Using High-Order Moments to Estimate Linear Independent Factor Models

Stéphane Bonhomme
CEMFI, Madrid

Jean-Marc Robin
Université de Paris I, Panthéon-Sorbonne,
University College London and
Institute for Fiscal Studies

November 2005

Abstract

We study the identification and estimation of linear factor models under the assumptions that factors and errors are independent and that factors are not normally distributed. Higher-order moments are shown to yield full identification of the matrix of factor loadings if factor distributions are sufficiently skewed or kurtic. We develop simple algorithms to estimate the matrix of factor loadings from the second, third and fourth-order moments of the data. We run Monte Carlo simulations and apply our methodology to estimating the returns to education.

JEL codes: C13.

Keywords: Factor models, high-order moments, independent component analysis.

1 Introduction

Linear factor models are routinely used in social sciences. Spearman's (1904) "g" factor is one of the earliest applications in psychology. Principal component analysis (PCA) is a leading technique in sociology to construct social indices and to uncover hidden causes of individual actions (like self-esteem). Econometric applications include measurement error models, error component models for panel data, structural VAR models in macroeconomics, and multifactor asset pricing models in empirical finance. They are also useful in microeconometrics, even though empirical microeconomic models are often nonlinear. For example, Carneiro, Hansen and Heckman's (2003) Roy model of educational choice is a successful attempt to motivate their usefulness for estimating treatment effects and other policy parameters using microdata.[1]

Despite these empirical successes, it is usually thought that the interest of linear multifactor models for structural applications is severely hampered by a fundamental lack of identification. Suppose that a vector of L observed measurements, Y, be related to a vector of K unobserved factors, X, by a noisy linear relationship: $Y = \Lambda X + U$, where and Λ is a matrix of parameters (factor loadings) and U is a vector of errors. In ordinary Factor Analysis, the identification of factor loadings rests on covariance restrictions, and it well known that matrix Λ is identified only up to a multiplicative orthogonal matrix (Anderson and Rubin, 1956). Parametric restrictions, often in the form of exclusion restrictions, are usually added for identification. In VAR models, for example, the identification of structural shocks is achieved by assuming a particular triangular form for Λ. Carneiro, Hansen and Heckman (2003) assume at least two specific measurements for each factor. In Principal Component Analysis, an implicit choice of rotation is involved in the minimisation of the sum of squared residuals.

In this paper, we show that most of these exclusion restrictions are unnecessary if two key conditions are satisfied: First, factors and errors are *independent*, not just uncorrelated. Second, the third and/or fourth-order moments of the vector of observed measurements are informative

[1] Continuous instruments with large supports allow to identify the distribution of latent variables and a linear factor structure is used to model the effect of unobserved heterogeneity on latent variables. See Cunha *et. al* (2005) and Heckman and Navarro (2005) for other applications of this idea.

(which implies that factors are *non normal*). We derive sufficient conditions on Λ (such as Λ is full column rank) which guaranty that, if $K \leq L$, then Λ is identified (up to a multiplication of each column by ± 1 and column permutations) from second, third and fourth-order moments, and if $K < L$, then Λ is identified from second and third-order moments.

The importance of the assumptions of independence and nonnormality for the identification of factor models is well known in the measurement-error literature. Since the seminal contributions of Geary (1942) and Reiersol (1950) a long series of papers have proposed different ways of using third and fourth-order moments to correct estimators for measurement errors in the regressors.[2] The class of estimators introduced in this paper can be seen as a generalisation of this approach to multifactor structures.

In a different branch of statistics, signal processing, linear factor models are commonly used to separate the components of linear mixtures of signals. Since its introduction at the beginning of the 1990's, Independent Component Analysis (ICA) has rapidly become a leading technique.[3] In this vast literature, one of the most popular methods is Cardoso and Souloumiac's (1993) JADE algorithm.[4] This is a joint diagonalisation algorithm of a set of well chosen matrices of fourth-order cumulants of measurements. In the past ten years, the ICA problem has also become a very important topic in the neural networks literature and Hyvärinen's (1999) FastICA algorithm has become another very popular algorithm.[5]

One serious drawback of ICA, at least for econometric applications, is that it rules out errors. Neglecting noise can be a source of very important biases, as we shall later show. All existing extensions of ICA allowing for noise make parametric assumptions on the distributions of errors (usually Gaussian) and factors (usually Gaussian mixtures).[6] As far as we know, our paper is

[2]Relevant contributions include Madanski (1959), Pal (1980), Dagenais and Dagenais (1997), Cragg (1997), Lewbel (1997), and Erickson and Whitted (2002). Less directly related to our work are the papers of Spiegelman (1979) and Van Montfort *et al.* (1989), using more of the information contained in the characteristic function of measurements but the value at zero of its first few derivatives. Lastly, Lewbel (2004) and Doz and Renault (2005) use heteroskedasticity as a source of identification.

[3]The name Independent Component Analysis was first proposed par Comon (1994). See Hyvärinen *et al.*, 2001, and Cardoso, 1999, for a survey.

[4]For an application of ICA and JADE to multivariate financial time series, see Back and Weigend (1997).

[5]See Xu, 2003, for a survey of Bayesian learning applications to ICA.

[6]For example, Moulines *et al.* (1997), Attias (1999) uses a ML approach and the EM algorithm. Xu (2000, 2001) allows for non-Gaussian errors and uses Bayesian learning algorithms. Ikeda and Toyama (2000) adopt a two-stage approach similar to ours except that they assume Gaussian noise. They use PCA to estimate the error

the first paper, out of a long list of contributions, to propose a fully semiparametric procedure to consistently estimate factor loadings from data moments in a noisy linear factor model with error distributions of unknown form. Our *quasi-JADE* algorithm proceeds in two stages: First, we estimate the second, third and fourth-order error moments, that we use to "remove" the noise component from the second, third and fourth-order moments of the data ("whitening" stage). Then, we straightforwardly apply Cardoso and Souloumiac's joint diagonalisation algorithm to "whitened" data.

