VI-1-3. Discriminant analysis.

We judge everything using information obtained in previous experiences in our daily life. We predict the weather tomorrow from today's temperature, air pressure, moisture, wind and so on comparing to previous data of one day before fine day or rainy day. We unconsciously have a threshold of a score compiling data of temperature, air pressure, moisture wind and so on. This a kind of discriminant analysis. We are discriminating todays data to one day before find day or on day before not fine day. Discriminant analysis is operation to make score for judgement for discrimination using previous data which we know the final event. Simplest discriminant analysis is linear discriminant analysis. In linear discriminant analysis, the discrimination score is made as linear combination of variables. It hypothesizes equality of variances between subpopulation such as one day before fine day and one day before not fine day among variables. Quadratic determinant analysis is expansion of linear discriminant analysis removing hypothesis of equality of variance among variables. Mixture discriminant analysis is computerized discriminant analysis using EM algorithm. It used sum of weighted probability of subpopulation as mixed normal distribution. Here, the author explains linear discriminant analysis and quadratic determinant analysis.

VI-1-3-2. Nature of discrimination score.

There is a set of previous data obtained from target population which include subpopulations 1, 2,$\cdots, k, \cdots, m$. We want to estimate subpopulation in which new sample belongs from variables of the sample. For this purpose, we make discrimination score as an index for judgement of subpopulation in which the sample belongs. Most simple method to make score is linear combination of variables.

$$a_1x_1 + a_2x_2 + \cdots + a_px_p = z$$

z: discrimination socore

$$\boldsymbol{A} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix}, \qquad \boldsymbol{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix},$$

$$Z = (a_1 \quad a_2 \quad \cdots \quad a_p)\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} = \boldsymbol{A}^T\boldsymbol{X}$$

Space geometrically, following equation is formula of hyperplane include origin of coordinate.

$$a_1x_1 + a_2x_2 + \cdots + a_px_p = 0$$

Denoting a data as d_{ki}

$$k = \text{subpopulation number } (1, \cdots, \text{m})$$

i: sample number in the group, $(1, \cdots, n_k)$

$$d_{ki} = \begin{pmatrix} x_{ki1} \\ \vdots \\ x_{kip} \end{pmatrix}$$

From the relation of plane and point

$$a_1 x_{ki1} + a_2 x_{ki2} + \cdots + a_p x_{ki1} = Z_{ki}$$

Z_{ki} is distance between the point and the hyperplane which includes origin of coordinate. In another word, $Z_{ki}$is projection of the point to the normal line of the hyperplane. (See figure 67).

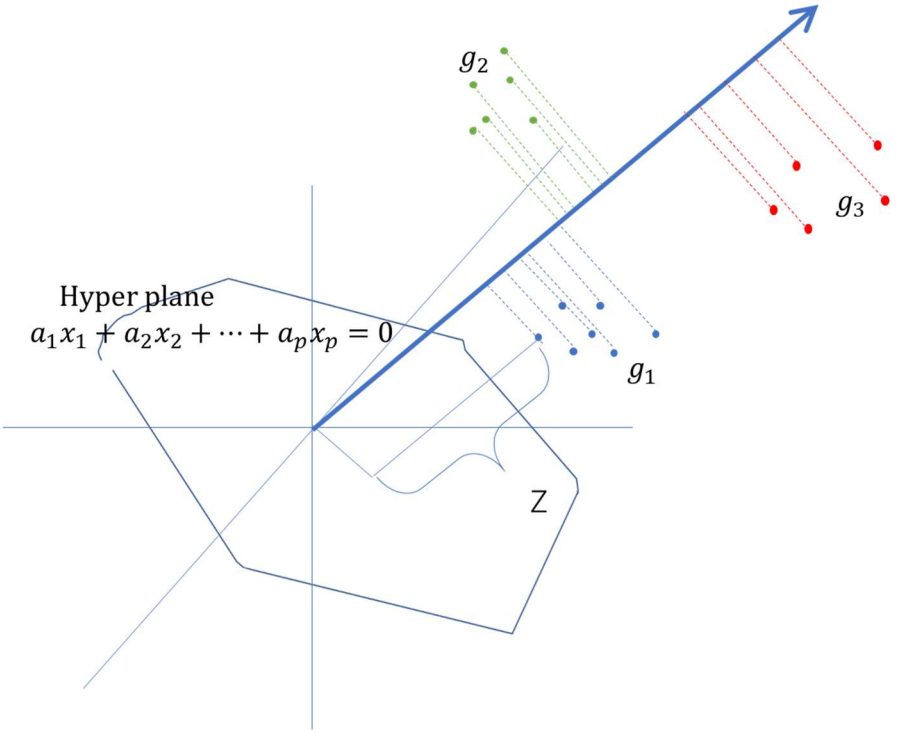

Fig.67 relation between discrimination score and normal line of hyperplane

The projections of data points distribute on the normal line in the same variance. This is the model of the linear discriminant analysis.

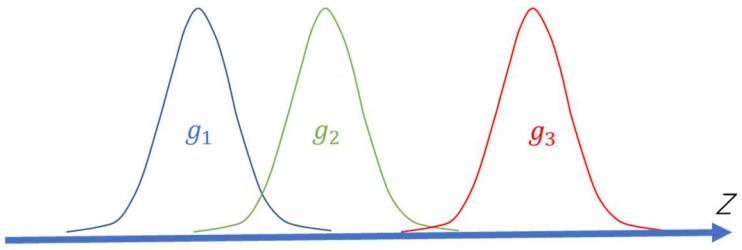

Fig.68 Distribution of Z (discrimination score) on the normal line.

VI-1-3-3. Analysis of variance of discrimination score.

There are various solutions in linear discriminant analysis, when we consider differences in detail. However, they can be categorized to two main approaches. One is analysis of variance of discrimination score. The other is linear algebraic procedure. Here, the author introduces method of ANOVA at first, and then introduces linear algebraic approach.

Actual data is as follow

Subpopulation	data number	data	variables		
			i	$\cdots$	p
1	1	$\boldsymbol{d_{11}}$	$= (\; d_{111}$	...	$d_{11p})^{\boldsymbol{T}}$
1	2	$\boldsymbol{d_{12}}$	$= (\; d_{121}$	...	$d_{12p})^{\boldsymbol{T}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
1	n_1	$\boldsymbol{d_{1n_1}}$	$= (\; d_{1n_11}$	...	$d_{1n_1p})^{\boldsymbol{T}}$
Sum subp 1		$\sum_{i=1}^{\boldsymbol{n_1}} \boldsymbol{d_{1i}}$	$=(\sum_{i=1}^{n_1} d_{1i1}$	$\cdots$	$\sum_{i=1}^{n_1} d_{1ip})^T$
Average subp1		$\overline{\boldsymbol{d_1}}$	$= \left(\frac{\sum_{i=1}^{n_1} d_{1i1}}{n_1} = \overline{d_{11}}\right.$	...	$\left.\frac{\sum_{i=1}^{n_1} d_{1ip}}{n_1} = \overline{d_{1p}}\right)^T$
2	1	$\boldsymbol{d_{21}}$	$= (\; d_{211}$	...	$d_{21p})^{\boldsymbol{T}}$
2	2	$\boldsymbol{d_{22}}$	$= (\; d_{221}$	...	$d_{22p})^{\boldsymbol{T}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
2	n_2	$\boldsymbol{d_{2n_1}}$	$= (\; d_{2n_11}$	...	$d_{2n_1p})^{\boldsymbol{T}}$
Sum subp 2		$\sum_{i=1}^{\boldsymbol{n_2}} \boldsymbol{d_{2i}}$	$=(\sum_{i=1}^{n_2} d_{2i1}$	$\cdots$	$\sum_{i=1}^{n_2} d_{2ip})^T$
Average subp2		$\overline{\boldsymbol{d_2}}$	$= \left(\frac{\sum_{i=1}^{n_2} d_{2i1}}{n_2} = \overline{d_{21}}\right.$	...	$\left.\frac{\sum_{i=1}^{n_2} d_{2ip}}{n_2} = \overline{d_{2p}}\right)^T$
			$\vdots$		
m	1	$\boldsymbol{d_{m1}}$	$= (\; d_{m11}$	...	$d_{m1p})^{\boldsymbol{T}}$
m	2	$\boldsymbol{d_{m2}}$	$= (\; d_{m21}$	...	$d_{m2p})^{\boldsymbol{T}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
m	n_m	$\boldsymbol{d_{mn_m}}$	$= (\; d_{mn_11}$	...	$d_{mn_1p})^{\boldsymbol{T}}$
Sum subp m		$\sum_{i=1}^{\boldsymbol{n_m}} \boldsymbol{d_{2i}}$	$=(\sum_{i=1}^{n_m} d_{mi1}$	$\cdots$	$\sum_{i=1}^{n_m} d_{mip})^T$
Average subp m		$\overline{\boldsymbol{d_m}}$	$= \left(\frac{\sum_{i=1}^{n_m} d_{mi1}}{n_m} = \overline{d_{m1}}\right.$	...	$\left.\frac{\sum_{i=1}^{n_m} d_{mip}}{n_m} = \overline{d_{mp}}\right)^T$
Total sum	$N = \sum_{k=1}^{m} n_m$	$\sum_{k=1}^{m}\sum_{i=1}^{\boldsymbol{n_k}} \boldsymbol{d_{ki}}$	$=(\sum_{k=1}^{m}\sum_{i=1}^{n_k} d_{ki}$	$\cdots$	$\sum_{k=1}^{m}\sum_{i=1}^{n_k} d_{ki1})^T$
Total average		$\frac{\sum_{k=1}^{m}\sum_{i=1}^{n_k} d_{ki}}{N} = \overline{\overline{\boldsymbol{d}}}$	$=\left(\frac{\sum_{k=1}^{m}\sum_{i=1}^{n_k} d_{ki1}}{N} = \overline{d_1}\right.$	...	$\left.\frac{\sum_{k=1}^{m}\sum_{i=1}^{n_k} d_{ki}}{N} = \overline{d_p}\right)^T$

