Consistent Noisy Independent Component Analysis

Jean-Marc Robin, Stéphane Bonhomme

Consistent Noisy Independent Component Analysis

Stéphane Bonhomme[1]
CEMFI, Madrid[2]

Jean-Marc Robin
Paris School of Economics,
Université Paris 1 Panthéon Sorbonne,[3]
and University College London[4]

Revised version: July 2008

[1]**Corresponding author:** CEMFI, Casado del Alisal, 5, 28014 Madrid, Spain. E-mail: bonhomme@cemfi.es

[2]Stéphane Bonhomme gratefully acknowledges the financial support from the Spanish Ministry of Science and Innovation through the Consolider-Ingenio 2010 Project "Consolidating Economics".

[3]Centre d'Economie de la Sorbonne, Université Paris 1 Panthéon Sorbonne, 106/112 bd de l'Hôpital, 75647 Paris Cedex 13, e-mail: jmrobin@univ-paris1.fr.

[4]Jean-Marc Robin gratefully acknowledges the financial support from the Economic and Social Research Council for the ESRC Centre for Microdata Methods and Practice, "Cemmap" (grant reference RES-589-28-0001).

Abstract

We study linear factor models under the assumptions that factors are mutually independent and independent of errors, and errors can be correlated to some extent. Under factor non-Gaussianity, second to fourth-order moments are shown to yield full identification of the matrix of factor loadings. We develop a simple algorithm to estimate the matrix of factor loadings from these moments. We run Monte Carlo simulations and apply our methodology to data on cognitive test scores, and financial data on stock returns.

JEL codes: C14.

Keywords: Independent Component Analysis, Factor Analysis, high-order moments, noisy ICA.

1 Introduction

A linear factor model relates a vector of L measurements to a vector of K unobserved sources, or factors, *via* a linear relationship:

$$\mathbf{Y} = \mathbf{\Lambda X} + \mathbf{U}, \tag{1}$$

where $\mathbf{\Lambda}$ is an L-by-K matrix of parameters (factor loadings) and $\mathbf{U}$ is a vector of L errors. A sample of N i.i.d. observations of $\mathbf{Y}$ is available for inference. In Factor Analysis (FA), it is assumed that $\text{Var}(\mathbf{X}) = \mathbf{I}_K$ (the identity matrix), and $\mathbf{\Lambda}$ is identified up to a rotation (Anderson and Rubin, 1956). Independent Component Analysis (ICA) strengthens the orthogonality assumption, and assumes that all the components of $\mathbf{X}$ and $\mathbf{U}$ are mutually independent. Then, if factors are *not* normally distributed, with a variance normalized to identity, $\mathbf{\Lambda}$ is generically identified up to sign and permutation normalizations (Comon, 1994, Eriksson and Koivunen, 2003). In the past ten years, ICA has become the standard approach to source separation, with numerous applications to signal processing, telecommunications, and medical imaging (Hyvärinen, Karhunen and Oja, 2001).

Independent factor models are also present in the econometrics literature. A well-known example is the measurement error model ($L = 2$, $K = 1$):

$$\left\{ \begin{array}{rcl} Y_1 & = & \lambda_{11} X + U_1 \\ Y_2 & = & \lambda_{21} X + U_2, \end{array} \right.$$

where X , U_1 and U_2 are assumed independent, and $\text{Var}(X) = 1$. Geary (1942) and Reiersol (1950) have shown that factor loadings are identified if X is not Gaussian. Since this seminal work, a long series of econometric contributions have proposed different ways to identify and estimate factor loadings in the measurement error model.[1] The class of

[1] A short list of contributions includes Pal (1980), Dagenais and Dagenais (1997), Lewbel (1997), Erickson and Whited (2002), and Schennach *et al.* (2007) for a recent nonparametric generalization of the model.

models that we consider in this paper can be seen as a generalization of this line of research to multi-factor structures.

Factor models are widely used in other areas of economics. Ross's (1976) Arbitrage Pricing Theory (APT) has profoundly influenced empirical finance. In microeconometrics, error component models for panel data can be understood as parsimonious linear factor models. Moreover, recent studies aiming at better understanding the sources of individual wage/productivity dispersion have used factor models to construct measures of "primary" mental abilities based on psychometric tests (e.g., Carneiro, Hansen and Heckman, 2003, and Heckman, Stixrud and Urzua, 2006). In macroeconometrics, structural VAR models (e.g., Blanchard and Quah, 1989) have evolved into more complex factor models (e.g., Forni and Reichlin, 1998, Pesaran, 2006, Chudik and Pesaran, 2007).

Factor independence is often assumed in these applications, as in Ross's APT, or in the aforementioned papers by James Heckman and coauthors. Independence, or higher-order uncorrelatedness, serves two goals. First, it provides another source of identification of the model, in the form of additional moment restrictions. Second, independence is a natural step beyond uncorrelatedness if one is interested in higher-order moments of the data. If the data are not Gaussian, and display skewness or kurtosis, then it is natural to seek to fit not only the variance but also third and/or fourth-order moments.[2]

In this kind of applications to social sciences, measurement error and/or specific factors are likely to be present. However, most ICA algorithms do not explicitly allow for noise. Indeed, in ICA applications, errors are usually assumed negligible ($\mathbf{U} \approx 0$), and noise-free methods work well if the signal-to-noise ratio is high enough (Cardoso and Pham, 2004). The methodological contribution of this paper is to fill this gap in the literature, and provide a close substitute to noise-free ICA algorithms that is consistent

[2]For example, there is evidence that the variance of returns is not an adequate measurement of risk as assumed in the CAPM model. For a recent extension of CAPM involving higher-order moments of asset returns, see Mencia and Sentana (2008).

in the presence of noise.

In the noise-free case, several efficient ICA algorithms are currently available to separate up to $K = L$ unobserved factors, FastICA (Hyvärinen and Oja, 1997) and JADE (Cardoso and Souloumiac, 1993) being especially popular. Most of these methods use a two-step approach to estimation.[3] In the first step (*prewhitening*), the data are transformed so that the covariance matrix is the identity, e.g. using Principal Component Analysis (PCA). In the second step (*source separation*) the rotation matrix is derived from higher-order information.

Two approaches have already been proposed to deal with noisy ICA models. In the first approach (Moulines *et al.*, 1997, Attias, 1999) a flexible parametric model is postulated for factor and error distributions. Maximum Likelihood is often used in estimation, together with the EM algorithm. This requires an appropriate parametric specification, e.g. a mixture model, and creates computational difficulties.

The second approach relies on a prewhitening step as in noise-free ICA methods, replacing PCA by Probabilistic PCA (Beckmann and Smith, 2004) or FA (Ikeda and Toyama, 2000, Stegeman and Mooijaart, 2007). This approach yields a fast semi-parametric estimation of $\mathbf{\Lambda}$. Yet, as only second-order moments of the data are used in the prewhitening step, the number K of common factors must be less than the Ledermann bound[4] for the procedure to be consistent. Moreover, it only deals with Gaussian errors. If errors are sizeable and the data are highly non-normal, this assumption can be problematic.

We also adopt a semi-parametric, two-step approach. In the first step, second to fourth-order moments of error variables are inferred from a set of linear restrictions,

[3] See Chen and Bickel (2005). Given the high number of moment restrictions implied by independence, optimal method-of-moments/minimum distance estimation is not tractable. We will provide simulation evidence to illustrate this point.

[4] The Ledermann bound is the maximal number of factors, K, such that the number of non redundant elements of $\mathrm{Var}(\mathbf{Y})$ is greater than the number of factor loadings plus the number of restrictions to pin down the rotation. The bound is $K = (2L + 1 - \sqrt{8L + 1})/2$ if errors are mutually uncorrelated.

and filtered out from the corresponding data moments. Importantly, unlike the previous literature we use all second, third and fourth-order data moments in the first estimation step. Then, the second step uses Cardoso and Souloumiac's (1993) JADE algorithm to estimate factor loadings. We call quasi-JADE this two-stage estimation procedure.

Quasi-JADE is consistent whether errors are Gaussian or not, and is almost as fast to run as JADE. An important property of the algorithm is that errors can be correlated to some extent. We show that, if J is the number of mutually independent error pairs, up to $K = \min\{J, L\}$ factors are generically identified. In the particular case of independent errors, we can thus relax the Ledermann bound and estimate up to L factors. This is because we use higher-order data moments in the prewhitening step of the algorithm.[5]

Finally, our approach is related to *overcomplete* ($K > L$) ICA models.[6] Indeed, our estimation procedure can be applied iteratively to estimate a model with L unrestricted factors, $L-1$ factors specific to measurements $\{2, ..., L\}$, $L-2$ factors specific to measurements $\{3, ..., L\}$, etc., and one last factor specific to the last two measurements, for a total of $L(L-1)/2$ factors with restrictions on factor loadings. Building on Cardoso (1991), De Lathauwer *et al.* (2007) have recently proposed an algorithm for overcomplete ICA based on fourth-order moments of the data. Compared to their approach, we do not impose error variables to be non-Gaussian, we use second (and higher) order moments in the estimation, and we allow for up to $L(L-1)/2$ factors when $\mathbf{\Lambda}$ is sparse enough. The results of De Lathauwer *et al.* (2007) are thus complementary to ours, as they develop a method to estimate overcomplete models without noise, while we propose a robust version of a standard ICA algorithm (JADE) in the presence of non-negligible noise.

Sections 2 and 3 present the model, derive the moment restrictions on which identi-

[5] The algorithm can also be applied to cases where factor loadings are restricted *ex-ante*, as in structural VARs. If there are sufficiently many restrictions for the rotation indeterminacy to disappear, factor loadings and error covariances can be jointly estimated from the first estimation step. The benefits of using higher-order information then translate into the possibility of allowing for a richer error structure.

[6] Algorithms for overcomplete ICA have been proposed by Comon (2004) for the case $L, K = 2, 3$, and by Albera *et al.* (2004), among others.

fication and estimation are based and show the identification of the number of factors, error cumulants and factor loadings. In Section 4 we discuss the estimation of the factor loadings, and develop the asymptotic distribution theory for JADE, surprisingly missing in the literature. In Section 5, we illustrate the finite-sample properties of our procedure by means of Monte-Carlo simulations, and in Section 6 we apply the method to two datasets: psychometric data on cognitive test scores, and financial data on stock returns. Lastly, Section 7 concludes.

2 Model and moment restrictions

2.1 The model

Let $\mathbf{Y} = (Y_1, ..., Y_L)^{\mathrm{T}}$ be a vector of $L \geq 2$ zero-mean, real-valued random variables (measurements), where $^{\mathrm{T}}$ denotes the transpose operator. Let $\mathbf{X} = (X_1, ..., X_K)^{\mathrm{T}}$ be a random vector of $K \geq 1$ real valued, non degenerate random variables (factors). Let also $\mathbf{U} = (U_1, ..., U_L)^{\mathrm{T}}$ be a vector of L real-valued random variables (errors). An observation sample is a collection of N independent draws of vector $\mathbf{Y}$.

Assumption A1 *There exists a L-by-K matrix of scalar parameters (factor loadings),* $\mathbf{\Lambda} = [\lambda_{\ell k}]$,[7] *such that* $\mathbf{Y} = \mathbf{\Lambda X} + \mathbf{U}$, *and* $\mathbf{\Lambda}$, $\mathbf{X}$ *and* $\mathbf{U}$ *satisfy the following conditions:*

1. $(\mathbf{X}^{\mathrm{T}}, \mathbf{U}^{\mathrm{T}})^{\mathrm{T}}$ *has zero mean and finite moments up to the fourth order.*

2. *The components of* $\mathbf{X}$ *are mutually independent, and independent of those of* $\mathbf{U}$.

3. *The components of* $\mathbf{X}$ *have unitary variance.*

A triple $(\mathbf{\Lambda}, \mathbf{X}, \mathbf{U})$, *satisfying these assumptions is called a representation of* $\mathbf{Y}$.

In the second statement, independence can be replaced by the weaker assumption of zero multivariate cumulants up to the fourth order. The third statement is a normaliza-

[7] A generic column of $\mathbf{\Lambda}$ is denoted $\boldsymbol{\lambda}_k$ and a generic row $\mathbf{\Lambda}_\ell$.

tion condition. If $(\mathbf{\Lambda}, \mathbf{X}, \mathbf{U})$ is a representation of $\mathbf{Y}$, then $\left(\mathbf{\Lambda D}^{-1}, \mathbf{DX}, \mathbf{U}\right)$ is another representation of $\mathbf{Y}$ for any diagonal matrix $\mathbf{D}$ with positive entries on the diagonal.

The normalization of the variance of $\mathbf{X}$ is not sufficient to grant identification. For any value of K, the number of factors, let us define the set of sign-permutation matrices as the set $\mathcal{S}_K$ of all products $\mathbf{DP}$, where $\mathbf{D}$ is a diagonal matrix with diagonal components equal to 1 or -1 and $\mathbf{P}$ is a permutation matrix. For given values of L and K, let $(\mathbf{\Lambda}, \mathbf{X}, \mathbf{U})$ be a representation of $\mathbf{Y}$. Clearly, for all $\mathbf{S} \in \mathcal{S}_K$, $\left(\mathbf{\Lambda S}, \mathbf{S}^{\mathrm{T}}\mathbf{X}, \mathbf{U}\right)$ is another representation. We say that the matrix of factor loadings $\mathbf{\Lambda}$ is identified if any representation of $\mathbf{Y}$, $(\widetilde{\mathbf{\Lambda}}, \widetilde{\mathbf{X}}, \widetilde{\mathbf{U}})$, is such that $\mathbf{\Lambda}$ and $\widetilde{\mathbf{\Lambda}}$ are equal modulo $\mathcal{S}_K$ (i.e. $\widetilde{\mathbf{\Lambda}} = \mathbf{\Lambda S}$ for some $\mathbf{S} \in \mathcal{S}_K$).

Given the linearity and independence assumptions, working with cumulants is especially convenient. Multivariate cumulants of centered random variables of order 2, 3 and 4 are defined as follows:

$$\begin{aligned}
\mathrm{Cum}\,(Z_1, Z_2) &= \mathbb{E}(Z_1 Z_2), \\
\mathrm{Cum}\,(Z_1, Z_2, Z_3) &= \mathbb{E}(Z_1 Z_2 Z_3), \\
\mathrm{Cum}\,(Z_1, Z_2, Z_3, Z_4) &= \mathbb{E}\left(Z_1 Z_2 Z_3 Z_4\right) - \mathbb{E}(Z_1 Z_2)\mathbb{E}(Z_3 Z_4) - \mathbb{E}(Z_1 Z_3)\mathbb{E}(Z_2 Z_4) \\
&\quad -\mathbb{E}(Z_1 Z_4)\mathbb{E}(Z_2 Z_3).
\end{aligned}$$

To ensure identification we impose the following restrictions on the first cumulants of error variables.

Assumption A2 *There exists a non empty set of indices $\mathcal{J} \subset \{(\ell, m) \in \{1, ..., L\}^2,\ \ell < m\}$ such that, for all $(\ell, m) \in \mathcal{J}$ and all measurement indices i and j, we have:*

$$\mathrm{Cum}\,(U_\ell, U_m) = \mathrm{Cum}\,(U_i, U_\ell, U_m) = \mathrm{Cum}\,(U_i, U_j, U_\ell, U_m) = 0.$$

Most of the ICA literature makes parametric assumptions on errors, usually assuming Gaussianity. However, Davies (2004) points out that error Gaussianity alone is not sufficient to provide identification of factor loadings in a noisy ICA model. For identification, one needs to restrict the dependence between errors, which is what Assumption A2 does.