The outline of the paper is as follows. In Section 2, we study the semiparametric identification of factor loadings. Section 3 deals with estimation issues: We discuss the estimation of the number of common factors using Robin and Smith's (2000) rank test; we present Cardoso and Souloumiac's (1993) JADE algorithm; we study its asymptotic properties; and we develop the quasi-JADE algorithm. In Section 4, we investigate the finite-sample properties of quasi-JADE by means of Monte-Carlo simulations. In Section 5, we apply our methodology to estimate returns to schooling in France. Our method allows to identify two factors in the wage-education relationship. Interestingly, while the first factor has a positive effect on wages, the second factor is positively related to education, yet negatively to wages. This second factor provides some evidence that overspecialisation in schooling is negatively valued in the labour market. Lastly, Section 6 concludes.

2 Identification of linear independent factor models

Let $Y = (Y_1, ..., Y_L)^T$ be a vector of $L \geq 2$ zero-mean, real-valued random variables (measurements). Let $X = (X_1, ..., X_K)^T$ be a random vector of $K \geq 1$ zero-mean, real valued, non degenerate random variables (factors). Let $U = (U_1, ..., U_L)^T$ be a vector of L zero-mean, real-valued random variables (errors). Both factors and errors are unobserved.

Assumption A1 ***(Linearity)*** *There exists a $L \times K$ matrix of scalar parameters (factor loadings), Λ, such that $Y = \Lambda X + U$.*

covariance matrix before plugging it into JADE.

The difference between factors and errors is a matter of definition. A given covariate is called a factor if it enters at least two measurement equations (i.e. every column $\boldsymbol{\lambda}_k$, $k = 1, ..., K$, of Λ has at least two non-zero entries). Otherwise, it is called an error.

In ordinary Factor Analysis (FA), factors and errors are uncorrelated and identification rests on the following covariance restrictions:

$$\Sigma_Y = \Lambda \Sigma_X \Lambda^T + \Sigma_U, \tag{1}$$

where Σ_Z denotes the variance-covariance matrix of any random vector Z. Obviously parameters Λ, Σ_X and Σ_U are not identified from second-order restrictions (see Anderson and Rubin, 1956). First, restrictions are needed on the correlations between errors and it is usually assumed that Σ_U is diagonal. Second, Σ_X is not separately identified from Λ. If (Λ, Σ_X) satisfies (1), then so does $(\Lambda\Omega, I_K)$, where $\Omega\Omega^T = \Sigma_X$. The variance-covariance matrix of X is therefore normalised to the identity matrix I_K. Thirdly, even if $\Sigma_X = I_K$ and Σ_U is diagonal, Λ is identified only up to an orthogonal matrix; that is, if Λ satisfies the covariance restrictions, then so does ΛP, for any orthonormal matrix P. Principal Component Analysis is the least-squares version of FA.[7]

In this paper, we maintain the assumptions that $\Sigma_X = I_K$ and Σ_U is diagonal and we intensify the absence of correlations between factors, between errors and between factors and errors by making them independent. Moreover we assume that factors are non Gaussian with finite third and fourth-order moments.

Assumption A2 ***(Normalization)*** *Factors have unit variances.*

Assumption A3 ***(Independence)*** *All factor and error variables are mutually independent.*

[7]Given a sample $\mathbf{Y} = (Y_1, ..., Y_N)$ of observations, Principal Component Analysis estimates both Λ and factor realisations $\mathbf{X} = (X_1, ..., X_N)$ by non linear least squares under the normalisation $\frac{1}{N}\mathbf{X}\mathbf{X}^T = I_K$. Principal components $\widehat{\mathbf{X}}$ are the first K eigenvectors of the $N \times N$ matrix $\mathbf{Y}^T\mathbf{Y}$ (corresponding to the K largest eigenvalues), and factor loadings are estimated by regressing $\mathbf{Y}$ on $\widehat{\mathbf{X}}$ by OLS. Common factors ΛX_n, $n = 1, ..., N$, are identified if matrix $\mathrm{Var}(Y)$ has no multiple eigenvalue but identifying factors X_n from factor loadings requires the arbitrary choice of a rotation P, even under the normalisation $\mathrm{Var}(X) = I_K$. The asymptotic theory of principal components usually assumes a fixed number of measurements L but a large sample size N (see Anderson, 1984, and Lawless and Maxwell, 1971). In a recent paper, Bai (2003) studies the case where both L and N tend to infinity.

Assumption A4 ***(Non-gaussianity)*** *Factor variables X_k, $k = 1, ..., K$, are non Gaussian with finite third and fourth-order moments.*

In addition, in order to prove the semi-parametric identification results below, we shall require characteristic functions and cumulant generating functions to exist and to be smooth on all $\mathbb{R}$ and not only locally around the origin, as implied by Assumption A4. This assumption is yet not necessary for the parametric estimation procedures that we shall later develop.

Assumption A5 ***(Characteristic functions)*** *The characteristic functions of factors and errors are of class $\mathcal{C}^2$ on $\mathbb{R}$, and are nonvanishing almost everywhere.*

We shall say that a representation (Λ, X, U) is regular if it satisfies all previously listed assumptions.

2.1 Definitions

For all K, let us define the set of sign-permutation matrices as the set $\mathcal{S}_K$ of all products DP, where D is a diagonal matrix with diagonal components equal to 1 or -1 and P is a permutation matrix. For given values of L and K, let (Λ, X, U) be a regular representation. Clearly, for all $S \in \mathcal{S}_K$, $\left(\Lambda S, S^T X, U\right)$ is another regular representation. Hence, identification has to be defined modulo the set $\mathcal{S}_K$.

Note that the group $\mathcal{S}_K$ is a finite subgroup of the infinite orthogonal group $\mathcal{O}_K$, up to which identification is defined in ordinary or orthogonal Factor Analysis. The quotient group $\mathcal{O}_K/\mathcal{S}_K$ is thus also infinite. Proving identification results modulo $\mathcal{S}_K$, instead of modulo $\mathcal{O}_K$, will result in a considerable reduction of the model's indeterminacy.

