VI-2-3. Factor analysis (FA)

VI-2-3-1. What is FA

In multiple regression, we express the phenomena as follow.

$$y = a_1x_1 + a_2x_2 + \cdots + a_px_p + e$$

This can be rewrite using matrix as follow.

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \boldsymbol{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \quad \boldsymbol{A} = (a_1 \quad \cdots \quad a_p)$$

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{A}^T + \boldsymbol{E}$$

This means that $\boldsymbol{Y}$ is objective variable and $\boldsymbol{X}$ is explanatory variables. $\boldsymbol{X}$ and $\boldsymbol{Y}$ are given and we estimate $\boldsymbol{A}$ in multiple regression. In factor analysis, $\boldsymbol{Y}$ is given, and we estimate $\boldsymbol{X}$ and $\boldsymbol{A}$. Needless to say, it is impossible. However, we can estimate both, when a number of objective variables are given.

$$\boldsymbol{Y} = \begin{pmatrix} y_{11} & \cdots & y_{1m} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nm} \end{pmatrix}$$

From this we estimate following matrixes.

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \quad \boldsymbol{A} = \begin{pmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mp} \end{pmatrix}$$

This is principle of factor analysis (FA). Multiple regression analysis (MRA) is a method to express relation between an explained variable (objective variable) and several explanatory variables. Principle component analysis (PCA) is a method to recognize the relation among variables. FA is a method to estimate latent variables (factor) which control observed variables. FA can be comparable to the estimation of mechanism in deep sea from observed phenomena in surface of the sea such as existence of unknown organisms and sea current. For this we need to use power of computer. Because of this, development in factor analysis was a key factor of modern statistics which discuss unobservable mechanism and structure. Factor analysis is sometimes compared with PCA. PCA is explanation data structure. It is not aiming to fined latent factor and random fluctuations are included in each component. FA excludes random fluctuation as an error term and finds latent factors.

VI-2-3-2. Theoretical model of FA

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{A}^T + \boldsymbol{E}$$

$$\boldsymbol{Z} = \begin{pmatrix} z_{11} & \cdots & z_{1m} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nm} \end{pmatrix}_{n \times m}$$

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}_{n\times p}$$

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mp} \end{pmatrix}$$

$$\boldsymbol{E} = \begin{pmatrix} e_{11} & \cdots & e_{1m} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nm} \end{pmatrix}_{n\times m}$$

$$\boldsymbol{A}^T = \begin{pmatrix} a_{11} & \cdots & a_{m1} \\ \vdots & \ddots & \vdots \\ a_{1p} & \cdots & a_{mp} \end{pmatrix}$$

$$\boldsymbol{E} = \begin{pmatrix} e_{11} & \cdots & e_{1m} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nm} \end{pmatrix} = \boldsymbol{Z} - \boldsymbol{X}\boldsymbol{A}^{\boldsymbol{T}} = \begin{pmatrix} z_{11} & \cdots & z_{1m} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nm} \end{pmatrix} - \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}\begin{pmatrix} a_{11} & \cdots & a_{m1} \\ \vdots & \ddots & \vdots \\ a_{1p} & \cdots & a_{mp} \end{pmatrix}$$

$$= \begin{pmatrix} z_{11} & \cdots & z_{1m} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nm} \end{pmatrix} - \begin{pmatrix} \sum_{k=1}^{p} a_{1k}x_{1k} & \sum_{k=1}^{p} a_{2k}x_{1k} & \cdots & \sum_{k=1}^{p} a_{mk}x_{1k} \\ \sum_{k=1}^{p} a_{1k}x_{2k} & \sum_{k=1}^{p} a_{2k}x_{2k} & \cdots & \sum_{k=1}^{p} a_{mj}x_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^{p} a_{1k}x_{nk} & \sum_{k=1}^{p} a_{2k}x_{nk} & \cdots & \sum_{k=1}^{p} a_{mk}x_{nk} \end{pmatrix}$$

$$= \begin{pmatrix} z_{11} - \sum_{k=1}^{p} a_{1k}x_{1k} & z_{12} - \sum_{k=1}^{p} a_{2k}x_{1k} & \cdots & z_{1m} - \sum_{k=1}^{p} a_{mk}x_{1k} \\ z_{21} - \sum_{k=1}^{p} a_{1k}x_{2k} & z_{22} - \sum_{k=1}^{p} a_{2k}x_{2k} & \cdots & z_{2m} - \sum_{k=1}^{p} a_{mk}x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} - \sum_{k=1}^{p} a_{1k}x_{nk} & z_{n2} - \sum_{k=1}^{p} a_{2k}x_{nk} & \cdots & z_{nm} - \sum_{k=1}^{p} a_{mk}x_{nk} \end{pmatrix}$$

$$e_{ij} = z_{ij} - \sum_{\boldsymbol{k=1}}^{\boldsymbol{p}} a_{jk}x_{ik} = z_{ij} - \begin{pmatrix} x_{i1} & \cdots & x_{ip} \end{pmatrix}\begin{pmatrix} a_{j1} \\ \vdots \\ a_{jp} \end{pmatrix} = z_{ij} - \boldsymbol{x}_i\,\boldsymbol{a}_j{}^T$$

In this equation, we are thinking that $\boldsymbol{x}_i = (x_{i1} \quad \cdots \quad x_{ip})$ *and* $\boldsymbol{a}_j = (a_{j1} \quad \cdots \quad a_{jp})$ are vectors. We also can express e_{ij} as function of $\boldsymbol{x}_i$ and $\boldsymbol{a}_j$ thinking z_{ij} is a definite, because z_{ij} is observed value and it does not include $\boldsymbol{x}_i$ or $\boldsymbol{a}_j$.

$$e_{ij} = f_{ij}(\boldsymbol{a}_j, \boldsymbol{x}_i)$$

$$\boldsymbol{EE^T} = \begin{pmatrix} e_{11} & \cdots & e_{1m} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nm} \end{pmatrix} \begin{pmatrix} e_{11} & \cdots & e_{n1} \\ \vdots & \ddots & \vdots \\ e_{1m} & \cdots & e_{nm} \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{j=1}^{m} (e_{1j})^2 & \sum_{j=1}^{m} e_{1j}e_{2j} & \cdots & \sum_{j=1}^{m} e_{1j}e_{nj} \\ \sum_{j=1}^{m} e_{2j}e_{1j} & \sum_{j=1}^{m} (e_{2j})^2 & \cdots & \sum_{j=1}^{m} e_{2j}e_{nj} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{j=1}^{m} e_{in}e_{i1} & \sum_{j=1}^{m} e_{in}e_{i2} & \cdots & \sum_{i=1}^{m} (e_{n2})^2 \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{j=1}^{m} \left(f_{1j}(\boldsymbol{a}_j, \boldsymbol{x}_1)\right)^2 & \sum_{j=1}^{m} \left(f_{1j}(\boldsymbol{a}_j, \boldsymbol{x}_1)\right)\left(f_{2j}(\boldsymbol{a}_j, \boldsymbol{x}_2)\right) & \cdots & \sum_{j=1}^{m} \left(f_{1j}(\boldsymbol{a}_j, \boldsymbol{x}_1)\right)\left(f_{nj}(\boldsymbol{a}_j, \boldsymbol{x}_n)\right) \\ \sum_{j=1}^{m} \left(f_{2j}(\boldsymbol{a}_j, \boldsymbol{x}_2)\right)\left(f_{1j}(\boldsymbol{a}_j, \boldsymbol{x}_1)\right) & \sum_{j=1}^{m} \left(f_{2j}(\boldsymbol{a}_j, \boldsymbol{x}_2)\right)^2 & \cdots & \sum_{j=1}^{m} \left(f_{2j}(\boldsymbol{a}_j, \boldsymbol{x}_2)\right)\left(f_{nj}(\boldsymbol{a}_j, \boldsymbol{x}_n)\right) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{j=1}^{m} \left(f_{nj}(\boldsymbol{a}_j, \boldsymbol{x}_n)\right)\left(f_{1j}(\boldsymbol{a}_j, \boldsymbol{x}_1)\right) & \sum_{j=1}^{m} \left(f_{nj}(\boldsymbol{a}_j, \boldsymbol{x}_n)\right)\left(f_{2j}(\boldsymbol{a}_j, \boldsymbol{x}_2)\right) & \cdots & \sum_{j=1}^{m} \left(f_{nj}(\boldsymbol{a}_j, \boldsymbol{x}_n)\right)^2 \end{pmatrix}$$

$$\boldsymbol{E^TE} = \begin{pmatrix} e_{11} & \cdots & e_{n1} \\ \vdots & \ddots & \vdots \\ e_{1m} & \cdots & e_{nm} \end{pmatrix} \begin{pmatrix} e_{11} & \cdots & e_{1m} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nm} \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^{n} (e_{i1})^2 & \sum_{i=1}^{n} e_{i1}e_{i2} & \cdots & \sum_{i=1}^{n} e_{i1}e_{im} \\ \sum_{i=1}^{n} e_{i2}e_{i1} & \sum_{i=1}^{n} (e_{i2})^2 & \cdots & \sum_{i=1}^{n} e_{i2}e_{im} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} e_{im}e_{i1} & \sum_{i=1}^{n} e_{im}e_{i2} & \cdots & \sum_{i=1}^{n} (e_{im})^2 \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^{n} (f_{i1}(\boldsymbol{a}_1, \boldsymbol{x}_i))^2 & \sum_{i=1}^{n} (f_{i1}(\boldsymbol{a}_1, \boldsymbol{x}_i))(f_{i2}(\boldsymbol{a}_2, \boldsymbol{x}_i)) & \cdots & \sum_{i=1}^{n} (f_{i1}(\boldsymbol{a}_1, \boldsymbol{x}_i))(f_{im}(\boldsymbol{a}_m, \boldsymbol{x}_1)) \\ \sum_{i=1}^{n} (f_{i2}(\boldsymbol{a}_2, \boldsymbol{x}_i))(f_{i1}(\boldsymbol{a}_1, \boldsymbol{x}_i)) & \sum_{i=1}^{n} (f_{i2}(\boldsymbol{a}_2, \boldsymbol{x}_i))^2 & \cdots & \sum_{i=1}^{n} (f_{i2}(\boldsymbol{a}_2, \boldsymbol{x}_i))(f_{im}(\boldsymbol{a}_m, \boldsymbol{x}_1)) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} (f_{im}(\boldsymbol{a}_m, \boldsymbol{x}_1))(f_{i1}(\boldsymbol{a}_1, \boldsymbol{x}_i)) & \sum_{i=1}^{n} (f_{im}(\boldsymbol{a}_m, \boldsymbol{x}_1))(f_{i2}(\boldsymbol{a}_2, \boldsymbol{x}_i)) & \cdots & \sum_{i=1}^{n} (f_{in}(\boldsymbol{a}_n, \boldsymbol{x}_i))^2 \end{pmatrix}$$

When we consider that $\boldsymbol{e}_i = (e_{i1} \quad \cdots \quad e_{im})$ and $\boldsymbol{\varepsilon}_j = \begin{pmatrix} e_{1j} \\ \vdots \\ e_{nj} \end{pmatrix}$ are vectors, the factors of matrix

$\boldsymbol{EE^T}$ and matrix $\boldsymbol{E^TE}$ are inner products of the vectors, and diagonal factors are squares of the length of vectors. One of possible definition of optimization is optimization of length of vectors to minimize the error.