The following lemma shows that Assumption A2 is satisfied by a broad class of error structures.

Lemma 1 *Let* $\mathbf{U} = \mathbf{\Pi}\boldsymbol{\varepsilon}$*, where* $\mathbf{\Pi}$ *is a* L*-by-*H *matrix of scalar parameters, and the components of* $\boldsymbol{\varepsilon}$ *are mutually independent and independent of those of* $\mathbf{X}$ *with finite moments up to the fourth order. Then* $\mathbf{U}$ *satisfies assumption A2, with*

$$\mathcal{J} = \{(\ell, m) \in \{1, ..., L\}^2,\ \ell \leq m,\ U_\ell \perp\!\!\!\perp U_m\},$$

where $\perp\!\!\!\perp$ *denotes statistical independence.*

The proof is in section A.1 of the mathematical Appendix.

Lemma 1 shows that several commonly used error dependence structures satisfy Assumption A2. A first example is provided by independent heteroskedastic errors. In this case:

$$\mathcal{J} = \{(\ell, m) \in \{1, ..., L\}^2, \ell < m\}, \quad \text{and} \quad J \equiv \#\mathcal{J} = \frac{L(L-1)}{2}.$$

If the data has a group structure, with r disjoint groups $\mathcal{M}_i$ $(i = 1, ..., r)$, and errors are independent between groups, then $\mathcal{J} = \{(\ell, m) \in \mathcal{M}_i \times \mathcal{M}_j, i \neq j, \ell < m\}$, and $J = \sum_{j=1}^{r-1} \#\mathcal{M}_j \left(L - \sum_{i=1}^{j} \#\mathcal{M}_i\right)$.

In addition, Assumption A2 allows for temporal or spatial correlation patterns. For instance, if errors are MA(q) then $\mathcal{J} = \{(\ell, m) \in \{1, ..., L\}^2, \ell < m - q\}$, and $J = (L-q)(L-q-1)/2$. Likewise, spatial MA models may also satisfy the assumption, with J depending on the zeros of the matrix of spatial weights (e.g., Anselin, 2003).

In contrast, autoregressive (or spatial autoregressive) error structures do not satisfy Assumption A2, as errors are correlated at all lags and leads. However, the methods of this paper are applicable in this case also. To see how one might proceed, let us consider a case where errors are ARMA(1,1). Then by taking quasi-differences $Y_\ell - \rho Y_{\ell-1}$, where ρ is the autoregressive parameter, we end up with MA(1) errors. Using the results below,

ρ can then be obtained in the first estimation step (prewhitening), together with error moments.

Throughout the paper, we will take the set $\mathcal{J}$ as given. In some cases, the choice of $\mathcal{J}$ may be very natural. For example, in the application to cognitive test scores we will allow for contemporaneous correlation in the errors, taking into account the group structure of the data. However, there may be cases where no obvious choice for $\mathcal{J}$ is available.

2.2 Moment restrictions

We start by deriving the moment restrictions implied by Assumption A1. Let $p \in \{2,3,4\}$ and $(\ell_1, ..., \ell_p) \in \{1, ..., L\}^p$. Assumption A1 implies

$$\text{Cum}\left(Y_{\ell_1}, ..., Y_{\ell_p}\right) = \sum_{k=1}^{K} \left(\prod_{i=1}^{p} \lambda_{\ell_i,k} \right) \kappa_p\left(X_k\right) + \text{Cum}\left(U_{\ell_1}, ..., U_{\ell_p}\right), \tag{2}$$

where we write $\kappa_p(Z) = \text{Cum}(Z, ..., Z)$ (repeat Z p times) for univariate cumulants of order $p \geq 1$.

Moment restrictions (2) have a common multilinear structure which can be conveniently expressed in matrix form, as in ordinary Factor Analysis. Define the following L-by-L, symmetric, square matrices:

$$\begin{aligned}
\mathbf{\Sigma_Y} &= \left[\text{Cum}\left(Y_i, Y_j\right)\right], \\
\mathbf{\Gamma_Y}\left(\ell\right) &= \left[\text{Cum}\left(Y_i, Y_j, Y_\ell\right)\right], \quad \ell \in \{1, ..., L\}, \\
\mathbf{\Omega_Y}\left(\ell, m\right) &= \left[\text{Cum}\left(Y_i, Y_j, Y_\ell, Y_m\right)\right], \quad \ell, m \in \{1, ..., L\},
\end{aligned}$$

with similar expressions for $\mathbf{\Sigma_U}$, $\mathbf{\Gamma_U}\left(\ell\right)$ or $\mathbf{\Omega_U}\left(\ell, m\right)$.

Restrictions (2) imply that

$$\mathbf{\Sigma_Y} = \mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}} + \mathbf{\Sigma_U}, \tag{3}$$

$$\mathbf{\Gamma_Y}\left(\ell\right) = \mathbf{\Lambda}\mathbf{D}_3 \,\text{diag}\left(\mathbf{\Lambda}_\ell\right)\mathbf{\Lambda}^{\mathrm{T}} + \mathbf{\Gamma_U}\left(\ell\right), \tag{4}$$

$$\mathbf{\Omega_Y}\left(\ell, m\right) = \mathbf{\Lambda}\mathbf{D}_4 \,\text{diag}\left(\mathbf{\Lambda}_\ell \odot \mathbf{\Lambda}_m\right)\mathbf{\Lambda}^{\mathrm{T}} + \mathbf{\Omega_U}\left(\ell, m\right), \tag{5}$$

where $\mathbf{\Lambda}_\ell \in \mathbb{R}^{1\times K}$ is the ℓth row of $\mathbf{\Lambda}$, $\mathbf{D}_3$ (resp. $\mathbf{D}_4$) is the diagonal matrix with cumulant $\kappa_3(X_k)$ (resp. $\kappa_4(X_k)$) in the kth entry of the diagonal, and $\odot$ is the Hadamard (element by element) matrix product.

Assumption A2 imposes additional restrictions. Combining the assumption with restrictions (5) yields:

$$\mathbf{\Omega_Y}(\ell,m) = \mathbf{\Lambda D}_4 \operatorname{diag}(\mathbf{\Lambda}_\ell \odot \mathbf{\Lambda}_m)\mathbf{\Lambda}^{\mathrm{T}}, \quad \forall (\ell,m) \in \mathcal{J}.$$

For a symmetric matrix $\mathbf{A} = [a_{ij}]$, we denote as vech the operator that stacks the elements of the upper triangular part of $\mathbf{A}$, extracted horizontally from left to right: $\operatorname{vech}(\mathbf{A}) = [a_{ij}]_{i\leq j}$. Applying the vech operator we obtain:

$$\boldsymbol{\omega}_{\mathbf{Y}}(\ell,m) \equiv \operatorname{vech}(\mathbf{\Omega_Y}(\ell,m)) = \mathbf{QD}_4(\mathbf{\Lambda}_\ell \odot \mathbf{\Lambda}_m), \quad \forall (\ell,m) \in \mathcal{J},$$

where $\mathbf{Q}$ is the $\frac{L(L+1)}{2}$-by-K matrix which generic (i,j) row, $i \leq j$, is $(\lambda_{i1}\lambda_{j1}, ..., \lambda_{iK}\lambda_{jK})$, i.e.

$$\mathbf{Q} \equiv \left[\operatorname{vech}\left(\boldsymbol{\lambda}_1\boldsymbol{\lambda}_1^{\mathrm{T}}\right), ..., \operatorname{vech}\left(\boldsymbol{\lambda}_K\boldsymbol{\lambda}_K^{\mathrm{T}}\right)\right],$$

where $\boldsymbol{\lambda}_k$ denotes the kth column of $\mathbf{\Lambda}$.

Next, construct the $\frac{L(L+1)}{2}$-by-J matrix $\mathbf{\Omega_Y}$ by concatenating columnwise all vectors $\operatorname{vech}(\mathbf{\Omega_Y}(\ell,m))$, $(\ell,m) \in \mathcal{J}$. Clearly:

$$\begin{aligned}\mathbf{\Omega_Y} &\equiv [\boldsymbol{\omega}_{\mathbf{Y}}(\ell,m)]_{(\ell,m)\in\mathcal{J}} \\ &= [\operatorname{Cum}(Y_i, Y_j, Y_\ell, Y_m)]_{(i\leq j)\times(\ell,m)\in\mathcal{J}}.\end{aligned}$$

Matrix $\mathbf{\Omega_Y}$ contains all fourth-order cumulants of measurements which are not contaminated by the presence of noise. Moreover, letting $\mathbf{Q}_{\mathcal{J}}$ be the J-by-K matrix obtained by selecting rows $(i,j) \in \mathcal{J}$ from $\mathbf{Q}$, we obtain:

$$\mathbf{\Omega_Y} = \mathbf{QD}_4\mathbf{Q}_{\mathcal{J}}^{\mathrm{T}}. \tag{6}$$

We can similarly construct the following matrix of third-order cumulants:

$$\mathbf{\Gamma_Y} = \left[\text{Cum}\left(Y_i, Y_\ell, Y_m\right)\right]_{i\times(\ell,m)},$$

where the rows of $\mathbf{\Gamma_Y}$ are indexed by $i \in \{1, ..., L\}$ and the columns are indexed by $(\ell, m) \in \mathcal{J}$. Then,

$$\mathbf{\Gamma_Y} = \mathbf{\Lambda D}_3 \mathbf{Q}_{\mathcal{J}}^{\mathrm{T}}. \tag{7}$$

In the next section, we take $\mathcal{J}$ as given and focus on the identification of factor loadings, using moment restrictions (3) to (7).

3 Identification results

In this section, we use the moment restrictions implied by the noisy ICA model to give sufficient conditions for the identification of factor loadings and error moments. We start with the number of factors, K.

3.1 Identification of the number of factors

The following theorem is an immediate consequence of (6) and (7).

Theorem 1 *The two following statements hold:*

i) Assume that all factor variables are kurtotic ($\kappa_4(X_k) \neq 0, \forall k$), and that matrix $\mathbf{Q}_{\mathcal{J}}$ has rank K, which in particular implies $K \leq J$. Then matrix $\mathbf{\Omega_Y}$ has rank K.

ii) Assume that all factors are skewed ($\kappa_3(X_k) \neq 0, \forall k$), and that both $\mathbf{\Lambda}$ and $\mathbf{Q}_{\mathcal{J}}$ have rank K, which implies that $K \leq \min\{J, L\}$. Then $\mathbf{\Gamma_Y}$ has rank K.

In Theorem 1, as in the other identification results below, we assume that $\mathbf{Q}_{\mathcal{J}}$ has rank K. Note that this is different from assuming that $\mathbf{\Lambda}$ has rank K. For example, if one column of $\mathbf{\Lambda}$, say the first one, has all elements equal to zero except one, then the

first column of $\mathbf{Q}_{\mathcal{J}}$ is identically zero, while $\mathbf{\Lambda}$ may still have rank K. Conversely, $\mathbf{Q}_{\mathcal{J}}$ may have rank K even if $\mathbf{\Lambda}$ has rank less than K.[8]

Theorem 1 shows that matrices $\mathbf{\Omega_Y}$ and $\mathbf{\Gamma_Y}$ allow to identify the number of common factors K. Notice that fourth-order cumulants can be used together with third-order cumulants. Define

$$\begin{aligned} \mathbf{\Omega_Y}(j) &= [\mathrm{Cum}(Y_i, Y_j, Y_\ell, Y_m)], \ j \in \{1, ..., L\}, \\ \text{and } \mathbf{\Phi_Y} &= [\mathbf{\Gamma_Y}, \mathbf{\Omega_Y}(1), ..., \mathbf{\Omega_Y}(L)]. \end{aligned}$$

Then, it is easily shown that, if factors are either skewed or kurtotic and $\mathbf{\Lambda}$ and $\mathbf{Q}_{\mathcal{J}}$ have rank K, then matrix $\mathbf{\Phi_Y}$ has rank K.

3.2 Identification of error moments

Applying operator vech to (3), (4) and (5) yields the following linear restrictions:

$$\begin{aligned} \mathrm{vech}(\mathbf{\Sigma_Y}) &= \mathbf{Q}\mathbf{1}_K + \mathrm{vech}(\mathbf{\Sigma_U}), \\ \mathrm{vech}(\mathbf{\Gamma_Y}(\ell)) &= \mathbf{Q}\mathbf{D}_3\mathbf{\Lambda}_\ell + \mathrm{vech}(\mathbf{\Gamma_U}(\ell)), \forall \ell, \\ \mathrm{vech}(\mathbf{\Omega_Y}(\ell, m)) &= \mathbf{Q}\mathbf{D}_4(\mathbf{\Lambda}_\ell \odot \mathbf{\Lambda}_m) + \mathrm{vech}(\mathbf{\Omega_U}(\ell, m)), \forall (\ell, m), \end{aligned}$$

where $\mathbf{1}_K$ is a K-dimensional vector of ones.

All factors are kurtotic. Let us begin by assuming that all factors are kurtotic, so that $\mathbf{D}_4$ has no zero on its main diagonal. Theorem 1 shows that, if matrix $\mathbf{Q}_{\mathcal{J}}$ has rank K, then $\mathrm{rank}(\mathbf{\Omega_Y}) = K$. So one can choose an orthogonal basis of the null space of $\mathbf{\Omega}_{\mathbf{Y}}^{\mathrm{T}}$, and construct a $\frac{L(L+1)}{2}$-by-$\left(\frac{L(L+1)}{2} - K\right)$ orthogonal matrix $\mathbf{B}$ that satisfies: $\mathbf{\Omega}_{\mathbf{Y}}^{\mathrm{T}}\mathbf{B} = 0$.

[8] For example, for $L = 3$, $K = 3$, $\mathcal{J} = \{(1,2),(1,3),(2,3)\}$: $\mathbf{\Lambda} = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 1 & 3 & 4 \end{pmatrix}$ has rank 2, but $\mathbf{Q}_{\mathcal{J}} = \begin{pmatrix} 1 & 2 & 6 \\ 1 & 3 & 8 \\ 1 & 6 & 12 \end{pmatrix}$ has rank 3.

Hence, as $\mathbf{Q}_{\mathcal{J}}\mathbf{D}_4$ has full column rank, it follows that $\mathbf{Q}^{\mathrm{T}}\mathbf{B} = 0$. So,

$$\mathbf{B}^{\mathrm{T}} \operatorname{vech}(\boldsymbol{\Sigma}_{\mathbf{Y}}) = \mathbf{B}^{\mathrm{T}} \operatorname{vech}(\boldsymbol{\Sigma}_{\mathbf{U}}), \tag{8}$$

$$\mathbf{B}^{\mathrm{T}} \operatorname{vech}(\boldsymbol{\Gamma}_{\mathbf{Y}}(\ell)) = \mathbf{B}^{\mathrm{T}} \operatorname{vech}(\boldsymbol{\Gamma}_{\mathbf{U}}(\ell)), \forall \ell, \tag{9}$$

$$\mathbf{B}^{\mathrm{T}} \operatorname{vech}(\boldsymbol{\Omega}_{\mathbf{Y}}(\ell, m)) = \mathbf{B}^{\mathrm{T}} \operatorname{vech}(\boldsymbol{\Omega}_{\mathbf{U}}(\ell, m)), \forall (\ell, m). \tag{10}$$

The following theorem shows that these linear restrictions identify error cumulants.

Theorem 2 *Assume that all factor variables have non zero excess kurtosis*[9] *and that matrix* $\mathbf{Q}_{\mathcal{J}}$ *has rank* K*. Then, second, third and fourth-order cumulants of error variables are uniquely defined by identifying restrictions (8), (9) and (10).*

The proof is in Section A.2 of the mathematical appendix.