Total average: $\bar{\bar{\boldsymbol{d}}}$

Average of subpopulation: $\bar{d}_{\iota}$, $(i = 1, \cdots m)$

For simplification, we trans form all the data as distance from total average

$$x_{kij} = d_{kij} - \bar{d}_j$$

Transformed data set is as follow

Subpopulation	data number	data	variables		
			i	$\cdots$	p
1	1	$\boldsymbol{x_{11}}$	$= (x_{111}$	$\cdots$	$x_{11p})^{\boldsymbol{T}}$
1	2	$\boldsymbol{x_{12}}$	$= (x_{121}$	$\cdots$	$x_{12p})^{\boldsymbol{T}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
1	n_1	$\boldsymbol{x_{1n_1}}$	$= (x_{1n_11}$	$\cdots$	$x_{1n_1p})^{\boldsymbol{T}}$
Sum subp 1		$\sum_{i=1}^{\boldsymbol{n_1}} \boldsymbol{x_{1i}}$	$=(\sum_{i=1}^{n_1} x_{1i1}$	$\cdots$	$\sum_{i=1}^{n_1} x_{1ip})^T$
Average subp1		$\overline{\boldsymbol{x_1}}$	$= \left(\frac{\sum_{i=1}^{n_1} x_{1i1}}{n_1} = \overline{x_{11}}\right.$	$\cdots$	$\left.\frac{\sum_{i=1}^{n_1} x_{1ip}}{n_1} = \overline{x_{1p}}\right)^T$
2	1	$\boldsymbol{x_{21}}$	$= (x_{211}$	$\cdots$	$x_{21p})^{\boldsymbol{T}}$
2	2	$\boldsymbol{x_{22}}$	$= (x_{221}$	$\cdots$	$x_{22p})^{\boldsymbol{T}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
2	n_2	$\boldsymbol{x_{2n_1}}$	$= (x_{2n_11}$	$\cdots$	$x_{2n_1p})^{\boldsymbol{T}}$
Sum subp 2		$\sum_{i=1}^{\boldsymbol{n_2}} \boldsymbol{x_{2i}}$	$=(\sum_{i=1}^{n_2} x_{2i1}$	$\cdots$	$\sum_{i=1}^{n_2} x_{2ip})^T$
Average subp2		$\overline{\boldsymbol{x_2}}$	$= \left(\frac{\sum_{i=1}^{n_2} x_{2i1}}{n_2} = \overline{x_{21}}\right.$	$\cdots$	$\left.\frac{\sum_{i=1}^{n_2} x_{2ip}}{n_2} = \overline{x_{2p}}\right)^T$
			$\vdots$		
m	1	$\boldsymbol{x_{m1}}$	$= (x_{m11}$	$\cdots$	$x_{m1p})^{\boldsymbol{T}}$
m	2	$\boldsymbol{x_{m2}}$	$= (x_{m21}$	$\cdots$	$x_{m2p})^{\boldsymbol{T}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
m	n_m	$\boldsymbol{x_{mn_m}}$	$= (x_{mn_11}$	$\cdots$	$x_{mn_1p})^{\boldsymbol{T}}$
Sum subp m		$\sum_{i=1}^{\boldsymbol{n_m}} \boldsymbol{x_{2i}}$	$=(\sum_{i=1}^{n_m} x_{mi1}$	$\cdots$	$\sum_{i=1}^{n_m} x_{mip})^T$
Average subp m		$\overline{\boldsymbol{x_m}}$	$= \left(\frac{\sum_{i=1}^{n_m} x_{mi1}}{n_m} = \overline{x_{m1}}\right.$	$\cdots$	$\left.\frac{\sum_{i=1}^{n_m} x_{mip}}{n_m} = \overline{x_{mp}}\right)^T$
Total sum	N	$\sum_{k=1}^{m}\sum_{i=1}^{\boldsymbol{n_k}} \boldsymbol{x_{ki}}$	$= (0$	$\cdots$	$0)^T$
Total average		$\bar{\bar{\boldsymbol{x}}}$	$= (0$	$\cdots$	$0)^T$

Here, we are assuming equality of variances among groups. What we are requested is to fined

optimum line to project data which emphasizes differences among averages in each subpopulation. This is maximization of F ratio variance among averages of subpopulation and variance in each subpopulation.

We consider average of all data and average of each group to make deviation of each data.

From

$$Z_{ki} = \boldsymbol{A}^T \boldsymbol{x_{ki}} = \left(a_1 x_{ki1} + a_2 x_{ki2} + \cdots + a_p x_{kip}\right)$$

Total average is

$$\bar{\bar{Z}} = \boldsymbol{A}^T \bar{\bar{\boldsymbol{x}}} = \boldsymbol{0}$$

Averages in each subpopulation are

$$\overline{Z_k} = \boldsymbol{A}^T \overline{\boldsymbol{x}_k} = \left(a_1 \overline{x_{k1}} + a_2 \overline{x_{k2}} + \cdots + a_p \overline{x_{kip}}\right) = \frac{1}{n_k} \sum_{i=1}^{n_k} \left(a_1 x_{ki1} + a_2 x_{ki2} + \cdots + a_p x_{kip}\right)$$

$$= (a_1 \quad \cdots \quad a_p) \begin{pmatrix} \frac{1}{n_k} \sum_{i=1}^{n_k} x_{ki} \\ \vdots \\ \frac{1}{n_k} \sum_{i=1}^{n_k} x_{kip} \end{pmatrix}$$

Then we separate data to two parts. One is average of each subpopulation and the other is difference from average of each subpopulation.

$$Z_{ki} = \overline{Z_k} + e_{ki}$$

$$e_{ki} = Z_{ki} - \overline{Z_k} = \boldsymbol{A}^T (\boldsymbol{x_{ki}} - \overline{\boldsymbol{x}_k}) = \boldsymbol{A}^T \begin{pmatrix} x_{ki} - \frac{1}{n_k} \sum_{i=1}^{n_k} x_{ki1} \\ \vdots \\ x_{kip} - \frac{1}{n_k} \sum_{i=1}^{n_k} x_{kip} \end{pmatrix}$$

We calculate Sum of square

$$SS_{total} = \sum_{k=1}^{m} \sum_{i=1}^{n_k} {Z_{ki}}^2 = \sum_{k=1}^{m} \sum_{i=1}^{n_k} (e_{ki} + \overline{Z_k})^2$$

$$= \sum_{k=1}^{m} \sum_{i=1}^{n_k} {e_{ki}}^2 + 2 \sum_{k=1}^{m} \overline{Z_k} \sum_{i=1}^{n_k} e_{ki} + \sum_{k=1}^{m} \sum_{i=1}^{n_k} \overline{Z_k}^2$$

$$= \sum_{k=1}^{m} \sum_{i=1}^{n_k} {e_{ki}}^2 + \sum_{k=1}^{m} \sum_{i=1}^{n_k} \overline{Z_k}^2$$

$$\left(\because \sum_{i=1}^{n_k} e_{ki} = 0 \right)$$

$$SS_{total} = SS_{residual} + SS_{subpopulatoin}$$

Sum of square of differences from average $\sum_{i=1}^{n_k} {e_{ki}}^2$ is sum of square in subpopulation. So, we can express that as SS_k.

