# FACTOR ROTATION AND ICA

*Yutaka KANO, Yusuke MIYAMOTO and Shohei SHIMIZU*

Graduate School of Human Sciences, Osaka University, Suita, Osaka 565-0871, Japan
kano@hus.osaka-u.ac.jp

## ABSTRACT

Similarities and distinctions have been pointed out between ICA and traditional multivariate methods such as factor analysis, principal component analysis and projection pursuit. In this paper, a new important connection between ICA and traditional factor analysis is made. The key of the connection is "factor rotation."

## 1. INTRODUCTION

Let $\boldsymbol{X}$ be an observed $p$-vector. The factor analysis model for $\boldsymbol{X}$ is written as

$$\boldsymbol{X} = \boldsymbol{\mu} + A\boldsymbol{s} + \boldsymbol{u}, \tag{1}$$

where $\boldsymbol{s}$ is an $m$-vector of latent (or hidden) factors or blind signals, $\boldsymbol{u}$ being a vector of unique factors or error factors, and $\boldsymbol{\mu}$ is a general mean vector and $A$ is a factor loading matrix or mixing matrix. When $\boldsymbol{u} = \boldsymbol{0}$, the model is said to be noise-free ICA or simply ICA (e.g., Hyvärinen 1999); otherwise, it is called noisy ICA. Here we do not seriously distinguish between noisy and noise-free ICA. In traditional factor analysis, $\boldsymbol{u}$ is an important term (to be zero hardly) and considered as a sum of a unique factor vector and an error vector.

Although the equation (1) represents both ICA and traditional factor analysis, there are substantial differences between them. In ICA, the components of $\boldsymbol{s}$ are distributed *independently and nonnormally.*[1] The independence and nonnormality are key assumptions. The traditional factor analysis does not use the independence and nonnormality assumptions, and rather than those, normality is often assumed. The second distinction is that ICA uses information other than second-order moments such as third-order and fourth-order moments in estimation, while the traditional factor analysis only uses second-order moments. When latent factors are not actually independent of one another, ICA is equivalent to projection pursuit (Friedman and Tukey 1974). See Hyvärinen and Kano (in press) for instance.[2]

Traditional factor analysis requires factor rotation. Here we simply consider the orthogonal factor model, $\mathrm{Var}(\boldsymbol{s}) = I_m$, so that the covariance structure of $\boldsymbol{X}$ is derived as

$$\mathrm{Var}(\boldsymbol{X}) = AA^T + \Psi, \tag{2}$$

where $\mathrm{Var}(\boldsymbol{u}) = \Psi$. If a matrix $A$ satisfies the equation (2), so does $AP$ for any orthogonal matrix $P$. The so-called factor rotation problem is caused by the fact that the factor analysis uses only second-order moments. On the other hand, ICA does not have this problem, and determine the rotation matrix $P$ by maximizing independency among latent factors $\boldsymbol{s}$. This is the third distinction.

As a result, ICA and traditional factor analysis are totally different procedures, and of course, they result in completely different outputs for the same data set.

## 2. QUICK REVIEW OF ESTIMATION PROCEDURE IN ICA

When noise-free ICA is considered (i.e., $\boldsymbol{u} = \boldsymbol{0}$ in (1)), it is useful to sphere $\boldsymbol{X}$ as a pre-analysis of ICA. We can assume that $\mathrm{Var}(\boldsymbol{s}) = I_m$ with no loss of generality. If the shpering is made for $\boldsymbol{X}$, then the mixing matrix $A$ is column-orthogonal. Sphering is often made by principal component analysis (PCA). PCA is also useful in dimensional reduction. If the dimension of $\boldsymbol{X}$ reduces to the number $m$ of blind signals, the $A$ may be restricted to be an orthogonal matrix of order $m$. Here

The author would like to express his special thanks to Professor Noboru Murata of Waseda University for giving an intensive lecture on ICA at Osaka University in 2002. The lecture note is available at `http://www.murata.elec.waseda.ac.jp/~mura/Lecture/ica/note/`

[1] One component is allowed to follow normally (see Comon (1994)).