Theorem 2 provides linear restrictions identifying error cumulants of order 2 to 4 irrespective of $\boldsymbol{\Lambda}$ and $\mathbf{X}$. The theorem shows that high-order moments of the data, appearing in (4) and (5), contain information on error moments that is not contained in second-order moments of the data. Exploiting this information allows to increase the number of common factors that can be identified in Factor Analysis, which relies exclusively on second-order restrictions (3).

The following corollary is immediate.

Corollary 1 *Assume that the conditions of Theorem 2 are satisfied. Then, the elements of* $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}$ *are uniquely defined by restrictions (3) and (8).*

If $K \leq L$, the corollary shows that if the conditions of Theorem 2 hold, then $\boldsymbol{\Lambda}$ is identified up to right-multiplication by an orthogonal matrix. The last part of the identification proof, that we derive in the next section, is devoted to the identification of this rotation.

[9] "Excess kurtosis" of a random variable refers to its standardized kurtosis minus three, which can be positive or negative. So, the assumptions in Theorem 2 allow for leptokurtic or platykurtic factors, the only requirement being that the kurtosis of all factor variables be different from that of the normal.

Corollary 1 can be of interest in its own right, if *ex-ante* restrictions are assumed on matrix $\mathbf{\Lambda}$. Indeed, if these restrictions are sufficient to identify $\mathbf{\Lambda}$ from the knowledge of $\mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}}$, then the rest of the identification proof is unnecessary.[10]

All factors are skewed. We can proceed similarly if every factor is skewed. If both $\mathbf{\Lambda}$ and $\mathbf{Q}_{\mathcal{J}}$ have full column rank K, Theorem 1 shows that $\mathbf{\Gamma_Y} = \mathbf{\Lambda D}_3\mathbf{Q}_{\mathcal{J}}^{\mathrm{T}}$ has rank K. Hence, there exists a L-by-$(L-K)$ orthogonal matrix $\mathbf{C}$ such that $\mathbf{\Gamma}_{\mathbf{Y}}^{\mathrm{T}}\mathbf{C} = 0$. So, as $\mathbf{D}_3$ has no zero on its diagonal, it must also be that $\mathbf{C}^{\mathrm{T}}\mathbf{\Lambda} = 0$.

The second, third and fourth-order cumulants of U_ℓ, for all $\ell \in \{1, ..., L\}$, thus satisfy the following linear restrictions:

$$\mathbf{C}^{\mathrm{T}}\mathbf{\Sigma_Y} = \mathbf{C}^{\mathrm{T}}\mathbf{\Sigma_U}, \tag{11}$$

$$\mathbf{C}^{\mathrm{T}}\mathbf{\Gamma_Y}(\ell) = \mathbf{C}^{\mathrm{T}}\mathbf{\Gamma_U}(\ell), \tag{12}$$

$$\mathbf{C}^{\mathrm{T}}\mathbf{\Omega_Y}(\ell, m) = \mathbf{C}^{\mathrm{T}}\mathbf{\Omega_U}(\ell, m). \tag{13}$$

Define, for all $\ell \in \{1, ..., L\}$, the sets

$$\mathcal{I}_\ell = \{m \in \{1, ..., L\},\ m < \ell \text{ or } (\ell, m) \in \mathcal{J}\},$$

with $I_\ell = \#\mathcal{I}_\ell$. Denote also $\mathbf{\Lambda}_{\mathcal{I}_\ell}$ the I_ℓ-by-K matrix obtained by selecting rows $i \in \mathcal{I}_\ell$ from $\mathbf{\Lambda}$. The following theorem gives conditions under which the system of linear restrictions (11), (12), and (13), has a unique solution.

Theorem 3 *Assume that every factor distribution is skewed, that $\mathbf{Q}_{\mathcal{J}}$ has rank K, and that $\mathbf{\Lambda}_{\mathcal{I}_\ell}$ has full column rank for all ℓ. Then, second, third and fourth-order cumulants of error variables are identified from restrictions (11), (12), and (13).*

[10] This is, for example, the case if $\mathbf{\Lambda}$ is assumed to be lower triangular, as in the following linear panel data model: $y_{it} = p_{it} + u_{it}$, $(i,t) \in \{1, ..., N\} \times \{1, ..., T\}$, where p_{it} is a random walk: $p_{it} = p_{i,t-1} + \varepsilon_{it}$, with $p_{i0}, \varepsilon_{i1}, ..., \varepsilon_{iT}$ independent. The transitory shocks u_{it} can be e.g. MA(q), or the sum of an MA(q) and an iid component (e.g., measurement error).

The proof is in Section A.3 of the mathematical appendix.

Theorem 3 implies that the number of factors is bounded by $\min\{I_\ell, \ell \in \{1,...,L\}\}$. In the particular case of independent errors this yields $K \leq L-1$. Focusing on the model with $L=2$ and $K=1$, Geary (1942) has shown that identification holds, provided that the factor is skewed. Theorem 3 provides a generalization of this result to multi-factor models.

Lastly, the discussion in this subsection can be generalized to the case where every factor is either skewed or kurtotic ($\kappa_3(X_k)\,\kappa_4(X_k) \neq 0$). One needs only replace matrix $\mathbf{\Gamma_Y}$ by matrix $\mathbf{\Phi_Y} = [\mathbf{\Gamma_Y}, \mathbf{\Omega_Y}(1), ..., \mathbf{\Omega_Y}(L)]$, and compute $\mathbf{C}$ such that $\mathbf{\Phi_Y^T C} = 0$.

3.3 Identification of factor loadings

In this section we assume that the cumulants of order 2, 3 and 4 of error components are known, the previous section giving sufficient conditions for their identification. Second, third and fourth-order restrictions (3), (4), (5) imply that matrix $\mathbf{\Lambda}$ satisfies, simultaneously,

$$\widetilde{\mathbf{\Sigma}}_\mathbf{Y} \equiv \mathbf{\Sigma_Y} - \mathbf{\Sigma_U} = \mathbf{\Lambda\Lambda}^\mathrm{T}, \tag{14}$$

$$\widetilde{\mathbf{\Gamma}}_\mathbf{Y}(\ell) \equiv \mathbf{\Gamma_Y}(\ell) - \mathbf{\Gamma_U}(\ell) = \mathbf{\Lambda D}_3 \operatorname{diag}(\mathbf{\Lambda}_\ell)\,\mathbf{\Lambda}^\mathrm{T}, \tag{15}$$

$$\widetilde{\mathbf{\Omega}}_\mathbf{Y}(\ell, m) \equiv \mathbf{\Omega_Y}(\ell, m) - \mathbf{\Omega_U}(\ell, m) = \mathbf{\Lambda D}_4 \operatorname{diag}(\mathbf{\Lambda}_\ell \odot \mathbf{\Lambda}_m)\,\mathbf{\Lambda}^\mathrm{T}. \tag{16}$$

Let us assume that $K \leq L$, and let $\mathbf{P}$ be an K-by-L matrix such that

$$\mathbf{P}\widetilde{\mathbf{\Sigma}}_\mathbf{Y}\mathbf{P}^\mathrm{T} = \mathbf{I}_K. \tag{17}$$

Matrix $\mathbf{P}$ can easily be constructed from eigenvectors and eigenvalues of $\widetilde{\mathbf{\Sigma}}_\mathbf{Y}$. Left and right-multiplying (14), (15) and (16) by $\mathbf{P}$ and $\mathbf{P}^\mathrm{T}$, respectively, we obtain:

$$\mathbf{P}\widetilde{\mathbf{\Gamma}}_\mathbf{Y}(\ell)\,\mathbf{P}^\mathrm{T} = \mathbf{V D}_3 \operatorname{diag}(\mathbf{\Lambda}_\ell)\,\mathbf{V}^\mathrm{T}, \quad \ell \in \{1,...,L\},$$

$$\mathbf{P}\widetilde{\mathbf{\Omega}}_\mathbf{Y}(\ell, m)\,\mathbf{P}^\mathrm{T} = \mathbf{V D}_4 \operatorname{diag}(\mathbf{\Lambda}_\ell \odot \mathbf{\Lambda}_m)\,\mathbf{V}^\mathrm{T}, \quad \ell \leq m,$$

where $\mathbf{V} = \mathbf{P}\mathbf{\Lambda}$ is orthonormal ($\mathbf{V}\mathbf{V}^{\mathrm{T}} = \mathbf{I}_K$). Therefore, $\mathbf{V}$ solves a joint diagonalization problem. Theorem 4 below gives conditions for the solution to this problem to be unique.

Theorem 4 *Assume that error cumulants are known, and that matrix $\mathbf{\Lambda}$ has full column rank K, so in particular $K \leq L$.*

(i) If at most one factor variable has zero excess kurtosis, then factor loadings are identified from second and fourth-order moment restrictions (14) and (16).

(ii) If at most one factor variable has zero skewness, then factor loadings are identified from second and third-order moment restrictions (14) and (15).

(iii) If for any couple of factors indices (k, k'), $(\kappa_3(X_k), \kappa_3(X_{k'}), \kappa_4(X_k), \kappa_4(X_{k'})) \neq 0$, then factor loadings are identified from second, third and fourth-order moment restrictions (14), (15) and (16).

The proof is in Section A.4 the mathematical appendix.

Combining Theorems 2, 3 and 4 we obtain that *(i)* at most $K = \min\{J, L\}$ factors can be identified if all factors are kurtotic, and *(ii)* at most $K = \min\{I_\ell, \ell \in \{1, ..., L\}\}$ if all factors are either skewed or kurtotic. In the case where errors are independent one can thus identify up to $K = L$ factors in the first case and $K = L - 1$ in the second case. By comparison, the number of factors in FA models is bounded by $K = \left(2L + 1 - \sqrt{8L+1}\right)/2$. This general identification result holds provided that sufficiently many errors are mutually independent.[11]

We end this section by remarking that Lemma 2, together with the previous identification theorems, imply that overcomplete ICA models are identified if there exist sufficiently many restrictions on factor loadings. To see that, let us consider the model:

$$\mathbf{Y} = \mathbf{\Lambda}_1\mathbf{X}_1 + ... + \mathbf{\Lambda}_S\mathbf{X}_S + \mathbf{U},$$

[11] To give an example, if errors follow an MA(q) process indexed by the measurement indices $\ell = 1, ..., L$, then one can generically identify L common factors if $J = (L-q)(L-q-1)/2 \geq L$, that is if $q \leq (2L - 1 - \sqrt{8L+1})/2$.

where, for all $s \in \{1, ..., S\}$, $\mathbf{X}_s$ has $K_s \leq L$ elements, $\mathbf{\Lambda}_s$ is L-by-K_s, and all factors and errors are assumed mutually independent. Let us suppose that all factors are kurtotic, the argument being similar when factors are skewed. Theorems 2 and 4 show that one can generically identify up to $K_1 = J_1$ factors $\mathbf{X}_1$, where J_1 is the number of components of $\mathbf{\Lambda}_2\mathbf{X}_2 + ... + \mathbf{\Lambda}_S\mathbf{X}_S + \mathbf{U}$ that are mutually independent. As an example, $K_1 = L - 1$ factors X_1 are identified, if the first row of all matrices $\mathbf{\Lambda}_2, ..., \mathbf{\Lambda}_S$ is identically zero. Applying this procedure sequentially shows identification in the case where $S = L - 1$, and, for all $s \in \{1, ..., S\}$, $K_s = L - s$, and the first $s - 1$ rows of $\mathbf{\Lambda}_s$ are zero. This corresponds to a block-triangular structure where the first $L - 1$ factors are common to all measurements, the next $L - 2$ factors are specific to $Y_2, ..., Y_L$, and so on. In this model there are $K = L(L-1)/2$ factors, and $L(L-1)^2/6$ restrictions on the $L^2(L-1)/2$ factor loadings.

4 Estimation

We start by discussing the issue of estimating factor loadings. Then, we provide the asymptotic theory of the JADE estimator, and discuss how to perform inference for JADE and quasi-JADE in practice.

4.1 Estimation of factor loadings

We assume that the number of factors K is known.[12] The two steps of the estimation algorithm are as follows.

[12] Assuming that $\mathbf{Q}_{\mathcal{J}}$ has full column rank and that factor variables show excess kurtosis, then matrix $\mathbf{\Omega}_{\mathbf{Y}}$ has rank $K \leq J$ (see Theorem 1). In the additional appendix to this paper, we propose a refined version of the sequential testing procedure developed in Robin and Smith (2000) to estimate the rank of $\mathbf{\Omega}_{\mathbf{Y}}$, and provide Monte Carlo simulations.

Prewhitening. In the first step error moments are estimated. In the case where all factors are kurtotic one may apply the following procedure:[13]

1. Construct matrix $\mathbf{\Omega_Y} = [\mathrm{Cum}\,(Y_i, Y_j, Y_\ell, Y_m)]$, where rows are indexed by couples (i,j), $i \leq j$, and columns are indexed by couples $(\ell, m) \in \mathcal{J}$.

2. Assuming that $\mathrm{rank}\,(\mathbf{\Omega_Y}) = K$, find the null space of $\mathbf{\Omega_Y^T}$, i.e. compute an orthogonal $\frac{L(L+1)}{2}$-by-$\left(\frac{L(L+1)}{2} - K\right)$ matrix $\mathbf{B}$ such that $\mathbf{\Omega_Y^T B} = 0$. A Singular Value Decomposition (SVD) can be used for this purpose.

3. Solve for the non-zero elements of $\mathbf{\Sigma_U}$ in the linear system (8). Proceed in the same way for third-order and fourth-order error cumulant matrices $\mathbf{\Gamma_U}\,(\ell)$ and $\mathbf{\Omega_U}\,(\ell, m)$.

In the algorithm, Step 3 can be performed by Least Squares. However, doing so does not necessarily deliver a positive-definite matrix $\mathbf{\Sigma_Y} - \mathbf{\Sigma_U}$. This is why it seems preferable to combine the linear restrictions (8) with the covariance restrictions (3), and perform a factor analysis of $\mathbf{\Sigma_Y}$ with linearly constrained error variances and covariances.

In practice, we simultaneously solve for the lower triangular matrices $\mathbf{W}$ (L-by-K) and $\mathbf{Z}$ (L-by-L) such that restrictions

$$\begin{aligned} \mathbf{\Sigma_Y} &= \mathbf{W}\mathbf{W}^{\mathrm{T}} + \mathbf{Z}\mathbf{Z}^{\mathrm{T}}, \\ \mathbf{B}^{\mathrm{T}}\,\mathrm{vech}\,(\mathbf{\Sigma_Y}) &= \mathbf{B}^{\mathrm{T}}\,\mathrm{vech}\left(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\right), \\ \left[\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\right]_{(\ell,m)} &= 0, \quad \forall (\ell, m) \in \mathcal{J}, \end{aligned}$$

approximately hold in a L^2 sense. This is a quadratic problem that can be solved using standard optimization routines.[14]

[13] Alternatively, if all factors are skewed, or either skewed or kurtotic, one can follow a similar procedure, basing the estimation on matrix $\mathbf{\Gamma_Y}$ or matrix $\mathbf{\Phi_Y}$, respectively.

[14] High-order moments are notoriously more difficult to estimate precisely than low-order moments (see the experiment in Table 1 below). One may thus want to weight second, third and fourth-order restrictions differently (see also Cragg, 1997).

