples.

Table 1. Values of m_i and λ_i in Order of $(\mathrm{div})_i$

Composite Predictors y_i	m_i	i	$(\mathrm{div})_i$
y_1	3.550	10.574	11.232
y_2	−0.050	0.166	2.100
y_3	−1.001	0.740	1.223
y_4	−0.203	0.260	1.156
y_5	0.088	2.182	0.326

Table 2. Accuracy of Discriminant Functions

	Quadratic Discriminant Functions	Accuracy	Ordinal Number of Wrongly-Discriminated Samples
Single	$D(y_1)$	19/22	2, 5, 21
	$D(y_2)$	18/22	7, 9, 4, 6
	$D(y_3)$	17/22	2, 9, 17, 21, 22
	$D(y_4)$	13/22	7, 8, 9, 10, 12, 4, 5, 13, 11
	$D(y_5)$	14/22	2, 8, 1, 5, 6, 13, 21
Composite	$D(y_1)+D(y_2)$	19/22	2, 5, 21
	$D(y_1)+D(y_2)+D(y_3)$	20/22*	5, 21
	$D(y_1)+D(y_2)+D(y_3)+D(y_4)$	20/22	5, 21
	$D(y_1)+D(y_2)+D(y_3)+D(y_4)+D(y_5)$	20/22	5, 21

* The optimum quadratic discriminant function.

Tables 1 and 2 show:

(1) The great difference between the two categories in both the mean and the variance of the composite predictor y_1 provides significant category information. The great differences between the two categories in the mean of the composite predictor y_2 and the small difference in the variance can also give significant category information. In contrast to y_2, y_3 shows the small difference in variance and the great difference in the mean.

(2) It is seen from the single discriminant function that the accuracy of $D(y_i)$ is progressively decreased from $i=1$ to 5.

(3) The discrimination accuracy is 19/22 when the composite discriminant function $D(y_1)+D(y_2)$ is used, and 20/22 when $D(y_1)+D(y_2)=D(y_3)$ used. Generally speaking, from then on, the accuracy does not increase any more, even though more y_i is added. Thus, $D(y_1)+D(y_2)+D(y_3)$ is considered to be the optimum quadratic discriminant function.

Comparison is made between this method and that of stepwise discrimination. Table 3 shows the result from the same data in North China during June through August by stepwise multiple discrimination.

Table 3. Accuracy of Stepwise Multiple Discrimination

F Standard	Positive Predictor Introduced	Accuracy	Ordinal Number of Wrongly-Discriminated Samples
3.0	$(+)x_5$, $(+)x_1$	16/22	1, 2, 5, 17, 21
1.0	$(+)x_2$	17/22	2, 3, 5, 21
0.0	$(+)x_4$ $(+)x_3$	18/22	1, 5, 17, 21

It is seen that:

(1) When the standard of significance tests F is equal to 3.0, 1.0, 0.0, the accuracy by stepwise discrimination is 16/22, 17/22, 18/22, respectively, worse than that of the optimum quadratic discriminant and even worse than that of $D(y_1)$.

(2) Nos. 5 and 21 in the sample are classified into the wrongly-discriminated samples either by this method or by that of stepwise discrimination. However, Nos. 1 and 17, which are wrongly-discriminated by stepwise one, can be corrected by the quadartic discriminant function. Thus, when a new sample is classified, the transformation discriminant function is better than stepwise discrimination.

We have used this method for predicting the rainfall in the middle and lower Changjiang River basin during June. The June mean monthly rainfall at the five stations of Shanghai Nanjing, Wuhu, Jiujiang and Hankou is selected as a predictant and three predictors selected from experience in forecasting and by statistical standard, i. e., x_1, mean monthly meridional circulation index in September of the previous year in Asia; x_2, mean monthly zonal circulation index in January of the current year in Eurasia; and x_3, mean height minus 5,000 geopotential meters at the five stations along 85°N from 150°E to 170°W in October of the previous year.