$$\sum_{i=1}^{n_k} {e_{ki}}^2 = SS_k$$

$$SS_{rsidual} = \sum_{k=1}^{m} SS_k = \sum_{k=1}^{m} (n_k - 1)\,{\sigma^2}_{residulal} = \sum_{k=1}^{m}\sum_{i=1}^{n_k} {e_{ki}}^2$$

$$\sum_{k=1}^{m} (n_k - 1) : degree\ of freedom\ of\ residual$$

$$SS_{total} = \sum_{k=1}^{m} SS_k + \sum_{k=1}^{m}\sum_{i=1}^{n_k} {\overline{Z_k}}^2$$

$$SS_{subpopulatoin} = \sum_{k=1}^{m}\sum_{i=1}^{n_k} {\overline{Z_k}}^2$$

$$SS_{total} = SS_{residual} + SS_{subpopulatoin}$$

$${\sigma^2}_{subpopulation} = \frac{1}{(m-1)}\sum_{k=1}^{m}\sum_{i=1}^{n_k} {\overline{Z_k}}^2$$

$${\sigma^2}_{residulal} = \frac{\sum_{k=1}^{m} SS_k}{\sum_{k=1}^{m}(n_k - 1)}$$

$${\sigma^2}_{subpopulation} = \frac{SS_{subpopulatoin}}{m-1}$$

Ratio of variance is F

$$F = \frac{{\sigma^2}_{subpopulation}}{{\sigma^2}_{residulal}} = \frac{SS_{subpopulatoin}}{m-1}\frac{\sum_{k=1}^{m}(n_k - 1)}{\sum_{k=1}^{m} SS_k} = \left(\frac{\sum_{k=1}^{m}(n_k - 1)}{m-1}\right)\left(\frac{SS_{subpopulatoin}}{SS_{residual}}\right)$$

$\left(\frac{\sum_{k=1}^{m}(n_k - 1)}{m-1}\right)$ is constant and has norelation with discriminationscore in this case

We consider maximization of $\left(\frac{SS_{subpopulatoin}}{SS_{residual}}\right)$ for maximization of F, and $\left(\frac{SS_{subpopulatoin}}{SS_{residual}}\right)$ is function of $\boldsymbol{A}$. So, we denote as follow

$$f(\boldsymbol{A}) = \left(\frac{SS_{subpopulatoin}}{SS_{residual}}\right)$$

Firstly, we consider sum of square of subpopulations.

$$SS_{subpopulatoin} = \sum_{k=1}^{m}\sum_{i=1}^{n_k} {\overline{Z_k}}^2$$

$$\overline{Z_k} = \boldsymbol{A}^T \overline{\boldsymbol{x_k}}$$

$$\overline{\boldsymbol{x_k}} = \begin{pmatrix} \overline{x_k}_1 \\ \overline{x_k}_2 \\ \vdots \\ \overline{x_k}_p \end{pmatrix} = \begin{pmatrix} \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ki1} \\ \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ki2} \\ \vdots \\ \frac{1}{n_k}\sum_{i=1}^{n_k} x_{kip} \end{pmatrix}$$

For simplification we as follow

$$\mu_{kj} = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{kij}$$

$$\boldsymbol{\mu_k} = \begin{pmatrix} \mu_{k1} \\ \mu_{k2} \\ \vdots \\ \mu_{kp} \end{pmatrix}$$

$$\overline{\boldsymbol{Z_k}} = \boldsymbol{A^T}\overline{\boldsymbol{x_k}} = \boldsymbol{A^T\mu_k}$$

$$(\overline{\boldsymbol{Z_k}})^2 = \left(\boldsymbol{A^T}(\overline{\boldsymbol{X_k}})\right)^2 = \boldsymbol{A^T\mu_k}(\boldsymbol{A^T\mu_k})^T = \boldsymbol{A^T\mu_k\mu_k{}^T A}$$

$$\boldsymbol{A^T\mu_k} = (a_1 \quad a_2 \quad \cdots \quad a_p)\begin{pmatrix} \mu_{k1} \\ \mu_{k2} \\ \vdots \\ \mu_{kp} \end{pmatrix}$$

$$\boldsymbol{\mu_k{}^T A} = (\mu_{k1} \quad \mu_{k2} \quad \cdots \quad \mu_{kp})\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix}$$

$$\boldsymbol{\mu_k\mu_k{}^T} = \begin{pmatrix} {\mu_{k1}}^2 & \mu_{k1}\mu_{k2} & \cdots & \mu_{k1}\mu_{k2} \\ \mu_{k2}\mu_{k1} & {\mu_{k2}}^2 & \cdots & \mu_{k2}\mu_{kp} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{kp}\mu_{k1} & \mu_{kp}\mu_{k2} & \cdots & {\mu_{kp}}^2 \end{pmatrix}$$

$$\boldsymbol{A^T\mu_k\mu_k{}^T A} = \boldsymbol{A^T}\begin{pmatrix} {\mu_{k1}}^2 & \mu_{k1}\mu_{k2} & \cdots & \mu_{k1}\mu_{k2} \\ \mu_{k2}\mu_{k1} & {\mu_{k2}}^2 & \cdots & \mu_{k2}\mu_{kp} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{kp}\mu_{k1} & \mu_{kp}\mu_{k2} & \cdots & {\mu_{kp}}^2 \end{pmatrix}\boldsymbol{A}$$

$$\boldsymbol{A^T\mu_k\mu_k{}^T A + A^T\mu_l\mu_l{}^T A} = \boldsymbol{A^T}(\boldsymbol{\mu_k\mu_k{}^T + \mu_l\mu_l{}^T})\boldsymbol{A}$$

$$\sum_{k=1}^{m}(\overline{Z_k})^2 = \boldsymbol{A^T}\left(\sum_{k=1}^{m}(\boldsymbol{\mu_k\mu_k{}^T + \mu_l\mu_l{}^T})\right)\boldsymbol{A}$$

$$= \boldsymbol{A}^T \begin{pmatrix} \sum_{k=1}^{m} {\mu_{k1}}^2 & \sum_{k=1}^{m} \mu_{k1}\mu_{k2} & \cdots & \sum_{k=1}^{m} \mu_{k1}\mu_{kp} \\ \sum_{k=1}^{m} \mu_{k2}\mu_{k1} & \sum_{k=1}^{m} {\mu_{k2}}^2 & \cdots & \sum_{k=1}^{m} \mu_{k2}\mu_{kp} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^{m} \mu_{kp}\mu_{k1} & \sum_{k=1}^{m} \mu_{kp}\mu_{k2} & \cdots & \sum_{k=1}^{m} {\mu_{kp}}^2 \end{pmatrix} \boldsymbol{A}$$

$$= \boldsymbol{A}^T \begin{pmatrix} \mu_{11} & \mu_{21} & \cdots & \mu_{m1} \\ \mu_{12} & \mu_{22} & \cdots & \mu_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{1p} & \mu_{2p} & \cdots & \mu_{mp} \end{pmatrix} \begin{pmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1p} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{m1} & \mu_{m2} & \cdots & \mu_{mp} \end{pmatrix} \boldsymbol{A}$$

From this, we can understand that $(\sum_{k=1}^{m}(\boldsymbol{\mu_k}\boldsymbol{\mu_k}^{\boldsymbol{T}} + \boldsymbol{\mu_l}\boldsymbol{\mu_l}^{\boldsymbol{T}}))$ is variance covariance matrix. We denote the variance covariance matrix as follow.

$$\boldsymbol{M} = \sum_{k=1}^{m} \boldsymbol{\mu_k}\boldsymbol{\mu_k}^{\boldsymbol{T}} = \begin{pmatrix} \sum_{k=1}^{m} {\mu_{k1}}^2 & \sum_{k=1}^{m} \mu_{k1}\mu_{k2} & \cdots & \sum_{k=1}^{m} \mu_{k1}\mu_{k2} \\ \sum_{k=1}^{m} \mu_{k2}\mu_{k1} & \sum_{k=1}^{m} {\mu_{k2}}^2 & \cdots & \sum_{k=1}^{m} \mu_{k2}\mu_{kp} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^{m} \mu_{kp}\mu_{k1} & \sum_{k=1}^{m} \mu_{kp}\mu_{k2} & \cdots & \sum_{k=1}^{m} {\mu_{kp}}^2 \end{pmatrix}$$

$$SS_{subpopulatoin} = \sum_{k=1}^{m} (\overline{Z_k})^2 = \boldsymbol{A}^T\boldsymbol{M}\boldsymbol{A}$$

Then we consider sum of square of residuals.

$$SS_k = \sum_{i=1}^{n_k} {e_{ki}}^2 = \sum_{i=1}^{n_k} \left(\boldsymbol{A}^T(\boldsymbol{x_{ki}} - \overline{\boldsymbol{x_k}})\right)\left(\boldsymbol{A}^T(\boldsymbol{x_{ki}} - \overline{\boldsymbol{x_k}})\right)^T = \sum_{i=1}^{n_k} \boldsymbol{A}^T(\boldsymbol{x_{ki}} - \overline{\boldsymbol{x_k}})(\boldsymbol{x_{ki}} - \overline{\boldsymbol{x_k}})^T\boldsymbol{A}$$

$$= \boldsymbol{A}^T \left(\sum_{i=1}^{n_k} (\boldsymbol{x_{ki}} - \overline{\boldsymbol{x_k}})(\boldsymbol{x_{ki}} - \overline{\boldsymbol{x_k}})^T \right) \boldsymbol{A}$$