The following remarks are in order. First, all second, third and fourth-order moments of the data are used to estimate $\mathbf{W}$. Higher-order moments appear in the matrix $\mathbf{B}$. This is for this reason why up to $K = L$ factors can be estimated, while, using second-order moments only, the maximal number of factors would be less than the following (generalized) Ledermann bound: $K \leq (2L + 1 - \sqrt{(2L+1)^2 - 8J})/2$.[15]

Second, if there are sufficiently many restrictions on $\mathbf{\Lambda}$, then one can estimate $\mathbf{\Lambda}$ together with $\mathbf{\Sigma_U}$ directly from this system. The source separation step below is not necessary (see the discussion following Corollary 1).

Source separation. Given whitened cumulant matrices $\widetilde{\mathbf{\Sigma}}_\mathbf{Y}$, $\widetilde{\mathbf{\Gamma}}_\mathbf{Y}(\ell)$ and $\widetilde{\mathbf{\Omega}}_\mathbf{Y}(\ell, m)$ (equations (14), (15) and (16)), we compute $\mathbf{V}$ as the K-by-K matrix of common orthonormal eigenvectors ($\mathbf{V}\mathbf{V}^\mathrm{T} = \mathbf{I}_K$) of matrices $\mathbf{P}\widetilde{\mathbf{\Gamma}}_\mathbf{Y}(\ell)\,\mathbf{P}^\mathrm{T}$ and $\mathbf{P}\widetilde{\mathbf{\Omega}}_\mathbf{Y}(\ell, m)\,\mathbf{P}^\mathrm{T}$, where $\mathbf{P}$ satisfies equation (17). For example, one can choose $\mathbf{P} = \mathbf{W}^-$ (the Moore-Penrose generalized inverse of $\mathbf{W}$), where $\mathbf{W}$ has been estimated in the prewhitening step. In this case, factor loadings are then obtained as $\mathbf{\Lambda} = \mathbf{W}\mathbf{V}$.

In practice, we replace theoretical moments by empirical analogs and use Cardoso and Souloumiac's (1993) Joint Approximate Diagonalization algorithm (JADE). This algorithm provides a fast way of minimizing with respect to an orthonormal matrix $\mathbf{V}$ the sum of squares of off-diagonal elements of matrices $\mathbf{V}^\mathrm{T}\mathbf{P}\widetilde{\mathbf{\Gamma}}_\mathbf{Y}(\ell)\,\mathbf{P}^\mathrm{T}\mathbf{V}$ and $\mathbf{V}^\mathrm{T}\mathbf{P}\widetilde{\mathbf{\Omega}}_\mathbf{Y}(\ell, m)\,\mathbf{P}^\mathrm{T}\mathbf{V}$. In practice, one may want to weight cumulant matrices according to their estimated precision.

The JADE algorithm is described in Appendix B. We call the resulting algorithm quasi-JADE, to emphasize the two-step nature of our procedure. It is only marginally more complicated to implement than JADE and almost as fast. However, unlike JADE,

[15] This bound is obtained by comparing the number of parameters to be estimated (that is: $K(L - (K-1)/2)$ unrotated factor loadings and $(L(L+1)/2 - J)$ error covariances) to the number of second-order moments (that is: $L(L+1)/2$). It coincides with the Ledermann bound when $\mathcal{J} = \{\ell < m\}$, and $J = L(L-1)/2$.

it is robust to the presence of (possibly correlated) noise.

Lastly, once factor loadings have been estimated, one can obtain the third and fourth-order cumulants of factor variables from the linear restrictions (6) and (7).[16]

4.2 Inference

As far as we know, there is no derivation of the asymptotic properties of JADE in the ICA literature. This section aims at filling this gap. At the end of the section, we discuss how to perform inference for the JADE and quasi-JADE estimates in practice.

To proceed, let $\widehat{\mathbf{A}}_1, ..., \widehat{\mathbf{A}}_S$ be root-N consistent and asymptotically normal estimators of S symmetric K-by-K matrices $\mathbf{A}_1, ..., \mathbf{A}_S$.[17] Construct $\widehat{\mathbf{A}} = \left[\widehat{\mathbf{A}}_1, ..., \widehat{\mathbf{A}}_S\right]$ and $\mathbf{A} = [\mathbf{A}_1, ..., \mathbf{A}_S]$ by concatenation. The JADE estimator is

$$\widehat{\mathbf{V}} = \arg\min_{\mathbf{V} \in \mathcal{O}_K} \sum_{s=1}^{S} \text{off}(\mathbf{V}^{\mathrm{T}} \widehat{\mathbf{A}}_s \mathbf{V}),$$

where $\text{off}(\mathbf{M}) = \sum_{i \neq s} m_{is}^2$ for a matrix $\mathbf{M} = [m_{is}]$, and $\mathcal{O}_K$ is the set of orthonormal K-by-K matrices.

Assume that there exists $\mathbf{V} \in \mathcal{O}_K$ such that, for all $s = 1, ..., S$, $\mathbf{V}^{\mathrm{T}} \mathbf{A}_s \mathbf{V} = \mathbf{D}_s$, where $\mathbf{D}_s$ is the diagonal matrix with diagonal elements $d_{s1}, ..., d_{sK}$. Define the K-by-K matrices:

$$\mathbf{R}_s = \left[\frac{(d_{sk} - d_{sm})}{\sum_{s'=1}^{S} (d_{s'k} - d_{s'm})^2} \right]_{k,m=1,...,K},$$

and $\mathbf{r}_s = \text{vec}(\mathbf{R}_s)$. Lastly, let $\mathbf{F}$ be the following K^2-by-SK^2 matrix:

$$\mathbf{F} = \left[\text{diag}\left(\mathbf{r}_1\right), ..., \text{diag}\left(\mathbf{r}_S\right) \right].$$

We show the following result in Appendix C.

[16] We use generalized inverses to do so. For example, using equation (6), the fourth-order cumulants of factor variables are estimated as the diagonal elements of an estimate of: $\mathbf{Q}^{-} \mathbf{\Omega}_{\mathbf{Y}} \left[\mathbf{Q}_{\mathcal{J}}^{-}\right]^{\mathrm{T}}$. We proceed similarly to estimate third-order cumulants.

[17] When applying JADE or quasi-JADE, $\widehat{\mathbf{A}}_1, ..., \widehat{\mathbf{A}}_S$ are whitened matrices of empirical cumulants, which are root-N consistent and asymptotically normal estimators of their population equivalents.

Theorem 5 *Assume that* $\sum_{s=1}^{S}(d_{sk}-d_{sm})^2 \neq 0$ *for all* $k \neq m$. *Then*

$$N^{\frac{1}{2}}\left(\mathrm{vec}(\widehat{\mathbf{V}})-\mathrm{vec}(\mathbf{V})\right) \to \mathcal{N}\left(0, \mathrm{Avar}\left(N^{\frac{1}{2}}\widehat{\mathbf{V}}\right)\right) \quad \textit{(weakly)},$$

where:

$$\mathrm{Avar}\left(N^{\frac{1}{2}}\widehat{\mathbf{V}}\right) = (\mathbf{I}_K \otimes \mathbf{V})\mathbf{F}(\mathbf{I}_S \otimes \mathbf{V}^{\mathrm{T}} \otimes \mathbf{V}^{\mathrm{T}})\,\mathrm{Avar}\left(N^{\frac{1}{2}}\widehat{\mathbf{A}}\right)(\mathbf{I}_S \otimes \mathbf{V} \otimes \mathbf{V})\mathbf{F}^{\mathrm{T}}(\mathbf{I}_K \otimes \mathbf{V}^{\mathrm{T}}), \quad (18)$$

where Avar *denotes the asymptotic variance.*

Let us consider the particular case of $S = 1$. In this case, (18) yields the well-known expression for the variance-covariance matrix of the eigenvectors of a symmetric matrix (e.g., Anderson, 1963). The diagonal coefficients of matrix $\mathbf{F}$ are equal to $1/(d_{1k} - d_{1m})$, for $k \neq m$. The variance of eigenvectors thus increases when two eigenvalues of $\mathbf{A}_1$ get close to each other.

In the general case of more than one matrix ($S > 1$), precise estimation requires $\sum_s (d_{sk} - d_{sm})^2$ to be away from zero, for all indices (k, m). Cardoso (1999) already noted that joint diagonalization algorithms seemed less sensitive to the presence of multiple roots than usual diagonalization techniques.[18] Theorem 5 permits to better understand the conditions granting a good precision.

In the quasi-JADE algorithm using fourth-order moments, indices are $s = (\ell, m)$, and matrices $\mathbf{D}_s$ are: $\mathbf{D}_4\,\mathrm{diag}\,(\mathbf{\Lambda}_\ell \odot \mathbf{\Lambda}_m)$. If there exist k, k' such that $d_{sk} = d_{sk'}$ for all s, it must be that

$$\lambda_{\ell k}\lambda_{mk}\kappa_4\left(X_k\right) = \lambda_{\ell k'}\lambda_{mk'}\kappa_4\left(X_{k'}\right), \forall(\ell, m).$$

This cannot happen if at most one factor has zero excess kurtosis and the columns of $\mathbf{\Lambda}$ are not proportional to each other.

This result is not surprising, as the variance of eigenvector estimators blows up when the model is not identified. Non identification arises in PCA when the variance of the

[18] See also the asymptotic distribution of estimators of Common Principal Components derived by Flury (1986).

vector of measurements has multiple eigenvalues (there are then obviously many possible choices for a basis of the corresponding eigenspace). In ICA this happens when two columns of the matrix of factor loadings are proportional or when factor distributions lack skewness and/or excess kurtosis. We shall produce Monte-Carlo simulations to illustrate this point.

Lastly, the asymptotic result for JADE given in Theorem 5 can be generalized to quasi-JADE, at the cost of introducing extra notation. As a result, the algorithm yields root-N consistent and asymptotically normal estimates of factor loadings and error moments, under the conditions of Theorem 2 (or Theorem 3, if using third-order moments) and Theorem 4. However, this generalization is not of direct interest to our purpose, as illustrated by the next remark.

Practical remark. In practice, we do *not* recommend to use formula (18) to compute standard errors. Instead, we suggest to compute standard errors or confidence intervals by standard bootstrap (maybe with appropriate recentering for finite sample improvements). The reason is that (18) involves variances of third and/or fourth-order moments of the data, i.e. sixth and eighth-order moments. These are difficult to estimate precisely (see Table 1 for an example with log-normal variables). In our simulation experiments, we obtained extremely imprecise estimates of matrix $\text{Avar}(\widehat{\mathbf{A}})$, even with very large samples (more than $10,000$ observations). In contrast, bootstrapping provided good approximations of the true variance-covariance matrix of the JADE estimator.

5 Monte-Carlo simulations

In this section, we study the finite-sample properties of our estimator with numerical simulations. Table 2 displays means and standard deviations of the Monte Carlo distributions of factor loadings estimates obtained from 1000 simulations of samples of various

Table 1: Empirical cumulants of the standard log-normal[1]

N	500	1000	5000	10000	true
κ_3[2]	4.49 (2.20)[3]	4.87 (2.47)	5.66 (2.54)	5.83 (3.17)	6.18
κ_4	35.9 (.88)	44.5 (.93)	72.9 (.72)	79.8 (.93)	110.9
κ_6	4,825 (.36)	8,698 (.35)	44,492 (.21)	55,505 (.28)	617,376
κ_8	856,819 (.22)	2,642,849 (.20)	59,108,559 (.12)	80,815,329 (.16)	1.647×10^{11}

[1] Estimated from 1000 independent draws of samples of size N.
[2] $\kappa_{3,4,6,8}$: skewness, excess kurtosis, 6th and 8th-order cumulants of a log-normal random variable.
[3] t-statistics in parentheses.

Table 2: Quasi-JADE for various sample sizes[1]

N	500	1000	5000	10000
λ_{11}	2.03 (.28)	2.03 (.17)	2.01 (.09)	2.01 (.06)
λ_{21}	.95 (.23)	.99 (.14)	1.00 (.07)	1.00 (.05)
λ_{31}	.95 (.23)	.99 (.15)	.99 (.07)	1.00 (.05)
$\mathrm{Var}(U_1)$	.77 (.59)	.87 (.43)	.96 (.20)	.98 (.16)

[1] Log-normal factors, standard normal errors, $\mathbf{\Lambda} = \mathbf{\Lambda_1}$.

sizes generated by standardized log-normal factors, standard normal errors and $\mathbf{\Lambda}$ equal to

$$\mathbf{\Lambda}_1 \equiv \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

We only report the estimates of the first column of $\mathbf{\Lambda}$ and the variance of the first error, the other estimates being qualitatively similar. Monte Carlo standard deviations of estimates are given between brackets. Estimation is based on all moments of order 2 and 4 of the data and uses the restrictions of Theorem 2. The error moments are estimated by least squares, based on restrictions (8) and (10).

Table 2 shows that finite sample biases are small and rapidly decrease as N increases. By comparison, small sample biases are much larger and convergence is much slower for empirical cumulants. The striking contrast between Tables 2 and 1 suggests that our algorithm does a good job at extracting the relevant information from high-order moments of the data, while being relatively immune to the imprecision of their estimation in finite samples.

Table 3: Robustness to noise[1]

JADE				
$\text{Var}(U_\ell)$	.01	.25	1	4
λ_{11}	2.00 (.07)	2.11 (.08)	2.36 (.12)	2.81 (.46)
λ_{21}	1.00 (.11)	1.00 (.12)	.95 (.24)	.72 (.86)
λ_{31}	1.00 (.11)	1.03 (.14)	1.08 (.22)	1.05 (.77)
Quasi-JADE				
$\text{Var}(U_\ell)$	.01	.25	1	4
λ_{11}	1.98 (.12)	2.01(.13)	2.03 (.17)	2.02 (.44)
λ_{21}	1.00 (.15)	.99 (.12)	.99 (.14)	.95 (.31)
λ_{31}	1.00 (.16)	.99 (.13)	.99 (.15)	.95 (.32)
$\text{Var}(U_1)$	.04 (.11)	.18 (.22)	.87 (.43)	3.77 (.98)
Minimum Distance				
$\text{Var}(U_\ell)$	.01	.25	1	4
λ_{11}	2.03 (.12)	2.04 (.14)	2.04 (.17)	2.02 (.43)
λ_{21}	.98 (.10)	.98 (.10)	.98 (.12)	.97 (.28)
λ_{31}	.98 (.10)	.98 (.11)	.99 (.13)	.98 (.28)
$\text{Var}(U_1)$	-.09 (.32)	.11 (.37)	.86 (.44)	3.75 (1.28)
% convergence	99.9%	100.0%	99.8%	84.3%

[1] Log-normal factors, standard normal errors, $\mathbf{\Lambda} = \mathbf{\Lambda_1}$, $N = 1000$.

Table 4: Near-Gaussianity biases[1]

κ_4	-6/5[2]	1/2	1	5	10	100	≈110[3]
λ_{11}	1.94 (.48)	1.66 (.78)	1.76 (.74)	2.03 (.33)	2.01 (.26)	2.01 (.19)	2.03 (20)
λ_{21}	.91 (.48)	.97 (.71)	.94 (.63)	.97 (.30)	.98 (.21)	.99 (.16)	.98 (.15)
λ_{31}	.92 (.48)	1.00 (.69)	.96 (.65)	.97 (.29)	.97 (.21)	.98 (.17)	.98 (.16)
$\text{Var}(U_1)$	.71 (.65)	.92 (.84)	.76 (.79)	.77 (.63)	.88 (.53)	.92 (.40)	.86 (.44)

[1] Factors are normal mixtures, standard normal errors, $\mathbf{\Lambda} = \mathbf{\Lambda_1}$, $N = 1000$.
[2] Uniform distribution.
[3] Log-normal distribution.

