Matrix Independent Component Analysis

Joni Virta[*1]

[1] *University of Turku, Finland*

Independent component analysis (ICA) is a popular means of dimension reduction for vector-valued random variables. In this short note we review its extension to arbitrary tensor-valued random variables by considering the special case of two dimensions where the tensors are simply matrices.

Keywords: FOBI, Kronecker structure, Kurtosis

1 Matrix independent component model

For an introduction to classical vector-valued independent component analysis (ICA) the reader is referred to [3]. The tensorial ICA theory we next review was first introduced in [6] and further investigated in [7].

Let $\mathbf{X} \in \mathbb{R}^{p_1 \times p_2}$ be a random matrix from the matrix location-scale model

$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\Omega}_1 \mathbf{Z} \boldsymbol{\Omega}_2^T, \tag{1}$$

where the location matrix $\boldsymbol{\mu} \in \mathbb{R}^{p_1 \times p_2}$ and the non-singular mixing matrices $\boldsymbol{\Omega}_1 \in \mathbb{R}^{p_1 \times p_1}$ and $\boldsymbol{\Omega}_2 \in \mathbb{R}^{p_2 \times p_2}$ are unknown parameters and $\mathbf{Z} \in \mathbb{R}^{p_1 \times p_2}$ is an unobserved random matrix with finite joint fourth moments. Defining $vec : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}^{p_1 p_2}$ as the function that stacks the columns of its argument into a vector, the model (1) can be written as

$$vec\left(\mathbf{X}\right) = vec\left(\boldsymbol{\mu}\right) + \left(\boldsymbol{\Omega}_2 \otimes \boldsymbol{\Omega}_1\right) vec\left(\mathbf{Z}\right), \tag{2}$$

where $\otimes$ is the Kronecker product. Thus (1) can also be thought as a structured location-scale model (Kronecker model) for random vectors.

We will next describe conditions under which the model (1) is well-defined. For any non-singular $\mathbf{A}_1 \in \mathbb{R}^{p_1 \times p_1}$ and $\mathbf{A}_2 \in \mathbb{R}^{p_2 \times p_2}$ it can be written as

$$\mathbf{X} = \boldsymbol{\mu} + \left(\boldsymbol{\Omega}_1 \mathbf{A}_1^{-1}\right) \left(\mathbf{A}_1 \mathbf{Z} \mathbf{A}_2^T\right) \left(\boldsymbol{\Omega}_2 \mathbf{A}_2^{-1}\right)^T = \boldsymbol{\mu} + \boldsymbol{\Omega}_1^* \mathbf{Z}^* \left(\boldsymbol{\Omega}_2^*\right)^T,$$

showing that the parameters are not identifiable as such. Note that we can never achieve full identifiability as for any non-zero scalar β the maps $\boldsymbol{\Omega}_1 \mapsto \beta \boldsymbol{\Omega}_1$ and

*Corresponding author: joni.virta@utu.fi

$\boldsymbol{\Omega}_2 \mapsto \beta^{-1}\boldsymbol{\Omega}_2$ preserve the model. In the following we will refer to identifiability up to this proportionality as *proportional identifiability*. As a first step towards proportional identifiability we set the following constraints for $\mathbf{Z}$.

$$E\left[vec\left(\mathbf{Z}\right)\right] = \mathbf{0}_{p_1p_2} \quad \text{and} \quad Cov\left[vec\left(\mathbf{Z}\right)\right] = \mathbf{I}_{p_1p_2}.$$

The first constraint fixes the location matrix $\boldsymbol{\mu}$ and the second makes both $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ proportionally identifiable up to orthogonal $\mathbf{A}_1$ and $\mathbf{A}_2$.

To impose more structure the model can be equipped with additional assumptions on the latent matrix $\mathbf{Z}$. The classical choice is to assume that $vec(\mathbf{Z}) \sim \mathcal{N}(\mathbf{0}_{p_1p_2}, \mathbf{I}_{p_1p_2})$, resulting in a general matrix normal distribution for $\mathbf{X}$. The normal model can further be generalized in two directions. Focusing on the orthogonal invariance of the standard normal distribution leads us to consider the class of spherical random matrices satisfying $\mathbf{Z} \sim \mathbf{U}_1\mathbf{Z}\mathbf{U}_2^T$ for all orthogonal $\mathbf{U}_1 \in \mathbb{R}^{p_1\times p_1}, \mathbf{U}_2 \in \mathbb{R}^{p_2\times p_2}$ and this in turn yields a matrix elliptical distribution for $\mathbf{X}$, see [4] for the previous two models.

The second generalization is based on another key characteristic of the standard multivariate normal distribution, the equivalence of uncorrelatedness and independence, and equips $\mathbf{Z}$ with the following assumption.