The data used in this study cover 29 years from 1952 through 1980, $n=29$, $p=3$. The first category is defined as drought and the second as flood. The June rainfall data are divided into two categories: drought (less than 180 mm) and flood (more than 180 mm). Through calculation we obtain the transformation matrix

$$\mathrm{A}=\begin{bmatrix} 4.1673 & 0.5290 & 0.3178 \\ 2.1476 & 9.6423 & 0.0765 \\ -8.8326 & 2.5951 & 0.0807 \end{bmatrix}.$$

Calculating $(\mathrm{div})_i$, $(i=1, 2, 3)$ from (17), we have $\lambda_1=0.403$, $m_1=1.290$; $\lambda_2=0.470$, $m_2=0.726$; and $\lambda_3=3.072$, $m_3=0.644$. From (13) and (14), we obtain the discriminant function

$$\begin{aligned} D(y_1)&=0.74y_1^2-3.20y_1+1.61, \\ D(y_2)&=0.563y_2^2-1.545y_2+0.183, \\ D(y_3)&=-0.3373y_3^2-0.210y_3+0.6287. \end{aligned} \tag{18}$$

By substituting the data from 1952 through 1980 into (18), the optimum composite discriminant function is obtained.

If $D(y)=D(y_1)+D(y_2)\geqslant -0.3$, drought (the first category) is expected in June and if $D(y)=D(y_1)+D(y_2)<-0.3$, flood (the second category) is expected in June. The accuracy of the sample fitting is 27/29. Error arise only for 1970 and 1973, when no severe drought or flood occurred. The results have proved to be quite satisfactory.

For 1981, $\mathrm{X}=(0.64\ \ 0.76\ \ 20)^T$. Using linear transformation (7) $\mathbf{Y}=\mathbf{A}(\mathbf{X}-\boldsymbol{\mu}(1))$, we

obtain $\mathbf{Y}=(-0.49\ 1.520\ 1.658)^T$. Substituting these values into the discriminant function, we have $D(y)=D(y_1)-D(y_2)=2.493$, which is larger than -0.3. Thus, we forecast drought in June and drought was indeed observed.

For 1982, $\mathbf{X}=(0.50\ 0.81\ 33.6)^T$. Through calculation we obtain $D(y)=-1.634$, which is smaller than -0.3. Thus, we forecast flood in June and flood was also observed.

For 1983, $\mathbf{X}=(0.48\ 0.71\ 23.6)^T$. Through calculation we obtain $D(y)=-1.7748$, which is smaller than -0.3. Thus, we forecast flood in June and flood was observed too.

V. SUMMARY AND DISCUSSION

(1) The advantage of using the quadratic discriminant function through orthogonal transformation is that the method has simple function forms and can be used to optimize the discriminant functions. The above examples show that it has better effects than the linear and stepwise discriminations.

(2) By using the criterion of the Kullback divergence, the discriminant efficiency of the predictor can be estimated in the case of nonlinearity. If there is only one predictor x_1, the value of divergence becomes

$$(\mathrm{div})_{x_1}=\frac{1}{2\sigma^2_{x_1}(1)\sigma^2_{x_1}(2)}\left\{[\sigma^2_{x_1}(1)-\sigma^2_{x_1}(2)]^2-[\mu_{x_1}(1)-\mu_{x_1}(2)]^2[\sigma^2_{x_1}(1)+\sigma^2_{x_1}(2)]\right\}, \tag{19}$$

which can be used in the single predictor analysis. On the supposition that $\sigma^2_{x_1}(1)=\sigma^2_{x_1}(2)$ in (19), the equation becomes the calculation formula of the Mahalalobis distance for the single predictor. It should be pointed out that the conclusion characterizing the discriminant efficiency of the predictor may not agree with the case of nonlinearity. For example, as $\mu_{x_1}(1)=\mu_{x_1}(2)$, the Mahalanobis distance of the predictor is equal to zero, and in the case of linearity, x_1 provides no information. However, when there is a marked difference in the variance of the predictor x_1 for the second category $(\mathrm{div})_{x_1}=\frac{1}{2}\frac{1}{\sigma^2_{x_1}(1)\sigma^2_{x_1}(2)}(\sigma^2_{x_1}(1)-\sigma^2_{x_1}(2))^2$, which is a large value, and x_1 is still considered to have the discriminant efficiency in the case of nonlinearity. It is seen that when $(x_1-\bar{x}_1)^2$ is large, the first category is easily expected; when $(x_1-\bar{x}_1)^2$ is small, the second category is easily expected. If the critical value is properly selected, $(x_i-\bar{x}_i)^2$ can be used as the discriminant function. Therefore, the criterion of the Kullback divergence has extended the Mahalanobis distance, a new index, which can be used for measuring the discriminant efficiency.