We then study the robustness of the (noise-free) JADE and quasi-JADE algorithms to noise (see Table 3). We run the simulations with normal errors, log-normal factors, a sample size of $N = 1000$ and $\mathbf{\Lambda} = \mathbf{\Lambda}_1$. The standard deviation of errors can take four values: 0.1, 0.5, 1 and 2. The performance of quasi-JADE deteriorates as the signal-to-noise ratio decreases. However, biases remain limited even for rather large error variances. By comparison, JADE, which does not allow explicitly for noise, produces large finite sample biases.[19]

In the last part of Table 3, we present the results of a Minimum Distance based on the complete set of moment restrictions (MD). The estimation is based on second and fourth-order restrictions (3) and (6). In all the simulations that we performed, MD proved to be highly unstable.[20] The bottom part of Table 3 presents simulation results with log-normal factors, normal errors and $\mathbf{\Lambda} = \mathbf{\Lambda}_1$. Conditional on numerical convergence,[21] MD yields only slightly more precise estimates of factor loadings than quasi-JADE. However, as error variances get larger, the MD algorithm fails to reach convergence more frequently (less than 1% of the time when $\text{Var}(U_\ell) \leq 1$ but 15% of the time when $\text{Var}(U_\ell) = 4$). In addition, the MD algorithm did not converge when we tried to estimate five factors or more, while quasi-JADE still delivered useful estimates in this case (see below). This comparison shows the usefulness of devising an estimation algorithm that efficiently combines the moments of the data while being numerically stable.

Next, we investigate the sensitivity of the quasi-JADE algorithm to factor Gaussianity.

[19] In the noise-free JADE algorithm, PCA was used to whiten the fourth-order cumulant matrices.

[20] Minimization with respect to the whole set of parameters $(\mathbf{\Lambda}, \mathbf{\Sigma_U}, \mathbf{D}_4)$ converged (numerically) in none of the cases that we considered. To obtain a more stable algorithm, admittedly with some efficiency loss, we treated the coefficients of $\mathbf{D}_4$ as nuisance parameters. Precisely, we minimized the MD norm, evaluated at $(\mathbf{\Lambda}, \mathbf{\Sigma_U}, \mathbf{D}_4(\mathbf{\Lambda}))$, with respect to $(\mathbf{\Lambda}, \mathbf{\Sigma_U})$ alone and where $\mathbf{D}_4(\mathbf{\Lambda})$ is the diagonal of $\mathbf{Q}^{-}\mathbf{\Omega_Y}\left[\mathbf{Q}_{\mathcal{J}}^{-}\right]^{\mathrm{T}}$. Note that using the optimal metric to estimate $\mathbf{D}_4(\mathbf{\Lambda})$ given $\mathbf{\Lambda}$, or incorporating third-order moment restrictions, yielded even greater instability.

[21] Starting values were chosen equal to the true parameters. We declared numerical convergence achieved when the gradient of the MD criterion was inferior to 10^{-3} in absolute value after 5000 Newton iterations.

Table 5: Comparison of Quasi-JADE and FA-JADE[(1)(2)]

Quasi-JADE				
Errors	normal		log-normal	
$\lambda_{11}, \lambda_{12}$	1.99 (.11)	1.00 (.18)	1.99 (.11)	1.00 (.19)
$\lambda_{21}, \lambda_{22}$	1.99 (.11)	1.00 (.18)	1.99 (.11)	1.00 (.19)
$\lambda_{31}, \lambda_{32}$	1.99 (.11)	1.00 (.18)	1.99 (.12)	1.00 (.19)
$\lambda_{41}, \lambda_{42}$	.99 (.19)	2.00 (.13)	.99 (.21)	1.99 (.16)
$\lambda_{51}, \lambda_{52}$	.99 (.19)	1.99 (.13)	.99 (.20)	2.00 (.15)

FA-JADE				
Errors	normal		log-normal	
$\lambda_{11}, \lambda_{12}$	2.00 (.09)	.98 (.20)	2.01 (.12)	.94 (.24)
$\lambda_{21}, \lambda_{22}$	2.00 (.09)	.98 (.20)	2.01 (.12)	.94 (.24)
$\lambda_{31}, \lambda_{32}$	2.00 (.09)	.98 (.20)	2.01 (.12)	.94 (.24)
$\lambda_{41}, \lambda_{42}$	1.00 (.22)	2.00 (.25)	1.03 (.27)	1.99 (.26)
$\lambda_{51}, \lambda_{52}$	1.00 (.23)	2.01 (.26)	1.04 (.26)	1.97 (.27)

[(1)] factors are normal mixtures with excess kurtosis $\kappa_4 = 5$, errors are normal or log-normal with unitary variance, $\mathbf{\Lambda} = \mathbf{\Lambda_2}$, $N = 1000$.

[(2)] "FA-JADE" is Stegeman and Mooijaart's (2007) sequential estimation algorithm.

The sample size is $N = 1000$. Errors are standard normal variables. We simulate symmetric, kurtotic factors as mixtures of two independent normals. Table 4 summarizes Monte Carlo distributions for excess kurtosis values in $\frac{1}{2}$, 2, 5, 10 and 100. In the first column of Table 4, we also report results for the case of uniformly distributed factors. The uniform distribution is platykurtic, with $\kappa_4 = -6/5$. The last column shows results for log-normal factors, with excess kurtosis equal to $e^4 + 2e^3 + 3e^2 - 6 \approx 110$. Overall, we find that the impact of kurtosis on the performance of the algorithm is far from negligible. The closer the excess kurtosis is to zero, the greater the estimator's bias and the lower its precision.[22]

It is also interesting to compare quasi-JADE to algorithms of noisy-ICA which use FA or PPCA in the prewhitening step (Beckmann and Smith, 2004, Ikeda and Toyama,

[22]We also experimented with errors being non-Gaussian, and found that quasi-JADE estimates were almost unchanged relative to the Gaussian case.

2000, Stegeman and Mooijaart, 2007). We use the algorithm of Stegeman and Mooijaart (2007, "FA-JADE" hereafter), which is based on an initial FA step. For this reason, the number of factors is limited by the Ledermann bound. For example, for $L = 3$ and independent errors, at most $K = 1$ factor can be estimated. To compare quasi-JADE and FA-JADE, we thus have to restrict the number of factors. We choose $L = 5, K = 2$, and set:

$$\mathbf{\Lambda}_2 \equiv \begin{pmatrix} 2 & 1 \\ 2 & 1 \\ 2 & 1 \\ 1 & 2 \\ 1 & 2 \end{pmatrix}.$$

Factors follow normal mixtures with excess kurtosis equal to $\kappa_4 = 5$, while errors have unitary variances and follow either a normal or a log-normal distribution. Lastly, $N = 1000$. Comparing the first four columns of Table 5 to the last four columns shows that allowing for higher-order moments in the prewhitening step of the algorithm (as in quasi-JADE) results in efficiency gains, at least for the second factor. Moreover, while FA-JADE is still consistent for $\mathbf{\Lambda}$ when errors are Gaussian, it fails to be when errors are log-normal. Indeed, factor loadings estimates are biased in the last two columns of the table. In comparison, quasi-JADE estimates remain almost unbiased.

Then, we investigate the finite-sample performance of our algorithm when the number of measurements and the number of factors increase. Table 6 illustrates the cases $L = K = 5$ and $L = K = 10$, respectively. In both cases, $\mathbf{\Lambda}$ has entries equal to 2 everywhere on the diagonal, and equal to one everywhere else. These simulations show that the performance of our algorithm is only moderately damped by the number of factors/measurements. We view this as quite remarkable a result as a hundred of factor loadings is certainly a significant number of parameters to estimate given that no explanatory variable is observed.[23]

[23]We performed three additional simulations, available in the additional appendix to this paper. We checked that, for skewed enough factors, using second and third-order moments only could result in precise estimates of factor loadings. We also found that quasi-JADE had good properties in the presence

Table 6: Increasing the number of factors and measurements[(1)]

	$L = K = 5$			$L = K = 10$		
N	500	1000	5000	500	1000	5000
λ_{11}	2.06 (.41)	2.03 (.28)	2.01 (.13)	1.85 (.72)	1.97 (.56)	2.00 (.27)
λ_{21}	.95 (.35)	.98 (.25)	.99 (.12)	.89 (.52)	.90 (.43)	.98 (.22)
λ_{31}	.95 (.34)	.98 (.24)	1.00 (.12)	.88 (.53)	.90 (.45)	.98 (.23)
λ_{41}	.95 (.35)	.98 (.24)	.99 (.11)	.88 (.53)	.92 (.43)	.98 (.22)
λ_{51}	.95 (.34)	.98 (.24)	.99 (.12)	.88 (.53)	.90 (.43)	.98 (.22)
λ_{61}				.88 (.54)	.91 (.43)	.98 (.22)
λ_{71}				.89 (.53)	.90 (.44)	.98 (.22)
λ_{81}				.88 (.52)	.90 (.44)	.98 (.23)
λ_{91}				.87 (.53)	.91 (.44)	.98 (.23)
$\lambda_{10,1}$				.88 (.52)	.89 (.44)	.98 (.22)
$\text{Var}(U_1)$	.58 (.56)	.81 (.44)	.95 (.20)	.40 (.55)	.49 (.53)	.88 (.28)

[(1)] Log-normal factors, standard normal errors.

6 Two applications

In this section we apply our methodology to data on cognitive test scores, and to financial data on stock returns. Details about the data and results are given in the additional appendix to this paper.

6.1 Test scores

We use data from the British National Child Development Study (NCDS), which is a longitudinal survey of a British birth cohort born in the same week of 1958. There are seven available test measures: mathematics and reading at age 7, 11 and 16, and a verbal test at age 11 only. We analyze the data with an independent factor model.[24] We allow the errors to be contemporaneously correlated (at age 7, age 11 and age 16). This is important and natural as the tests were taken on the same day. The data present sufficient non-normality for three factors to be estimable.

Table 7 shows the estimation results.[25] The first three columns show the factor

of correlated errors, and in a sparse overcomplete model.

[24] See Jennrich and Trendafilov (2005) for an application of noise-free ICA methods to psychometrics.

[25] In the estimation we use second, third and fourth-order moments jointly. Relative to second-order moments, third-order moments are weighted by a factor .178, and fourth-order ones by .091. These numbers correspond to the average of the (bootstrapped) variances of the components of $\widehat{\mathbf{\Sigma}}_{\mathbf{Y}}$ divided

Table 7: Test scores data: model estimates[(1)]

	Factor loadings			Error covariances						
Math (7)	13.6 (.25)	5.43 (.60)	−4.03 (.92)	363 (7.0)	16.4 (4.4)	0	0	0	0	0
Reading (7)	10.9 (.26)	13.4 (.90)	−6.64 (1.5)	16.4 (4.4)	141 (8.0)	0	0	0	0	0
Math (11)	22.4 (.21)	5.30 (.45)	−1.83 (1.1)	0	0	128 (5.1)	37.1 (2.9)	60.1 (3.1)	0	0
Reading (11)	11.0 (.25)	8.77 (.27)	2.06 (1.2)	0	0	37.1 (2.9)	104 (1.9)	37.2 (2.2)	0	0
Verbal (11)	14.8 (.28)	10.7 (.39)	−1.47 (1.4)	0	0	60.1 (3.1)	37.2 (2.2)	175 (3.7)	0	0
Math (16)	18.8 (.28)	4.07 (.41)	3.79 (.94)	0	0	0	0	0	114 (3.6)	−41.6 (2.1)
Reading (16)	12.2 (.37)	10.9 (.65)	7.05 (1.4)	0	0	0	0	0	−41.6 (2.1)	15.2 (1.7)
Skewness	.552 (.036)	−1.65 (.069)	.009 (.91)	−.0764 (.034)	−5.23 (.50)	−1.82 (.22)	−.0635 (.082)	−1.03 (.11)	1.18 (.16)	−68.9 (9.5)
Ex. kurtosis	−1.28 (.040)	1.28 (.21)	.520 (1.04)	−1.71 (.066)	3.52 (.61)	−8.83 (.62)	−3.07 (.17)	−6.51 (.32)	−.927 (.49)	185 (57)

[(1)] Source: NCDS 1965, 1969 and 1974. Details on the data are given in the additional appendix to this paper.

loadings estimates that correspond to each of the seven test score measures. The last seven columns give the estimates of the variance-covariance matrix of error variables. The last two rows give the skewness and excess kurtosis of the factor and error variables. Lastly, bootstrap standard errors are given in parentheses (100 iterations). We see that errors are sizeable in this application, the ratio of the sum of squares of factor loadings to total variance being 60%. This suggests that overlooking error variables in the model can have severe consequences on the results. To check that, we re-estimated factor loadings using JADE. We found that the second and third factors were essentially driven by the math and reading test scores at age 7, respectively. This is likely to be because the large errors in the test score at age 7 are wrongly interpreted as extra factors.

The first factor in Table 7 is correlated with scores in reading and mathematics, the correlation being stronger with the latter. In contrast, the second factor is correlated to reading and verbal test scores, but has small or zero correlation with the scores in mathematics. Lastly, the third factor puts positive weight on scores of tests taken at age

by the average of the variances of the components of $\widehat{\mathbf{\Gamma}}_{\mathbf{Y}}$ (of $\widehat{\mathbf{\Omega}}_{\mathbf{Y}}$, respectively).

16, and negative weight on tests taken at age 7. These three factors account for 45%, 19% and 4% of the total variance, respectively, while errors account for 32%. We interpret the first and second factors as mathematical and verbal abilities, and the third one as reflecting heterogeneous learning slopes, characterizing children who perform better early on (at age 7) or later in their academic career (at age 16).

6.2 Stock returns

In an influential paper, Fama and French (1993) identify two factors, in addition to the market return, explaining a large proportion of the variance of time-series of U.S. excess stock returns. "Small Minus Big" (SMB) is the difference between the average of the returns on two stock portfolios: one containing firms with market value (price time number of shares) less than the median, and one containing firms with size above the median. "High Minus Low" (HML) is the difference between the average of the returns on two stock portfolios: one gathering firms with book-to-market ratio (book value of capital divided by market value) less that the 30th percentile and another one containing all firms with a ratio above the 70th percentile. Fama and French show that these three factors explain monthly data on 25 portfolios formed by intersecting size and book-to-market quintiles remarkably well.

We apply quasi-JADE to estimate a linear independent factor model with three factors. We use daily data on the returns to 25 stock portfolios formed on size and book-to-market, collected between 01/07/1963 and 31/08/2005 by Fama and French.[26] The size and book-to-market breakpoints are NYSE quintiles. There are 10,616 observations. Returns are net of the risk-free rate.

Table 8 shows the correlations between the three factors that we estimate using quasi-JADE and Fama and French's factors. We see that the three factors estimated

[26]These data can be downloaded from Kenneth French's website:
http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Table 8: Stock returns data: Fama French factors *versus* quasi-JADE estimates

Factors	First	Second	Third
Market return	.84	.24	-.41
"Small minus big"	-.49	.85	-.09
"High minus low"	-.11	.23	.90

(1) Source: 25 stock portfolios formed on size and book-to-market, daily US data. Downloaded from Kenneth French's website.

by quasi-JADE are strongly correlated with the market, size and book-to-market factors constructed by Fama and French. The correlations are .84, .85 and .90, respectively. This indicates that our blind source separation procedure yields factor estimates which have sound economic sense.[27]

7 Conclusion

The recent literature on Independent Component Analysis (ICA) has produced several methods able to deal with noise-free, linear independent factor models with up to $K = L$ factors. In this paper we have developed an algorithm that robustifies one of the most popular ICA algorithms, Cardoso and Souloumiac's (1993) JADE, when measurement error cannot be neglected. We have constructed a two-stage consistent estimator for noisy ICA with clustered errors, quasi-JADE. In the prewhitening step, error moments are estimated from second to fourth-order moments of the data, while in the source separation step JADE is applied to the whitened cumulant matrices.