In least square method, optimization means minimization of sum of the diagonal factors. We call sum of diagonal factors as trace

$$tr(\boldsymbol{EE^T}) = \boldsymbol{e_1 e_1}^{\boldsymbol{T}} + \boldsymbol{e_2 e_2}^{\boldsymbol{T}} + \cdots + \boldsymbol{e_n e_n}^{\boldsymbol{T}} = \sum_{i=1}^{n} \boldsymbol{e_i e_i}^{\boldsymbol{T}} = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(f_{ij}(\boldsymbol{a_j}, \boldsymbol{x_i}) \right)^2$$

$$tr(\boldsymbol{E^T E}) = \boldsymbol{\varepsilon_1}^{\boldsymbol{T}} \boldsymbol{\varepsilon_1} + \boldsymbol{\varepsilon_2}^{\boldsymbol{T}} \boldsymbol{\varepsilon_2} + \cdots + \boldsymbol{\varepsilon_m}^{\boldsymbol{T}} \boldsymbol{\varepsilon_m} = \sum_{j=1}^{m} \boldsymbol{\varepsilon_j}^{\boldsymbol{T}} \boldsymbol{\varepsilon_j} = \sum_{j=1}^{m} \sum_{i=1}^{n} \left(f_{ij}(\boldsymbol{a_j}, \boldsymbol{x_i}) \right)^2$$

In most likelihood method, optimization means maximization of infinite products of provability of diagonal factors.

$$Prob(\boldsymbol{EE^T}) = \prod_{i=1}^{n} Prob(\boldsymbol{e_i e_i}^{\boldsymbol{T}}) = \prod_{i=1}^{n} Prob \sum_{j=1}^{m} \left(f_{ij}(\boldsymbol{a_j}, \boldsymbol{x_i}) \right)^2$$

$$Prob(\boldsymbol{E^T E}) = \prod_{j=1}^{m} Prob(\boldsymbol{\varepsilon_i \varepsilon_i}^{\boldsymbol{T}}) = \prod_{j=1}^{m} Prob \sum_{i=1}^{n} \left(f_{ij}(\boldsymbol{a_j}, \boldsymbol{x_i}) \right)^2$$

We separate variables by taking logarithm of probability for differentiation.

$$\log_e Prob(\boldsymbol{EE^T}) = \log_e \prod_{i=1}^{n} Prob \sum_{j=1}^{m} \left(f_{ij}(\boldsymbol{a_j}, \boldsymbol{x_i}) \right)^2 = \sum_{i=1}^{n} \log_e Prob \sum_{j=1}^{m} \left(f_{ij}(\boldsymbol{a_j}, \boldsymbol{x_i}) \right)^2$$

$$\log_e Prob(\boldsymbol{E^T E}) = \log_e \prod_{j=1}^{m} Prob \sum_{i=1}^{n} \left(f_{ij}(\boldsymbol{a_j}, \boldsymbol{x_i}) \right)^2 = \sum_{j=1}^{m} \log_e Prob \sum_{i=1}^{n} \left(f_{ij}(\boldsymbol{a_j}, \boldsymbol{x_i}) \right)^2$$

Here, we consider $\boldsymbol{a_j}$ as given coefficients, $tr(\boldsymbol{EE^T})$ and $\log_e \text{Prob}(\boldsymbol{EE^T})$ are polynomial equation composed from n terms, and i^{th} term is function of $\boldsymbol{x_i}$. Each term is function of only a vector and not include other vector. We can obtain differentiation of $tr(\boldsymbol{EE^T})$ and $\log_e \text{Prob}(\boldsymbol{EE^T})$ as sum of differentiation of each term. Similarly, When we consider $\boldsymbol{x_i}$ as given coefficients, $tr(\boldsymbol{E^T E})$ and $\log_e \text{Prob}(\boldsymbol{E^T E})$ are polynomial equation composed from m terms, and j^{th} term is function of $\boldsymbol{a_j}$. Each term is function of only a vector and not include other vector. We can obtain differentiation of $tr(\boldsymbol{E^T E})$ and $\log_e \text{Prob}(\boldsymbol{E^T E})$ as sum of differentiation of each term.

$$\frac{\partial \left(tr(\boldsymbol{EE}^T) \right)}{\partial \boldsymbol{X}} = \sum_{i=1}^{n} \frac{\partial \left(\sum_{j=1}^{m} \left(f_{ij}(a_j, \boldsymbol{x_i}) \right)^2 \right)}{\partial \boldsymbol{x_i}}$$

$$\frac{\partial (\log_e \text{Prob}(\boldsymbol{EE}^T))}{\partial \boldsymbol{X}} = \sum_{i=1}^{n} \frac{\partial \left(\log_e \sum_{j=1}^{m} \left(f_{ij}(\boldsymbol{a_j}, \boldsymbol{x_i}) \right)^2 \right)}{\partial \boldsymbol{x_i}}$$

$$\frac{\partial \left(tr(\boldsymbol{E}^T \boldsymbol{E}) \right)}{\partial \boldsymbol{A}} = \sum_{j=1}^{m} \frac{\partial \left(\sum_{j=1}^{m} \left(f_{ij}(\boldsymbol{a_j}, \boldsymbol{x_i}) \right)^2 \right)}{\partial \boldsymbol{a_j}}$$

$$\frac{\partial(\log_e \mathrm{Prob}(\boldsymbol{E}^T\boldsymbol{E}))}{\partial \boldsymbol{A}} = \sum_{j=1}^{m} \frac{\partial\left(\log_e \sum_{j=1}^{m}\left(f_{ij}(\boldsymbol{a}_j,\, \boldsymbol{x}_i)\right)^2\right)}{\partial \boldsymbol{a}_j}$$

This means that when we define optimization of $\boldsymbol{E}$ as optimization of diagonal factors of matrix $\boldsymbol{E}\boldsymbol{E}^T$, we can obtain optimal $\boldsymbol{X}$ by giving arbitrary $\boldsymbol{A}$, and when we define optimization of $\boldsymbol{E}$ as optimization of diagonal factors of matrix $\boldsymbol{E}^T\boldsymbol{E}$, we can obtain optimal $\boldsymbol{A}$ by giving arbitrary $\boldsymbol{X}$. Another idea optimization of the matrix $\boldsymbol{E}\boldsymbol{E}^T$ and $\boldsymbol{E}\boldsymbol{E}^T$ is orthogonalization of vectors. This is possible and factors except diagonal factors of the matrixes become 0 by the orthogonalization. However, optimization of diagonal factors and orthogonalization of other factors are not always compatible. In factor analysis, we consider only optimization of diagonal factors.

$$\boldsymbol{E} = \begin{pmatrix} e_{11} & \cdots & e_{1m} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nm} \end{pmatrix} = \boldsymbol{Z} - \boldsymbol{X}\boldsymbol{A}^{\boldsymbol{T}}$$

$$e_{ij} = f_{ij}(\boldsymbol{a}_j,\, \boldsymbol{x}_i)$$

$$\boldsymbol{\psi} = \begin{pmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_n \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{m}\left(f_{1j}(\boldsymbol{a}_j,\, \boldsymbol{x}_1)\right)^2 & 0 & \cdots & 0 \\ 0 & \sum_{j=1}^{m}\left(f_{2j}(\boldsymbol{a}_j,\, \boldsymbol{x}_2)\right)^2 & \cdots & 0 \\ & & \ddots & \vdots \\ \vdots & \vdots & & \\ 0 & 0 & \cdots & \sum_{j=1}^{m}\left(f_{nj}(\boldsymbol{a}_j,\, \boldsymbol{x}_n)\right)^2 \end{pmatrix}$$

$$\boldsymbol{\varphi} = \begin{pmatrix} \varphi_1 & 0 & \cdots & 0 \\ 0 & \varphi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \varphi_m \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n}(f_{i1}(\boldsymbol{a}_1,\, \boldsymbol{x}_i))^2 & 0 & \cdots & 0 \\ 0 & \sum_{i=1}^{n}(f_{i2}(\boldsymbol{a}_2,\, \boldsymbol{x}_i))^2 & \cdots & 0 \\ & & \ddots & \vdots \\ \vdots & \vdots & & \\ 0 & 0 & \cdots & \sum_{i=1}^{n}(f_{im}(\boldsymbol{a}_m,\, \boldsymbol{x}_i))^2 \end{pmatrix}$$

In least square method, we minimize $\psi_1, \cdots, \psi_m$ and $\varphi_1, \cdots, \varphi_n$. In most likelihood method, we maximize $\log_e Prob\left(\psi_1\right), \cdots, \log_e Prob\left(\psi_m\right)$ and $\log_e Prob\left(\varphi_1\right), \cdots, \log_e Prob\left(\varphi_n\right)$. In this model, we have to give proper $\boldsymbol{A}$ or $\boldsymbol{X}$ from outsides for calculation. There are various ingenuities and methods. However, optimizing of diagonal factor is a basic model of FA. Conclusively, FA is method to find major latent factors neglecting the orthogonality of factors. However, the factors are independent each other and we can expect orthogonality among major factors in many cases as the results.

VI-2-3-3. Factor extraction

We can understand the model of FA and form of solution. However, we cannot obtain a single solution in one step by deterministic way such as optimization of residual vector by

orthogonalization between linearly combined explanatory vector (factor vector) and observed vector, because there are multiple observed vector (objective variables). We tentatively give weight for linear combination (coefficient matrix) of explanatory variables (factors) and optimize explanatory variables (E-step). Then, we optimize the weight using obtained explanatory variables (M-step). Then we implement next E-step using obtained weight in last M-step. Repeating this, we can obtain optimal weight (coefficient) and variable vectors (factor), if the weight and variable vector converge in set up ranges.

Basic strategy of analysis in FA is as above. There are several issues to be solved in the implementation of calculation. This is the issue of factor extraction in FA. There are various calculation methods. Explanations of calculation methods are not systematized, and we are confused in selection of factor extraction method when we read such explanations. However, there are only two major ways. One is least square method and the other is most likelihood method. There are several methods other than least square method and most likelihood method to bypass the repeated computation such as principle component method and so on. Those methods are alternative methods when we cannot proper solution by least square method or most likelihood method and are not fit essential concept of FA. Selection of least square method or most likelihood method is depending on the nature of data (form of distribution of data) and essential issue.

Least square method and most likelihood method

Least square method is targeting data distributing unimodally without bias. Center of the distribution should be the peak and frequency of the data decrease symmetrically to both sides. Because of this nature, we can use distance from center as alternative variable of variant, and minimization of the distance means maximization of possibility.

We calculate possibility directly in most likelihood method. For this reason, we can treat biased data, when the data were standardized. However, the calculation procedure is troublesome. The explanatory value gives maximum possibility is the explanatory value which gives extreme value of possibility function. We differentiate infinite product of possibility by explanatory variables to obtain derivatives. We cannot differentiate definite product directly. Consequently, we take logarithm of definite product, and differentiate logarithmic probability as alternative function. In this process, a term includes sums are included in logarithm. We can differentiate sum of logarithm by general differentiation method, though we cannot differentiate logarithm of sum. This issue can be solved by transformation of this term to the form of sum of logarithm using Jensen's inequality. Jensen's inequation is approximate calculation based on a precondition. When the data are not satisfied the precondition, the calculation often gives improper solution. In this case, improper solution means solution in

which sum of the variance of variables exceed 1. When sum of the variance exceed 1, the variance of the error is minus. Minus variance is not possible. In such case, the solution is improper solution. This is the essence of issue of factor extraction.

Explanatory variables and coefficients are estimated from number of objective variables in FA. It is impossible to obtain determinative single solution in one step approach. We consider asymptotic approach by repeated estimation starting from set of tentatively given data. This approach is called EM algorism (Estimation Maximization algorithm). EM algorithm is a calculation method depending on machine power of computer. Commonly, EM algorithm is explained relating to most likelihood method. It can be effectively used in other cases for estimation. There exist other ideas and calculation methods to avoid EM algorithm. Those methods can provide solution robustly. However, such methods are neglecting basic structure of factor analysis, in which given data is objective variables. The author is thinking that such methods is alternative method when we obtain improper solution, and he explains only asymptotic methods here.