(3) The prerequisite for the quadratic discrimination through orthogonal transformation is that the covariance matrix should be positively definite, that is $|\boldsymbol{\Sigma}(1)|\neq 0$. This condition can generally be satisfied except that a predictor is a linear combination of some other predictors. In the latter case, $|\boldsymbol{\Sigma}(1)|=0$. We have made experiments on this case by putting in a predictor composed of the linear combination. In the process of calculation an overflow occurs. By removing the overflowed predictor, the transformation matrix can be easily determined. However, it should be noted that when a great number of predictors are used and there are not many examples in the sample, $\boldsymbol{\Sigma}(1)$ may approximate to 0, thus making calculation unstable. This situation, of course, does not happen when stepwise linear discrimination is used. When a large number of predictors have to be used ($p>20$), it is better to roughly select the predictors by stepwise regression or stepwise discrimination under very low standard F of significance tests so as to sift out some predictors.

of the linear combination. In this way the above-mentioned situation can certainly be avoided when this method is used.

(4) The critical value of the discriminant function is theoretically 0 (see Eqs. 2 and 4). However, as $P(\mathbf{X}/\omega_1)$, $P(\mathbf{X}/\omega_2)$ may randomly deviate from normal distribution, we may select a value approaching to 0 as the critical value on the principle of the fewest erroneous examples.

(5) This method can be extended to the discriminant analysis of the G ($G>2$) category. For example, in order to discriminate $G=3$, the discriminant function should be

$$D_{ij}(\mathbf{X})=\ln\frac{P(\mathbf{X}/\omega_i)}{P(\mathbf{X}/\omega_j)},\quad i\neq j,\ i,j\in G. \tag{20}$$

By the analogous transformation

$$\mathbf{X}-\boldsymbol{\mu}(i);\quad \mathbf{A}\boldsymbol{\Sigma}(i)\mathbf{A}^T=\mathbf{I};\quad \mathbf{A}\boldsymbol{\Sigma}(j)\mathbf{A}^T=\boldsymbol{\Lambda};$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix, whose diagonal elements are given by the roots of the following equation

$$|\boldsymbol{\Sigma}(j)-\lambda\boldsymbol{\Sigma}(i)|=0, \tag{21}$$

the discriminant function has the form

$$D_{ij}(\mathbf{Y})=\sum_{k=1}^{p}D(y_k)=\frac{1}{2}\sum_{k=1}^{p}\left[\ln\lambda_k-y_k^2+\frac{1}{\lambda_k}(y_k^2+m_k^2-2m_ky_k)\right], \tag{22}$$

and $D_{ij}(y)=-D_{ji}(y)$. When $D_{12}\geqslant 0$, and $D_{13}\geqslant 0$, $\mathbf{X}$ is expected in the first category; when $D_{21}\geqslant 0$ and $D_{23}\geqslant 0$, X is expected in the second category; and when $D_{31}\geqslant 0$ and $D_{32}\geqslant 0$, $\mathbf{X}$ is expected in the third category.

Sincere thanks are due to Mr. Liu Guifu for his kind help in writing this article and also for his demonstration of some formulas.

REFERENCES

[1] Тер-Мкртчян М.Г. и.д., *Использование Дискриминатного Анализа для Прогноза Гололеда*, Тр. ГМЦ, вып. 90, 1971.
[2] 施能，气象分类预报中的二次判别法，气象，1979：7.
[3] Fukunaga, K., *Introduction to Statistical Pattern Recognition*, Academic Press, 1972.
[4] Абшаев М.А., и д., *Способ Прогноза Града на Северном Кавказе Методом квадратичного Дискриминатного Анализа*, Метерология и Гидрология, 1975, 8.
[5] 复旦大学，概率论，第二册，数理统计（第二分册），人民教育出版社，1979.
[6] 中国科学院地质研究所，数学地质引论，地质出版社，1977.