We denote as follow.

$$\boldsymbol{x_{ki}} - \overline{\boldsymbol{x_k}} = \begin{pmatrix} x_{ki1} - \dfrac{1}{n_k}\sum_{i=1}^{n_k} x_{ki} \\ \vdots \\ x_{k_{iip}} - \dfrac{1}{n_k}\sum_{i=1}^{n_k} x_{kip} \end{pmatrix} = \begin{pmatrix} \phi_{ki} \\ \vdots \\ \phi_{ki1} \end{pmatrix}$$

$$(\boldsymbol{x_{ki}} - \overline{\boldsymbol{x}_k})(\boldsymbol{x_{ki}} - \overline{\boldsymbol{x}_k})^T = \begin{pmatrix} \phi_{ki1} \\ \vdots \\ \phi_{kip} \end{pmatrix} (\phi_{ki1} \quad \cdots \quad \phi_{kip}) = \begin{pmatrix} {\phi_{ki1}}^2 & \phi_{ki1}\phi_{ki2} & \cdots & \phi_{ki1}\phi_{kip} \\ \phi_{ki2}\phi_{ki1} & {\phi_{ki2}}^2 & \cdots & \phi_{ki2}\phi_{kip} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{kip}\phi_{ki1} & \phi_{kip}\phi_{ki2} & \cdots & {\phi_{kip}}^2 \end{pmatrix}$$

$$\sum_{k=1}^{m}\sum_{i=1}^{n_k}(\boldsymbol{x_{ki}} - \overline{\boldsymbol{x}_k})(\boldsymbol{x_{ki}} - \overline{\boldsymbol{x}_k})^T = \begin{pmatrix} \sum_{k=1}^{m}\sum_{i=1}^{n_k}{\phi_{ki1}}^2 & \sum_{k=1}^{m}\sum_{i=1}^{n_k}\phi_{ki1}\phi_{ki2} & \cdots & \sum_{k=1}^{m}\sum_{i=1}^{n_k}\phi_{ki}\ \phi_{ki} \\ \sum_{k=1}^{m}\sum_{i=1}^{n_k}\phi_{ki2}\phi_{ki1} & \sum_{k=1}^{m}\sum_{i=1}^{n_k}{\phi_{ki2}}^2 & \cdots & \sum_{k=1}^{m}\sum_{i=1}^{n_k}\phi_{ki2}\phi_{kip} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^{m}\sum_{i=1}^{n_k}\phi_{kip}\phi_{ki1} & \sum_{k=1}^{m}\sum_{i=1}^{n_k}\phi_{kip}\phi_{ki2} & \cdots & \sum_{k=1}^{m}\sum_{i=1}^{n_k}{\phi_{kip}}^2 \end{pmatrix}$$

This is variance covariance matrix, and we dente the variance and covariance matrix as $\boldsymbol{V}$.

$$\boldsymbol{V} = \sum_{k=1}^{m}\sum_{i=1}^{n_k}(\boldsymbol{x_{ki}} - \overline{\boldsymbol{x}_k})(\boldsymbol{x_{ki}} - \overline{\boldsymbol{x}_k})^T$$

$$SS_{residual} = \sum_{k=1}^{m} SS_k = \boldsymbol{A}^T\boldsymbol{V}\boldsymbol{A}$$

$$f(\boldsymbol{A}) = \left(\frac{SS_{subpopulatoin}}{\sum_{k=1}^{m} SS_k}\right) = \frac{\boldsymbol{A}^T\boldsymbol{M}\boldsymbol{A}}{\boldsymbol{A}^T\boldsymbol{V}\boldsymbol{A}}$$

We solve maximization of $f(\boldsymbol{A})$ by partial differential equation.

$$f(\boldsymbol{A}) = \frac{\boldsymbol{A}^T\boldsymbol{M}\boldsymbol{A}}{\boldsymbol{A}^T\boldsymbol{F}\boldsymbol{A}}$$

$$\frac{df(\boldsymbol{A})}{d\boldsymbol{A}} = 0$$

Formula 75

We can separate variances depending of source of fluctuation, and we can get $\boldsymbol{A}$ as solution of simultaneous equation of the partial differential equation. $\boldsymbol{A}$ is a normal vector expressing gradient of a hyperplane. When we put coordinate of a data in the $\boldsymbol{x}$ of following equation. obtainable scalar means the distance (d) from origin of coordinate to the hyperplane which include point of $\boldsymbol{x}$. This is discriminant score, which is distance from the reference hyperplane, which include coordinate origin.

$$d = \boldsymbol{A}^T\boldsymbol{x} = \text{DS (discriminant score)}$$

Then, we should determine threshold of DS. However, there is no general theory for determination of the threshold, because preferable risk rate is different among issues depending on the judgment of analysts. In some case such as toxicity of chemicals, we have to judge in safety side nearly 0 risk. In another case such as betting of horse race, we need to be challenging to accept risks to lose money. One of neutral setting up of threshold is to select

hyperplane which include midpoint of centers of two subpopulation.

$$threshold\ between\ subpopulation\ A\ and\ B(p(A) = p(B)) = \frac{d_A + d_A}{2} = \boldsymbol{A^T}\left(\frac{\boldsymbol{x_A} + \boldsymbol{x_B}}{\boldsymbol{2}}\right)$$

In linear discriminant analysis, we hypothesize homoscedasticity among subpopulation. This means data distribution of subpopulation is the same shape and size. Thus, we can make the risk of miss-judgement when we select A and when we select B the same by this selection.
We could understand theory of discriminant analysis. However, operation of differential is often troublesome work particularly when $f(\boldsymbol{A})$ is complicated. We need to consider more simple operation. More essentially, we can understand discriminant analysis space geometrically through creation of simple and quick operation of discriminant analysis.

Exercise

We have dataset composed from 8 data. Among them, 4 data are belonging to subpopulation 1, and the others are belonging to subpopulation 2. Each data has 2 variables. We will produce discrimination score for identification of subpopulation of new data.

Dataset

		data	
subpopulation	sample No	d_1	d_2
1	1	5	8
1	2	7	4
1	3	8	5
1	4	8	7
2	1	5	5
2	2	7	2
2	3	4	3
2	4	4	6

Total average

$$\overline{\overline{d_1}} = \frac{5+7+8+8+5+7+4+4}{8} = \frac{48}{8} = 6$$

$$\overline{\overline{d_2}} = \frac{8+4+5+7+5+2+3+6}{8} = \frac{40}{8} = 5$$

$$\overline{\overline{\boldsymbol{d}}} = \begin{pmatrix} 6 \\ 5 \end{pmatrix}$$

Average of subpopulation1

$$\overline{d_{11}} = \frac{5+7+8+8}{4} = \frac{28}{4} = 7$$

$$\overline{d_{12}} = \frac{8+4+5+7}{4} = \frac{24}{4} = 6$$

$$\overline{\boldsymbol{d}_1} = \begin{pmatrix} 7 \\ 6 \end{pmatrix}$$

Average of subpopulation2

$$\overline{d_{21}} = \frac{5+7+4+4}{4} = \frac{20}{4} = 5$$

$$\overline{d_{22}} = \frac{5+2+3+6}{4} = \frac{16}{4} = 4$$

$$\overline{d_2} = \begin{pmatrix} 5 \\ 4 \end{pmatrix}$$

Deviation of each data from total average

$$\boldsymbol{x_{ki}} = \boldsymbol{d_{ki}} - \overline{\overline{\boldsymbol{d}}} = \begin{pmatrix} d_{ki1} \\ d_{ki2} \end{pmatrix} - \begin{pmatrix} 6 \\ 5 \end{pmatrix} = \begin{pmatrix} d_{ki1} - 6 \\ d_{ki} - 5 \end{pmatrix}$$

Deviation of average of subpopulation from total average.

$$\overline{\boldsymbol{d}_k} - \overline{\overline{\boldsymbol{d}}} = \begin{pmatrix} \overline{d_{k1}} \\ \overline{d_{k2}} \end{pmatrix} - \begin{pmatrix} 6 \\ 5 \end{pmatrix}$$

$$\boldsymbol{d_{ki}} - \overline{\boldsymbol{d}_k} = \begin{pmatrix} d_{ki1} \\ d_{ki} \end{pmatrix} - \begin{pmatrix} \overline{d_{k1}} \\ \overline{d_{k2}} \end{pmatrix}$$

$$\overline{\boldsymbol{d}_1} - \overline{\overline{\boldsymbol{d}}} = \begin{pmatrix} 7 \\ 6 \end{pmatrix} - \begin{pmatrix} 6 \\ 5 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \overline{\boldsymbol{d}_2} - \overline{\overline{\boldsymbol{d}}} = \begin{pmatrix} 5 \\ 4 \end{pmatrix} - \begin{pmatrix} 6 \\ 5 \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$$