A1. The components of $\mathbf{Z}$ are mutually independent.

While assumption A1 is rather strong, actually strong enough to guarantee the proportional identifiability of $\mathbf{Z}$ in (1) up to some trivialities when paired with A2 below, it is still a natural choice in applications where the components of $\mathbf{Z}$ can each be thought to model one separate aspect of the phenomenon which then combine independently to produce the observation $\mathbf{X}$.

The Skitovich-Darmois theorem [5] states that if a set of independent random variables can be combined to yield non-trivial linear combinations that are itself independent they must all be normally distributed. Thus we must further restrict the presence of multivariate normal distribution in the latent matrix to avoid $\mathbf{A}_1\mathbf{Z}\mathbf{A}_2^T$ having independent components for non-trivial $\mathbf{A}_1$ and $\mathbf{A}_2$.

A2. At most one row of $\mathbf{Z}$ has a multivariate normal distribution and at most one column of $\mathbf{Z}$ has a multivariate normal distribution.

Assumptions A1 and A2 now jointly guarantee that $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ are proportionally identifiable up to $\mathbf{A}_1$ and $\mathbf{A}_2$ containing a single ± 1 in each of their rows and columns. Consequently the matrix $\mathbf{Z}$ can be estimated up to the order and signs of its rows and columns, a defect that is usually of no consequence in practice.

Definition 1. We say that $\mathbf{X} \in \mathbb{R}^{p_1\times p_2}$ obeys the matrix independent component model (MICM) if it satisfies (1) along with assumptions A1 and A2.

To wrap everything up, in matrix independent component analysis we assume that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ is a random sample from the distribution of $\mathbf{X}$ obeying MICM and our objective is the estimation of the matrices $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$.

2 The estimation of Z

Let $\mathbf{X}$ obey MICM. Centering the random matrix as $\mathbf{X} \mapsto \mathbf{X} - E[\mathbf{X}]$ shows that without loss of generality we may assume in the following that $vec(\boldsymbol{\mu}) = \mathbf{0}_{p_1 p_2}$.

A key notion in the model is that, instead of treating the elements of $\mathbf{Z}$ separately, we consider them in an aggregate sort of way via their corresponding rows and columns. As an example take assumptions A1 and A2, the first of which can be written equivalently as "the rows of $\mathbf{Z}$ are mutually independent and the columns of $\mathbf{Z}$ are mutually independent". The same thought is also reflected in our definitions of the row and column covariance matrices,

$$\boldsymbol{\Sigma}_1(\mathbf{X}) = \frac{1}{p_2} E\left[\mathbf{X}\mathbf{X}^T\right] \quad \text{and} \quad \boldsymbol{\Sigma}_2(\mathbf{X}) = \frac{1}{p_1} E\left[\mathbf{X}^T\mathbf{X}\right].$$

The matrices $\boldsymbol{\Sigma}_1(\mathbf{X})$ and $\boldsymbol{\Sigma}_2(\mathbf{X})$ can be interpreted as the average covariance matrices of the p_2 columns and p_1 rows of $\mathbf{X}$, respectively. Under the independent component model they further enjoy the "equivariance property" described by the next lemma.

Lemma 1. *Let $\boldsymbol{X}$ obey MICM. Then the inverse square roots of the row and column covariance matrix satisfy*

$$\boldsymbol{\Sigma}_1(\boldsymbol{X})^{-1/2} = \frac{p_2^{1/2}}{\|\boldsymbol{\Omega}_2\|_F} \boldsymbol{U}_1 \boldsymbol{\Omega}_1^{-1} \quad \textit{and} \quad \boldsymbol{\Sigma}_2(\boldsymbol{X})^{-1/2} = \frac{p_1^{1/2}}{\|\boldsymbol{\Omega}_1\|_F} \boldsymbol{U}_2 \boldsymbol{\Omega}_2^{-1},$$

for some orthogonal matrices $\boldsymbol{U}_1 \in \mathbb{R}^{p_1 \times p_1}$ and $\boldsymbol{U}_2 \in \mathbb{R}^{p_2 \times p_2}$, where $\|\cdot\|_F$ is the Frobenius (Euclidean) norm.

For the proof of Lemma 1 and all other results in this review see [6]. Lemma 1 immediately yields the first step towards the estimation of $\mathbf{Z}$:

Lemma 2. *Let $\boldsymbol{X}$ obey MICM. Then we have*

$$\left(\boldsymbol{\Sigma}_1(\boldsymbol{X})^{-1/2}\right) \boldsymbol{X} \left(\boldsymbol{\Sigma}_2(\boldsymbol{X})^{-1/2}\right)^T = \gamma \boldsymbol{U}_1 \boldsymbol{Z} \boldsymbol{U}_2^T,$$

with orthogonal $\boldsymbol{U}_1 \in \mathbb{R}^{p_1 \times p_1}$ and $\boldsymbol{U}_2 \in \mathbb{R}^{p_2 \times p_2}$ and $\gamma = (p_1 p_2)^{1/2} \|\boldsymbol{\Omega}_2 \otimes \boldsymbol{\Omega}_1\|_F^{-1}$.