Monte Carlo results are encouraging. For sufficiently non symmetric and/or kurtotic data, we obtain small biases and precise estimates, even in relatively small samples. Moreover, the application to test scores shows that allowing for noise can be very important in practical situations. This suggests that quasi-JADE can be a valid alternative to

[27]We also experimented on stock data grouped by industries, finding much lower correlations. This casts some doubts on the ability of Fama and French's factors to explain disaggregate data on stock returns with the same success.

existing methods in traditional applications of ICA, like signal processing, where it can be used in place of noise-free methods. Moreover, in situations where factor analysis is widely used (macroeconomics, finance, psychometrics) quasi-JADE provides a consistent way to fix the rotation matrix.

In the future, we plan to pursue two directions of research. First, it would be interesting to extend quasi-JADE to deal with *noisy* overcomplete ICA models ($K > L$). The second direction of research concerns the extension of the method of this paper to the case of a very large number of measurements. Bai and Ng (2002) and Bai (2003) provide extensive analyses of the PCA estimator in this case. Financial and macroeconomic applications motivate the need to extend ICA methods in this direction.

Finally, once factor loadings have been estimated, it remains to estimate the distribution of factors and errors. This is done in a companion paper (see Bonhomme and Robin, 2008).

APPENDIX

A Mathematical proofs

We start with some notation. For a n-by-m matrix $\mathbf{A}$, we denote as $\mathbf{A}[\mathcal{R}, \mathcal{C}]$ the submatrix of rows $i \in \mathcal{R}$ and columns $j \in \mathcal{C}$, for $\mathcal{R} \subseteq \{1, ..., n\}$ and $\mathcal{C} \subseteq \{1, ..., m\}$. If $\mathcal{R} = \{1, ..., n\}$ or $\mathcal{C} = \{1, ..., m\}$, we write $\mathbf{A}[\cdot, \mathcal{C}]$ and $\mathbf{A}[\mathcal{R}, \cdot]$.

So, in particular, matrix $\mathbf{Q}_{\mathcal{J}}$ can be equivalently written as $\mathbf{Q}[\mathcal{J}, \cdot]$, and $\mathbf{\Lambda}_{\mathcal{I}_\ell}$ as $\mathbf{\Lambda}[\mathcal{I}_\ell, .]$.

A.1 Proof of Lemma 1

Let $(\ell, m) \in \mathcal{J}$. As $U_\ell \perp\!\!\!\perp U_m$, we have $\text{Cov}(U_\ell, U_m) = 0$. In addition: $U_\ell = \mathbf{\Pi}_\ell^{\mathrm{T}} \boldsymbol{\varepsilon} \perp\!\!\!\perp U_m = \mathbf{\Pi}_m^{\mathrm{T}} \boldsymbol{\varepsilon}$, where $\mathbf{\Pi}_\ell^{\mathrm{T}}$ is the ℓth row of matrix $\mathbf{\Pi}$. It follows from Darmois' theorem (e.g., Comon, 1994, p. 306) that for all $h \in \{1, ..., H\}$ either ε_h is Gaussian or $\pi_{\ell h}\pi_{mh} = 0$. In either case:

$$\pi_{\ell h}\pi_{mh}\kappa_3(\varepsilon_h) = \pi_{\ell h}\pi_{mh}\kappa_4(\varepsilon_h) = 0.$$

The conclusion comes from the cumulant identities:

$$\begin{aligned} \text{Cum}(U_i, U_\ell, U_m) &= \sum_{h=1}^{H} \pi_{ih}\pi_{\ell h}\pi_{mh}\kappa_3(\varepsilon_h), \\ \text{Cum}(U_i, U_j, U_\ell, U_m) &= \sum_{h=1}^{H} \pi_{ih}\pi_{jh}\pi_{\ell h}\pi_{mh}\kappa_4(\varepsilon_h). \end{aligned}$$

A.2 Proof of Theorem 2

To simplify the exposition, let us define $\boldsymbol{\sigma}_{\mathbf{Y}} \equiv \text{vech}(\mathbf{\Sigma}_{\mathbf{Y}})$, $\boldsymbol{\gamma}_{\mathbf{Y}}(\ell) \equiv \text{vech}(\mathbf{\Gamma}_{\mathbf{Y}}(\ell))$, and $\boldsymbol{\omega}_{\mathbf{Y}}(\ell, m) \equiv \text{vech}(\mathbf{\Omega}_{\mathbf{Y}}(\ell, m))$, with similar notation for $\boldsymbol{\sigma}_{\mathbf{U}}$, $\gamma_{\mathbf{U}}(\ell)$ and $\boldsymbol{\omega}_{\mathbf{U}}(\ell, m)$. Let also

$$\mathcal{J}^c = \left\{(\ell, m) \in \{1, ..., L\}^2, \ell \leq m\right\} \setminus \mathcal{J}.$$

Remark that $\boldsymbol{\sigma}_{\mathbf{U}}$, $\boldsymbol{\gamma}_{\mathbf{U}}(\ell)$ and $\boldsymbol{\omega}_{\mathbf{U}}(\ell, m)$ have zero entries in positions $(i, j) \in \mathcal{J}$. Construct vectors $\boldsymbol{\sigma}_{\mathbf{U}}[\mathcal{J}^c]$, $\boldsymbol{\gamma}_{\mathbf{U}}(\ell)[\mathcal{J}^c]$ and $\boldsymbol{\omega}_{\mathbf{U}}(\ell, m)[\mathcal{J}^c]$ by dropping the zero entries. Let also $\mathbf{B}[\mathcal{J}^c, \cdot]$ be the submatrix obtained by selecting the rows of $\mathbf{B}$ indexed by couples $(\ell, m) \notin \mathcal{J}$. Equations (8), (9) and (10) imply

$$\mathbf{B}^{\mathrm{T}}\boldsymbol{\sigma}_{\mathbf{Y}} = \mathbf{B}[\mathcal{J}^c, \cdot]^{\mathrm{T}}\boldsymbol{\sigma}_{\mathbf{U}}[\mathcal{J}^c], \tag{A1}$$

$$\mathbf{B}^{\mathrm{T}}\boldsymbol{\gamma}_{\mathbf{Y}}(\ell) = \mathbf{B}[\mathcal{J}^c, \cdot]^{\mathrm{T}}\boldsymbol{\gamma}_{\mathbf{U}}(\ell)[\mathcal{J}^c], \forall \ell, \tag{A2}$$

$$\mathbf{B}^{\mathrm{T}}\boldsymbol{\gamma}_{\mathbf{Y}}(\ell, m) = \mathbf{B}[\mathcal{J}^c, \cdot]^{\mathrm{T}}\boldsymbol{\omega}_{\mathbf{U}}(\ell, m)[\mathcal{J}^c], \forall (\ell, m). \tag{A3}$$

We shall show that matrix $\mathbf{B}[\mathcal{J}^c, \cdot]$ has full row rank, which will prove the identification of error moments. To proceed, remark that $\mathbf{B}[\mathcal{J}^c, \cdot]$ has $\frac{L(L+1)}{2} - J$ rows and $\frac{L(L+1)}{2} - K$ columns. If $J \geq K$, $\mathbf{B}[\mathcal{J}^c, \cdot]$ has more columns than rows. Let $r = \text{rank}(\mathbf{B}[\mathcal{J}^c, \cdot])$.

Suppose that $r < \frac{L(L+1)}{2} - J$. There exists a $\left(\frac{L(L+1)}{2} - K\right)$-by-$\left(\frac{L(L+1)}{2} - K - r\right)$ matrix $\mathbf{A}$, full column rank, such that $\mathbf{B}[\mathcal{J}^c, \cdot]\mathbf{A} = 0$. As both $\mathbf{B}$ and $\mathbf{A}$ have full column rank, $\mathbf{BA}$ has full column rank, hence $\mathbf{B}[\mathcal{J}, \cdot]\mathbf{A}$ necessarily has full column rank $\frac{L(L+1)}{2} - K - r$, with

$$\frac{L(L+1)}{2} - K - r > J - K. \tag{A4}$$

Moreover, as $\mathbf{Q}^{\mathrm{T}}\mathbf{B} = 0$ by construction,

$$0 = \mathbf{Q}^{\mathrm{T}}\mathbf{BA} = \mathbf{Q}[\mathcal{J}, \cdot]^{\mathrm{T}}\mathbf{B}[\mathcal{J}, \cdot]\mathbf{A},$$

Now, $\mathbf{Q}[\mathcal{J}, \cdot]$ has J rows and K columns. It has full column rank, so its null space has dimension $J - K$. This contradicts condition (A4) on the rank of $\mathbf{B}[\mathcal{J}, \cdot]\mathbf{A}$. Hence, $r = \frac{L(L+1)}{2} - J$ and matrix $\mathbf{B}[\mathcal{J}^c, \cdot]$ therefore must have full row rank.

This ends the proof of Theorem 2.

A.3 Proof of Theorem 3

Let us define $\mathcal{I}_\ell^c = \{m \in \{1, ..., L\}, m \notin \mathcal{I}_\ell\}$, for all $\ell \in \{1, ..., L\}$, that is,

$$\mathcal{I}_\ell^c = \{m \in \{1, ..., L\}, \ell \leq m \text{ and } (\ell, m) \in \mathcal{J}^c\}.$$

1. We first show that, for all ℓ, $\mathbf{C}[\mathcal{I}_\ell^c, \cdot]$ has full row rank in the same way as in the proof of Theorem 2.

Matrix $\mathbf{C}[\mathcal{I}_\ell^c, \cdot]$ has $L - I_\ell$ rows and $L - K$ columns. As, by assumption, $\mathbf{\Lambda}[\mathcal{I}_\ell, \cdot]$ has rank K and dimensions I_ℓ-by-K, $I_\ell \geq K$. Suppose that $r = \text{rank}(\mathbf{C}[\mathcal{I}_\ell^c, \cdot]) < L - I_\ell$. There exists a full column rank, $(L - K)$-by-$(L - K - r)$ matrix $\mathbf{A}$, such that $\mathbf{C}[\mathcal{I}_\ell^c, \cdot]\mathbf{A} = 0$. Both $\mathbf{A}$ and $\mathbf{C}$ having full column rank, $\mathbf{CA}$ has also full column rank. Hence, $\mathbf{C}[\mathcal{I}_\ell, \cdot]\mathbf{A}$ has full column rank $L - K - r$.

Moreover, $\mathbf{C}^{\mathrm{T}}\mathbf{\Lambda} = 0$. Hence,

$$0 = \mathbf{\Lambda}^{\mathrm{T}}\mathbf{CA} = \mathbf{\Lambda}[\mathcal{I}_\ell, \cdot]^{\mathrm{T}}\mathbf{C}[\mathcal{I}_\ell, \cdot]\mathbf{A}.$$

By assumption, $\mathbf{\Lambda}[\mathcal{I}_\ell, \cdot]$ is full column rank K and its null space has dimension $I_\ell - K$. Therefore, $\mathbf{C}[\mathcal{I}_\ell, \cdot]\mathbf{A}$ cannot have a rank greater than $I_\ell - K$:

$$L - K - r \leq I_\ell - K.$$

Hence $r \geq L - I_\ell$, which contradicts the assumption.

2. Now, applying the vech operator to (11), (12), (13) shows that error cumulants satisfy the linear system (A1), (A2), (A3) with, in place of $\mathbf{B}[\mathcal{J}^c, \cdot]$, the block diagonal matrix

$$\mathbf{D} \equiv \text{diag}\left(\mathbf{C}[\mathcal{I}_1^c, \cdot], ..., \mathbf{C}[\mathcal{I}_L^c, \cdot]\right).$$

As $\mathbf{C}[\mathcal{I}_\ell^c, \cdot]$ has full row rank for all ℓ, it follows that $\mathbf{D}$ has also full row rank.

This ends the proof of Theorem 3.

A.4 Proof of Theorem 4

To prove Theorem 4, we first prove the following lemma giving conditions under which the joint eigenvectors of a set of matrices is uniquely defined (up to sign and permutation).[28]

Lemma 2 *Let K and L be any integers. Let $\mathbf{A}_1, ..., \mathbf{A}_L$ be K-by-K matrices. Suppose that there exist $\mathbf{x}^k = \left(x_1^k, ..., x_L^k\right)^{\mathrm{T}} \in \mathbb{R}^L$ and $\mathbf{v}^k \in \mathbb{R}^K$, $\mathbf{v}^k \neq 0$, $k = 1, ..., K+1$, solutions to the joint diagonalization problem:*

$$x_\ell^k \mathbf{v}^k = \mathbf{A}_\ell \mathbf{v}^k, \quad \forall \ell = 1, ..., L.$$

Assume that the set $\left\{\mathbf{v}^1, ..., \mathbf{v}^K\right\}$ is linearly independent, that all $\mathbf{v}^k$, $k = 1, ..., K+1$, have norm one, and that $\mathbf{x}^k \neq \mathbf{x}^{k'}$ for all $(k, k') \in \{1, ..., K\}^2$, $k \neq k'$. Then, there exists $k \in \{1, ..., K\}$ such that $\mathbf{v}^{K+1} = \pm \mathbf{v}^k$.

Proof. Since $\left\{\mathbf{v}^1, ..., \mathbf{v}^K\right\}$ is a basis of $\mathbb{R}^K$, there exists $\mathbf{c} = (c_1, ..., c_K) \neq 0$ such that $\mathbf{v}^{K+1} = c_1 \mathbf{v}^1 + ... + c_K \mathbf{v}^K$. Then, for all $\ell = 1, ..., L$,

$$\sum_{k=1}^K c_k x_\ell^k \mathbf{v}^k = \sum_{\ell=1}^K c_k \mathbf{A}_\ell \mathbf{v}^k = \mathbf{A}_\ell \sum_{\ell=1}^K c_k \mathbf{v}^k = \mathbf{A}_\ell \mathbf{v}^{K+1} = x_\ell^{K+1} \mathbf{v}^{K+1} = x_\ell^{K+1} \left(\sum_{k=1}^K c_k \mathbf{v}^k\right).$$

As $(\mathbf{v}^1, ..., \mathbf{v}^K)$ is linearly independent, it follows from the last equality that:

$$c_k x_\ell^k = c_k x_\ell^{K+1},$$

for all (k, ℓ). Hence, for all k:

$$c_k \mathbf{x}^k = c_k \mathbf{x}^{K+1}.$$

As $\mathbf{c} \neq 0$, there exists k such that $c_k \neq 0$. For this k: $\mathbf{x}^k = \mathbf{x}^{K+1}$. Moreover, as $\mathbf{x}^k \neq \mathbf{x}^{k'}$ for all $k' \neq k$ in $\{1, ..., K\}$, it follows that $c_{k'} = 0$ for all $k' \neq k$. Hence

$$\mathbf{v}^{K+1} = c_k \mathbf{v}^k.$$

[28] The result of Lemma 2 can also be found in De Lathauwer *et al.* (2004), p. 305.

As both $\mathbf{v}^k$ and $\mathbf{v}^{K+1}$ have norm one, $c_k = \pm 1$. The result follows. ■

The proof of Theorem 4 easily follows.