We define semi-parametric identification as follows.

Definition 1 ***(Semiparametric identification)*** *A regular representation (Λ, X, U) is said identifiable if for every other regular representation $(\widetilde{\Lambda}, \widetilde{X}, \widetilde{U})$ there exists a matrix S in $\mathcal{S}_K$ such that: $\widetilde{\Lambda} = \Lambda S$, $\widetilde{X} \overset{d}{=} S^T X$, and $\widetilde{U} \overset{d}{=} U$, where $\overset{d}{=}$ means "equal in distribution."*

Semiparametric identification draws information on the finite-dimensional parameter Λ and the infinite-dimensional parameters that are the distributions of X and U from the whole dis-

tribution of observed measurements. For practical reasons, it is useful to understand how much of the model's structure can be identified from a finite set of parameters of this distribution. We thus also define parametric identification as follows.

Definition 2 ***(Parametric identification)*** *A regular representation* (Λ, X, U) *is said to be parametrically identified if their exists a finite vector of moments* $\mathbb{M}(Z)$*, defined for a vector of r.v.* Z*, such that, for every other regular representation* $(\widetilde{\Lambda}, \widetilde{X}, \widetilde{U})$*, moment equality:* $\mathbb{M}(\Lambda X + U) = \mathbb{M}(\widetilde{\Lambda}\widetilde{X} + \widetilde{U})$*, implies that there exists a sign-permutation matrix* $S \in \mathcal{S}_K$ *such that* $\widetilde{\Lambda} = \Lambda S$*,* $\mathbb{M}(\widetilde{X}) = \mathbb{M}\left(S^T X\right)$*, and* $\mathbb{M}(\widetilde{U}) = \mathbb{M}(U)$*.*

Unless otherwise specified, we shall simply say that the factor model is parametrically identified if it is identified from second, third and fourth-order moments of observed measurements.

2.2 Identifying restrictions

In this subsection, we develop some implications of the regularity assumptions in terms of cumulant generating functions and their derivatives.

Cumulant generating function. Denote the cumulant generating functions (the log of characteristic functions) of Y, X_k and U_ℓ as κ_Y, κ_{X_k} and κ_{U_ℓ}. The independence assumptions and the linear factor structure imply that, for almost all $t = (t_1, ..., t_L)^T \in \mathbb{R}^L$,

$$\kappa_Y(t) \equiv \ln\left[\mathbb{E}\exp\left(\sqrt{-1}\cdot t^T Y\right)\right] = \sum_{k=1}^{K} \kappa_{X_k}\left(\boldsymbol{\lambda}_k^T t\right) + \sum_{\ell=1}^{L} \kappa_{U_\ell}(t_\ell), \tag{2}$$

where ln denotes the principal branch of the logarithm.

Then, define the following sets of multi-indices:

$$\begin{aligned}
\overline{\Delta}_{L,p} &= \left\{\alpha = (\alpha_1, ..., \alpha_L) \in \{0, ..., p\}^L : |\alpha| = \alpha_1 + ... + \alpha_L = p\right\}, \\
\Delta_{L,p} &= \left\{\alpha = (\alpha_1, ..., \alpha_L) \in \{0, ..., p-1\}^L : |\alpha| = \alpha_1 + ... + \alpha_L = p\right\}.
\end{aligned}$$

Let $\#\overline{\Delta}_{L,p}$ (resp. $\#\Delta_{L,p}$) be the number of elements in $\overline{\Delta}_{L,p}$ (resp. $\Delta_{L,p}$). For $p = 2$: $\#\overline{\Delta}_{L,2} = \frac{L(L+1)}{2}$ and $\#\Delta_{L,2} = \frac{L(L-1)}{2}$.[8]

[8]Multi-indices are convenient ways to select p components of a vector of size L with repetition, *via* the

Let $\alpha = (\alpha_1, ..., \alpha_L) \in \overline{\Delta}_{L,p}$. For any vector $x = (x_1, ..., x_L) \in \mathbb{R}^L$, define the monomial $x^\alpha = x_1^{\alpha_1} \cdots x_L^{\alpha_L}$. Then, assuming that derivatives exist, we have

$$\kappa_Y^{(\alpha)}(t) \equiv \partial_\alpha \kappa_Y(t) \equiv \frac{\partial^{|\alpha|} \kappa_Y(t)}{\partial t_1^{\alpha_1} \cdots \partial t_L^{\alpha_L}} = \sum_{k=1}^{K} \boldsymbol{\lambda}_k^\alpha \kappa_{X_k}^{(p)} \left(\boldsymbol{\lambda}_k^T t\right) + \sum_{\ell=1}^{L} \delta_{\alpha_\ell, p} \kappa_{U_\ell}^{(p)} (t_\ell), \tag{3}$$

where $\kappa_{X_k}^{(p)}$ and $\kappa_{U_\ell}^{(p)}$ are the pth derivative of κ_{X_k} and κ_{U_ℓ}, and δ_{ij} is the Kronecker delta (= 1 if $i = j$ and = 0 if $i \neq j$).

Cumulants. For any multi-index α, one defines a multivariate cumulant as

$$\kappa_\alpha (Y) = \frac{\kappa_Y^{(\alpha)}(0)}{\left(\sqrt{-1}\right)^{|p|}}.$$

Let $\mathbf{i}_\ell$, $\ell \in \{0, ..., L\}$, be the ℓth column of the $L \times L$ identity matrix. For any $p$tuple $(\ell_1, ..., \ell_p) \in \{0, ..., L\}^p$, we denote as $\text{Cum}\left(Y_{\ell_1}, ..., Y_{\ell_p}\right) \equiv \kappa_{\mathbf{i}_{\ell_1} + ... + \mathbf{i}_{\ell_p}} (Y)$ the multivariate cumulant of $\left(Y_{\ell_1}, ..., Y_{\ell_p}\right)$.