[2] ICA provides an important characterization of projection pursuit that if an observed vector consists of linearly-mixed independent latent factors, the projection pursuit identifies the independent factors.

we consider the simple case where $p = m$, $\boldsymbol{u} = \boldsymbol{0}$ and $\boldsymbol{X}$ is sphered.

Let $\hat{\boldsymbol{s}} = W^T\boldsymbol{X}$ and let us estimate $W = [\boldsymbol{w}_1, \dots, \boldsymbol{w}_m]$, an $m \times m$ (orthogonal) matrix. A traditional criterion in estimation of ICA is to maximize the sum of squared fourth-order cumulants or kurtoses of $\hat{\boldsymbol{s}}$, that is,

$$\max_{W:m\times m} \sum_{r=1}^{m} \mathrm{Cum}^2(\hat{s}_r) = \max_{W:m\times m} \sum_{i=1}^{m} \mathrm{Cum}^2(\boldsymbol{w}_r^T\boldsymbol{X}). \tag{3}$$

See Comon (1994, Theorem 16) for instance. If one has sphered $\boldsymbol{X}$, $W$ can be restricted to be orthogonal as noted in the previous section.

Let $\kappa_{r_1\dots r_k}$ be the $k$-th order cumulants of $\hat{\boldsymbol{s}}$. The criterion above is written as $\sum_{r=1}^{m}\kappa^2_{rrrr}$. Comon has proposed many types of criteria in estimation of ICA, which includes

$$\frac{1}{48}\sum_{r=1}^{m}\left(4\kappa^2_{rrr} + \kappa^2_{rrrr} + 7\kappa^4_{rrr} - 6\kappa^2_{rrr}\kappa_{rrrr}\right). \tag{4}$$

This comes from the minimization of the mutual information approximated by the Edgeworth expansion of a density function. A simpler version of this was presented as $\frac{1}{48}\sum_{r=1}^{m}\left(4\kappa^2_{rrr} + \kappa^2_{rrrr}\right)$ (e.g., Hyvärinen et al. 2001, page 115). This is what Jones and Sibson (1987) studied in the context of projection pursuit. Comon (1994) also studied a criterion using all the $k$-th order cumulants: $\sum_{r_1,\dots,r_k}\kappa^2_{r_1\dots r_k}$.

Cardoso and Souloumiac (1996) made an alternative approach using the cumulants of the data vector $\boldsymbol{X}$. Recall that $\boldsymbol{X} = A\boldsymbol{s}$ and let $N = (n_{ij})$ be an arbitrary given matrix of order $m$. The $m\times m$ matrix $C(N)$ with the $(i, j)$ element $\sum_{k,l=1}^{m}\mathrm{Cum}(X_i, X_j, X_k, X_l)n_{kl}$ can be expressed as

$$AD_{N,A}A^T, \tag{5}$$

where $D_{N,A}$ is a diagonal matrix possibly depending on $N$, $A$ and the kurtoses of $\boldsymbol{s}$. Note that $A$ is independent of $N$. When one can take $A$ to be an orthogonal matrix, (5) represents the diagonalization of the symmetric matrix $C(N)$. The so-called JADE procedure is to find an orthogonal matrix $A$ that makes the joint-diagonalization of $C(N)$ for several fixed matrices $N$'s.

Although the idea of ICA using cumulants is simple, it may not be stable nor robust against outliers because neither is estimation of the higher-order moments. Hyvärinen (1999b) suggested

$$\max_{A\in\mathcal{O}(m)}\Big[E(G(A\boldsymbol{s})) - E(G(A\boldsymbol{Z}))\Big]^2, \tag{6}$$

where $G(\cdot)$ is a nonlinear and nonquadratic function and $\boldsymbol{Z}$ follows according to the standard multivariate normal distribution. Hyvärinen suggests the hyperbolic tangent function as the $G(\cdot)$.