Approach by least square method

E step

Mathematic model of FA is as follow.

$$\boldsymbol{E} = \boldsymbol{Z} - \boldsymbol{X}\boldsymbol{A}^{\boldsymbol{T}} = \begin{pmatrix} e_{11} & \cdots & e_{1m} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nm} \end{pmatrix}$$

$$e_{ij} = f_{ij}(\boldsymbol{a}_j, \boldsymbol{x}_i)$$

$$\boldsymbol{\psi} = \begin{pmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_n \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{m} \left(f_{1j}(\boldsymbol{a}_j, \boldsymbol{x}_1)\right)^2 & 0 & \cdots & 0 \\ 0 & \sum_{j=1}^{m} \left(f_{2j}(\boldsymbol{a}_j, \boldsymbol{x}_2)\right)^2 & \cdots & 0 \\ & & \ddots & \vdots \\ \vdots & \vdots & \cdots & \sum_{j=1}^{m} \left(f_{nj}(\boldsymbol{a}_j, \boldsymbol{x}_n)\right)^2 \\ 0 & 0 & & \end{pmatrix}$$

In the E step, we give tentative $\boldsymbol{A}$ to the equation to make $\boldsymbol{E}$ to function of $\boldsymbol{X}$. Then optimize diagonal matrix $\boldsymbol{\psi}$ by minimizing ψ_i. When we express function of $f_{ij}(\boldsymbol{a}_j, \boldsymbol{x}_i)$ in least square method,

$$f_{ij}(\boldsymbol{a}_j, \overline{\boldsymbol{x}_l}) = z_{ij} - \overline{\boldsymbol{x}_l}\boldsymbol{a}_j{}^T$$

Hear, $\overline{\boldsymbol{x}_l}$ is expectation value of $\boldsymbol{x}_i$. Vector $\boldsymbol{x}_i$ and $\boldsymbol{a}_j$ are as follows.

$$\overline{\boldsymbol{x}_l} = (\overline{x_{l1}} \quad \cdots \quad \overline{x_{lp}})$$

$$\boldsymbol{a}_j = (a_{j1} \quad \cdots \quad a_{jp})$$

When these vectors expressed as coordinate in orthogonal coordinate system,

$$f_{ij}(\boldsymbol{a}_j,\ \boldsymbol{x}_i) = z_{ij} - \sum_{k=1}^{p} \overline{x_{ip}}\, a_{jp}$$

In E step, $\boldsymbol{a}_j$is given as row of matrix $\boldsymbol{A}$, and、 $\boldsymbol{x}_i = (x_{i1} \quad \cdots \quad x_{ip})$ is expectation value. When we consider $\boldsymbol{X}$ as row of column vector as follow.

$$\boldsymbol{X} = (\boldsymbol{x}_1 \quad \cdots \quad \boldsymbol{x}_p)$$

$$\boldsymbol{x}_k = \begin{pmatrix} x_{1k} \\ \vdots \\ x_{pk} \end{pmatrix}$$

In factor analysis $\boldsymbol{x}_k$ and $\boldsymbol{x}_\kappa (k \neq \kappa)$ are not necessarily orthogonal. We can express the function as $f(\boldsymbol{a}_j,\ \boldsymbol{x}_i) = z_{ij} - \boldsymbol{x}_i \boldsymbol{a}_j{}^T$, only when all combinations of $\boldsymbol{x}_k$ and $\boldsymbol{x}_\kappa$ are orthogonal each other. We are not given angles among $\boldsymbol{x}_k$ and $\boldsymbol{x}_\kappa$. Consequently, we have to express formula of our calculation as follow.

$$f_{ij}(\boldsymbol{a}_j,\ \boldsymbol{x}_i) = z_{ij} - \sum_{k=1}^{p} x_{ip} a_{jp}$$

Here, we consider the relation between $\overline{\boldsymbol{x}_\iota} \boldsymbol{a}_j{}^T$ and $\sum_{k=1}^{p} x_{ip} a_{jp}$.
The author gives the conclusion of our consideration at first.

$$\overline{\boldsymbol{x}_\iota} \boldsymbol{a}_j{}^T \geq \sum_{k=1}^{p} x_{ip} a_{jp}$$

Condition of equation is orthogonality of all combinations of $\boldsymbol{x}_k$ and $\boldsymbol{x}_\kappa$ （$\boldsymbol{x}_k{}^T \boldsymbol{x}_\kappa = 0$）. Generally, this relation is explained by Jensen's inequality. Explanation by Jensen's inequality is logical, though the author tries to make sensuous explanation at first. When we express $\boldsymbol{z}_{ij}$ and $\overline{\boldsymbol{x}_\iota}$ as points in $p+1$ dimension hyperspace.

$$\boldsymbol{z}_{ij} = (z_{ij} \quad \overline{x_{\iota 1}} a_{j1} \quad \cdots \quad \overline{x_{\iota p}} a_{jp})$$

$$\overline{\boldsymbol{x}_{\iota j}} = (L_0 \quad \overline{x_{\iota 1}} a_{j1} \quad \cdots \quad \overline{x_{\iota p}} a_{jp})$$

Function $f_{ij}(\boldsymbol{a}_j, \overline{\boldsymbol{x}_\iota}) = z_{ij} - \overline{\boldsymbol{x}_\iota} \boldsymbol{a}_j{}^T$is the length of vector in figure 87.

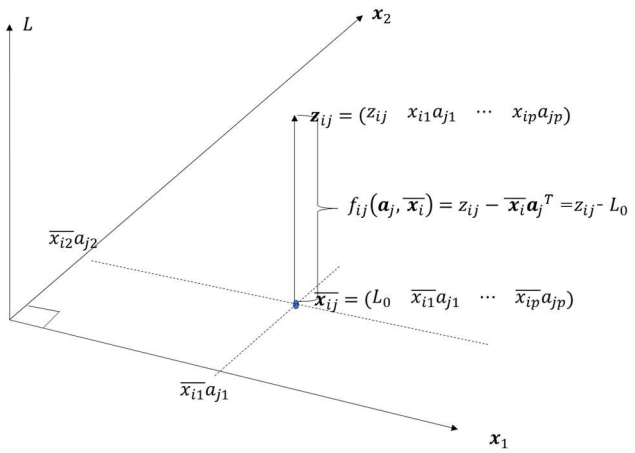

Fig. 87. The relation between $\boldsymbol{z}_{ij}$ in $p+1$ dimensional hyperspace and $\overline{\boldsymbol{x}_{\iota j}}$ on p次 dimensional hyperplane.

When all combinations of $\boldsymbol{x}_k$ are orthogonal each other, $\overline{\boldsymbol{x}_{\imath J}} = (L_0 \quad \overline{x_{\imath 1}}a_{j1} \quad \cdots \quad \overline{x_{\imath p}}a_{jp})$ is exists on p dimensional hyperplane as linear combination of $(\overline{x_{\imath 1}} \quad \cdots \quad \overline{x_{\imath p}})$. Vector $\overrightarrow{\overline{\boldsymbol{x}_{\imath J}}\boldsymbol{z}_{\imath J}}$ is normal vector to the hyperplane from $\boldsymbol{z}_{ij} = (z_{ij} \quad \overline{x_{\imath 1}}a_{j1} \quad \cdots \quad \overline{x_{\imath p}}a_{jp})$, and $\overline{\boldsymbol{x}_{\imath J}} = (L_0 \quad \overline{x_{\imath 1}}a_{j1} \quad \cdots \quad \overline{x_{\imath p}}a_{jp})$ is foot of the normal vector. The length of the vector is the distance between the hyper plane and point $\boldsymbol{z}_{ij}$. The length of the normal vector is definition of distance between hyperplane and point and minimal length from the point of out of the hyperplane to points on the hyperplane. $\left|\overrightarrow{\overline{\boldsymbol{x}_{\imath J}}\boldsymbol{z}_{\imath J}}\right|$ is minimal length among $\left|\overrightarrow{\boldsymbol{x}_{\imath J}\boldsymbol{z}_{\imath J}}\right|$. Here, $\boldsymbol{x}_{ij}$ is point on the hyperplane. We do not consider orthogonality of element vectors of $(x_{i1} \quad \cdots \quad x_{ip})$, in expectation of $(x_{i1} \quad \cdots \quad x_{ip})$. For this reason, $(x_{i1} \quad \cdots \quad x_{ip})$ is not generally equal to $(\overline{x_{\imath 1}} \quad \cdots \quad \overline{x_{\imath p}})$.

When

$$(x_{i1} \quad \cdots \quad x_{ip}) \neq (\overline{x_{\imath 1}} \quad \cdots \quad \overline{x_{\imath p}}),$$

$$\boldsymbol{x}_{ij} = (L_0 \quad x_{i1}a_{j1} \quad \cdots \quad x_{ip}a_{jp}) \neq (L_0 \quad \overline{x_{\imath 1}}a_{j1} \quad \cdots \quad \overline{x_{\imath p}}a_{jp}) = \overline{\boldsymbol{x}_{\imath J}}$$

Consequently, $\boldsymbol{x}_{ij} = (L_0 \quad x_{i1}a_{j1} \quad \cdots \quad x_{ip}a_{jp})$ is not foot of normal line from $\boldsymbol{z}_{ij}$, and

$$\left|\overrightarrow{\overline{\boldsymbol{x}_{\imath J}}\boldsymbol{z}_{\imath J}}\right| \leq \left|\overrightarrow{\boldsymbol{x}_{\imath J}\boldsymbol{z}_{\imath J}}\right|$$

$$z_{ij} - \overline{\boldsymbol{x}_{\imath}}\boldsymbol{a}_j{}^T \leq z_{ij} - \boldsymbol{x}_i\boldsymbol{a}_j{}^T$$

$$\overline{\boldsymbol{x}_{\imath}}\boldsymbol{a}_j{}^T \geq \boldsymbol{x}_i\boldsymbol{a}_j{}^T$$

Conclusively,

$$\overline{\boldsymbol{x}_{\imath}}\boldsymbol{a}_j{}^T \geq \sum_{k=1}^{p} x_{ip}a_{jp}$$

This means that $\sum_{k=1}^{p} x_{ip}a_{jp}$ is infimum variation of $\overline{\boldsymbol{x}_{\imath}}\boldsymbol{a}_j{}^T$. In another word, $\sum_{k=1}^{p} x_{ip}a_{jp}$ is minimum limit of $\overline{\boldsymbol{x}_{\imath}}\boldsymbol{a}_j{}^T$.

Inversely,

$$z_{ij} - \overline{\boldsymbol{x}_{\imath}}\boldsymbol{a}_j{}^T \leq z_{ij} - \sum_{k=1}^{p} x_{ip}a_{jp}$$

$z_{ij} - \sum_{k=1}^{p} x_{ip}a_{jp}$ is supremum variation of $z_{ij} - \overline{\boldsymbol{x}_{\imath}}\boldsymbol{a}_j{}^T$. Goal of lest square method is to obtain minimum value of absolute value of $z_{ij} - \overline{\boldsymbol{x}_{\imath}}\boldsymbol{a}_j{}^T$. We cannot obtain this directly, because we cannot estimate $\overline{\boldsymbol{x}_{\imath}}$. However, we can obtain supremum variation of $z_{ij} - \overline{\boldsymbol{x}_{\imath}}\boldsymbol{a}_j{}^T$ as $z_{ij} - \sum_{k=1}^{p} x_{ip}a_{jp}$. We can obtain $(x_{i1} \quad \cdots \quad x_{ip})$ which gives minimum value of abosolute value of $z_{ij} - \sum_{k=1}^{p} x_{ip}a_{jp}$ by solving following differential equations.

$$\frac{d\sum_{j=1}^{m}\left(f_{1j}(\boldsymbol{a}_j, \boldsymbol{x}_1)\right)^2}{d\,\boldsymbol{x}_1} = 0$$

$$\frac{d\sum_{j=1}^{m}\left(f_{1j}(\boldsymbol{a}_j, \boldsymbol{x}_2)\right)^2}{d\,\boldsymbol{x}_2} = 0$$

$$\vdots$$

$$\frac{d\sum_{j=1}^{m}\left(f_{1j}(\boldsymbol{a}_j,\ \boldsymbol{x}_n)\right)^2}{d\ \boldsymbol{x}_n}=0$$

We reform this differential by $\boldsymbol{x}_i$ to following partial differential.

$$\frac{\partial\sum_{j=1}^{m}\left(f_{1j}(\boldsymbol{a}_j,\ \boldsymbol{x}_i)\right)^2}{\partial\ x_{i1}}=0$$

$$\frac{\partial\sum_{j=1}^{m}\left(f_{1j}(\boldsymbol{a}_j,\ \boldsymbol{x}_i)\right)^2}{\partial\ x_{i2}}=0$$

$$\vdots$$

$$\frac{\partial\sum_{j=1}^{m}\left(f_{1j}(\boldsymbol{a}_j,\ \boldsymbol{x}_i)\right)^2}{\partial\ x_{ip}}=0$$

Solving each simultaneous equation, we can obtain expectation value of $\boldsymbol{x}_i$. Repeating this for $\boldsymbol{x}_1 \cdots \boldsymbol{x}_n$, expectation value of matrix $\boldsymbol{X}$ can be obtained.

Jensen's inequality

"In convex function, function of expectation value is larger than expectation vale of function." This is Jensen's equation. Jensen's inequality is satisfied only in convex function. Average is one of the expectation values. Sumo of sampled data weighted by proper weights based on some logic is called expectation value including simple average, weighted average, estimation of return and so on. Expectation value is representative of all sampled data and value which of which probability to obtain is highest by random sampling from mother population. すIn this meaning we named such values as expectation values. In simple average, the weight of all sampled data is $\frac{1}{n}$.

Convex function is function of which second order differential is exclusively positive or negative. When $f\ (x)=\log_e x$,

$$\frac{df(x)}{dx}=\frac{d\log_e x}{dx}=\frac{1}{x}$$

$$\frac{d^2f(x)}{dx^2}=\frac{d^2\log_e x}{dx^2}=-\frac{1}{x^2}<0$$

We can conclude that $f\ (x)=\log_e x$ is convex function. The author shows an example of Jensen's inequality using $f\ (x)=\log_e x$.