Fourth-order moments. Second and fourth-order cumulant restrictions (3)-(5) yield:

$$\widetilde{\mathbf{\Omega}}_{\mathbf{Y}}(\ell, m) = \mathbf{\Lambda}\mathbf{D}_4 \operatorname{diag}(\mathbf{\Lambda}_\ell \odot \mathbf{\Lambda}_m)\mathbf{\Lambda}^{\mathrm{T}}, \quad \ell \leq m, \tag{A5}$$

$$\widetilde{\mathbf{\Sigma}}_{\mathbf{Y}} = \mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}}. \tag{A6}$$

Let $\widetilde{\mathbf{\Lambda}}$ be another value satisfying restrictions (A5)-(A6). We show that under the conditions of Theorem 4, there necessarily exists a sign-permutation matrix $\mathbf{S}$ such that $\widetilde{\mathbf{\Lambda}} = \mathbf{\Lambda}\mathbf{S}$.

$\mathbf{\Lambda}$ having full column rank K, and $\widetilde{\mathbf{\Sigma}}_{\mathbf{Y}}$ being positive definite, there exists a unique orthonormal L-by-K matrix $\mathbf{O}$ ($\mathbf{O}^{\mathrm{T}}\mathbf{O} = \mathbf{I}_K$) and a unique K-by-K diagonal, positive matrix $\mathbf{D}$ such that $\widetilde{\mathbf{\Sigma}}_{\mathbf{Y}} = \mathbf{O}\mathbf{D}\mathbf{O}^{\mathrm{T}}$. Let $\mathbf{P} = \mathbf{D}^{-1/2}\mathbf{O}^{\mathrm{T}}$. Then $\mathbf{V} = \mathbf{P}\mathbf{\Lambda}$ is a matrix of joint orthonormal eigenvectors ($\mathbf{V}\mathbf{V}^{\mathrm{T}} = \mathbf{I}_K$) of

$$\mathbf{P}\widetilde{\mathbf{\Omega}}_{\mathbf{Y}}(\ell, m)\mathbf{P}^{\mathrm{T}} = \mathbf{P}\mathbf{\Lambda}\mathbf{D}_4 \operatorname{diag}(\mathbf{\Lambda}_\ell \odot \mathbf{\Lambda}_m)\mathbf{\Lambda}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}, \quad \ell \leq m.$$

In general, there can be infinitely many joint eigenvectors to a set of matrices if all matrices have multiple roots. However, Lemma 2 shows that the problem of diagonalizing matrices $\mathbf{P}\widetilde{\mathbf{\Omega}}_{\mathbf{Y}}(\ell, m)\mathbf{P}^{\mathrm{T}}$ has a unique solution up to column sign and permutation if for all $(k, k') \in \{1, ..., K\}^2$, $k \neq k'$, there exists $\ell \leq m$ such that

$$\lambda_{\ell k}\lambda_{mk}\kappa_4(X_k) \neq \lambda_{\ell k'}\lambda_{mk'}\kappa_4(X_{k'}).$$

As either $\kappa_4(X_k) \neq 0$ or $\kappa_4(X_{k'}) \neq 0$, and as any two columns of $\mathbf{\Lambda}$ are linearly independent, this condition is always satisfied. It follows that $\mathbf{V}$ is uniquely defined, up to column sign and permutation.

Now, the true $\mathbf{\Lambda}$ necessarily verifies:

$$\mathbf{\Lambda} = \mathbf{\Lambda}(\mathbf{P}\mathbf{\Lambda})^{\mathrm{T}}(\mathbf{P}\mathbf{\Lambda}) = \mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}\mathbf{P}\mathbf{\Lambda} = \widetilde{\mathbf{\Sigma}}_{\mathbf{Y}}\mathbf{P}^{\mathrm{T}}\mathbf{P}\mathbf{\Lambda} = \mathbf{O}\mathbf{D}^{1/2}\mathbf{P}\mathbf{\Lambda} = \mathbf{O}\mathbf{D}^{1/2}\mathbf{V}.$$

It is thus unique as $\mathbf{V}$ is unique.

Third-order moments. The same argument applies to third-order cumulant matrices $\widetilde{\mathbf{\Gamma}}_{\mathbf{Y}}(\ell)$. Indeed, in the noise-free case third-order restrictions (4) become

$$\widetilde{\mathbf{\Gamma}}_{\mathbf{Y}}(\ell) = \mathbf{\Lambda}\mathbf{D}_3 \operatorname{diag}(\mathbf{\Lambda}_\ell)\mathbf{\Lambda}^{\mathrm{T}}, \quad \ell \in \{1, ..., L\}.$$

In this case, Lemma 2 shows that the common eigenvectors corresponding to eigenvalues $\mathbf{D}_3 \operatorname{diag}(\mathbf{\Lambda}_\ell)$ are uniquely determined up to column sign and permutation if for all $(k, k') \in \{1, ..., K\}^2$, $k \neq k'$, there exists $\ell \in \{1, ..., L\}$ such that

$$\lambda_{\ell k} \kappa_3(X_k) \neq \lambda_{\ell k'} \kappa_3(X_{k'}).$$

As before, this condition is always satisfied.

Third and fourth-order moments. The proof is almost identical to the two previous ones. With $\widetilde{\mathbf{\Omega}}_{\mathbf{Y}}(\ell, m)$ and $\widetilde{\mathbf{\Gamma}}_{\mathbf{Y}}(\ell)$ together, eigenvectors are identified if for all $(k, k') \in \{1, ..., K\}^2$, $k \neq k'$, there exists (ℓ, m) such that

$$\lambda_{\ell k} \lambda_{mk} \kappa_4(X_k) \neq \lambda_{\ell k'} \lambda_{mk'} \kappa_4(X_{k'}),$$

or there exists $\ell \in \{1, ..., L\}$ such that

$$\lambda_{\ell k} \kappa_3(X_k) \neq \lambda_{\ell k'} \kappa_3(X_{k'}).$$

As one of the four moments $\kappa_3(X_k)$, $\kappa_3(X_{k'})$, $\kappa_4(X_k)$ and $\kappa_4(X_{k'})$ is non zero, it follows from the assumptions on $\mathbf{\Lambda}$ that this condition is always satisfied.

B The JADE algorithm

Let $\mathcal{A} = \{A_k, k = 1...K\}$ a set of real, symmetric, L-by-L matrices. Let us define the function:

$$\operatorname{off}(\mathbf{A}) = \sum_{i \neq j} a_{ij}^2,$$

for all $\mathbf{A} = [a_{ij}]$. Then joint approximate diagonalization of $\mathcal{A}$ is achieved by minimizing

$$\sum_{k=1}^{K} \operatorname{off}(\mathbf{U} \mathbf{A}_k \mathbf{U}^{\mathrm{T}}), \tag{B7}$$

with respect to $\mathbf{U}$ orthogonal.

Let $\theta \in [-\pi, \pi]$, let $(i, j) \in \{1, ..., L\}^2$ and let $\mathbf{R}_{ij}(\theta)$ be the L-by-L matrix equal to the identity matrix except at the (i, i), (i, j), (j, i) and (j, j) entries where it is equal to:

$$\begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix}.$$

Let $i \neq j$, and let us define:

$$O_{i,j}(\theta) = \sum_{k=1}^{K} \operatorname{off}\left(\mathbf{R}_{ij}(\theta) \mathbf{A}_k \mathbf{R}_{ij}(\theta)^{\mathrm{T}}\right).$$

Lastly, let $\mathbf{h}_{i,j}(\mathbf{A}) = (a_{ii} - a_{jj}, a_{ij} + a_{ji})$, and let:

$$\mathbf{G}_{i,j} = \sum_{k=1}^{K} \mathbf{h}_{i,j}^{\mathrm{T}}(\mathbf{A}_k)\mathbf{h}_{i,j}(\mathbf{A}_k) = (g_{ij})_{i,j=1,2}\,.$$

Cardoso and Souloumiac (1996) show that θ_0 such that:

$$\cos(\theta_0) = \sqrt{\frac{x+r}{2r}}, \quad \sin(\theta_0) = \sqrt{\frac{y}{2r(x+r)}},$$

where $x = g_{11} - g_{22}$, $y = g_{12} + g_{21}$ and $r = \sqrt{x^2 + y^2}$, minimizes $O_{i,j}(\theta)$.

This closed-form expression for θ_0 allows to minimize (B7) by the following algorithm:

1. Start with $\mathbf{U}(0) = \mathbf{I}_L$.
2. Begin loop on step s.
3. Begin loop on (i, j).
4. Compute $\mathbf{G}_{i,j}$.
5. Compute θ_0.
6. If θ_0 is different enough from zero, continue. Else stop.
7. Compute $\mathbf{R}_{ij}(\theta_0)\mathbf{A}_k\mathbf{R}_{ij}(\theta_0)^{\mathrm{T}}$ and modify $\mathcal{A}$ consequently.
8. Update $\mathbf{U}(s)$ as $\mathbf{U}(s+1) = \mathbf{R}_{ij}(\theta_0)\mathbf{U}(s)$.
9. End loop on (i, j).
10. End loop on s.

C Asymptotic theory of the JADE estimator

First-order conditions. The JADE estimator solves

$$\widehat{\mathbf{V}} = \arg\min_{\mathbf{V}\in\mathcal{O}_K} \sum_{s=1}^{S} \mathrm{off}(\mathbf{V}^{\mathrm{T}}\widehat{\mathbf{A}}_s\mathbf{V}).$$

The Lagrangian associated with the minimization problem is:

$$\begin{aligned}\mathcal{L}(\mathbf{V}, \boldsymbol{\gamma}) &= \sum_{s=1}^{S} \mathrm{off}(\mathbf{V}^{\mathrm{T}}\widehat{\mathbf{A}}_s\mathbf{V}) + \boldsymbol{\gamma}^{\mathrm{T}} \operatorname{vec}\left(\mathbf{V}^{\mathrm{T}}\mathbf{V} - \mathbf{I}_K\right), \\ &= \sum_{s} \sum_{m\neq k} (\mathbf{v}_k^{\mathrm{T}}\widehat{\mathbf{A}}_s\mathbf{v}_m)^2 + \sum_{k} \gamma_{kk}(\mathbf{v}_k^{\mathrm{T}}\mathbf{v}_k - 1) + \sum_{m\neq k} \gamma_{mk}\mathbf{v}_k^{\mathrm{T}}\mathbf{v}_m,\end{aligned}$$

where $\boldsymbol{\gamma}$ is a vector of length K^2 containing Lagrange multipliers γ_{mk}, and $\mathbf{v}_k$ is the kth column of matrix $\mathbf{V}$.

Differentiating the Lagrangian with respect to $\mathbf{v}_\ell$, for $\ell = 1...K$, yields:

$$\frac{\partial \mathcal{L}\left(\widehat{\mathbf{V}}, \widehat{\boldsymbol{\gamma}}\right)}{\partial \mathbf{v}_\ell} = 2\sum_s \sum_{k \neq \ell} (\widehat{\mathbf{v}}_k^{\mathrm{T}} \widehat{\mathbf{A}}_s \widehat{\mathbf{v}}_\ell) \widehat{\mathbf{A}}_s \widehat{\mathbf{v}}_k + 2\widehat{\gamma}_{\ell\ell} \widehat{\mathbf{v}}_\ell + \sum_{k \neq \ell} \widehat{\gamma}_{k\ell} \widehat{\mathbf{v}}_k = 0.$$

Then, multiplying this equation by $\widehat{\mathbf{v}}_m^{\mathrm{T}}$, for $m \neq \ell$, gives:

$$2\sum_s \sum_{k \neq \ell} (\widehat{\mathbf{v}}_k^{\mathrm{T}} \widehat{\mathbf{A}}_s \widehat{\mathbf{v}}_\ell) \widehat{\mathbf{v}}_m^{\mathrm{T}} \widehat{\mathbf{A}}_s \widehat{\mathbf{v}}_k + \widehat{\gamma}_{m\ell} = 0.$$

Using that $\widehat{\gamma}_{m\ell} = \widehat{\gamma}_{\ell m}$ by symmetry, it follows that

$$\sum_s \sum_{k \neq \ell} (\widehat{\mathbf{v}}_k^{\mathrm{T}} \widehat{\mathbf{A}}_s \widehat{\mathbf{v}}_\ell) \widehat{\mathbf{v}}_m^{\mathrm{T}} \widehat{\mathbf{A}}_s \widehat{\mathbf{v}}_k = \sum_s \sum_{k \neq m} (\widehat{\mathbf{v}}_k^{\mathrm{T}} \widehat{\mathbf{A}}_s \widehat{\mathbf{v}}_m) \widehat{\mathbf{v}}_\ell^{\mathrm{T}} \widehat{\mathbf{A}}_s \widehat{\mathbf{v}}_k,$$

or, equivalently, as $\widehat{\mathbf{A}}_s$ is symmetric for all s:

$$\sum_s \widehat{\mathbf{v}}_\ell^{\mathrm{T}} \widehat{\mathbf{A}}_s \left(\sum_{k \neq \ell} \widehat{\mathbf{v}}_k \widehat{\mathbf{v}}_k^{\mathrm{T}} \right) \widehat{\mathbf{A}}_s \widehat{\mathbf{v}}_m = \sum_s \widehat{\mathbf{v}}_m^{\mathrm{T}} \widehat{\mathbf{A}}_s \left(\sum_{k \neq m} \widehat{\mathbf{v}}_k \widehat{\mathbf{v}}_k^{\mathrm{T}} \right) \widehat{\mathbf{A}}_s \widehat{\mathbf{v}}_\ell.$$

Then, as $\sum_{k=1}^K \widehat{\mathbf{v}}_k \widehat{\mathbf{v}}_k^{\mathrm{T}} = \widehat{\mathbf{V}} \widehat{\mathbf{V}}^{\mathrm{T}} = \mathbf{I}_K$ we obtain

$$\sum_s \widehat{\mathbf{v}}_\ell^{\mathrm{T}} \widehat{\mathbf{A}}_s \left(\mathbf{I}_K - \widehat{\mathbf{v}}_\ell \widehat{\mathbf{v}}_\ell^{\mathrm{T}} \right) \widehat{\mathbf{A}}_s \widehat{\mathbf{v}}_m = \sum_s \widehat{\mathbf{v}}_m^{\mathrm{T}} \widehat{\mathbf{A}}_s \left(\mathbf{I}_K - \widehat{\mathbf{v}}_m \widehat{\mathbf{v}}_m^{\mathrm{T}} \right) \widehat{\mathbf{A}}_s \widehat{\mathbf{v}}_\ell,$$

which we write after rearranging:

$$\sum_s \widehat{\mathbf{v}}_\ell^{\mathrm{T}} \widehat{\mathbf{A}}_s \widehat{\mathbf{v}}_m \left(\widehat{\mathbf{v}}_m^{\mathrm{T}} \widehat{\mathbf{A}}_s \widehat{\mathbf{v}}_m - \widehat{\mathbf{v}}_\ell^{\mathrm{T}} \widehat{\mathbf{A}}_s \widehat{\mathbf{v}}_\ell \right) = 0. \tag{C8}$$

Equation (C8) holds for all $\ell < m$. The JADE estimator $\widehat{\mathbf{V}}$ solves these $K(K-1)/2$ non redundant equations, together with the $K(K+1)/2$ orthogonality constraints:

$$\widehat{\mathbf{v}}_\ell^{\mathrm{T}} \widehat{\mathbf{v}}_m = \delta_{\ell m}, \text{ for all } \ell \leq m.$$

Identification and consistency. Let $\widetilde{\mathbf{V}} = (\widetilde{\mathbf{v}}_1, ..., \widetilde{\mathbf{v}}_K) \in \mathcal{O}_K$ be such that

$$\widetilde{\mathbf{V}} = \arg \min_{\mathbf{V} \in \mathcal{O}_K} \sum_{s=1}^S \mathrm{off}(\mathbf{V}^{\mathrm{T}} \mathbf{A}_s \mathbf{V}).$$

Then, as: $\min_{\mathbf{V} \in \mathcal{O}_K} \sum_{s=1}^S \mathrm{off}(\mathbf{V}^{\mathrm{T}} \mathbf{A}_s \mathbf{V}) = 0$ at the true value, it follows that $\widetilde{\mathbf{V}}^{\mathrm{T}} \mathbf{A}_s \widetilde{\mathbf{V}} = \widetilde{\mathbf{D}}_s$ is diagonal for all s. As for all $k \neq m$ there exists $s \in \{1...S\}$ such that $d_{sk} \neq d_{sm}$, one can apply Lemma 2 to show that $\widetilde{\mathbf{V}}$ is equal to the true $\mathbf{V}$, up to column sign and permutation. This shows the identification of $\mathbf{V}$. Consistency follows from classical arguments, as the parameter space $\mathcal{O}_K$ is compact.