$\boldsymbol{d_{11}} - \overline{\boldsymbol{d}_1} = \begin{pmatrix} 5 \\ 8 \end{pmatrix} - \begin{pmatrix} 7 \\ 6 \end{pmatrix} = \begin{pmatrix} -2 \\ 2 \end{pmatrix}$, $\boldsymbol{d_{12}} - \overline{\boldsymbol{d}_1} = \begin{pmatrix} 7 \\ 4 \end{pmatrix} - \begin{pmatrix} 7 \\ 6 \end{pmatrix} = \begin{pmatrix} 0 \\ -2 \end{pmatrix}$, $\boldsymbol{d_{13}} - \overline{\boldsymbol{d}_1} = \begin{pmatrix} 8 \\ 5 \end{pmatrix} - \begin{pmatrix} 7 \\ 6 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$, =
$\boldsymbol{d_{12}} - \overline{\boldsymbol{d}_1} \begin{pmatrix} 8 \\ 7 \end{pmatrix} - \begin{pmatrix} 7 \\ 6 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

$\boldsymbol{d_{21}} - \overline{\boldsymbol{d}_2} = \begin{pmatrix} 5 \\ 5 \end{pmatrix} - \begin{pmatrix} 5 \\ 4 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \boldsymbol{d_{22}} - \overline{\boldsymbol{d}_2}, = \begin{pmatrix} 7 \\ 2 \end{pmatrix} - \begin{pmatrix} 5 \\ 4 \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \end{pmatrix}, \boldsymbol{d_{23}} - \overline{\boldsymbol{d}_2} = \begin{pmatrix} 4 \\ 3 \end{pmatrix} - \begin{pmatrix} 5 \\ 4 \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$,

$\boldsymbol{d_{24}} - \overline{\boldsymbol{d}_2} = \begin{pmatrix} 4 \\ 6 \end{pmatrix} - \begin{pmatrix} 5 \\ 4 \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$

Variance and covariance matrix of deviation of center of subpopulation from total average

$$\boldsymbol{M} = \left(\overline{\boldsymbol{d}_1} - \overline{\overline{\boldsymbol{d}}} \quad \overline{\boldsymbol{d}_2} - \overline{\overline{\boldsymbol{d}}}\right)\left(\overline{\boldsymbol{d}_1} - \overline{\overline{\boldsymbol{d}}} \quad \overline{\boldsymbol{d}_2} - \overline{\overline{\boldsymbol{d}}}\right)^{\boldsymbol{T}}$$

$$= \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}$$

Variance and covariance matrix of deviation from center of subpopulation

$$\boldsymbol{V} = \begin{pmatrix} -2 & 0 & 1 & 1 & 0 & 2 & -1 & -1 \\ 2 & -2 & -1 & 1 & 1 & -2 & -1 & 2 \end{pmatrix} \begin{pmatrix} -2 & 2 \\ 0 & -2 \\ 1 & -1 \\ 1 & 1 \\ 0 & 1 \\ 2 & -2 \\ -1 & -1 \\ -1 & 2 \end{pmatrix}$$

$$= \begin{pmatrix} 4+0+1+1+0+4+1+1 & -4+0-1+1+0-4+1-2 \\ -4+0-1+1+0-4+1-2 & 4+4+1+1+1+4+1+4 \end{pmatrix} = \begin{pmatrix} 12 & -9 \\ -9 & 20 \end{pmatrix}$$

$$\boldsymbol{V} = SS_{ressidual} = \begin{pmatrix} 12 & -9 \\ -9 & 20 \end{pmatrix}$$

We consider maximization of $\frac{(\boldsymbol{A}^T \boldsymbol{M} \boldsymbol{A})}{(\boldsymbol{A}^T \boldsymbol{V} \boldsymbol{A})}$

$$g(\boldsymbol{A}) = \boldsymbol{A}^T \boldsymbol{M} \boldsymbol{A} = \boldsymbol{A}^t \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix} \boldsymbol{A} = 2\boldsymbol{A}^t \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \boldsymbol{A}$$

$$= 2(a_1 \quad a_2) \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = 2(a_1 + a_2 \quad a_1 + a_2) \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = 2({a_1}^2 + 2a_1 a_2 + {a_2}^2)$$

$$h(\boldsymbol{A}) = \boldsymbol{A}^T \boldsymbol{F} \boldsymbol{A} = \boldsymbol{A}^T \begin{pmatrix} 12 & -9 \\ -9 & 20 \end{pmatrix} \boldsymbol{A}$$

$$= (a_1 \quad a_2) \begin{pmatrix} 12 & -9 \\ -9 & 20 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

$$= (12a_1 - 9a_2 \quad -9a_1 + 20a_2) \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

$$= 12{a_1}^2 - 9a_1 a_2 - 9a_1 a_2 + 20{a_2}^2$$

$$= 12{a_1}^2 - 18a_1 a_2 + 20{a_2}^2$$

$$= 2(6{a_1}^2 - 9a_1 a_2 + 10{a_2}^2)$$

$$f(\boldsymbol{A}) = \frac{2({a_1}^2 + 2a_1 a_2 + {a_2}^2)}{2(6{a_1}^2 - 9a_1 a_2 + 10{a_2}^2)} = \frac{{a_1}^2 + 2a_1 a_2 + {a_2}^2}{6{a_1}^2 - 9a_1 a_2 + 10{a_2}^2}$$

$$\frac{\partial f(\boldsymbol{A})}{\partial a_1} = \frac{\partial \left(\frac{g(a_1)}{h(a_1)}\right)}{\partial a_1} = \frac{\frac{\partial g(a_1)}{\partial a_1} h(a_1) - g(a_1) \frac{\partial h(a_1)}{\partial a_1}}{h(a_1)^2}$$

$$g(a_1) = {a_1}^2 + 2a_1 a_2 + {a_2}^2$$

$$h(a_1) = 6{a_1}^2 - 9a_1 a_2 + 10{a_2}^2$$

$$\frac{\partial g(a_1)}{\partial a_1} = 2a_1 + 2a_2$$

$$\frac{\partial h(a_1)}{\partial a_1} = 12a_1 - 9a_2$$

$$\frac{\partial f(\boldsymbol{A})}{\partial a_1} = \frac{(2a_1 + 2a_2)(6{a_1}^2 - 9a_1a_2 + 10{a_2}^2) - ({a_1}^2 + 2a_1a_2 + {a_2}^2)(12a_1 - 9a_2)}{(6{a_1}^2 - 9a_1a_2 + 10{a_2}^2)^2}$$

$$= \frac{(12{a_1}^3 - 18{a_1}^2a_2 + 20a_1{a_2}^2 + 12{a_1}^2a_2 - 18a_1{a_2}^2 + 20{a_2}^3) - (12{a_1}^3 + 24{a_1}^2a_2 + 12a_1{a_2}^2 - 9{a_1}^2a - 18a_1{a_2}^2 - 9{a_2}^3)}{\left(6{a_1}^2 - 9a_1a_2 + 10{a_2}^2\right)^2}$$

$$= \frac{(12{a_1}^3 - 6{a_1}^2a_2 + 2a_1{a_2}^2 + 20{a_2}^3) - (12{a_1}^3 + 15{a_1}^2a_2 - 6a_1{a_2}^2 - 9{a_2}^3)}{(6{a_1}^2 - 9a_1a_2 + 10{a_2}^2)^2}$$

$$= \frac{-21{a_1}^2a_2 + 8a_1{a_2}^2 + 29{a_2}^3}{(6{a_1}^2 - 9a_1a_2 + 10{a_2}^2)^2}$$

$$\frac{\partial}{\partial a_2} f(\boldsymbol{A}) = \frac{\partial \left(\frac{g(a_2)}{h(a_2)}\right)}{\partial a_1} = \frac{\frac{\partial g(a_2)}{\partial a_2} h(a_2) - g(a_2)\frac{\partial h(a_2)}{\partial a_2}}{h(a_1)^2}$$

$$\frac{\partial g(a_2)}{\partial a_2} = 2a_1 + 2a_2$$

$$\frac{\partial h(a_2)}{\partial a_2} = -9a_1 + 20a_2$$

$$\frac{\partial f(\boldsymbol{A})}{\partial a_2} = \frac{(2a_1 + 2a_2)(6{a_1}^2 - 9a_1a_2 + 10{a_2}^2) - ({a_1}^2 + 2a_1a_2 + {a_2}^2)(-9a_1 + 20a_2)}{(6{a_1}^2 - 9a_1a_2 + 10{a_2}^2)^2}$$

$$= \frac{(12{a_1}^3 - 18{a_1}^2a_2 + 20a_1{a_2}^2 + 12{a_1}^2a_2 - 18a_1{a_2}^2 + 20{a_2}^3) - (-9{a_1}^3 - 18{a_1}^2a_2 - 9a_1{a_2}^2 + 20{a_1}^2a_2 + 40a_1{a_2}^2 + 20{a_2}^2)}{(6{a_1}^2 - 9a_1a_2 + 10{a_2}^2)^2}$$

$$= \frac{(12{a_1}^3 - 6{a_1}^2a_2 + 2a_1{a_2}^2 + 20{a_2}^3) - (-9{a_1}^3 + 2{a_1}^2a_2 + 31a_1{a_2}^2 + 20{a_2}^2)}{(6{a_1}^2 - 9a_1a_2 + 10{a_2}^2)^2}$$

$$= \frac{21{a_1}^3 - 8{a_1}^2a_2 - 29a_1{a_2}^2}{(6{a_1}^2 - 9a_1a_2 + 10{a_2}^2)^2}$$