In the context of factor analysis, Mooijaart (1985) studied the generalized least squares estimation using the second- and third-order moments for the model with nonnormal independent factors, and found that the rotation problem does not take place if the skewness' of the factors are all different. Thus, this is equivalent to an estimation procedure in the noisy ICA using skewness. However, Mooijaart and his followers have not noticed that his procedure is able to seperate independnet blind signals.

## 3. FACTOR ROTATION AND ICA ESTIMATION

For the rotation problem as described in Section 1, the factor analysis chooses *most interpretable* rotation. High interpretability often achieves when the elements of $A$ have high contrast and many zeros. This idea connects with several techniques of ICA, for example sparse coding. Many mathematical ways for implementing the idea have been developed and they are installed as a standard option in the programs of factor analysis. The varimax rotation by Kaiser (1958) among others is most often applied. The orthomax procedure is an extension of the varimax.

Let $\mathcal{O}(m)$ be the class of all orthogonal matrices of order $m$. Let $B = (b_{ir}) = AP$ for a $P \in \mathcal{O}(m)$. The varimax rotation procedure determines the rotation matrix $P$ as a solution to the maximization problem:

$$\max_{P\in\mathcal{O}(m)}\sum_{r=1}^{m}\sum_{i=1}^{p}(b^2_{ir} - \bar{b^2_r})^2, \tag{7}$$

where $\bar{b^2_r}$ is the average of the squared elements of the $r$-th column of $B$. The orthomax procedure (Crawford and Ferguson 1970) is given as

$$\max_{P\in\mathcal{O}(m)}\left[\sum_{r=1}^{m}\sum_{i=1}^{p}b^4_{ir} - \frac{\omega}{p}\sum_{r=1}^{m}\Big(\sum_{i=1}^{p}b^2_{ir}\Big)^2\right]. \tag{8}$$

The orthomax criterion provides a family of rotation methods by choosing values of $\omega$. The orthomax with $\omega = 1$ gives the varimax procedure. The idea behind them is that when many of $b^2_{ir}$ are close to zero or large, these criteria, which are basically variance of $b^2_{ir}$, will have a large value.

The idea that the variance of $b^2_{ir}$ $(i = 1, \dots, p)$ be maximized is closely related to the maximization of the cumulant of $b_{ir}$ $(i = 1, \dots, p)$. In fact, if we take $\omega = 3$

in (8), the criterion of the orthomax becomes

$$p\sum_{r=1}^{m}\left[\frac{1}{p}\sum_{i=1}^{p}b_{ir}^{4}-3\Big(\frac{1}{p}\sum_{i=1}^{p}b_{ir}^{2}\Big)^{2}\right]. \tag{9}$$

Let $\kappa_{rrrr}$ be the fourth-order cumulant of the elements of the $r$-th column of $B$. If we centralize $b_{ir}$ within the $r$-th column for each $r$, the quantity in (9) is expressible (the multiplier $p$ is omitted) as

$$\sum_{r=1}^{m}\kappa_{rrrr}. \tag{10}$$

We can regard (10) as a criterion that can be used in ICA because $\kappa_{rrrr}$'s are measures of nonnormality. As a consequence, if one considers a centered $m$-dimensional data set as a factor loading matrix, the factor rotation implements ICA.

As noted in Section 2, Comon (1994) has studied a variety of criteria using cumulants in order to perform ICA, among which the criterion $\sum_{r=1}^{m}\kappa_{rrrr}^{2}$ is close to (10).

The criterion in (10) has a problem for a case where there are both leptokurtic (super-Gaussian) and platykurtic (sub-Gaussian) distributions in the $m$-dimensional distribution. Comon takes the square of the cumulants for the reason. Thus, for use of (10), all the marginal distributions must be leptokurtic.

If the data vectors $\boldsymbol{X}$'s are sphered, one can restrict the transformation matrix $W$ to be orthogonal, so the maximization in (8) can do the ICA. If they are not orthogonal but just centralized, we can use oblique rotations to implement the ICA.