The second-order cumulants of zero-mean random variables are equal to their covariances:

$$\text{Cum}\left(Y_{\ell_1}, Y_{\ell_2}\right) = \mathbb{E}(Y_{\ell_1} Y_{\ell_2}), \tag{4}$$

Third-order cumulants are:

$$\text{Cum}\left(Y_{\ell_1}, Y_{\ell_2}, Y_{\ell_3}\right) = \mathbb{E}(Y_{\ell_1} Y_{\ell_2} Y_{\ell_3}). \tag{5}$$

And fourth-order cumulants:

$$\begin{aligned} \text{Cum}\left(Y_{\ell_1}, Y_{\ell_2}, Y_{\ell_3}, Y_{\ell_4}\right) &= \mathbb{E}(Y_{\ell_1} Y_{\ell_2} Y_{\ell_3} Y_{\ell_4}) - \mathbb{E}(Y_{\ell_1} Y_{\ell_2}) \mathbb{E}(Y_{\ell_3} Y_{\ell_4}) \\ &\quad - \mathbb{E}(Y_{\ell_1} Y_{\ell_3}) \mathbb{E}(Y_{\ell_2} Y_{\ell_4}) - \mathbb{E}(Y_{\ell_2} Y_{\ell_3}) \mathbb{E}(Y_{\ell_1} Y_{\ell_4}). \end{aligned} \tag{6}$$

Taking $t = 0$ in equation (3) yields a set of restrictions on cumulants of factors and mea-

---

following bijection

$$\begin{array}{rccl} \Psi_{L,p}: & \overline{\Delta}_{L,p} & \longrightarrow & \{1, ..., L\}^p \\ & \alpha & \longmapsto & (\ell_1, ..., \ell_p) = \Psi_{L,p}(\alpha) \end{array}$$

where $(\ell_1, ..., \ell_p)$ is such that $\ell_1 \leq ... \leq \ell_p$ and $\mathbf{i}_{L,\ell_1} + ... + \mathbf{i}_{L,\ell_p} = \alpha$, vector $\mathbf{i}_{L,\ell}$ denoting the ℓth column of the identity matrix of dimension L. For example, $\alpha = (2, 1, 1) \in \overline{\Delta}_{3,4}$ corresponds to variable indices $(\ell_1, \ell_2, \ell_3, \ell_4) = (1, 1, 2, 3)$, $\alpha = (0, 2, 2)$ to $(\ell_1, \ell_2, \ell_3, \ell_4) = (2, 2, 3, 3)$, etc. To simplify the notation, we shall also denote as $\overline{\Delta}_{L,p}$ and $\Delta_{L,p}$ their image by $\Psi_{L,p}$.

surements:

$$\begin{aligned}\text{Cum}\left(Y_{\ell_1},...,Y_{\ell_p}\right) &= \sum_{k=1}^{K}\left(\prod_{i=1}^{p}\lambda_{\ell_i,k}\right)\kappa_p\left(X_k\right)+\text{Cum}\left(U_{\ell_1},...,U_{\ell_p}\right)\\ &= \sum_{k=1}^{K}\left(\prod_{i=1}^{p}\lambda_{\ell_i,k}\right)\kappa_p\left(X_k\right)+\delta_{\ell_1,...,\ell_p}\kappa_p\left(U_{\ell_1}\right), \qquad (7)\end{aligned}$$

where $\delta_{\ell_1,...,\ell_p}=1$ if $\ell_1=...=\ell_p$ and $=0$ otherwise, and $\kappa_p\left(Z\right)$ denotes the pth cumulant of a univariate random variable Z. If Z has zero mean,

$$\begin{aligned}\kappa_2\left(Z\right) &= \text{Cum}\left(Z,Z\right)=\text{Var}\left(Z\right)=\mathbb{E}Z^2,\\ \kappa_3\left(Z\right) &= \text{Cum}\left(Z,Z,Z\right)=\mathbb{E}Z^3,\\ \kappa_4\left(Z\right) &= \text{Cum}\left(Z,Z,Z,Z\right)=\mathbb{E}\left(Z^4\right)-3\mathbb{E}\left(Z^2\right)^2.\end{aligned}$$

Moment restrictions. It will prove convenient to write moment restrictions in matrix form, provided that the corresponding moments exist. Using (7) with $p=2$, second-order restrictions are equivalently rewritten as

$$\Sigma_Y=\Lambda\Lambda^T+\Sigma_U, \qquad (8)$$

where Σ_Y and Σ_U denote the variance-covariances matrices of Y and U.

Next, define the following matrices of third-order cumulants

$$\Gamma_Y\left(\ell\right)=\left[\text{Cum}\left(Y_i,Y_\ell,Y_j\right);\left(i,j\right)\in\left\{1,...,L\right\}^2\right]\in\mathbb{R}^{L\times L},\quad \ell\in\left\{1...L\right\}. \qquad (9)$$

Third-order restrictions ($p=3$) imply that

$$\Gamma_Y\left(\ell\right)=\Lambda D_3\,\text{diag}\left(\Lambda_\ell\right)\Lambda^T+\kappa_3\left(U_\ell\right)\text{Sp}_{L,\ell}, \qquad (10)$$

where $\Lambda_\ell^T\in\mathbb{R}^{K\times 1}$ is the ℓth row of Λ, D_3 is the diagonal matrix with $\kappa_3\left(X_k\right)$ in the kth entry of the diagonal, and $\text{Sp}_{L,\ell}$ is the $L\times L$ sparse matrix with only one 1 in position (ℓ,ℓ).