$\mathrm{Ex}(x)$ is expectation value of x. Here we define expectation value as average.

$$\mathrm{Ex}(x)=\frac{\sum_{i=1}^{n}x_i}{n}$$

$$f\big(Ex(x)\big)=\log_e\frac{\sum_{i=1}^{n}x_i}{n}$$

$$\mathrm{Ex}\big(f(x)\big) = \frac{1}{n}\sum_{i=1}^{n} \log_e x_i$$

When x_i=1, 2, 4, 7

$$\mathrm{Ex}(x) = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1+2+4+7}{4} = 3.5$$

$$f\big(Ex(x)\big) = \log_e 3.5 = 1.252763$$

On the other hand,

$f(1) = \log_e 1 = 0,\ f(2) = \log_e 2 = 0.693147,\ f(4) = \log_e 4 = 1.386294, f(4) = \log_e 7 = 1.94591$

$$\mathrm{Ex}\big(f(x)\big) = \frac{0 + 0.693147 + 1.386294 + 1.94591}{4} = 1.006338$$

$$f\big(Ex(x)\big) > \mathrm{Ex}\big(f(x)\big)$$

Convex function in multidimensional space is domical function which has only one peak. Here, the author proves Jensen's inequality in two dimensional space.

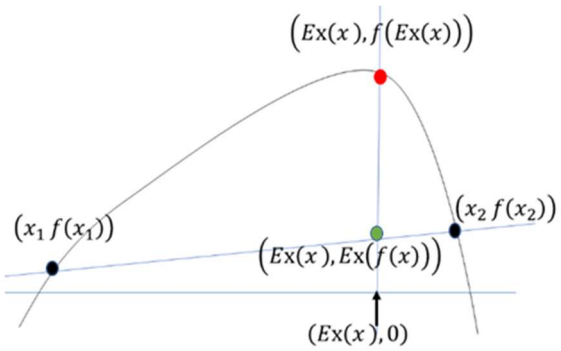

Fig. 88-1. Relation between function of expectation value and expectation value of function in case of 2 data.

.

Curve line in this figure 88 is expressing $f(x)$, and horizontal line express variable x. There are two points on the curve line. The points are $(x_1, f(x_1))$ and $(x_2, f(x_2))$. We calculate expectation value form the two points. In the case when points $(x_2, f(x_2))$ express 1.5 times strongly express the characteristics of mother population in some reason, the expectation values are as follow.

$$\left(\frac{1x_1 + 1.5x_2}{1 + 1.5}, \frac{1f(x_1) + 1.5f(x_2)}{1 + 1.5}\right) = (0.4x_1 + 0.6x_2, 0.4f(x_1) + 0.6f(x_2))$$

In the case when both points express the characteristics of mother population similarly, the expectation value is simple average, and the coordinate of the expectation value is as follow.

$$\left(\frac{1x_1+1x_2}{1+1},\frac{1f(x_1)+1f(x_2)}{1+1}\right)=(0.5x_1+0.5x_2,0.5f(x_1)+0.5f(x_2))$$

The point of the expectation value is internally dividing point by λ_2: λ_1 (green segment) of line segment $(x_1,f(x_1))$-$(x_2,f(x_2))$. The internally dividing point can be expressed as $\left(Ex(x),Ex(f(x))\right)$. Function of $Ex(x)$ is $f(Ex(x))$ and the red point on the curve line is $\left(Ex(x),f(Ex(x))\right)$. The red point is obviously higher than green point $\left(Ex(x),Ex(f(x))\right)$. When we add a new data $(x_2,f(x_2))$, the point of expectation value moves to blue point (figure 88-2).

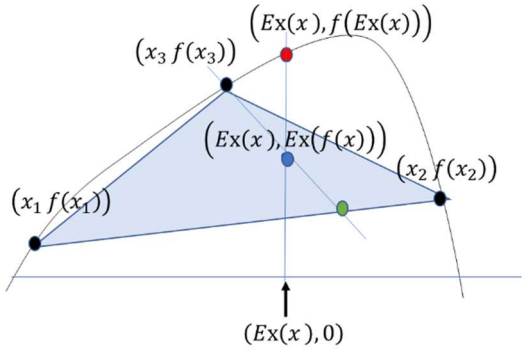

Fig. 88-2. Relation between function of expectation value and expectation value of function in case of 3 data.

Blue point exist on the line between new point and green point. The blue point is exist in the light blue triangle and red point is obviously higher the blue point.

When we increase the number of samples, samples form polygon and the point of expectation value (yellow point) exists inside of the polygon (figure 88-3).

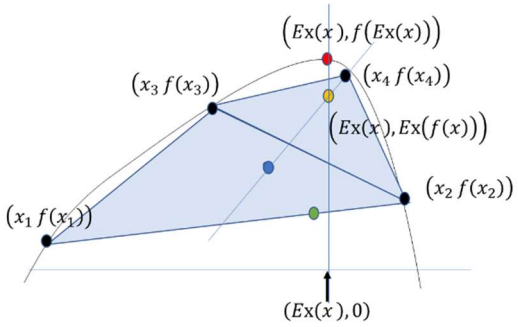

Fig. 88-3. Relation between function of expectation value and expectation value of function in case of 4 data.

The polygon is inscribed by the line of the function. The polygon is convex hull. Convex hull means polygon without depressed part (shape of rubber film covering polygon). Data obtained from convex function form convex hull and expectation values exist inside of the

convex hull. Consequently, the point of function of expectation value is higher than expectation value of function. Intuitive expression of Jensen's inequality is "height of structure in a dome shorter than height of the ceiling of the dome".

$$f\big(Ex(\boldsymbol{x})\big) \geq Ex\big(f(x)\big)$$

Formula78

However, when the dome has multi peaks length of several pillars are shorter than average height of ceiling of the dome. This is the reason why Jensen's inequality is true only in convex function. Multiple sub populations often exist in sampled population. Improper solution will sometimes be obtained in such cases.

In E step in least square method, calculation of function $\boldsymbol{x}\boldsymbol{a}_j{}^T$is $\sum_{k=1}^{p} \overline{x_{ip}} a_{jp}$. This is weighted average of $\overline{x_{ip}}$. On the other hand, $\sum_{k=1}^{p} x_{ip} a_{jp}$ is weighted average of $(x_{i1} \quad 0 \quad \cdots \quad 0), (0 \quad x_{i2} \quad \cdots \quad 0), \cdots, (0 \quad \cdots \quad 0 \quad x_{ip})$. Each x_{ik} is optimized separately. This can be expressed as follow.

$$f\big(Ex(\boldsymbol{x})\big) = \boldsymbol{x}\boldsymbol{a}_j{}^T$$

$$Ex\big(f(x)\big) = \sum_{k=1}^{p} x_{ip} a_{jp}$$

Therefore, using Jensen's inequality,

$$\overline{\boldsymbol{x}_i}\boldsymbol{a}_j{}^T \geq \sum_{k=1}^{p} x_{ip} a_{jp}$$

M step

We can make pseudo inverse matrix $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}$ from $\boldsymbol{X}$, and we can obtain $\boldsymbol{A}$ by multiplying pseudo inverse matrix $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}$ to $\boldsymbol{Z}$.

$$\boldsymbol{E} = \boldsymbol{Z} - \boldsymbol{X}\boldsymbol{A}^{\boldsymbol{T}}$$

$$\boldsymbol{Z} \cong \boldsymbol{X}\boldsymbol{A}^T$$

$$(\boldsymbol{X}^T\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}\boldsymbol{Z} = \boldsymbol{A}^T$$

When the data $\boldsymbol{Z}$ is expressed as distances from mean, $\boldsymbol{Z}^T\boldsymbol{Z}$is variance covariance matrix of observed data and $\boldsymbol{A}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{A}^T$is variance covariance of estimated value.

$$\boldsymbol{Z}^T\mathbf{Z} = \begin{pmatrix} z_{11} & \cdots & z_{n1} \\ \vdots & \ddots & \vdots \\ z_{1m} & \cdots & z_{nm} \end{pmatrix} \begin{pmatrix} z_{11} & \cdots & z_{1m} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nm} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} {z_{i1}}^2 & \sum_{i=1}^{n} z_{i1}z_{i2} & \cdots & \sum_{i=1}^{n} z_{i1}z_{im} \\ \sum_{i=1}^{n} z_{i2}z_{i1} & \sum_{i=1}^{n} {z_{i2}}^2 & \cdots & \sum_{i=1}^{n} z_{i2}z_{im} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} z_{in}z_{i1} & \sum_{i=1}^{n} z_{in}z_{i2} & \cdots & \sum_{i=1}^{n} {z_{im}}^2 \end{pmatrix}$$

$$\boldsymbol{XA}^T = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} a_{11} & \cdots & a_{m1} \\ \vdots & \ddots & \vdots \\ a_{1p} & \cdots & a_{mp} \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{p} x_{1k}a_{1k} & \cdots & \sum_{k=1}^{p} x_{1k}a_{mk} \\ \vdots & \ddots & \vdots \\ \sum_{k=1}^{p} x_{nk}a_{1k} & \cdots & \sum_{k=1}^{p} x_{nk}a_{mk} \end{pmatrix}$$

$$\boldsymbol{AX}^T\boldsymbol{XA}^T = \begin{pmatrix} \sum_{k=1}^{p} x_{1k}a_{1k} & \cdots & \sum_{k=1}^{p} x_{nk}a_{1k} \\ \vdots & \ddots & \vdots \\ \sum_{k=1}^{p} x_{1k}a_{mk} & \cdots & \sum_{k=1}^{p} x_{nk}a_{mk} \end{pmatrix} \begin{pmatrix} \sum_{k=1}^{p} x_{1k}a_{1k} & \cdots & \sum_{k=1}^{p} x_{1k}a_{mk} \\ \vdots & \ddots & \vdots \\ \sum_{k=1}^{p} x_{nk}a_{1k} & \cdots & \sum_{k=1}^{p} x_{nk}a_{mk} \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^{n}\left(\left(\sum_{k=1}^{p} x_{ik}a_{1k}\right)^2\right) & \sum_{i=1}^{n}\left(\left(\sum_{k=1}^{p} x_{ik}a_{1k}\right)\left(\sum_{k=1}^{p} x_{ik}a_{2k}\right)\right) & \cdots & \sum_{i=1}^{n}\left(\left(\sum_{k=1}^{p} x_{ik}a_{1k}\right)\left(\sum_{k=1}^{p} x_{ik}a_{mk}\right)\right) \\ \sum_{i=1}^{n}\left(\left(\sum_{k=1}^{p} x_{ik}a_{2k}\right)\left(\sum_{k=1}^{p} x_{ik}a_{1k}\right)\right) & \sum_{i=1}^{n}\left(\left(\sum_{k=1}^{p} x_{ik}a_{2k}\right)^2\right) & \cdots & \sum_{i=1}^{n}\left(\left(\sum_{k=1}^{p} x_{ik}a_{2k}\right)\left(\sum_{k=1}^{p} x_{ik}a_{mk}\right)\right) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n}\left(\left(\sum_{k=1}^{p} x_{ik}a_{mk}\right)\left(\sum_{k=1}^{p} x_{ik}a_{1k}\right)\right) & \sum_{i=1}^{n}\left(\left(\sum_{k=1}^{p} x_{ik}a_{mk}\right)\left(\sum_{k=1}^{p} x_{ik}a_{2k}\right)\right) & \cdots & \sum_{i=1}^{n}\left(\left(\sum_{k=1}^{p} x_{ik}a_{mk}\right)^2\right) \end{pmatrix}$$

Diagonal factor in these matrixes are sum of squares and total SS is sum of explainable SS and SS of error.

$$\sum_{i=1}^{n} {z_{ij}}^2 = \sum_{i=1}^{n}\left(\left(\sum_{k=1}^{p} x_{ik}a_{jk}\right)^2\right) + \sum_{i=1}^{n} {e_{ij}}^2$$

$$\sum_{i=1}^{n} {e_{ij}}^2 = \sum_{i=1}^{n} {z_{ij}}^2 - \sum_{i=1}^{n}\left(\left(\sum_{k=1}^{p} x_{ik}a_{jk}\right)^2\right) \geq 0$$

When $\sum_{i=1}^{n} {z_{ij}}^2 < \sum_{i=1}^{n}\left(\left(\sum_{k=1}^{p} x_{ik}a_{jk}\right)^2\right)$, $\boldsymbol{X}$ or $\boldsymbol{A}$ is improper solution. When $\boldsymbol{X}$ and $\boldsymbol{A}$ are not improper we go back to E step and calculate expectation value of $\boldsymbol{X}$ giving obtained $\boldsymbol{A}$ as second tentative $\boldsymbol{A}$. Then got next M step to obtain optimal $\boldsymbol{A}$ using new $\boldsymbol{X}$. We repeat this to reach stable variance of error in narrow range.