According to Lemma 2 the two-sided standardization of $\mathbf{X}$ reduces the problem of estimating $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ to the easier task of estimating two orthogonal

matrices. In the following we denote by $\mathbf{X}^{st}$ the standardized matrix on the left-hand side of Lemma 2.

Our method for estimating $\mathbf{U}_1$ and $\mathbf{U}_2$ is based on an extension of a multivariate ICA method called *fourth order blind identification* (FOBI) [1] and will hereafter be referred to as MFOBI. Heuristically ICA can be thought of as the maximization of non-normality and FOBI achieves it via considering a matrix measuring kurtosis, a classical indicator of non-normality. Sure enough, we define both row and column versions of the matrix.

$$\mathbf{B}_1(\mathbf{X}) = \frac{1}{p_2} E\left[\mathbf{X}\mathbf{X}^T\mathbf{X}\mathbf{X}^T\right] \quad \text{and} \quad \mathbf{B}_2(\mathbf{X}) = \frac{1}{p_1} E\left[\mathbf{X}^T\mathbf{X}\mathbf{X}^T\mathbf{X}\right].$$

A key property of $\mathbf{B}_1(\mathbf{X})$ and $\mathbf{B}_2(\mathbf{X})$ with respect to our problem, diagonality under independence, is described in the next lemma.

Lemma 3. *Let the random matrix $\boldsymbol{Z} \in \mathbb{R}^{p_1 \times p_2}$ have mutually independent components with zero means, unit variances and finite joint fourth moments. Then we have*

$$\begin{aligned}\boldsymbol{B}_1(\boldsymbol{Z}) &= (p_1 + p_2 + 1)\,\boldsymbol{I}_{p_1} + diag(\kappa_{1\bullet}, \ldots, \kappa_{p_1\bullet}) \\ \boldsymbol{B}_2(\boldsymbol{Z}) &= (p_1 + p_2 + 1)\,\boldsymbol{I}_{p_2} + diag(\kappa_{\bullet 1}, \ldots, \kappa_{\bullet p_2}),\end{aligned}$$

where $\kappa_{i\bullet}$ is the ith row mean and $\kappa_{\bullet j}$ is the jth column mean of the kurtosis matrix $\boldsymbol{\kappa} = \left(E\left[z_{ij}^4 - 3\right]\right)_{ij}$.

Both matrices $\mathbf{B}_1(\mathbf{X})$ and $\mathbf{B}_2(\mathbf{X})$ are orthogonally equivariant and we obtain the following.

Lemma 4. *Let $\boldsymbol{X}$ obey MICM. Then we have*

$$\boldsymbol{B}_1(\boldsymbol{X}^{st}) = \gamma^4\,\boldsymbol{U}_1\boldsymbol{B}_1(\boldsymbol{Z})\,\boldsymbol{U}_1^T \quad \textit{and} \quad \boldsymbol{B}_2(\boldsymbol{X}^{st}) = \gamma^4\,\boldsymbol{U}_2\boldsymbol{B}_2(\boldsymbol{Z})\,\boldsymbol{U}_2^T,$$

where $\boldsymbol{B}_1(\boldsymbol{Z})$ and $\boldsymbol{B}_2(\boldsymbol{Z})$ are diagonal by Lemma 3.

The two equations in Lemma 4 are the eigendecompositions of $\mathbf{B}_1(\mathbf{X}^{st})$ and $\mathbf{B}_2(\mathbf{X}^{st})$ and to guarantee the consistent estimation of $\mathbf{U}_1$ and $\mathbf{U}_2$, the corresponding eigenspectra must be distinct. In the light of Lemma 3 this requirement takes the following form.

A3. The row means of $\boldsymbol{\kappa}$ are distinct and the column means of $\boldsymbol{\kappa}$ are distinct, where $\boldsymbol{\kappa} = \left(E\left[z_{ij}^4 - 3\right]\right)_{ij}$ is the kurtosis matrix of the latent $\mathbf{Z}$.

Assumption A3 is a stronger version of assumption A2 and in particular says that no two rows or columns of $\mathbf{Z}$ may consist solely of random variables with identical distributions. Our main result is then the following.