Let us also define the following matrices of fourth-order cumulants

$$\Omega_Y\left(\ell,m\right)=\left[\text{Cum}\left(Y_i,Y_\ell,Y_m,Y_j\right);\left(i,j\right)\in\left\{1,...,L\right\}^2\right]\in\mathbb{R}^{L\times L},\quad \left(\ell,m\right)\in\overline{\Delta}_{L,2}. \qquad (11)$$

Fourth-order restrictions ($p=4$) imply that

$$\Omega_Y\left(\ell,m\right)=\Lambda D_4\,\text{diag}\left(\Lambda_\ell\odot\Lambda_m\right)\Lambda^T+\delta_{\ell m}\kappa_4\left(U_\ell\right)\text{Sp}_{L,\ell}, \qquad (12)$$

where D_4 is the diagonal matrix with $\kappa_4(X_k)$ in the kth entry of the diagonal, and $\odot$ is the Hadamard (element by element) matrix product.

2.3 Semiparametric identification

We now use restrictions (3) to derive necessary and sufficient conditions for the semi-parametric identification of independent factor models. The next theorem, proved in the mathematical appendix, gives sufficient conditions for identification.

Theorem 1 ***(Sufficient conditions for semiparametric identification)*** *Let (Λ, X, U) be a regular representation. Let $(\widetilde{\Lambda}, \widetilde{X}, \widetilde{U})$ be an alternative regular representation. The following two propositions hold true:*

1. *Every column of Λ is a scalar multiple of a column of $\widetilde{\Lambda}$.*
2. *If the $\frac{L(L-1)}{2} \times K$ matrix $Q(\Lambda) = [\lambda_{\ell 1}\lambda_{m1}, ..., \lambda_{\ell K}\lambda_{mK}; (\ell, m) \in \Delta_{L,2}]$, where rows are stacked by increasing order of (ℓ, m), $\ell < m$, is full column rank, then (Λ, X, U) is identified.*

If factor variables are not normally distributed, then the matrix of factor loadings is identified whatever the number of factors. This result is well-known in the ICA literature, at least since Comon (1994). Moreover, it suffices that $\text{rank}(Q(\Lambda)) = K$ for the distributions of factors and errors to be identified. A model with $\frac{L(L-1)}{2}$ factors is thus potentially identifiable when factors and errors are assumed independent instead of uncorrelated.

Theorem 1 is a straightforward generalisation of Eriksson and Koivunen's (2003) identification result for nonnoisy ICA. Proposition (i) of Theorem 1 follows from factor nongaussianity by a straightforward application of a result due to Kagan, Linnik and Rao (1973) that is stated in the mathematical appendix. Proposition (ii) of Theorem 1 easily follows from proposition (i).

We now show that the rank condition in proposition (ii) is generically necessary, that is: for a class of distribution functions dense in the set of continuous distribution functions. As far as

we know, this is a new result. Let us first define the class of distributions divisible by a normal distribution.

Definition 3 ***(Distribution divisible by a normal)*** *Let X be a continuous random variable with density f and characteristic function φ. The distribution of X is divisible by a normal distribution if there exists $\sigma^2 > 0$ such that $\widetilde{\varphi}(t) = \varphi(t)\exp\left(\frac{\sigma^2 t^2}{2}\right)$ is the characteristic function of a random variable $\widetilde{X}$.*

The distribution of a variable X is divisible by a normal if and only if $X \overset{d}{=} \widetilde{X} + \mathcal{N}\left(0, \sigma^2\right)$, where $\mathcal{N}\left(0, \sigma^2\right)$ is a normal r.v. with mean 0 and variance σ^2. The set of distributions divisible by a normal is dense in the set of continuous distribution functions.[9]

For a representation (Λ, X, U) to be identifiable, the next theorem shows that either $Q(\Lambda)$ is full-column-rank or it is not, but then at least some of the factor and error variables must not be divisible by a normal distribution.

Theorem 2 ***(Necessary condition for semiparametric identification)*** *Let (Λ, X, U) be a representation. Suppose that $Q(\Lambda)$ is not full-column-rank and that the distributions of factors and errors are divisible by normal distributions. Then (Λ, X, U) is not identifiable.*

We refer the reader to the mathematical appendix for a proof of Theorem 2. We show that under the assumptions of Theorem 2, for any representation (Λ, X, U) such that $Q(\Lambda)$ is not full-column-rank and the distributions of factors X and errors U are divisible by normal distributions, then one can construct another representation $(\widetilde{\Lambda}, \widetilde{X}, \widetilde{U})$ that is not equal to (Λ, X, U) up to a sign-permutation matrix and that still verifies the equality $\Lambda X + U \overset{d}{=} \widetilde{\Lambda}\widetilde{X} + \widetilde{U}$.

2.4 Parametric identification of factor loadings in the noise-free case ($U = 0$)

We here derive parametric identification results based on the first four moments of the data. The identification proofs are constructive, and will be used for estimation in the next section.

[9] Let X be a continuous random variable. Let $X_n = X + \mathcal{N}(0, \sigma^2)$. Then $X_n \xrightarrow{d} X$ when $\sigma^2 \to 0$.

We first consider the case of factor models without errors. In this case, second, third and fourth-order restrictions (8), (10), (12) imply that matrix Λ satisfies simultaneously

$$\Sigma_Y = \Lambda\Lambda^T, \tag{13}$$

$$\Gamma_Y(\ell) = \Lambda D_3 \operatorname{diag}(\Lambda_\ell)\Lambda^T, \quad \ell \in \{1...L\}, \tag{14}$$

$$\Omega_Y(\ell, m) = \Lambda D_4 \operatorname{diag}(\Lambda_\ell \odot \Lambda_m)\Lambda^T, \quad (\ell, m) \in \overline{\Delta}_{L,2}. \tag{15}$$

Left and right-multiplying (13), (14) and (15) by $\Sigma_Y^{-1/2}$ and $\Sigma_Y^{-T/2}$, respectively, where $\Sigma_Y^{-1/2}\Sigma_Y\Sigma_Y^{-T/2} = I_K$, one obtains:

$$\begin{aligned}
\Sigma_Y^{-1/2}\Gamma_Y(\ell)\Sigma_Y^{-T/2} &= VD_3 \operatorname{diag}(\Lambda_\ell)V^T, \quad \ell \in \{1...L\}, \\
\Sigma_Y^{-1/2}\Omega_Y(\ell, m)\Sigma_Y^{-T/2} &= VD_4 \operatorname{diag}(\Lambda_\ell \odot \Lambda_m)V^T, \quad (\ell, m) \in \overline{\Delta}_{L,2}.
\end{aligned}$$

where $V = \Sigma_Y^{-1/2}\Lambda$ is orthogonal; that is: $VV^T = I_K$. Therefore, V solves a joint diagonalization problem. Theorem 3 below gives conditions for the solution to this joint diagonalization problem to be unique.