Approach by most likelihood method

Probability of multidimensional normal distribution is as follow.

$$P(\boldsymbol{Z}) = \frac{1}{\left(\sqrt{2\pi}\right)^n \sqrt{|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{Z})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{Z})}$$

$\boldsymbol{Z}$ is standardized data by subtracting of expectation value.

When we breakdown the formula to probability of each data,

$$P(z_{ij}) = \frac{1}{(\sqrt{2\pi})\sigma_j} e^{-\frac{1}{2}\left(\frac{(z_{ij}-\boldsymbol{x}_i\boldsymbol{a}_j{}^T)}{\sigma_j}\right)^2}$$

$$\boldsymbol{x}_i = (x_{i1} \quad \cdots \quad x_{ip})$$
$$\boldsymbol{a}_j = (a_{j1} \quad \cdots \quad a_{jp})$$
$$\sigma_j{}^2\text{: variance of } \boldsymbol{z}_j = (z_{1j} \quad \cdots \quad z_{nj})$$

When all z_{ij} are independent each other, $P(\boldsymbol{Z})$ is infinite of $P(z_{ij})$.

$$P(\boldsymbol{Z}) = \prod_{i=1}^{n}\prod_{j=1}^{m} P(z_{ij}) = \prod_{i=1}^{n}\prod_{j=1}^{m} \frac{1}{(\sqrt{2\pi})\sigma_j} e^{-\frac{1}{2}\left(\frac{(z_{ij}-\boldsymbol{x}_i\boldsymbol{a}_j{}^T)}{\sigma_j}\right)^2}$$

We want to differentiate $P(\boldsymbol{Z})$ giving tentative $\boldsymbol{a}_j$ to obtain optimal expectation value of $\boldsymbol{x}_i$. However, we cannot differentiate $P(\boldsymbol{Z})$ directly because the variables are not separated. Thus we tale logarithmic probability for the separation.

$$\log_e P(\boldsymbol{Z}) = \sum_{i=1}^{n}\sum_{j=1}^{m} \log_e\left(\frac{1}{(\sqrt{2\pi})\sigma_j} e^{-\frac{1}{2}\left(\frac{(z_{ij}-\boldsymbol{x}_i\boldsymbol{a}_j{}^T)}{\sigma_j}\right)^2}\right)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m}\left(-\log_e(\sqrt{2\pi})\sigma_j - \frac{1}{2}\left(\frac{(z_{ij}-\boldsymbol{x}_i\boldsymbol{a}_j{}^T)}{\sigma_j}\right)^2\right)$$

$\sum_{j=1}^{m}\left(-\log_e(\sqrt{2\pi}\sigma_j) - \frac{1}{2}\left(\frac{(z_{ij}-\boldsymbol{x}_i\boldsymbol{a}_j{}^T)}{\sigma_j}\right)^2\right)$ includes only a $\boldsymbol{x}_i$, and $-\log_e(\sqrt{2\pi}\sigma_j)$ is a constant given by observation value of $\boldsymbol{Z}$.

$$\frac{d\left(\sum_{j=1}^{m}\left(\frac{z_{ij}-\boldsymbol{x}_i\boldsymbol{a}_j{}^T}{\sigma_j}\right)^2\right)}{d\boldsymbol{x}_i} = 0$$

We can obtain $\boldsymbol{x}_i$ by solving upper differential equation. However, we cannot fix $\boldsymbol{x}_i$ at specified point in the hyperspace. This is the same as the explanation in E step by least square method. Thus, we consider infimum variation and optimization of infimum variation. In most likelihood method, distance between observed value and expected value is standardized by standard deviation as $\left(\frac{z_{ij}-\boldsymbol{x}_i\boldsymbol{a}_j{}^T}{\sigma_j}\right)$. This is the difference between least square method and most likelihood method. Consequently, the results from variance covariance matrix and from correlation matrix is the same in most likelihood method. The results from variance covariance matrix and from correlation matrix is different in least square method and the result by most likelihood and by least square method are different from variance covariance matrix and the difference is same from correlation matrix.

In M step of we make following differential equation and optimize $\boldsymbol{A}$ giving $\boldsymbol{X}$ from the result of E step.

$$\frac{dLL}{d\boldsymbol{A}} = \frac{d\sum_{j=1}^{m}\sum_{i=1}^{n}\left(z_{ij}{}^{2}\boldsymbol{\Sigma^{-1}}\right)}{d\boldsymbol{A}} = 0$$

This is repeat of multiple regression and calculation is the same between most likelihood method and least square method. Most simple and easy calculation method is to make pseudo inverse matrix $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ from $\boldsymbol{X}$ obtained in E step.

$$\boldsymbol{Z} \cong \boldsymbol{X}\boldsymbol{A}^T$$

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z} \cong (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{A}^T$$

$$\boldsymbol{A}^T = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z}$$

Optimization of infimum variation premising convex function of probability distribution is the same in least square method and most likelihood method. Most likelihood method is natural and logical, though we cannot always apply Jensen's inequation to function of probability distribution. As an example, normal distribution has two flexion points and is not convex hull. When actual data distributes strictly in normal distribution, we cannot apply Jensen's inequation, and we cannot consider infimum variation. However, theoretical probability distribution and actual data distribution is different issue. We can apply Jensen's equation when actual data distribution is approximately convex hull. On the other hand, quadratic function is obviously convex hull. From this reason, frequency of improper solution higher in most likelihood method than in lest square method. Low frequency of improper solution is a merit of lest square method. When improper solution is obtained one possible, one possible countermeasure is to try least square method using unstandardized data. This is commonly recommended strategy in text books, because when proper solution in obtained, no reviewer will not claim revise of the methodology. More essentially, when we need to analyze unstandardized data, we cannot select most likelihood method. When subpopulations are existing in the data, data distribution often has plural peaks and we cannot apply Jensen's inequation. We need careful consideration for elimination of subpopulation. Adequacy of the elimination depends on the purpose of analysis and nature of data set. When improper solution is obtained, we can select other methods such as principle factor method. However, it is better to check the characteristics of data set such as bias of data, existence of sub population and so on. The author implements principle component analysis before factor analysis to confirm existence of major factor. Purpose and mathematical model are different between factor analysis and principle component analysis. Principle component analysis is easier than factor analysis and we can obtain certain result in principle component analysis and we can forecast proper number of factors by the result of principle component analysis.

Recent trend in extraction of factor extraction (Bayesian method by MCMC)

Development of computer software is remarkable recently. Bayesian approaches are used as the method to empirically approach to the solution using machine power in various fields. In such method, random numbers are generated in certain condition and parameters of probability distribution are optimized using the random number at first step. Then second random number is generated independently to first step and optimize parameters of probability distribution. Repeating such processes, we can obtain stable solution. This is basic idea. Generally, we need huge replication of for calculation. For the promotion of efficiency of calculation, various algorisms are proposed. Detailed explanation of such algorism is beyond capacity of the author. Please read manual of such software when reader need to use such software. MCMCpack in R has function of MCMCfactnal. The function is system to perform factor analysis by MCMC.

VI-2-3-4. Rotation

We can extract factors by upper methods. An observed variable is explained by plural number of factors. However, it is difficult to interpret meaning of factors when the factor has weak relation to various observed variables. Interpretation becomes easier when number of variables which related to the factor are limited. This is simplification. When direction of vector of observed variable overlap to the axis of factor. The factor is easily interpreted by the variable. The purpose rotation is maximization of absolute value of coefficient of several observed variables and minimize absolute value of the coefficient of other variables. We need mathematical skill for rotation, though purpose of rotation is interpretation of factors and there is no mathematical logic. For this reason, we have to rotate to make interpretation easier. Rotation of axis of coordinate is not difficult when we transform the data by polar coordinate. Readers who has no knowledge of polar coordinate, please read III-3-4 coordinate conversion. We consider transformation from polar coordinate (r, θ) to 2 dimensional orthogonal coordinate

$$x = r\cos\theta$$

$$y = r\sin\theta$$

When we rotate θ the polar coordinate to anticlockwise direction, $\theta = 0$、

$$x = r\cos 0 = r$$

$$y = r\sin 0 = 0$$

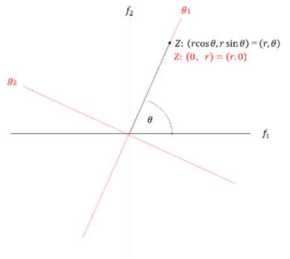

Fig 88. Rotation of axis of polar coordinate

We consider this transformation by rotation matrix $\boldsymbol{T}$.

$$\boldsymbol{T} = \begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{pmatrix}$$

$$(r \quad 0) = (r\cos\theta \quad r\sin\theta)\boldsymbol{T} = (r\cos\theta \quad r\sin\theta)\begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{pmatrix}$$

$$0 = r\cos\theta\, t_{11} + r\sin\theta t_{21} = r(\cos\theta\, t_{11} + \sin\theta t_{21})$$

$$r = r\cos\theta\, t_{12} + r\sin\theta t_{22} = \mathrm{r}(\cos\theta\, t_{12} + \sin\theta t_{22})$$

$$\cos\theta\, t_{11} + \sin\theta t_{21} = 1$$

$$\cos\theta\, t_{12} + \sin\theta t_{22} = 0$$

$$\boldsymbol{T} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

In 3 dimensional space, we consider 2 successive rotation around Z axis and X axis.

Rotation around Z axis is

$$\begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Rotation around X axis is

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\varphi & -\sin\varphi \\ 0 & \sin\varphi & cos\,\varphi \end{pmatrix}$$

Transformation by successive rotation is

$$\begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\varphi & -\sin\varphi \\ 0 & \sin\varphi & cos\,\varphi \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta\cos\varphi & \sin\theta\sin\psi \\ \sin\theta & \cos\theta\cos\psi & -\cos\theta\sin\varphi \\ 0 & \sin\psi & \cos\psi \end{pmatrix}$$

We can make rotation matrix in higher dimensional space though the formula is complicated and sensuous understanding of rotation in higher dimensional space is difficult. Heare, we consider nature of rotation matrix in linear algebra.

When we multiply transposed matrix of rotation matrix to the rotation matrix, the product is identity matrix.

Example 1 (2 dimension)

$$\boldsymbol{T} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

$$\boldsymbol{T}^T = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

$$\boldsymbol{T}\boldsymbol{T}^T = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} = \begin{pmatrix} \cos^2\theta + \sin^2\theta & \sin\theta\cos\theta - \sin\theta\cos\theta \\ \sin\theta\cos\theta - \sin\theta\cos\theta & \sin^2\theta + \cos^2\theta \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I$$

Example 2. (3 dimension)

$$\boldsymbol{T} = \begin{pmatrix} \cos\theta & -\sin\theta\cos\varphi & \sin\theta\sin\psi \\ \sin\theta & \cos\theta\cos\psi & -\cos\theta\sin\varphi \\ 0 & \sin\psi & \cos\psi \end{pmatrix}$$

$$\boldsymbol{T}^T = \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta\cos\varphi & \cos\theta\cos\psi & \sin\psi \\ \sin\theta\sin\psi & -\cos\theta\sin\varphi & \cos\psi \end{pmatrix}$$

$$\boldsymbol{T}\boldsymbol{T}^T = \begin{pmatrix} \cos\theta & -\sin\theta\cos\varphi & \sin\theta\sin\psi \\ \sin\theta & \cos\theta\cos\psi & -\cos\theta\sin\varphi \\ 0 & \sin\psi & \cos\psi \end{pmatrix}\begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta\cos\varphi & \cos\theta\cos\psi & \sin\psi \\ \sin\theta\sin\psi & -\cos\theta\sin\varphi & \cos\psi \end{pmatrix} =$$

$$\begin{pmatrix} \cos^2\theta + \sin^2\theta\cos^2\varphi + \sin^2\theta\sin^2\varphi & \sin\theta\cos\theta + \sin\theta\cos\theta\cos^2\psi + \sin\theta\cos\theta\sin^2\psi & -\sin\theta\sin\psi\cos\psi + \sin\theta\sin\psi\cos\psi \\ \cos\theta\sin\psi - \cos\theta\sin\psi\cos^2\psi - \cos\theta\sin\psi\sin^2\psi + & \sin^2\theta + \cos^2\theta\cos^2\psi + \sin^2\theta\sin^2\psi + & \cos\theta\cos\psi\sin\psi - \cos\theta\sin\psi\cos\psi \\ -\sin\psi\sin\theta\cos\psi + \cos\psi\sin\theta\sin\psi & \sin\psi\cos\theta\cos\psi - \cos\psi\cos\theta\sin\psi & \sin^2\psi + \cos^2\psi \end{pmatrix}$$

$$= \begin{pmatrix} \cos^2\theta + \sin^2\theta\,(\cos^2\psi + \sin^2\psi) & 0 & 0 \\ 0 & \sin^2\theta + \cos^2\theta\,(\cos^2\psi + \sin^2\psi) & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \cos^2\theta + \sin^2\theta & 0 & 0 \\ 0 & \sin^2\theta + \cos^2\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = I$$

This result is trivial, because undo after rotation gives original matrix. However, this is very useful nature. We can rotate keeping orthogonality by multiplying regular matrix of sihc transposed matrix is inverse matrix.