Theorem 1. *Let* $\boldsymbol{X}$ *obey MICM and satisfy assumption A3. Further let* $\boldsymbol{V}_1 \in \mathbb{R}^{p_1 \times p_1}$ *and* $\boldsymbol{V}_2 \in \mathbb{R}^{p_2 \times p_2}$ *contain the eigenvectors of* $\boldsymbol{B}_1(\boldsymbol{X}^{st})$ *and* $\boldsymbol{B}_2(\boldsymbol{X}^{st})$*, respectively, as their columns. Then we have*

$$\boldsymbol{V}_1^T \boldsymbol{X}^{st} \boldsymbol{V}_2 = \gamma \boldsymbol{Z} \propto \boldsymbol{Z}.$$

The MFOBI solution of Theorem 1 enables the estimation of $\mathbf{Z}$ up to the scaling factor γ which is usually satisfactory enough, the shape and other higher-order properties of the components being of greater interest than their scales. In practice the MFOBI solution is obtained by replacing the expected values by the corresponding sample estimates. After the estimation of $\mathbf{Z}$ a further problem is the choosing of the most "interesting" components among the $p_1 p_2$ elements of $\mathbf{Z}$. Our kurtosis-based approach immediately leads to consider the components with extremal kurtosis, or to stay more in line with the spirit of the method, the rows and columns with the highest and lowest mean kurtoses. However, as the classical kurtosis is a very non-robust statistic the choice of a suitable criterion is still an open question.

3 Discussion

The naïve approach to model (1) is to vectorize it, resulting into (2), and proceed with standard methods of vector-valued ICA. However, this completely ignores the Kronecker structure of the mixing matrix $\mathbf{\Omega}_2 \otimes \mathbf{\Omega}_1$ and the price we pay for our negligence further comes in the form of stronger assumptions and increased computational cost. As an example, consider applying MFOBI to (1) versus applying FOBI to (2). Assumption A1 takes the same form for both methods but the counterpart of assumption A2 for FOBI is much more strict. Namely, it requires that at most one element of $vec(\mathbf{Z})$ has a normal distribution while in MFOBI the majority of the elements of $\mathbf{Z}$ can be normal if conveniently located. Similarly our assumption A3 and its vector-valued analogy, stating that the kurtoses of the elements of $vec(\mathbf{Z})$ are distinct, share the same relationship.

In order to compare the computational costs of the two methods we focus for simplicity only on the computationally most intensive part of the algorithms, the eigendecompositions. For $\mathbf{X} \in \mathbb{R}^{p_1 \times p_2}$ FOBI has to perform two eigendecompositions of a $p_1 p_2 \times p_1 p_2$ matrix while MFOBI requires the eigendecompositions of two $p_1 \times p_1$ and two $p_2 \times p_2$ matrices. In essence MFOBI "divides" the computational load into a larger number of smaller operations, lessening the overall complexity.

In [7] an extension of a second classical ICA method, *joint approximate diagonalization of eigen-matrices* (JADE) [2] for tensor-valued data was intro-

duced. Called TJADE, the method shares the standardization step of MFOBI (or more accurately, of TFOBI, its general tensor-valued extension) but approaches the estimation of the orthogonal matrices differently. In TJADE, instead of diagonalizing a single kurtosis matrix, we diagonalize several of them at once, essentially using more information in the estimation (and consequently increasing the computational burden as well). The implementations of both methods along with several other tensor extensions of classical methods can be found in the R-package *tensorBSS* [8].

Interestingly, restricting to matrix-valued observations only in this review serves more than just instructional purposes. In [6], [7] it is shown that the general tensor versions of the methods can be reduced to the matrix case. Similarly it can be shown that for the limiting distributions of the corresponding estimators it is sufficient to consider only the matrix case.

This note has been left devoid of examples and applications of the discussed methodology and instead several can be found, for example, in [6, 7].

References

[1] J.-F. Cardoso. Source separation using higher order moments. In *International Conference on Acoustics, Speech, and Signal Processing, 1989. ICASSP-89.*, pages 2109–2112. IEEE, 1989.

[2] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non-Gaussian signals. In *IEE Proceedings F (Radar and Signal Processing)*, volume 140, pages 362–370. IET, 1993.

[3] P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

[4] A. K. Gupta and D. K. Nagar. *Matrix variate distributions*, volume 104. CRC Press, 1999.

[5] I. Ibragimov. On the Ghurye-Olkin-Zinger theorem. *Journal of Mathematical Sciences*, 199(2), 2014.

[6] J. Virta, B. Li, K. Nordhausen, and H. Oja. Independent component analysis for tensor-valued data. *Preprint in arXiv:1602.00879*, 2016.

[7] J. Virta, B. Li, K. Nordhausen, and H. Oja. JADE for tensor-valued observations. *Preprint in arXiv:1603.05406*, 2016.

[8] J. Virta, B. Li, K. Nordhausen, and H. Oja. *tensorBSS: Blind Source Separation Methods for Tensor-Valued Observations*, 2016. R package version 0.3.