Theorem 3 ***(Parametric identification in the noise-free case)*** *Assume (i) $U = 0$, (ii) $K \leq L$ and (iii) Λ has rank K.*

If (iv) at most one factor variable has zero kurtosis excess, then factor loadings are identified from second and fourth-order moment restrictions (13) and (15).

If (iv') at most one factor variable has zero skewness, then factor loadings are identified from second and third-order moment restrictions (13) and (14).

If (iv") 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 (13), (14) and (15).

The proof is in the mathematical appendix. Theorem 3 shows that high order moments are a source of identification in noise-free factor models. This insight has been widely used in the ICA literature. For instance, Cardoso and Souloumiac (1993) use restrictions (13) and (15) as the basis of their JADE algorithm. The ICA literature does not want to rely on factor

dissymmetry, and thus neglects third-order information. However, there is no strong argument in favour of discarding third-order moments of the data in econometrics. It is yet true that the variables of interest are often transformed to make them as much Gaussian as possible. For instance, by taking the logarithm of income, one obtains a distribution which is close to being normal, at least as far as the skewness and kurtosis are concerned. However, there can still be enough non-normality in the multivariate distribution of the data for factor loadings and factor moments to be well identified. The application in Section 5 will provide an illustration of this remark.

2.5 Parametric identification of error moments

In the "noisy" case ($U \neq 0$), the previous identification results apply, provided that the first moments of error variables are identified. We here give conditions under which these moments are identified. Two cases are distinguished, depending on whether all fourth-order cumulants of factors are zero or not.

2.5.1 First case: all factor distributions are kurtotic

Let Ω_Y be the $\frac{L(L+1)}{2} \times \frac{L(L-1)}{2}$ matrix of *all* fourth-order cumulants of the data, defined by

$$\Omega_Y = \left[\mathrm{Cum}\left(Y_i, Y_j, Y_\ell, Y_m\right); (i,j) \in \overline{\Delta}_{L,2}, (\ell,m) \in \Delta_{L,2}\right] \in \mathbb{R}^{\frac{L(L+1)}{2} \times \frac{L(L-1)}{2}}. \tag{16}$$

The rows of Ω_Y are indexed by $(i,j) \in \overline{\Delta}_{L,2}$ and the columns are indexed by $(\ell,m) \in \Delta_{L,2}$, i.e. $(i,j) \in \{1,...,L\}^2$, $i \leq j$, and $(\ell,m) \in \{1,...,L\}^2$, $\ell < m$. The factor structure implies that

$$\Omega_Y = \overline{Q} D_4 Q^T, \tag{17}$$

where

$$Q \equiv Q(\Lambda) = [\lambda_{\ell k}\lambda_{mk}; (\ell,m) \in \Delta_{L,2}, k \in \{1,...,K\}] \in \mathbb{R}^{\frac{L(L-1)}{2} \times K}, \tag{18}$$

$$\overline{Q} \equiv \overline{Q}(\Lambda) = [\lambda_{\ell k}\lambda_{mk}; (\ell,m) \in \Delta_{L,2}, k \in \{1,...,K\}] \in \mathbb{R}^{\frac{L(L+1)}{2} \times K}. \tag{19}$$

We first show that, under the assumption that all factors have kurtosis excess, it suffices that Q be full column rank for the first four error moments to be identified from the first four moments of the data.

Lemma 1 *Assume that (i)* $K \leq \frac{L(L-1)}{2}$*, (ii)* Q *has rank* K *and (iii) factor variables have non zero kurtosis excess. Then the following propositions hold true.*

1. *Matrix* Ω_Y *has rank* K*.*

2. *Let* $\overline{C} \in \mathbb{R}^{\frac{L(L+1)}{2} \times \left(\frac{L(L+1)}{2} - K\right)}$ *be a basis of the null space of* Ω_Y^T*; that is: the columns of* $\overline{C}$ *are linearly independent and* $\Omega_Y^T \overline{C} = 0$*. The first four moments of* U_ℓ*,* $\ell \in \{1, ..., L\}$*, satisfy the linear restrictions:*

$$\overline{C}^T \operatorname{vech}(\Sigma_Y) = \sum_{\ell=1}^{L} \operatorname{Var}(U_\ell) \overline{C}_{(\ell,\ell)}, \tag{20}$$

$$\overline{C}^T \operatorname{vech}(\Gamma_Y(\ell)) = \kappa_3(U_\ell) \overline{C}_{(\ell,\ell)}, \tag{21}$$

$$\overline{C}^T \operatorname{vech}(\Omega_Y(\ell,\ell)) = \kappa_4(U_\ell) \overline{C}_{(\ell,\ell)}, \tag{22}$$

   where $\overline{C}_{(\ell,\ell)}^T$ *denotes the* (ℓ,ℓ)*th row of* $\overline{C}$*, when the* $\frac{L(L+1)}{2}$ *rows of* $\overline{C}$ *are indexed by* $\overline{\Delta}_{L,2}$*, and where* vech *is the linear matrix operator stacking all* $\frac{L(L+1)}{2}$ *non redundant elements of a symmetric matrix.*[10]

3. *Matrix* $\left[\overline{C}_{(1,1)}, ..., \overline{C}_{(L,L)}\right]$ *is full rank and* $\operatorname{Var}(U_\ell)$*,* $\kappa_3(U_\ell)$ *and* $\kappa_4(U_\ell)$ *are uniquely defined by identification restrictions (20), (21) and (22).*

The proof is in Section A.4 of the mathematical appendix. The following theorem then follows straightforwardly.