We can obtain matrix $\boldsymbol{\Lambda}$ keeping orthogonality by multiplying rotation matrix $\boldsymbol{T}$.

$$\boldsymbol{A}\boldsymbol{T} = \boldsymbol{\Lambda}$$

$$\boldsymbol{T}^T\boldsymbol{A}^T = \boldsymbol{\Lambda}^T$$

We apply this transformation to the model of factor analysis

$$\boldsymbol{Z} = \boldsymbol{F}\boldsymbol{A}^T + \boldsymbol{U}\boldsymbol{D}$$

We do not make change in matrix $\boldsymbol{Z}$ and $\boldsymbol{U}\boldsymbol{D}$.

$$\boldsymbol{Z = FA^T + UD}$$
$$\boldsymbol{= FIA^T + UD}$$
$$\boldsymbol{= FTT^TA^T + UD}$$
$$\boldsymbol{= FT\Lambda^T + UD}$$

Here, $\boldsymbol{FT = G}$

$$\boldsymbol{Z = G\Lambda^T + UD}$$

We can explain $\boldsymbol{Z}$ by new factor $\boldsymbol{G}$ keeping $\boldsymbol{Z}$ and $\boldsymbol{UD}$.

We can rotate factors keeping distances among data (norm) in orthogonal rotation for easier. However, there are discussion of needs of keeping orthogonality, because model of factor analysis neglects orthogonality and considering only optimization of diagonal elements. Neglection of covariance is not means that covariance is 0. We can conjecture high independency among factors, because each factor is extracted independently. However, we cannot conclude no correlation among factors only by independency. Keeping orthogonality has no meaning in many cases in rotation. In such case, we select oblique rotation for rotation. We have to keep variance of data even in oblique rotation. Thus, diagonal elements of product of multiplication of rotation matrix and it transposed matrix rotation matrix should be 1.

$$\mathrm{diag}\boldsymbol{TT^T = I}$$

We put this relation to $\boldsymbol{Z = FA^T + UD}$

$$\boldsymbol{Z = FA^T + UD}$$
$$\boldsymbol{= FIA^T + UD}$$
$$\boldsymbol{= FTT^{-1}A^T + U}$$
$$\boldsymbol{= FT(T^T)^{T^{-1}}A^T + UD}$$
$$\boldsymbol{= FT(A(T^T)^{-1})^T + UD}$$

Consequently,

$$\boldsymbol{G = FT}$$
$$\boldsymbol{A(T^T)^{-1} = \Lambda}$$
$$\boldsymbol{Z = G\Lambda^T + UD}$$

This is process of calculation. We could not understand goal of calculation from this formula. We discuss the criterion of rotation in next paragraph.

Criterion of rotation

Purpose of rotation is to make easy interpretation of factors. In another word, the purpose is simplification.

The author gives an example of rotation, though he has no data in his file. Therefore, he use pick up a dataset cited in a text book（豊田秀樹(2012): 因子分析入門―R で学ぶ最新データ

解析。東京図書）and explain the function of rotation by his method. The dataset is result of questionnaire survey of evaluation of ski resort. It is not in his area of his specialty and he cannot understand the contents of the analysis. According to the text book, elements of the evaluation of resort include preference, activeness and magnitude. Respondents choose a option from Seven levels Likert items (1.Strongly disagree, 2.disagree, 3. slightly disagree, 4. neither agree nor disagree, 5. Slightly agree, 6. Agree, 7. Strongly agree) or Five levels Likert items (1.Strongly disagree, 2.disagree, 3. neither agree nor disagree, 4. Agree, 5. Strongly agree). Question items are like-dislike, fancy-unrefined, unique-ordinary, dynamic-static, bright-somber, strong- weak, hard-soft, stable-instable and small-large. Initial solution before rotation is as in following table.

Matrix of factor loadings before rotation

	Factor 1	Factor 2	Factor 3	Communality
Like	0.87	-0.36	-0.01	0.88
Fancy	0.94	-0.31	0.01	0.99
Unique	0.86	-0.37	-0.02	0.88
Dynamic	0.62	0.71	0.28	0.97
Bright	0.59	0.69	0.21	0.88
Strong	0.53	0.70	0.26	0.84
Hard	-0.37	-0.52	0.72	0.93
Stable	-0.22	-0.46	0.72	0.78
Small	0.30	0.29	-0.85	0.90

Communality is sum of square of factor loadings, and it indicate portion of variance which can be explained by factors. All communalities are satisfactory, and we can explain each question items by factors. We can vaguely recognize that factor 1 is relating to preference (like, fancy, unique), factor 2 is relating to activeness (dynamic, bright, strong), and factor 3 is relating to magnitude (hard, stable, small). However, factor loadings of factor 1 to dynamic, bright and strong are also relatively high. We cannot deny the relation between factor 1 and activeness from initial solution. Following table is results of orthogonal rotation (varimax method)

Factor loading matrix after orthogonal rotation (varimax method)

	Factor 1	Factor 2	Factor 3	Communality
Like	0.93	0.12	-0.05	0.88
Fancy	0.97	0.19	-0.08	0.99
Unique	0.93	0.11	-0.05	0.88
Dynamic	0.17	0.96	0.13	0.97
Bright	0.16	0.90	-0.18	0.88

Strong	0.10	0.90	-0.12	0.84
Hard	-0.07	-0.28	0.92	0.93
Stable	0.03	-0.17	0.87	0.78
Small	0.13	0.01	-0.94	0.90

We can clearly recognize that factor 1 is relating to preference, factor 2 is relating to activeness, and factor 3 is relating to magnitude after varimax rotation without making any changes in communality. In orthogonal rotation, we keep orthogonality of factors. Thus, we cannot make 0 factor loading. This is limitation of orthogonal rotation. Ideally, we want make following table by rotation.

An Example of perfect cluster solution

	Factor 1	Factor 2	Factor 3
Like	0.93	0	0
Fancy	0.97	0	0
Unique	0.93	0	0
Dynamic	0	0.96	0
Bright	0	0.90	0
Strong	0	0.90	0
Hard	0	0	0.92
Stable	0	0	0.87
Small	0	0	-0.94

We call this type of solution as perfect cluster solution. In perfect cluster solution, each question item is explained only by one factor. We are not always able to obtain perfect cluster solution even by oblique rotation, though perfect cluster solution is one of the goals of oblique rotation. Simple structure is an attitude in process of rotation. However, simple structure does not always have absolute value. A factor sometimes has relation to more than two observed variables, actually. Well known example is factors relating to achievement of national language, mathematics, science, social study and English. Those performances are relating to 2 factors. One is logical thought and the other is competence and memory. Achievement of mathematics and science is relating to logical thought and national language and social study is relating to competence and memory. However, English (capacity for learning language) has both relation to logical thought and competence and memory. Levels of aiming simplification should be considered depending on the nature of data, purpose of analysis and previous knowledge. This is the reason why various rotation standard are proposed.

Commonly used text books lack explanation of rotation standard. The author makes holistic easy explanation of rotation standards. We consider rotation in n dimensional space. The simplest approach is rotate around one axis giving rotating angle θ_1 and repeat this one axis

by one axis $n-1$ times. We consider example of two factor and 4 observed variables.

Factor loading matrix of initial solution before rotation

	Factor 1	Factor 2
Z1	0.87	-0.36
Z2	0.94	-0.31
Z3	0.62	0.71
Z4	0.59	0.69

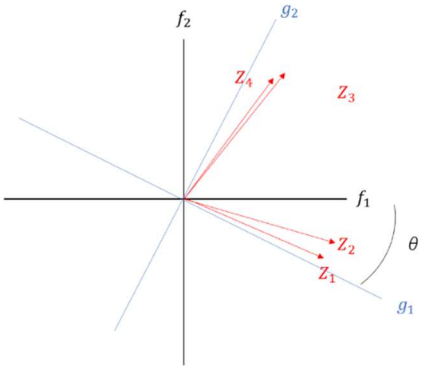

Fig. 89. Vector of observed variables when factor 1 and factor 2 are 1.

Figure 89 is plot of factor loading matrix when factor 1 and factor 2 are 1. Simplification by orthogonal rotation means rotation of axes $f_1 - f_2$ to overlap axes $g_1 - g_2$ by rotating θ keeping orthogonality of axes, in this case rotation angle is negative. We consider range of rotation angle θ

$$-\frac{\pi}{2} < \theta < \frac{\pi}{2}$$

Purpose of rotation is to approximate angles between g_1 and Z_1, Z_2, $(\psi_{11}, \varphi_{12})$, to 0, angles between g_1 and $Z_3, Z_4, (\psi_{13}, \varphi_{14})$, to $\frac{\pi}{2}$, angles between g_2 and $Z_1, Z_2, (\psi_{21}, \varphi_{22})$, to $-\frac{\pi}{2}$, and angles between g_2 and Z_3, Z_4, $(\psi_{23}, \varphi_{24})$, to 0. Rotation matrix in two dimensional space, $\boldsymbol{T}$, is as follow.

$$\boldsymbol{T} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

Rotation of factor loading matrix $\boldsymbol{A}$ is as follow.

$$\boldsymbol{A} = \begin{pmatrix} 0.87 & -0.36 \\ 0.94 & -0.31 \\ 0.62 & 0.71 \\ 0.59 & 0.69 \end{pmatrix}$$

$$\boldsymbol{AT} = \begin{pmatrix} 0.87 & -0.36 \\ 0.94 & -0.31 \\ 0.62 & 0.71 \\ 0.59 & 0.69 \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

We guess $\theta \cong -\frac{\pi}{6}$ from figure 89.

$$\boldsymbol{AT} = \begin{pmatrix} 0.87 & -0.36 \\ 0.94 & -0.31 \\ 0.62 & 0.71 \\ 0.59 & 0.69 \end{pmatrix} \begin{pmatrix} \cos\left(-\frac{\pi}{6}\right) & -\sin\left(-\frac{\pi}{6}\right) \\ \sin\left(-\frac{\pi}{6}\right) & \cos\left(-\frac{\pi}{6}\right) \end{pmatrix} = \begin{pmatrix} 0.87 & -0.36 \\ 0.94 & -0.31 \\ 0.62 & 0.71 \\ 0.59 & 0.69 \end{pmatrix} \begin{pmatrix} \cos\left(\frac{\pi}{6}\right) & \sin\left(\frac{\pi}{6}\right) \\ -\sin\left(\frac{\pi}{6}\right) & \cos\left(\frac{\pi}{6}\right) \end{pmatrix}$$

$$\boldsymbol{AT} = \begin{pmatrix} 0.87 & -0.36 \\ 0.94 & -0.31 \\ 0.62 & 0.71 \\ 0.59 & 0.69 \end{pmatrix} \begin{pmatrix} 0.866025 & 0.5 \\ -0.50 & 0.866025 \end{pmatrix} = \begin{pmatrix} 0.9247815 & 0.118231 \\ 0.9690635 & 0.20153225 \\ 0.1819355 & 0.92487775 \\ 0.16595475 & 0.89255725 \end{pmatrix}$$

$$\boldsymbol{\Lambda} = \begin{pmatrix} 0.9247815 & 0.118231 \\ 0.9690635 & 0.20153225 \\ 0.1819355 & 0.92487775 \\ 0.16595475 & 0.89255725 \end{pmatrix}$$

We can simplify even by rotation using approximate guess. We may say that approximate guess is useful in a sense, though it cannot give optimal solution. Loadings of factor 1 before rotation are $0.87,\ 0.94,\ 0.62$ and 0.59. The loadings after rotation are 0.92, 0.97, 0.18 and 0.17. Variation in loadings is increased by rotation. Several readers may feel the variance in factor 2 is not increased by rotation. However, factor loading is regression coefficient of factor and slope of variance to the axis. Factor loading fluctuate between -1 to 1. It can be negative and we have to consider the absolute value of factor loading. We need to consider variance in square of loading, For factor 2 we have to consider changes in sum of square of loading.