The orthomax criterion can produce a variety of factor rotation methods with changing values of $\omega$, and the rotation methods for several specific values of $\omega$ are given specific names, for example, quartimax ($\omega = 0$), biquartimax ($\omega = 1/2$), varimax ($\omega = 1$), equamax ($\omega = m/2$), persimax ($\omega = p(m-1)/(p+m-2)$), and factor parsimony ($\omega = p$). No specific name has been given for $\omega = 3$.

The varimax criterion is known to have the following alternative expression:

$$\min_{P\in\mathcal{O}(m)}\sum_{r\neq s}\left[\sum_{i=1}^{p}b_{ir}^{2}b_{is}^{2}-\frac{1}{p}\Big(\sum_{i=1}^{p}b_{ir}^{2}\Big)\Big(\sum_{i=1}^{p}b_{is}^{2}\Big)\right], \tag{11}$$

which is the sum of the covariances between the squares of elements of the $r$-th and $s$-th columns of $B$. The corresponding correlation is called energy correlation in the context of the topographic ICA (see e.g., Hyvärinen, Hoyer and Inki 2000).

The objective functions as in (8) is called the *simplicity* functions for rotations in factor analysis. The program SAS(2001) uses the generalized Crawford-Ferguson family (Jennrich 1973).

## 4. SIMULATION

A small simulation experiment was conducted to study performance of the varimax-based ICA described above. We generated five-dimensional random variates $\boldsymbol{s}$ of size $T = 500(= p)$ as source signals where their components are independently distributed according to the Gamma distribution with parameters yielding kurtoses from 6 to 2. The population kurtosis and simulated-data kurtosis are shown in Table 1. We took as a mixing matrix

$$A=\begin{pmatrix}1&1&0&0&0\\1&1&1&0&0\\1&0&1&1&0\\1&0&0&1&1\\1&0&0&0&1\end{pmatrix}.$$

We implemented the varimax-based ICA with applying the procedure PROC FACTOR in SAS (2001), Release 8.1. The varimax rotation without Kaiser's normalization for each row of a factor loading matrix was made. As a pre-analysis, principal component analysis was made to sphere the observed data $\boldsymbol{X} = A\boldsymbol{s}$ so that the resultant variates have zero means and unit variances and they are orthogonal. Note that in the case $\sum_{i=1}^{p}b_{ir}^{2} = 1$ for any $r$ in (8), and hence, any orthomax procedure for an arbitrarily chosen value of $\omega$ are equivalent to each other.

To compare with it, we used the FastICA developed by Hyvärinen and Oja (1997) and Hyvärinen(1999a) to analyze the same data set, where the symmetric orthogonalization was applied. Table 2 shows the correlation matrix between the source signals $\boldsymbol{s}$ and estimated signals $\hat{\boldsymbol{s}}$ estimated by the ICA methods. It is seen that the estimated correlations are all very close to 1 and those are identical over the two methods up to the second decimal places. Table 1 shows achieved kurtosis estimates by the both methods. Here again, they are identical over the methods up to the second decimal places. Thus the both ICA's perform equally nicely for the data set.

We have also applied oblique rotations for the centralized data. However, the use of oblique rotations does not perform as well as the FastICA and the varimax-based ICA. The average correlation estimate between the source signals and estimated signals is approximately 0.95. So it is not reported here. There have been proposed so many oblique rotation procedures,

Table 1: Population and estimated kurtosis

| | s1 | s2 | s3 | s4 | s5 |
|---|---|---|---|---|---|
| population | 6.000 | 5.000 | 4.000 | 3.000 | 2.000 |
| simulated data | 6.685 | 4.433 | 4.110 | 4.314 | 0.920 |
| varimax | 7.279 | 4.488 | 4.484 | 4.481 | 0.827 |
| FastICA | 7.279 | 4.488 | 4.484 | 4.481 | 0.828 |