Condition of extreme value

$$\frac{\partial f(\boldsymbol{A})}{\partial a_1} = \frac{\partial f(\boldsymbol{A})}{\partial a_2} = 0$$

$$(6{a_1}^2 - 9a_1a_2 + 10{a_2}^2)^2 > 0$$

We solve following simultaneous equation

$$-21{a_1}^2a_2 + 8a_1{a_2}^2 + 29{a_2}^3 = 0$$

$$21{a_1}^3 - 8{a_1}^2a_2 - 29a_1{a_2}^2 = 0$$

$$a_1 \neq 0$$

Divide both sides by ${a_1}^3$

$$21 - 8\frac{a_2}{a_1} - 29\left(\frac{a_2}{a_1}\right)^2 = 0$$

$$\frac{a_2}{a_1} = t$$

$$-21t + 8t^2 + 29t^3 = 0$$

$$21 - 8t - 29t^2 = 0$$

$$29t^2 + 8t - 21 = 0$$

$$(29t - 21)(t + 1) = 0$$

$$t = \frac{21}{29}, \qquad t = -1$$

From $t = \frac{21}{29}$, $a_1 = 29, a_2 = 21$

From $t = -1$, $a_1 = 1, a_2 = -1$

$$f(\boldsymbol{A}) = \frac{{a_1}^2 + 2a_1a_2 + {a_2}^2}{6{a_1}^2 - 9a_1a_2 + 10{a_2}^2}$$

$$f(t) = \frac{1 + 2t + t^2}{6 - 9t + 10t^2} = \frac{\frac{1}{t^2} + 2\frac{1}{t} + 1}{6\frac{1}{t^2} - 9\frac{1}{t} + 10}$$

t	$-\infty$		-1		$\frac{21}{29}$		$-\infty$
$\frac{\partial V}{\partial t}$		−	0	+	0	−	
V	0.1		0		0.6289		0.1

Conclusively, following $\boldsymbol{A}$ gives maximum value of $f(\boldsymbol{A})$.

$$f(\boldsymbol{A}) = \frac{{a_1}^2 + 2a_1a_2 + {a_2}^2}{6{a_1}^2 - 9a_1a_2 + 10{a_2}^2}$$

$$\boldsymbol{A} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 29 \\ 21 \end{pmatrix}$$

$\boldsymbol{A}$ is the normal vector of the reference hyperplane. When we sellect parallel hyperplane to the reference hyperplane which includes midpoint of center of subpopulation 1 and subpopulation 2.

$$\text{DS} \quad threshold = \begin{pmatrix} 29 \\ 21 \end{pmatrix}^T \frac{1}{2}\left(\begin{pmatrix} 7 \\ 6 \end{pmatrix} + \begin{pmatrix} 5 \\ 4 \end{pmatrix}\right) = (29 \quad 21)\begin{pmatrix} 6 \\ 5 \end{pmatrix} = 29 \times 6 + 21 \times 5 = 289$$

VI-1-3-4. Linear algebraic procedure.

Method of Lagrange multipliers is a method to find extreme value and condition of extreme value in constrained conditions. The method is often used in maximization or minimization of function. Principle of the method is osculation two hyper-solid locus. This means osculation of two curved surface or curved surface with flat surface. In any case, two surfaces share a tangent flat and normal line. One locus is the function to maximize or minimize and the

other locus is constrained condition. The locus of constrained condition is fixed, and we expand the function to maximize or maximize. When two locues contact each other at first in the process of the expansion, the value of the function to minimize is minimum value when the center of the target function exists in the area of locus of function of constrained condition. the target function reaches maximum value a t the last point of contact of two locus (See V-2-6. Maximum and minimum, Method of Lagrange multipliers.).

Linear algebraic procedure in this paragraph uses similar approach, though the locus of target function is not expanded but rotated in same size. Constrained condition is hyperplane, but not fixed. It moves in a restriction.

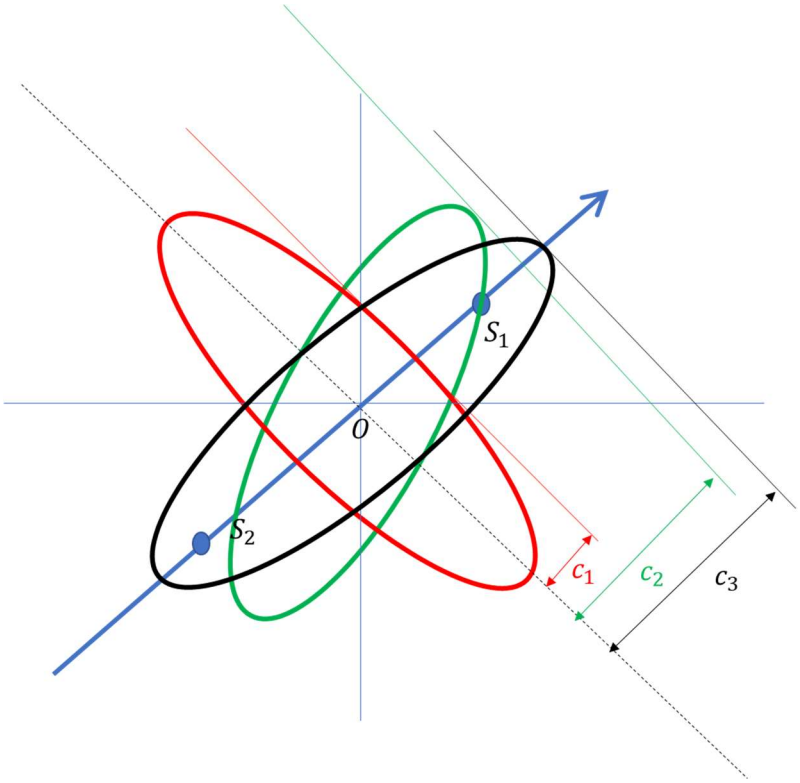

Fig.69 Rotation of ellipse and its projection to a vector

Figure 69 illustrates relation between the gradient of ellipse and the length of the mapping on the vector across the origin of the coordinate. It is obvious that the length of the projection is shortest when the vector of minimum radius is parallel to the vector for mapping. When there are two subpopulations, the vector for mapping is on the line connected center of two subpopulations. The ratio of length of mapped shadow and the distance of center of two subpopulation is minimum when the ellipse is rotated to fit the vector of minimum radius to the vector connecting center of two subpopulations.