Factor 2 before rotation

$$\frac{0.36^2 + 0.31^2 + 0.71^2 + 0.69^2}{4} = 0.301475$$

$$(0.36^2 - 0.301475)^2 + (0.31^2 - 0.301475)^2 + (0.71^2 - 0.301475)^2 + (0.36^2 - 0.301475)^2 = 0.143271$$

Factor 2 after rotation

$$\frac{0.12^2 + 0.20^2 + 0.92^2 + 0.89^2}{4} = 0.423225$$

$$(0.12^2 - 0.423225)^2 + (0.20^2 - 0.423225)^2 + (0.92^2 - 0.423225)^2 + (0.89^2 - 0.423225)^2 = 0.629145$$

We generalize the case in 2 factor and 4 variables

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & \lambda_{32} \\ \lambda_{41} & \lambda_{42} \end{pmatrix}$$

$$\mu_1 = \frac{1}{m}\sum_{j=1}^{m} {\lambda_{j1}}^2$$

$$SS_1 = \sum_{j=1}^{m} \left({\lambda_{j1}}^2 - \mu_1\right)^2$$

$$= \sum_{j=1}^{m} {\lambda_{j1}}^4 - 2\mu_1 \sum_{j=1}^{m} {\lambda_{j1}}^2 + m{\mu_1}^2$$

$$\sum_{j=1}^{m} {\lambda_{j1}}^4 - 2\frac{1}{m}\sum_{j=1}^{m} {\lambda_{j1}}^2 \sum_{j=1}^{m} {\lambda_{j1}}^2 + m\left(\frac{1}{m}\sum_{j=1}^{m} {\lambda_{j1}}^2\right)\left(\frac{1}{m}\sum_{j=1}^{m} {\lambda_{j1}}^2\right)$$

$$= \sum_{j=1}^{m} {\lambda_{j1}}^4 - \frac{1}{m}\left(\sum_{j=1}^{m} {\lambda_{j1}}^2\right)\left(\sum_{j=1}^{m} {\lambda_{j1}}^2\right)$$

$$= \sum_{j=1}^{m} {\lambda_{j1}}^4 - \frac{1}{m}\left(\sum_{j=1}^{m} {\lambda_{j1}}^2\right)^2$$

There are p factors.

$$\mathrm{Q} = \sum_{k=1}^{p}\left(\sum_{j=1}^{m} {\lambda_{jk}}^4 - \frac{1}{m}\left(\sum_{j=1}^{m} {\lambda_{jp}}^2\right)^2\right)$$

$$= \sum_{k=1}^{p}\sum_{j=1}^{m} {\lambda_{jp}}^4 - \frac{1}{m}\sum_{k=1}^{p}\left(\sum_{j=1}^{m} {\lambda_{jp}}^2\right)^2$$

This Q is varimax rotation standard. We select θ to maximize Q. There are several formula of $\mathrm{Q}(\lambda)$, and they are called rotation standard. We calculate extreme vale for selection of rotation angle. Orthomax standard is unified expression of orthogonal rotation standard.

Orthomax standard

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_{11} & \cdots & \lambda_{1p} \\ \vdots & \ddots & \vdots \\ \lambda_{m1} & \cdots & \lambda_{mp} \end{pmatrix}_{m\times p}$$

$$\mathrm{Q}_{or} = \sum_{j=1}^{p}\sum_{k=1}^{m} {\lambda_{jk}}^4 - \frac{\omega}{m}\sum_{k=1}^{p}\left(\sum_{j=1}^{m} {\lambda_{jk}}^2\right)^2$$

ω: weight

Formula 79

Orthomax standard include 6 orthogonal rotation, namely quartimax rotation, biquartimax rotation, varimax rotation, equamax rotation, parsimax rotation, factor parsimony rotation. Difference of rotations are difference of ω.

Orthomax standard and rotation method

Rotation method	ω

Quartimax	0
Biquartimax	1/2
Varimax	1
Equamax	p/2
Parsimax	m(p-1)/(m+p-2)
Factor parsimony	m

p is number of factor. m is number of observed variables.

We developed formula of varimax standard theoretically.

$$\mathrm{Q}_v = \sum_{k=1}^{p}\sum_{j=1}^{m} {\lambda_{jk}}^4 - \frac{1}{m}\sum_{k=1}^{p}\left(\sum_{j=1}^{m} {\lambda_{jk}}^2\right)^2$$

This is a form of simplified calculation of sum of square and second term is expectation value of ${\lambda_j}^2$. In varimax standard, representative (center of distribution, expectation value) of column of factor loading matrix is average. We should not necessarily consider expectation value is average. In quatimax, standard, expectation value is 0 . λ_j is slope. When we consider average of slope is 0 originally, the logic of quatimax standard has valid point. In biquatimax standard, expectation values exist at midpoint between o and average. Vale of ω is weight in calculation. Addition to this, ω determines expectation of slope. Larger ω makes smaller the difference of explanation power among factors and smaller ω gives strong explanation power to particular factors.

In this explanation λ_{kj} is function of θ. This means that Q can be expressed as function of θ, and $Q(\boldsymbol{\theta})$ can be differentiated by vector $\boldsymbol{\theta}$.

$$\frac{dQ(\boldsymbol{\theta})}{d\boldsymbol{\theta}} = 0$$

We can obtain $\boldsymbol{\theta}$ by solving upper differential equation. Theoretically, this explanation is perfect, though it is not orthodox explanation in common text books. One of the week ness is that we have to make formula of rotation in multiple dimensional space. It is not impossible though tedious. This is not essential weakness. More essentially, we cannot expand this idea to oblique rotation, because rotation by fixed angle keeps to norm and we can apply this method only for orthogonal rotation. For the unified expression, we put constraint condition to the differentiation equation.

The constraint condition for orthogonal rotation is

$$\boldsymbol{T}\boldsymbol{T}^T = I$$

The constraint condition for oblique rotation is

$$\mathrm{diag}\boldsymbol{T}\boldsymbol{T}^T = I$$

Then we solve following differential equation

$$\frac{dQ(\boldsymbol{T})}{d\boldsymbol{T}} = 0$$

For this, we need the skill to solve extreme value with constraint condition. This is a kind of trade off. Solving function of rotation angles has fewer unknowns and we the range of unknowns are limited in $-\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}$. However, we have to make rotation matrix to each dimension. When we calculate rotation matrix by solution of extreme value with constraint condition problem, number of unknown variables increase though we do not need to consider rotation in multi-dimensional space and solution of extreme value with constraint condition problem has broad utility. We use method of Lagrange's multiplier for solution of extreme value with constraint condition which is already explained in V-2-6 maximum and minimum. In V-2-6, the author explained showing flat image,

In oblique rotation, there are two major approaches. One is rotation under constraint condition, and the other is rotation to approach to particular goals to follow previous knowledges. Author is thinking personally that approaches to follow previous knowledge is not rotation but adjustment.

Oblimin standard is standard for oblique rotation.

Oblimin standard

$$\mathrm{Q}_{ob} = \sum_{k<l=1}^{p} \left\{ \sum_{j=1}^{m} {\lambda_{jk}}^2 {\lambda_{jl}}^2 - \frac{\omega}{m} \left(\sum_{j=1}^{m} {\lambda_{jk}}^2 \right) \left(\sum_{j=1}^{m} {\lambda_{jl}}^2 \right) \right\}$$

Oblimin standard and rotation method

Rotation method	ω
Qurtimin	0
Biquartimin	1/2
Covarimin	1

Standard of covarimin rotation is the form of simplified calculation of covariance.

We solve following differential equation to obtain $\boldsymbol{T}$ under constraint condition of $\mathrm{diag}\boldsymbol{T}\boldsymbol{T}^T = I$.

$$\frac{dQ(\boldsymbol{T})}{d\boldsymbol{T}} = 0$$

Orthogonal rotation is keeping orthogonality of initial solution. However, model of factor analysis neglect orthogonality of factors originally. We cannot confirm orthogonality among factors in initial solution. Second term in oblimin standard is relating to correlation among

factors. We can consider several oblique rotations are method to give proper orthogonality among factors ex-post facto. When ω increase the oblique rotation close to orthogonal rotation. There are various methods and standard in oblique rotation other than oblimin standard, namely Crawford-Ferguson family standard, Geomin standard, Harris-Kaiser's independent cluster rotation. The author has enough knowledge and experiences to explain the availability and characteristics of each standard and methods.

Procrustes rotation and Promax rotation.

The author cannot understand why we call oblique rotation as rotation. Oblique rotation is not rotation but adjustment, and we can accept applicability of adjustment in the sense of the author, because model of factor analysis neglects rigorous orthogonality originally. When we accept applicability of adjustment, it can be realistic method to determine $\boldsymbol{T}$ to the semblance of previous information. There are counter opinions to such realistic approach from the viewing point of exploratory data analysis. However, when there exists reliable previous research in some degree, we can accept such approaches as hypothetical verification. We call this Procrustes rotation. In Procrustes rotation $\boldsymbol{T}$ proposed in a previous work is used as target and the difference to the target is minimized. When we accept this method, we can use result of orthogonal rotation as the target. This method is called promax rotation (Procrustes +varimax). However when we approximate $\boldsymbol{T}$ in oblique rotation to the result of varimax rotation it becomes $\boldsymbol{T}$ obtained in varimax rotation. For this reason third power or forth power of factor loading by varimax rotation are used as target. High loading factors close to 1 are kept in high value and low loading factors became smaller by this adjustment.

Actually, selection of rotation method is difficult and troublesome. Many researchers use promax rotation at first. They may accept promax rotation as adjustment of varimax rotation, because varimax rotation is logically natural. Depending on the results of promax rotation, we consider whether we can accept the result of promax rotation, or we need try other rotations. As an example, when we estimate existence of observed variables which can be explained by plural factors, we consider trial of goemin rotation which gives results of low simplification level. Or when the orthogonality among factors is high, we back to varimax rotation, because varimax rotation is method of which arbitrariness is low. The author does not have enough experiences to discuss the selection of proper rotation method. There are various discussions for the selection of rotation methods. Please refer such discussions and select proper method flexibly.

Method of Lagrange multiplier. Solution of extreme value with constraint condition.

Method of Lagrange multiplier is explained in V-2-6 maximum and minimum. In that chapter, the author used flat image, though method of Lagrange multiplier is generally used in multidimensional space.

Problem of extreme value with constraint condition is as follows.

There is a function of $\boldsymbol{X}$

$$f(\boldsymbol{X}) = f(x_1, \cdots, x_n)$$

Variables of $\boldsymbol{X}$ has constraint condition such as $g(\boldsymbol{X}) = c$

Find extreme value of $f(\boldsymbol{X})$ under the constraint condition.

An example of constraint condition is

$$x_1{}^2 + \cdots + x_n{}^2 = k$$

$$x_1{}^2 + \cdots + x_n{}^2 - k = 0$$

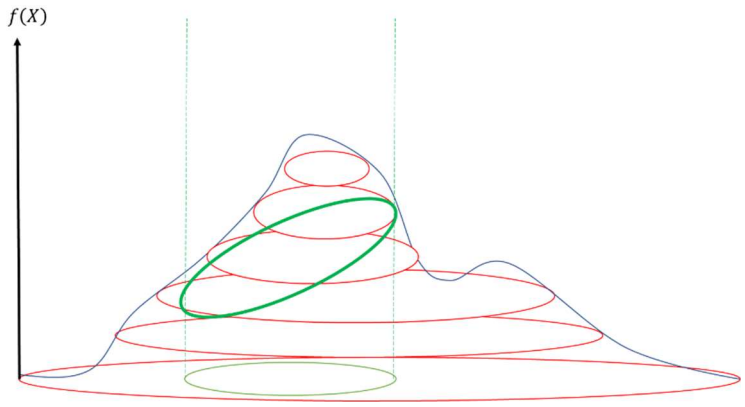

Fig. 90. Function $f(\boldsymbol{X})$ and constraint condition $g(\boldsymbol{X})$ in multidimensional space.

We draw shape of constraint condition on $x_1, \cdots, x_n$ multidimensional space (ellipse drown with fine green line), and then we project it to the curved surface of $f(\boldsymbol{X})$ in n+1dimensional space. (ellipse drawn by heavy green line, actually it is not always ellipse) the extreme value is the top of the heavy green line. When we can project the constraint condition to the surface of function, it becomes problem of extreme value without constraint condition. Most important information obtainable from the figure is that the tangent line of contour line of $f(\boldsymbol{X})$ (showing by red ellipse) is the same to tangent line of projected ellipse. In the case when two ellipses do not share the same tangent line, the point on the green line remove from the counter line by shortest translation on the green line. This means increase or decrease of the function. Thus, tangent line and normal line of function and constraint condition are the same at extreme point.

$$\frac{df(\boldsymbol{X})}{dX} = \frac{dg(\boldsymbol{X})}{dX}$$

Lagrange function is expressing upper relation.