Asymptotic distribution. A first-order Taylor expansion of (C8) around the true value $\mathbf{V}$ yields:

$$\sum_s^S \mathbf{v}_m^{\mathrm{T}} \widehat{\mathbf{A}}_s \mathbf{v}_k \left(\mathbf{v}_k^{\mathrm{T}} \widehat{\mathbf{A}}_s \mathbf{v}_k - \mathbf{v}_m^{\mathrm{T}} \widehat{\mathbf{A}}_s \mathbf{v}_m \right) + \sum_s^S \left(\mathbf{v}_k^{\mathrm{T}} \widehat{\mathbf{A}}_s \mathbf{v}_k - \mathbf{v}_m^{\mathrm{T}} \widehat{\mathbf{A}}_s \mathbf{v}_m \right) \left(\mathbf{v}_m^{\mathrm{T}} \widehat{\mathbf{A}}_s (\widehat{\mathbf{v}}_k - \mathbf{v}_k) + \mathbf{v}_k^{\mathrm{T}} \widehat{\mathbf{A}}_s (\widehat{\mathbf{v}}_m - \mathbf{v}_m) \right) + \sum_s^S \mathbf{v}_m^{\mathrm{T}} \widehat{\mathbf{A}}_s \mathbf{v}_k \left(\mathbf{v}_k^{\mathrm{T}} \widehat{\mathbf{A}}_s (\widehat{\mathbf{v}}_k - \mathbf{v}_k) - \mathbf{v}_m^{\mathrm{T}} \widehat{\mathbf{A}}_s (\widehat{\mathbf{v}}_m - \mathbf{v}_m) \right) = o_p \left(N^{-1/2} \right).$$

As $\operatorname*{plim}_{N \to \infty} \widehat{\mathbf{A}}_s = \mathbf{A}_s$ for all s, and as $\mathbf{v}_k^{\mathrm{T}} \mathbf{A}_s \mathbf{v}_m = 0$ for all $k \neq m$, this yields:

$$\sum_s^S (d_{sk} - d_{sm})\, \mathbf{v}_m^{\mathrm{T}} (\widehat{\mathbf{A}}_s - \mathbf{A}_s) \mathbf{v}_k + \sum_s^S (d_{sk} - d_{sm}) \left(\mathbf{v}_m^{\mathrm{T}} \mathbf{A}_s (\widehat{\mathbf{v}}_k - \mathbf{v}_k) + \mathbf{v}_k^{\mathrm{T}} \mathbf{A}_s (\widehat{\mathbf{v}}_m - \mathbf{v}_m) \right) = o_p \left(N^{-1/2} \right),$$

where $d_{sk} = \mathbf{v}_k^{\mathrm{T}} \mathbf{A}_s \mathbf{v}_k$ are the diagonal elements of $\mathbf{V}^{\mathrm{T}} \mathbf{A}_s \mathbf{V}$.

At this stage, it is convenient to define $\widehat{x}_{mk} \equiv \mathbf{v}_m^{\mathrm{T}} (\widehat{\mathbf{v}}_k - \mathbf{v}_k)$. As $\mathbf{v}_m^{\mathrm{T}} \mathbf{A}_s = d_{sm} \mathbf{v}_m^{\mathrm{T}}$, one has:

$$\sum_s^S (d_{sk} - d_{sm})\, \mathbf{v}_m^{\mathrm{T}} (\widehat{\mathbf{A}}_s - \mathbf{A}_s) \mathbf{v}_k + \sum_s^S (d_{sk} - d_{sm}) (d_{sm} \widehat{x}_{mk} + d_{sk} \widehat{x}_{km}) = o_p \left(N^{-1/2} \right).$$

Now, a Taylor expansion of the orthogonality constraints yields:

$$\widehat{x}_{mk} + \widehat{x}_{km} = \mathbf{v}_m^{\mathrm{T}} (\widehat{\mathbf{v}}_k - \mathbf{v}_k) + \mathbf{v}_k^{\mathrm{T}} (\widehat{\mathbf{v}}_m - \mathbf{v}_m) = 0, \text{ for all } m, k.$$

Thus we have:

$$\sum_s^S (d_{sk} - d_{sm})^2 \widehat{x}_{mk} = -\sum_s^S (d_{sk} - d_{sm})\, \mathbf{v}_m^{\mathrm{T}} (\widehat{\mathbf{A}}_s - \mathbf{A}_s) \mathbf{v}_k + o_p \left(N^{-1/2} \right). \tag{C9}$$

Let $\widehat{\mathbf{X}} = \mathbf{V}^{\mathrm{T}} \left(\widehat{\mathbf{V}} - \mathbf{V} \right)$. Then, equation (C9) is equivalently written, in matrix form, as:

$$\mathrm{vec}(\widehat{\mathbf{X}}) = -\mathbf{F} \left(\mathbf{I}_S \otimes \mathbf{V}^{\mathrm{T}} \otimes \mathbf{V}^{\mathrm{T}} \right) \left(\mathrm{vec}(\widehat{\mathbf{A}}) - \mathrm{vec}(\mathbf{A}) \right) + o_p \left(N^{-1/2} \right),$$

where $\mathbf{F}$, $\mathbf{A}$ and $\widehat{\mathbf{A}}$ have been defined in the text. Note that $\mathbf{F}$ is well defined provided that $\sum_s^S (d_{sk} - d_{sm})^2 \neq 0$ for all $k \neq m$.

Then, as:

$$\mathrm{vec}(\widehat{\mathbf{X}}) = \left(\mathbf{I}_K \otimes \mathbf{V}^{\mathrm{T}} \right) \left(\mathrm{vec}(\widehat{\mathbf{V}}) - \mathrm{vec}\,(\mathbf{V}) \right),$$

it follows that

$$N^{\frac{1}{2}} \left(\mathrm{vec}(\widehat{\mathbf{V}}) - \mathrm{vec}\,(\mathbf{V}) \right) = -\left(\mathbf{I}_K \otimes \mathbf{V} \right) \mathbf{F} \left(\mathbf{I}_S \otimes \mathbf{V}^{\mathrm{T}} \otimes \mathbf{V}^{\mathrm{T}} \right) N^{\frac{1}{2}} \left(\mathrm{vec}(\widehat{\mathbf{A}}) - \mathrm{vec}\,(\mathbf{A}) \right) + o_p\,(1).$$

This achieves to prove the theorem.

References

[1] ALBERA, L., A. FERREOL, P. COMON, and P. CHEVALIER (2004): "Blind Identification of Overcomplete Mixtures of Sources (BIOME)," *Lin. Alg. Appl.,* vol. 391, 1-30.

[2] ANDERSON, T.W. (1963): "Asymptotic Theory for Principal Component Analysis," *Ann. Math. Stat.,* 34, 122-148.

[3] ANDERSON, T.W., and H. RUBIN (1956): "Statistical Inference in Factor Analysis," in *Proceedings of the Third Symposium in Mathematical Statistics and Probability*, Vol. 5. University of California press.

[4] ANSELIN, L. (2003): "Spatial Externalities, Spatial Multipliers, and Spatial Econometrics," *International Regional Science Review,* 26(2), 153-166.

[5] ATTIAS, H. (1999): "Independent Factor Analysis," *Neural Computation*, ;11:803-851.

[6] BAI, J. (2003): "Inferential Theory for Factor Models of Large Dimensions," *Econometrica,* 71, 135-171.

[7] BAI, J., and S. NG (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica,* 70, 191-221.

[8] BECKMANN, C.F., and S.M. SMITH (2004): "Probabilistic Independent Component Analysis for Functional Magnetic Resonance Imaging, " *IEEE Trans. on Medical Imaging*, 23(2), 137-152.

[9] BLANCHARD, O.J., and D. QUAH (1989): "The Dynamic Effects of Aggregate Demand and Supply Disturbances, " *American Economic Review*, 79(4), 655-673.

[10] BONHOMME, S., and J.M. ROBIN (2008): "Generalized Nonparametric Deconvolution with an Application to Earnings Dynamics," *mimeo.*

[11] CARDOSO, J.-F. (1991): "Super-Symmetric Decomposition of the Fourth-Order Cumulant Tensor. Blind Identification of More Sources than Sensors," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP-91),* Toronto, 3109-3112.

[12] CARDOSO, J.-F. (1999): "High-order contrasts for independent Component Analysis," *Neural Computation,* 11, 157-192.

[13] CARDOSO, J.-F., and A. SOULOUMIAC (1993): "Blind Beamforming for Non-Gaussian Signals," *IEE-Proceedings-F,* 140, 362-370.

[14] CARDOSO, J.-F., and A. SOULOUMIAC (1996): "Jacobi Angles for Simultaneous Diagonalization," *SIAM J. Mat. An. Appl.,* 17, 161-164.

[15] CARDOSO, J.-F., and D.-T. PHAM (2004): "Optimization Issues in Noisy Gaussian ICA," *Proc ICA 2004,* Granada, Spain.

[16] CARNEIRO, P., K. HANSEN, and J.J. HECKMAN (2003): "Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice," *International Economic Review*, 44(2): 361-422.

[17] CHEN, A., and P.J. BICKEL (2005): "Consistent Independent Component Analysis and Prewhitening," *IEEE Trans. on Signal Processing,* 53(10), 3625-3632.

[18] CHUDIK, A., and M.H. PESARAN (2007): "Infinite Dimensional VARs and Factor Models," Working Paper, University of Cambridge.

[19] COMON, P. (1994): "Independent Component Analysis, a New Concept?," *Signal Processing*, 36(3), 287-314.

[20] COMON, P. (2004): "Blind Identification and Source Separation in 2×3 Under-determined Mixtures," *IEEE Trans. Signal Processing*, 11-22.

[21] CRAGG, J.G. (1997): "Using Higher Moments to Estimate the Simple Errors-in-Variables Model," *RAND Journal of Economics*, 28, S71-S91.

[22] DAGENAIS, M.G., and D.L. DAGENAIS (1997): "Higher Moment Estimators for Linear Regression Models with Errors in Variables," *Journal of Econometrics*, 76, 193-221.

[23] DAVIES, M. (2004): "Identifiability Issues in Noisy ICA," *IEEE Signal Processing Letters*, 11(5), 470-473.

[24] DE LATHAUWER, L., J. CASTAING, and J.F. CARDOSO (2007): "Fourth-Order Cumulant-Based Blind Identification of Underdetermined Mixtures," *IEEE Trans. Signal Processing*, 55(6), 2965-2972.

[25] DE LATHAUWER, L., B. DE MOOR, and J. VANDEWALLE (2004): "Computation of the Canonical Decomposition by Menas of a Simultaneous Generalized Schur Decomposition," *SIAM Journal of Matrix Analysis and Applications*, 26(2), 295-327.

[26] ERICKSON, T., and T. WHITED (2002): "Two-Step GMM Estimation of the Error-in-Variables Model Using High-Order Moments," *Econometric Theory*, 18, 776-799.

[27] ERIKSSON, J., and V. KOIVUNEN (2003): "Identifiability and separability of linear ICA models revisited," *4th International Symposium on ICA and Blind Signal Separation*, 23-27.

[28] FAMA, E., and K. FRENCH (1993): "Common Risk Factors in the Returns on Stocks and Bonds," *Journal of Financial Economics*, 33(1), 3-56.

[29] FLURY, B. (1986): "Asymptotic Theory for Common Principal Component Analysis," *Annals of Statistics*, 14, 418-430.

[30] FORNI, M., and L. REICHLIN (1998): "Let's Get Real: A Factor Analytical Approach to Disaggregated Business Cycle Dynamics," *Review of Economic Studies*, 65, 453-473.

[31] GEARY, R.C. (1942): "Inherent Relations Between Random Variables," *Proc. Royal Irish Academy*, 47, 63-76.

[32] HECKMAN, J.J., J. STIXRUD, and S. URZUA (2006): "The Effect of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior," *Journal of Labor Economics*, 24(39), 411-482.

[33] HYVARINEN, A., J. KARHUNEN, and E. OJA (2001): *Independent Component Analysis*, John Wiley & Sons, New York.

[34] HYVARINEN, A., and E. OJA (2001): "A Fast Fixed Point Algorithm for Independent Component Analysis," *Neural Computation*, 9(7), 1483-1492.

[35] IKEDA, S., and K. TOYAMA (2000): "Independent Component Analysis for Noisy Data–MEG Data Analysis," *Neural Networks*, 13(10), 1063-1074.

[36] JENNRICH, R.I., and N. TRENDAFILOV (2005): "Independent Component Analysis as a Rotation Method: A Very Different Solution to Thurstone's Box Problem," *British Journal of Mathematical and Statistical Psychology*, 58, 199-208.

[37] LEWBEL, A. (1997): "Constructing Instruments for Regressions with Measurement Error When No Additional Data are Available, with an Application to Patents and R&D," *Econometrica*, 65, 1201-1213.

[38] MENCIA, J., and E. SENTANA (2008): "Multivariate Location-Scale Mixtures of Normals and Mean-Variance-Skewness Portfolio Allocation," CEMFI Working Paper 0805.

[39] MOULINES, E., J.-F. CARDOSO, and E. GASSIAT (1997): "Maximum Likelihood for Blind Separation and Deconvolution of Noisy Signals Using Mixture Models," Proc. ICASSP'97 Munich, vol. 5, 3617-20.

[40] PAL, M. (1980): "Consistent Moment Estimators of Regression Coefficients in the Presence of Errors-in-Variables," *Journal of Econometrics*, 14, 349-364.

[41] PESARAN, M.H. (2006): "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure," *Econometrica*, 74 (4), 967-1012.

[42] REIERSOL, O. (1950): "Identifiability of a Linear Relation Between Variables which are Subject to Error," *Econometrica*, 9, 1-24.

[43] ROBIN, J.M., and R.J. SMITH (2000): "Tests of Rank," *Econometric Theory*, 16, 151-175.

[44] ROSS, S.A. (1976): "The Arbitrage Pricing Theory of Capital Asset Pricing," *Journal of Economic Theory*, 16, 341-360.

[45] SCHENNACH, S.M., HU, Y., and A. LEWBEL (2007): "Nonparametric Identification of the Classical Errors-in-Variables Model Without Side Information," *Cemmap* working paper 14/07.

[46] STEGEMAN, A., and A. MOOIJAART (2007): "Independent Factor Analysis by Least Squares," *mimeo*.