Table 2: Correlation between source signals and estimated signals

| estimated | | source signal | | | | |
|---|---|---|---|---|---|---|
| signal | | s1 | s2 | s3 | s4 | s5 |
| varimax | $\hat{s}_1$ | 0.984 | 0.055 | 0.186 | 0.054 | -0.079 |
| | $\hat{s}_2$ | -0.099 | 0.993 | -0.050 | 0.067 | 0.055 |
| | $\hat{s}_3$ | -0.145 | -0.022 | 0.981 | 0.004 | -0.092 |
| | $\hat{s}_4$ | 0.029 | -0.089 | 0.019 | 0.995 | -0.091 |
| | $\hat{s}_5$ | 0.029 | -0.058 | 0.001 | 0.056 | 0.987 |
| FastICA | $\hat{s}_1$ | 0.984 | 0.055 | 0.186 | 0.054 | -0.079 |
| | $\hat{s}_2$ | -0.099 | 0.992 | -0.050 | 0.068 | 0.057 |
| | $\hat{s}_3$ | -0.145 | -0.022 | 0.981 | 0.004 | -0.092 |
| | $\hat{s}_4$ | 0.029 | -0.090 | 0.019 | 0.995 | -0.091 |
| | $\hat{s}_5$ | 0.029 | -0.060 | 0.001 | 0.055 | 0.987 |

among which there might be a certain oblique rotation that can work nicely.

## 5. DISCUSSION

In this paper we showed that the factor rotation options in traditional factor analysis can be used to implement the ICA for the case where every blind signal of $\boldsymbol{s}$ has excess kurtosis. We could mention that factor analysis developers had noticed many important notions in ICA such as maximization of kurtosis, sparse coding and energy correlation when they developed factor rotation procedures. They did not use these terminologies though. However, they did not pay any attention to applications to the blind source separation problem.

In this paper, we do not have any intention to argue that factor analysis can do the ICA as well as many ICA methods (e.g., FastICA and JADE) developed in the informatics, and/or that factor rotation programs are no longer necessary which may be replaced with the ICA in near future. Apparently, the current varimax-based ICA can not analyze data sets where some components are super-Gaussian and the others are sub-Gaussian. In addition, it could not handle data sets of extremely large sample size, and it does not offer any on-line algorithm.

We would like to promote crossing useful information between each other by pointing out the connection between the ICA and traditional factor analysis.

## 6. REFERENCES

[1] Crawford, C. B. and Ferguson, G. A. (1970). A general rotation criteria and its use in orthogonal rotation. *Psychometrika*, **35**, 321-332.

[2] Cardoso, F.-F. and Souloumiac A. (1996). Jacobi angles for simultaneous diagonalization. *SIAM J. Math. Analysis Appl.*, **17**, 161-164.

[3] Comon, P. (1994). Independent component analysis, A new concept? *Signal Processing*, **36**, 287-314.

[4] Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Tranzactions on Computers*, **C-23**, 881-890.

[5] Hyvärinen, A. (1999a). Survey on independent component analysis. *Neural Computing Surveys*, **2**, 94-128.

[6] Hyvärinen, A. (1999b). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks.*

[7] Hyvärinen, A. and Kano, Y. (in press). Independent component analysis for non-normal factor analysis. In *New Developments in Psychometrics* (Yanai, H. et al., Eds.). Springer Verlag: Tokyo.

[8] Hyvärinen, A., Karhunen, J. and Oja, E. (2001). *Independent Component Analysis.* John Wiley & Sons.

[9] Hyvärinen, A., Hoyer, P. O. and Inki, M. (2000). Topographic independent component analysis: Visualizing the independence structure. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)* pp. 591-596, Helsinki, Finland.

[10] Hyvärinen, A. and Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation*, **9(7)**, 1483-1492

[11] Jennrich, R. I. (1973). Standard errors for obliquely rotated factor loadings. *Psychometrika*, **38**, 593-604.

[12] Jones, M. C. and Sibson, R. (1987). What is projection pursuit? (with discussion). *Journal of the Royal Statistical Society, Series A*, **150**, 1-36.

[13] Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, **23**, 187-200.

[14] Mooijaart, A. (1985). Factor analysis for nonnormal variables. *Psychometrika*, **50**, 323-342.

[15] SAS Online Documents (2001). SAS/STAT Software: Changes and Enhancements, Release 8.1. `http://www.sas.com/rnd/app/da/new/801ce/stat/index.htm`