Here, we remember Cauchy-Schwarz inequation explained in V-2-6-2.

$$(\boldsymbol{a}^t\boldsymbol{a})(\boldsymbol{b}^t\boldsymbol{b}) \geq (\boldsymbol{a}^t\boldsymbol{b})^2$$

We modified this basic inequation as follow.

$$\boldsymbol{a} = \boldsymbol{B}^{\frac{1}{2}}\boldsymbol{\alpha}$$

$$\boldsymbol{b} = \boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{\beta}$$

($\boldsymbol{B}$ is symmetric)

$$\left(\left(\boldsymbol{B}^{\frac{1}{2}}\boldsymbol{\alpha}\right)^T\left(\boldsymbol{B}^{\frac{1}{2}}\boldsymbol{\alpha}\right)\right)\left(\left(\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{\beta}\right)^T\left(\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{\beta}\right)\right) \geq \left(\left(\boldsymbol{B}^{\frac{1}{2}}\boldsymbol{\alpha}\right)^T\left(\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{\beta}\right)\right)^2$$

$$\left(\boldsymbol{\alpha}^t\boldsymbol{B}^{\frac{1}{2}^{\boldsymbol{T}}}\boldsymbol{B}^{\frac{1}{2}}\boldsymbol{\alpha}\right)\left(\boldsymbol{\beta}^t\boldsymbol{B}^{-\frac{1}{2}^{\boldsymbol{T}}}\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{\beta}\right) \geq \left(\boldsymbol{\alpha}^t\boldsymbol{B}^{\frac{1}{2}^{\boldsymbol{T}}}\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{\beta}\right)^2$$

$$\left(\boldsymbol{\alpha}^T\boldsymbol{B}^{\frac{1}{2}}\boldsymbol{B}^{\frac{1}{2}}\boldsymbol{\alpha}\right)\left(\boldsymbol{\beta}^T\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{\beta}\right) \geq \left(\boldsymbol{\alpha}^T\boldsymbol{B}^{\frac{1}{2}}\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{\beta}\right)^2$$

$$(\boldsymbol{\alpha}^T\boldsymbol{B}\boldsymbol{\alpha})(\boldsymbol{\alpha}^T\boldsymbol{B}^{-1}\boldsymbol{\alpha}) \geq (\boldsymbol{\alpha}^T\boldsymbol{\beta})^2$$

Condition of equality

$$\boldsymbol{a} = c\boldsymbol{b}$$

This means that $(\boldsymbol{\alpha}^T\boldsymbol{B}\boldsymbol{\alpha})(\boldsymbol{\alpha}^T\boldsymbol{B}^{-1}\boldsymbol{\alpha})$ is minimum value $(\boldsymbol{\alpha}^T\boldsymbol{\beta})^2$, when the vector $\boldsymbol{a}$ and $\boldsymbol{b}$ exist on same direction including inverse direction.

We modify the inequation furthermore.

$$\boldsymbol{B}^{\frac{1}{2}}\boldsymbol{\alpha} = c\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{\beta}$$

$$\boldsymbol{\alpha} = c\boldsymbol{B}^{-1}\boldsymbol{\beta}$$

We divide both side by $(\boldsymbol{\alpha}^T\boldsymbol{B}\boldsymbol{\alpha})$

$$\frac{(\boldsymbol{\alpha}^T\boldsymbol{\beta})^2}{\boldsymbol{\alpha}^T\boldsymbol{B}\boldsymbol{\alpha}} \leq \boldsymbol{\alpha}^T\boldsymbol{B}^{-1}\boldsymbol{\alpha}$$

$$(\because \boldsymbol{\alpha}^T\boldsymbol{B}\boldsymbol{\alpha} > \boldsymbol{0})$$

$$\frac{\boldsymbol{\alpha}^T\boldsymbol{B}\boldsymbol{\alpha}}{(\boldsymbol{\alpha}^T\boldsymbol{\beta})^2} \geq \boldsymbol{\alpha}^T\boldsymbol{B}^{-1}\boldsymbol{\alpha}$$

$$\left(\because \frac{(\boldsymbol{\alpha}^T\boldsymbol{B}\boldsymbol{\alpha})}{(\boldsymbol{\alpha}^T\boldsymbol{\beta})^2} > \boldsymbol{0},\ \ \boldsymbol{\alpha}^T\boldsymbol{B}^{-1}\boldsymbol{\alpha} > \boldsymbol{0},\right)$$

$$\frac{(\boldsymbol{\alpha}^T\boldsymbol{B}\boldsymbol{\alpha})}{(\boldsymbol{\alpha}^T\boldsymbol{\beta})(\boldsymbol{\alpha}^T\boldsymbol{\beta})} \geq \boldsymbol{\alpha}^T\boldsymbol{B}^{-1}\boldsymbol{\alpha}$$

$$\frac{(\boldsymbol{\alpha}^T\boldsymbol{B}\boldsymbol{\alpha})}{\boldsymbol{\alpha}^T\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{\alpha}} \geq \boldsymbol{\alpha}^T\boldsymbol{B}^{-1}\boldsymbol{\alpha}$$

$$(\because \boldsymbol{\alpha}^T\boldsymbol{\beta} = \boldsymbol{\beta}^T\boldsymbol{\alpha})$$

Denoting、$\boldsymbol{\alpha} = \boldsymbol{A},\ \boldsymbol{\beta} = \boldsymbol{g},\ \boldsymbol{B} = \boldsymbol{U}$

$\boldsymbol{A}$ is normal vector of reference hyperplane. $\boldsymbol{g}$ is vector connecting centers of two

subpopulation. $\boldsymbol{B^{\frac{1}{2}}}$ is symmetric matrix for rotation of the reference hyper vector.

$$\frac{\boldsymbol{A^T U A}}{\boldsymbol{A^T g g^T A}} \geq \boldsymbol{A^T U^{-1} A}$$

Condition of equality is obvious, because $\boldsymbol{A^T g}$ and $\boldsymbol{g^T A}$ are both inner product of vectors $\boldsymbol{A}$ and $\boldsymbol{g}$, the inner product is maximum when the vectors are on same direction. We do not need to perform following calculation.

$$\boldsymbol{\alpha} = c\boldsymbol{B^{-1}\beta}$$
$$\boldsymbol{A} = c\boldsymbol{U^{-1}g}$$
$$\boldsymbol{UA} = c\boldsymbol{g}$$

Conclusion

$h(\boldsymbol{A}) = \frac{\boldsymbol{A^T U A}}{\boldsymbol{A^T g g^T A}}$ reaches its minimum value $\boldsymbol{A^T U^{-1} A}$ when $\boldsymbol{A} = c\boldsymbol{U^{-1}g}$.

In other words, when vector $\boldsymbol{UA}$ is parallel to vector $\boldsymbol{g}$, $h(\boldsymbol{A})$ is minimum.
For application this to discrimination analysis, we need to allocate actual data to $\boldsymbol{g}$ and $\boldsymbol{U}$
We can use unit vector on the right line between centers of two subpopulation in the case there are only two subpopulations. For $\boldsymbol{U}$, one possible idea is to use $\mathbf{V}$ as $\boldsymbol{U}$, because we want to minimize the ratio of variance of residuals to variance among average of subpopulations.

$$\boldsymbol{V} = \sum_{k=1}^{m}\sum_{i=1}^{n_k}(\boldsymbol{x_{ki}} - \overline{\boldsymbol{x_k}})(\boldsymbol{x_{ki}} - \overline{\boldsymbol{x_k}})^T$$

However, in this case, we do not need to calculate deviation of each data from center of subpopulation because there only two factor composing total variances. One is differences among subpopulation. The other is residual.

$$SS_{total} = SS_{residual} + SS_{subpopulation}$$

When we divide both side by $SS_{subpopulation}$

$$\frac{SS_{total}}{SS_{subpopulation}} = \frac{SS_{residual}}{SS_{subpopulation}} + 1$$

From this we can understand that when $\frac{SS_{total}}{SS_{subpopulation}}$ is minimum, $\frac{SS_{residua}}{SS_{subpopulation}}$ is minimum.

This means that we can use total variance and covariance matrix instead of $\boldsymbol{V}$=$SS_{residual}$.
This is the explanation of liner discrimination analysis by rotation of hyper ellipse.

Exercise

We implement discriminant analysis of example dataset in previous paragraph.

Dataset

		data	
subpopulation	sample No	d_1	d_2
1	1	5	8
1	2	7	4
1	3	8	5
1	4	8	7
2	1	5	5
2	2	7	2
2	3	4	3
2	4	4	6

Total average

$$\overline{\overline{d_1}} = \frac{5+7+8+8+5+7+4+4}{8} = \frac{48}{8} = 6$$

$$\overline{\overline{d_2}} = \frac{8+4+5+7+5+2+3+6}{8} = \frac{40}{8} = 5$$

$$\overline{\overline{\boldsymbol{d}}} = \begin{pmatrix} 6 \\ 5 \end{pmatrix}$$

$$\overline{d_{11}} = \frac{5+7+8+8}{4} = \frac{28}{4} = 7$$

$$\overline{d_{12}} = \frac{5+7+4+4}{4} = \frac{20}{4} = 5$$

$$\overline{d_{21}} = \frac{8+4+5+7}{4} = \frac{24}{4} = 6$$

$$\overline{d_{22}} = \frac{5+2+3+6}{4} = \frac{16}{4} = 4$$

$$\overline{\boldsymbol{d_1}} = \begin{pmatrix} 7 \\ 5 \end{pmatrix}$$

$$\overline{\boldsymbol{d_2}} = \begin{pmatrix} 6 \\ 4 \end{pmatrix}$$

Vector connecting centers of subpopulations.

$$\boldsymbol{g} = c\left(\overline{\boldsymbol{d_1}} - \overline{\boldsymbol{d_2}}\right) = \mathrm{c}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

We calculate deviation from total average.