Theorem 4 ***(Sufficient conditions for parametric identification when*** $K \leq L$***)*** *Assume that (i)* $K \leq \min\left\{L, \frac{L(L-1)}{2}\right\}$*, (ii)* Λ *is full column rank, (iii)* Q *has rank* K*, and (iv) factor variables have non zero kurtosis excess. Then, factor loadings are identified from second and fourth-order moments.*

Theorem 3 shows that the maximal number of factors for which Λ can be identified (up to column sign and permutation) is $K = 1$ if $L = 2$, and $K = L$ if $L \geq 3$.

[10] Let $A = [a_{ij}]$ be a $L \times L$ matrix. Then $\operatorname{vech}(A) = [a_{ij}; i \leq j] \in \mathbb{R}^{\frac{L(L+1)}{2} \times 1}$, ordering couples (i,j) by increasing order.

2.5.2 Second case: all factor distributions are either skewed or kurtotic

We now consider the problem of identifying factor loadings, in the "noisy" factor model, when some or all factor distributions have zero kurtosis excess.

Let

$$\Omega_Y(j) = \left[\mathrm{Cum}\left(Y_i, Y_j, Y_\ell, Y_m\right); i \in \{1, ..., L\}, (\ell, m) \in \Delta_{L,2}\right] \in \mathbb{R}^{L \times \frac{L(L-1)}{2}}. \tag{23}$$

The rows of Ω_Y are indexed by $i \in \{1, ..., L\}$ and the columns are indexed by $(\ell, m) \in \Delta_{L,2}$, i.e. $\ell < m$. The factor structure implies that

$$\Omega_Y(j) = \Lambda \,\mathrm{diag}\left(\Lambda_j\right) D_4 Q^T. \tag{24}$$

Let also Γ_Y be the $L \times \frac{L(L-1)}{2}$ matrix of third-order cumulants of the data defined by

$$\Gamma_Y = \left[\mathrm{Cum}\left(Y_i, Y_\ell, Y_m\right); i \in \{1, ..., L\}, (\ell, m) \in \Delta_{L,2}\right] \in \mathbb{R}^{L \times \frac{L(L-1)}{2}}, \tag{25}$$

The rows of Γ_Y are indexed by $i \in \{1, ..., L\}$ and the columns are indexed by $(\ell, m) \in \Delta_{L,2}$, i.e. $(\ell, m) \in \{1, ..., L\}^2$, $\ell < m$. The factor structure implies that

$$\Gamma_Y = \Lambda D_3 Q^T. \tag{26}$$

Lastly, let Ξ_Y be the $L \times \frac{L(L-1)(L+1)}{2}$ matrix of *all* third and fourth-order cumulants of the data, obtained by stacking matrices Γ_Y, $\Omega_Y(1)$,..., $\Omega_Y(L)$ columnwise:

$$\Xi_Y = \left[\Gamma_Y, \Omega_Y(1), ..., \Omega_Y(L)\right]. \tag{27}$$

We first establish a set of linear restrictions on error moments.

Lemma 2 *Assume that (i)* $K \leq \min\left\{L, \frac{L(L-1)}{2}\right\}$*, (ii)* Λ *and* Q *are full column rank* K *and (iii) every factor distribution is either skewed or kurtotic. Then the following propositions hold true.*

1. Ξ_Y *has rank* K*.*

2. *Let $C \in \mathbb{R}^{L\times(L-K)}$ be a basis of the null space of Ξ_Y^T; that is: the columns of C are linearly independent, and $\Xi_Y^T C = 0$. Let C_ℓ^T denote the ℓth row of C. The second, third and fourth-order moments of U_ℓ, for all $\ell \in \{1, ..., L\}$, satisfy the linear restrictions:*

$$C^T \begin{pmatrix} \mathbb{E}(Y_1 Y_\ell) \\ \vdots \\ \mathbb{E}(Y_L Y_\ell) \end{pmatrix} = \mathrm{Var}(U_\ell) C_\ell, \tag{28}$$

$$C^T \begin{pmatrix} \mathbb{E}(Y_1 Y_\ell^2) \\ \vdots \\ \mathbb{E}(Y_L Y_\ell^2) \end{pmatrix} = \kappa_3(U_\ell) C_\ell. \tag{29}$$

and

$$C^T \begin{pmatrix} \mathbb{E}(Y_1 Y_\ell^3) - 3\mathbb{E}(Y_1 Y_\ell)\mathbb{E}(Y_\ell^2) \\ \vdots \\ \mathbb{E}(Y_L Y_\ell^3) - 3\mathbb{E}(Y_L Y_\ell)\mathbb{E}(Y_\ell^2) \end{pmatrix} = \kappa_4(U_\ell) C_\ell. \tag{30}$$

Lemma 2 is not sufficient to identify error moments if $K = L$, as in this case matrix C is zero. We thus require additional assumptions on Λ.

Lemma 3 *Assume, in addition to the conditions of Lemma 2, that (i) $K \leq L-1$, and (ii) every submatrix of Λ made of a selection of $L-1$ rows has rank K. Then, no column of C is nil ($C_\ell \neq 0$, $\forall \ell$) and $\mathrm{Var}(U_\ell)$, $\kappa_3(U_\ell)$ and $\kappa_4(U_\ell)$ are identified.*

The proofs are in Section A.5 of the mathematical appendix. The following theorem then follows immediately.

Theorem 5 ***(Sufficient conditions for parametric identification when** $K \leq L-1$**)*** *Assume that (i) $K \leq L-1$, (ii) every submatrix of Λ made of a selection of $L-1$ rows has rank K, (iii) matrix Q has rank K, (iv) every factor distribution is either skewed or kurtotic. Then, factor loadings are parametrically identified from second, third and fourth-order moments.*

As a special case, if all factors are skewed then factor loadings are parametrically identified from second and third-order moments.