Lagrange function.

$$L(x_1, \cdots x_n, \lambda) = f(x_1, \cdots x_n) - \lambda g(x_1, \cdots x_n)$$

We can obtain $x_1, \cdots x_n,$ by solving following differential equation.

$$\frac{\partial L}{\partial x_1} = \cdots = \frac{\partial L}{\partial x_n} = \frac{\partial L}{\partial \lambda} = 0$$

We call λ as Lagrange's multiplier.

At first, we consider meaning of $\frac{\partial L}{\partial \lambda} = 0$

$$\frac{\partial L}{\partial \lambda} = \frac{\partial f(x_1, \cdots x_n)}{\partial \lambda} - g(x_1, \cdots x_n)\frac{\partial \lambda}{\partial \lambda} = 0$$

$f(x_1, \cdots x_n)$ does not include λ. This means

$$\frac{\partial f(x_1, \cdots x_n)}{\partial \lambda} = 0$$

$$\frac{\partial \lambda}{\partial \lambda} = 1$$

Consequently, $-g(x_1, \cdots x_n) = 0$

$$g(x_1, \cdots x_n) = 0$$

Therefore, $\frac{\partial L}{\partial \lambda} = 0$ means constraint condition.

We consider meaning of $\frac{\partial L}{\partial x_1}, = \cdots = \frac{\partial L}{\partial x_n} = 0$

We line up all differential equation vertically.

$$\frac{\partial L}{\partial x_1} = \frac{\partial f(x_1, \cdots x_n)}{\partial x_1} - \lambda\frac{\partial g(x_1, \cdots x_n)}{\partial x_1} = 0$$

$$\frac{\partial L}{\partial x_2} = \frac{\partial f(x_1, \cdots x_n)}{\partial x_2} - \lambda\frac{\partial g(x_1, \cdots x_n)}{\partial x_2} = 0$$

$$\vdots$$

$$\frac{\partial L}{\partial x_n} = \frac{\partial f(x_1, \cdots x_n)}{\partial x_n} - \lambda\frac{\partial g(x_1, \cdots x_n)}{\partial x_n} = 0$$

Transposition

$$\frac{\partial f(x_1, \cdots x_n)}{\partial x_1} = \lambda\frac{\partial g(x_1, \cdots x_n)}{\partial x_1}$$

$$\frac{\partial f(x_1, \cdots x_n)}{\partial x_2} = \lambda\frac{\partial g(x_1, \cdots x_n)}{\partial x_2}$$

$$\vdots$$

$$\frac{\partial f(x_1,\cdots x_n)}{\partial x_n}=\lambda\frac{\partial g(x_1,\cdots x_n)}{\partial x_n}$$

We consider $\frac{\partial f(x_1,\cdots x_n)}{\partial x_i},\frac{\partial g(x_1,\cdots x_n)}{\partial x_1}$ as elements of vectors.

$$\begin{pmatrix}\frac{\partial f(x_1,\cdots x_n)}{\partial x_1}\\ \vdots \\ \frac{\partial f(x_1,\cdots x_n)}{\partial x_n}\end{pmatrix}=\lambda\begin{pmatrix}\frac{\partial g(x_1,\cdots x_n)}{\partial x_1}\\ \vdots \\ \frac{\partial g(x_1,\cdots x_n)}{\partial x_n}\end{pmatrix}$$

$$\vec{f}=\lambda\vec{g}$$

Length of $\vec{f}\ and\ \vec{g}$ is different though the same in the direction. Conclusively, method of Lagrange multiplier is saying that total differentials of $f(x_1,\cdots x_n)$ and $g(x_1,\cdots x_n)$ are parallel at extreme value.

Unnecessary addition,

$$f(x_1,\cdots x_n)=\mathrm{c}$$

We translate the point shortest distance along tangent line. Tangent line is the same as contour line. Thus,

$$f(x_1+\Delta x_1,\cdots x_n+\Delta x_n)=c$$

On the other hand, we translate along the slope of the tangent flat.

$$f(x_1+\Delta x_1,\cdots x_n+\Delta x_n)=c+\frac{\partial f(x_1,\cdots x_n)}{\partial x_1}\Delta x_1+\cdots+\frac{\partial f(x_1,\cdots x_n)}{\partial x_n}\Delta x_n$$

$$\frac{\partial f(x_1,\cdots x_n)}{\partial x_1}\Delta x_1+\cdots+\frac{\partial f(x_1,\cdots x_n)}{\partial x_n}\Delta x_n=0$$

This can be expressed as follow.

$$\left(\frac{\partial f(x_1,\cdots x_n)}{\partial x_1}\quad\cdots\quad\frac{\partial f(x_1,\cdots x_n)}{\partial x_n}\right)\begin{pmatrix}\Delta x_1\\ \vdots \\ \Delta x_n\end{pmatrix}=0$$

This is inner product of vector $\left(\frac{\partial f(x_1,\cdots x_n)}{\partial x_1}\quad\cdots\quad\frac{\partial f(x_1,\cdots x_n)}{\partial x_n}\right)$ and vector$(\Delta x_1\quad\cdots\quad\Delta x_n)$.

$(\Delta x_1\quad\cdots\quad\Delta x_n)$ is tangent flat. Consequently, $\left(\frac{\partial f(x_1,\cdots x_n)}{\partial x_1}\quad\cdots\quad\frac{\partial f(x_1,\cdots x_n)}{\partial x_n}\right)$is orthogonal to the tangent flat. We call this as gradient and expressed as ∇f. Using this, Method of Lagrange multiplier is expressed as follow.

$$\nabla f=\lambda\nabla g$$

This means normal lines of f and g are the same at a hyperplane.

VI-2-3-5. Proper number of factors

Addition to method of factor extraction and rotation method, number of factor is very important in factor analysis. In factor analysis, we presume existence of independent factor which have orthogonality among them in some degree, and we calculate the strength of the relation (factor loading) between the factors and observed variables. When number of factors is improper, reliability of analysis decrease. Thus, number of factors is essential issue in factor analysis.

Scree plot

Model and purpose of principle component is completely different from factor analysis. However, principle component analysis shows position of data in orthogonal hyper place and it provides useful information for the selection of method and number of factors in factor analysis. It is natural to consider that direction of vector of factors similar to the major principle compositions. Actually, early software for factor analysis include method to use result of principle component analysis as initial solution of factor analysis. In principle component analysis, we make following figure taking contribution ratio at vertical axis 案 line up from left side in the order of contribution ratio for decision of number principle component. The name of the figure is scree plot. We look for point where the contribution ratio is steeply decrease. We call the point scree, and we select component in left side of the scree as principle components.

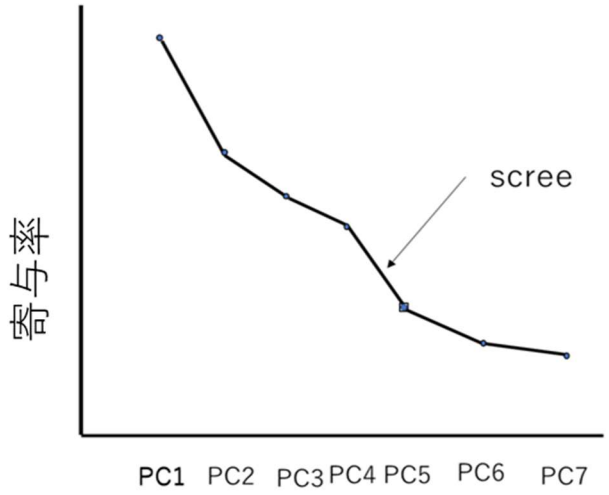

Fig. 91 Scree plot

We suppose that proper number of factors is similar to the number of principle components. This is an reasonable idea for the selection of number of factors. Contribution ratio is variance of component. When the measurement units of observed value are extremely different, the comparison of variance has no meaning, though when the data was standardized by standard deviation the variance means expansion of the data to the direction and the scree plot provide

information for the decision of number of factors, and we can judge components which have small variance have no meaning.

Guttman standard

When we standardize the data by standard deviation of started principle component analysis, the variance of variables are unified to 1. The principle components of which eigen values are more than 1 are stronger power in explanation of the phenomenon than average. From this, we can consider there are same number of factor with the number of principle components of which eigenvalues are larger than 1. This is called Guttman standard.

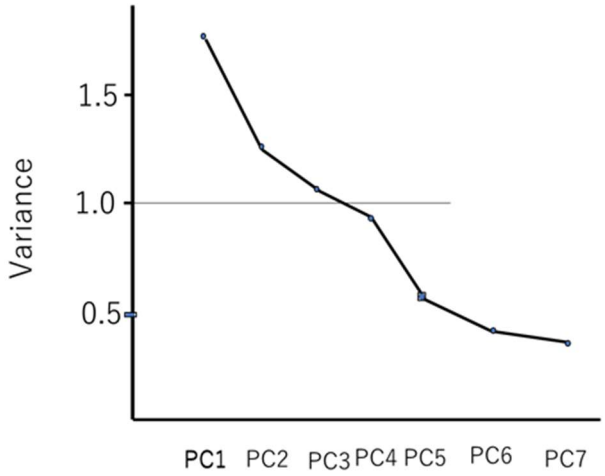

Fig. 92 Judgement by Guttman standard

Parallel analysis

When the variance is small, there is a possibility that the reason of small variance is influence of existence of components with strong explanation power. This means standard of judgement is not constant with the number of components. One possible method is to make same size random dataset with actual data and implement principle component analysis. We compare the eigenvector of both data. Random dataset does not include particular principle component its eigenvalues are expected eigenvalues of error. However, it include random fluctuation and eigenvalues of several components are higher than 1 and eigenvalues of the others are lower than 1. The slope of random dataset is gentle (red line). We select number of components existing left side of the intersection point. This is parallel analysis.

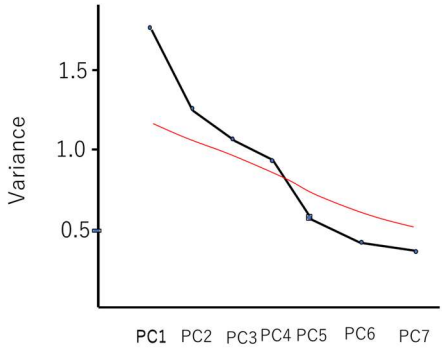

Fig. 93. Judgement by parallel analysis

MAP(Minimum Average Partial) test

Judgement of scree plot, Guttman standard and parallel analysis are method to consider candidate of number of factors depending on the result of scree plot of results of principle component analysis. The weakness of this method is that principle component analysis is completely different from factor analysis. The prior information provided by principle component is useful, however principle component analysis is not factor analysis. In principle component analysis, number of principle component is the problem of how we explain the phenomenon in detail. The results of principle component analysis do not change depending on the number of principle component. However, result of factor analysis changes with the number of factors in factor analysis. For this reason, we have to discuss adequacy of number of factors depending on the results of factor analysis. Minimum average partial test (MPA test) is a method to discuss the adequacy of number of factors.

In MPA, we consider factor as control variables. Control variable was explained in ***VI-1-2. Multicollinearity and partial correlation analysis.*** Here, we consider that the correlations between observed variables are made by the strong correlation with control variables (factor) and there are no correlation originally. If so, when we remove the influence of the control variable, the correlation between the observed apparent variables will disappear. In some case, there are true correlation between the observed variables. When partial correlation is small, apparent correlation is caused by strong influence of control variable (factor). When there is little partial correlation between 2 observed variables to a factor, the sum of square of partial correlation is small. This means that the phenomenon is explained exclusively by control variables(factors). In the case when number of observed variables is 6 and number of factors is 3. We select 2 variables from 6 observed variables to each factor. The number of combinations in each factor is

$$ {}_6C_2 = \frac{6 \times 5}{2} = 15 $$

Total number is product of multiplying number of factors to the number of combination.

$$ 15 \times 3 = 45 $$

We sum up square of partial correlation of 45 combinations. We compare this value among different number of factors.

This is a method to discuss adequacy of number of factors ex-post facto. This method is logical and applicable. Of course, it is preferable to include variables which has higher correlation in multiple variance analysis generally. However, we sometimes use variables which is not completely independent unavoidably. Hypothesis of no relationship among observed variables which have higher factor loadings is not realistic. In such case, MPA test make underestimation. However, partial correlation analysis among observed variables in each factor provide useful information for interpretation of results.

Conclusively, MPA test recommends smaller number of factors and parallel analysis recommend larger number of factors. We select proper number of factors between two recommended numbers.