$$\boldsymbol{x}_{11} = \begin{pmatrix} 5 \\ 8 \end{pmatrix} - \begin{pmatrix} 6 \\ 5 \end{pmatrix} = \begin{pmatrix} -1 \\ 3 \end{pmatrix}, \boldsymbol{x}_{12} = \begin{pmatrix} 7 \\ 4 \end{pmatrix} - \begin{pmatrix} 6 \\ 5 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \ \boldsymbol{x}_{13} = \begin{pmatrix} 8 \\ 5 \end{pmatrix} - \begin{pmatrix} 6 \\ 5 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}, \ \boldsymbol{m}_{14} = \begin{pmatrix} 8 \\ 7 \end{pmatrix} - \begin{pmatrix} 6 \\ 5 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

$$\boldsymbol{x_{21}} = \begin{pmatrix}5\\5\end{pmatrix} - \begin{pmatrix}6\\5\end{pmatrix} = \begin{pmatrix}-1\\0\end{pmatrix},\ \boldsymbol{x_{22}} = \begin{pmatrix}7\\2\end{pmatrix} - \begin{pmatrix}6\\5\end{pmatrix} = \begin{pmatrix}1\\-3\end{pmatrix},\ \boldsymbol{x_{23}} = \begin{pmatrix}4\\3\end{pmatrix} - \begin{pmatrix}6\\5\end{pmatrix} = \begin{pmatrix}-2\\-2\end{pmatrix}, \boldsymbol{m_{24}} = \begin{pmatrix}4\\6\end{pmatrix} - \begin{pmatrix}6\\5\end{pmatrix} = \begin{pmatrix}-2\\1\end{pmatrix}$$

$$\boldsymbol{u}^T = \begin{pmatrix}-1 & 1 & 2 & 2 & -1 & 1 & -2 & -2\\ 3 & -1 & 0 & 2 & 0 & 3 & -2 & 1\end{pmatrix}$$

$$\boldsymbol{U} = \boldsymbol{u}\boldsymbol{u}^T = \begin{pmatrix}20 & -1\\ -1 & 28\end{pmatrix}$$

$$\boldsymbol{U^{-1}} = \frac{1}{559}\begin{pmatrix}28 & 1\\ 1 & 20\end{pmatrix}$$

$$\boldsymbol{A} = c\boldsymbol{U^{-1}}\boldsymbol{g}$$

$$\boldsymbol{A} = c\frac{1}{559}\begin{pmatrix}28 & 1\\ 1 & 20\end{pmatrix}\begin{pmatrix}1\\1\end{pmatrix} = c\frac{1}{559}\begin{pmatrix}29\\21\end{pmatrix}$$

We select $c = 559$

$$\boldsymbol{A} = \begin{pmatrix}29\\21\end{pmatrix}$$

This is conclusion. We confirmed that we can obtain same result by operation by ANOVA and linear algebraic procedure. Advantage of linear algebraic procedure is simplicity of operation process. We do not need separation of variances and differential. One of the weakness of the linear algebraic procedure is difficulty of understanding of theoretical explanation particularly for readers who are unfamiliar with linear algebra. However, the theory itself is rather simple and not difficult. The author recommends reading of V-2-6. Maximum and minimum, Method of Lagrange multipliers for getting background knowledge. Theoretical weakness of linear algebraic procedure is selection of $\boldsymbol{g}$. In the case when there are only two subpopulations vector $\boldsymbol{g}$ is on the straight line connecting two centers. This is because, the distance on the line is expressed the difference between the two populations. When there are more than three subpopulations, we should consider how to draw the line, On of the method is to draw the multiple regression line among center of subpopulation. However, in the case when number of variables are more than number of subpopulation, we cannot estimate the multiple regression line. In the first place, the author does not know there exist such case in reality. When there exist many variables, the relation between variables are complex and we should not select linear discrimination analysis for prediction of subpopulation. There are many other methods to categorize data such as principle component analysis (PCA), factor analysis (FA), multidimensional scaling method (MDS)、cluster analysis and so on. The merit of liner discriminant analysis is flexibility of threshold. We can put our policy and philosophy in the selection of threshold, because discriminant score is simple length from reference hyperplane.

VI-1-3-5. Threshold of discrimination score

We can understand how to get optimal gradient of hyperplane. As an example, the author draws a line of threshold to include the point of total mean in Figure 70, this is because the author has no information of distribution of each subpopulation. The author hypothesizes homoscedasticity between two subpopulations from visual information from the scatter graph of the data. The author has no confidence for his judgement. However, the threshold line of discrimination score Z separate data to correct subpopulation as sown in figure 70.
The value of the threshold is obtained by putting center of total distribution $d_1 = 6, d_2 = 5$ in formula of discrimination score. All the data are completely separated to correct sub population as following figure.

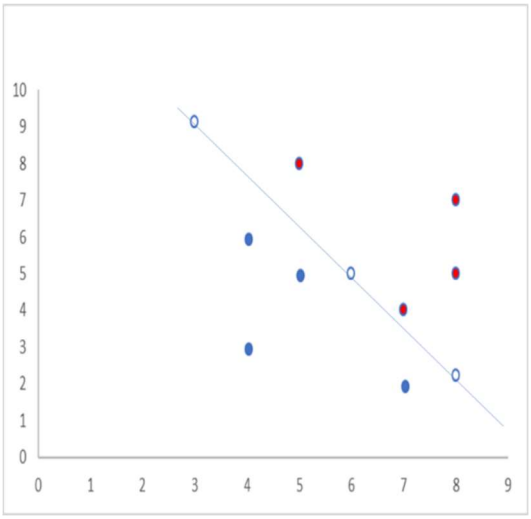

Fig. 70. Separation by threshold line of discrimination score $Z = 29d_1 + 21d_2 = 279$
Red circle: subpopulation1, blue circle: subpopulation2. Bleu line. Threshold line.

Data

subpopulation	$d_{1,}$	$d_{2,}$	Z	$Z -$ threshold(279)
1	5	8	313	34
1	7	4	287	8
1	8	5	337	58
1	8	7	379	100
2	5	5	250	-29
2	7	2	245	-34
2	4	3	179	-100
2	4	6	242	-37

VI-1-3-6. Additional discussion for application of discriminant analysis.

When we hypothesize normal distribution in each subpopulation, we can estimate probability of error by the judgement by normalized distance by standard deviation from the center of subpopulation. In upper case, the threshold exists at same distance from each center of subpopulation. The probability of error is the same in both subpopulations. We need not to make the probability the same. In many cases meaning of error judgement of A for correct B and error judgement of B for correct A is different. As an example, in simple rapid test of disease, the purpose of the test is to find out infected individuals. The seriousness of diagnostic error that judges true infected individual as healthy individual should be strictly avoided. For this, we need to accept error that judges healthy individual as infected individual. In that case, we should set the threshold to minimize the possibility of misjudgement that diagnose infected individual as healthy individuals. We set the threshold line closer to center of subpopulation of healthy individuals as shown in figure 71.

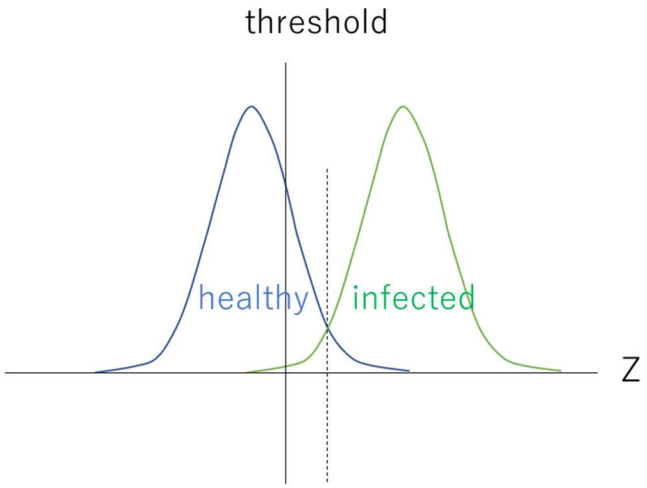

Fig. 71. Configuration of threshold of discrimination score.

We understand that we can implement discriminant analysis without knowledge ANOVA and calculation of differential. This is same as multiple linear regression analysis. In discriminant analysis, we need to optimize number of variables, because we can simplify the operation in case of diagnosis. This is also important in other applications. Too much variables are inefficient and makes confusion. This is also same as multiple linear regression. In the explanation of this text, the author presumes equal variances among subpopulation. This is sometimes unnatural. On the other side, he does not presume equal variances among variables. However, in several case, we need normalization of data by dividing raw data by standard

deviation of each variables to make equality of variance among variables as in multiple linear regression. However, when we normalize the data, liner discriminant analysis became no meaning, because data distribution forms hyper sphere. When we emphasize orthogonality of variables, discriminant analysis using principle component scores may be a possible approach. However, it may make unnecessary confusion relating to the meaning of components, and it may be inefficient. There are many discussions relating to discrimination analyses. We should select proper method considering purpose of analysis and nature of data.