Corollary 6 ***(Sufficient conditions for parametric identification from second and third-order moments when** $K \leq L-1$**)*** *Assume that (i) $K \leq L-1$, (ii) every submatrix of Λ made of a selection of $L-1$ rows has rank K, (iii) matrix Q has rank K, and (iv) all*

factor distributions are skewed. Then, factor loadings are parametrically identified from second and third-order moments.

For example, consider the case of $L = 2$ and $K = 1$ and factor X_1 has a non symmetric distribution:

$$\begin{cases} Y_1 = \lambda_{11} X_1 + U_1, \\ Y_2 = \lambda_{21} X_1 + U_2, \end{cases}$$

and $\mathbb{E}\left(X_1^3\right) \neq 0$. One easily finds:

$$\begin{aligned} \lambda_{11} &= \sqrt{\mathbb{E}(Y_1 Y_2) \frac{\mathbb{E}(Y_1 Y_1 Y_2)}{\mathbb{E}(Y_1 Y_2 Y_2)}}, \\ \lambda_{21} &= \sqrt{\mathbb{E}(Y_1 Y_2) \frac{\mathbb{E}(Y_1 Y_2 Y_2)}{\mathbb{E}(Y_1 Y_1 Y_2)}}. \end{aligned}$$

Interestingly, the ratio of the two factor loadings is then

$$\frac{\lambda_{21}}{\lambda_{11}} = \frac{\mathbb{E}(Y_1 Y_2 Y_2)}{\mathbb{E}(Y_1 Y_1 Y_2)}. \tag{31}$$

Replacing expectations by sample means, we obtain a consistent estimator of $\frac{\lambda_{21}}{\lambda_{11}}$ which is the coefficient of the regression Y_2 on Y_1 with no intercept, by 2SLS, using $Y_1 Y_2$ as an instrument for Y_1. This is the estimator of the measurement error model that was proposed by Geary (1942). Interestingly, the quasi-JADE estimator that we shall propose in the next section also satisfies equation (31). The estimators introduced in this paper can thus be interpreted as a generalization of Geary's IV estimator.

3 Estimation

We start by discussing the issue of estimating the number of factors.

3.1 Estimating the number of factors K

Estimating K when $K \leq \frac{L(L-1)}{2}$ and all factors are kurtotic. Assuming that Q is full column rank and that factor variables show kurtosis excess, then matrix Ω_Y has rank K (see Lemma 1). For any i.i.d. sample, let $\widehat{\Omega}_Y$ be the empirical counterpart of Ω_Y, obtained by replacing expectations by sample means. We use the sequential testing procedure developed by Robin and Smith (2000) to estimate the rank of Ω_Y.[11]

[11] Robin and Smith's rank test is described in Appendix D.

Monte Carlo simulations show that the rank test, applied to matrix Ω_Y alone, suffers from substantial size distortions (see the simulations in the next section). Assuming $K \leq L$, the factor structure provides additional rank conditions that can be used to improve the test's properties. We propose the following refinement.

Consider matrices $\Omega_Y(\ell, m)$ for all $(\ell, m) \in \Delta_{L,2}$ $(\ell < m)$. They satisfy the restrictions:

$$\Omega_Y(\ell, m) = \Lambda D_4 \operatorname{diag}(\Lambda_\ell \odot \Lambda_m) \Lambda^T.$$

Let $w = (w_{1,2}, ..., w_{L-1,L})$ be a vector of $\frac{L(L-1)}{2}$ positive weights. Then,

$$\Omega_{Y,w} \equiv \sum_{(\ell,m)\in\Delta_{L,2}} w_{\ell,m}\Omega_Y(\ell, m) = \Lambda D_4 \operatorname{diag}\left(Q^T w\right) \Lambda^T. \tag{32}$$

As no column of Q is identically zero, matrix $\Omega_{Y,w}$ has rank K for almost all w.

It seems natural to weight cumulant matrices more if they are more precise. We therefore suggest to choose $w_{\ell,m}$ equal to the inverse of the average of asymptotic variances of the components of the empirical estimate $\widehat{\Omega}_Y(\ell, m)$ of $\Omega_Y(\ell, m)$. These variances can be computed by standard bootstrap.

Estimating K when $K \leq L$ and all factors are skewed or kurtotic. Assuming that Λ and Q are full column rank and that factor variables have non zero skewness, then matrix Γ_Y has rank K (see Lemma 2). One can thus apply the rank test to any root-N estimator $\widehat{\Gamma}_Y$.

More generally, one can use the following version of the rank test, which uses third and fourth-order information of the data. Lemma 2 shows that, assuming that Λ and Q are full column rank (so that $K \leq L$) and that each factor distribution is either skewed or kurtotic, matrix Ξ_Y has rank K. One can thus test the rank of any root-N consistent estimator $\widehat{\Xi}_Y$.

Again, one can refine the test and account for moments' variability, as matrix

$$\Xi_{Y,w} = \Gamma_Y + \sum_{j=1}^{L} w_j \Omega_Y(j) = \Lambda \left[D_3 + D_4 \operatorname{diag}\left(\Lambda^T w\right)\right] Q^T \tag{33}$$

has rank K, for almost all weight $w = (w_1, ..., w_L)^T \in \mathbb{R}^L$.[12] We suggest to set w_j equal to the

[12] This is because the set

$$\{w \in \mathbb{R}^L, \kappa_3(X_k) + \kappa_4(X_k)\left(\sum_{j=1}^{L} w_j \lambda_{jk}\right) = 0\}$$

has measure zero in $\mathbb{R}^L$, for all $k = 1...K$.

average of the variances of the components of $\widehat{\Gamma}_Y$ divided by the average of the variances of the components of $\widehat{\Omega}_Y(j)$.

3.2 Cardoso and Souloumiac's JADE procedure

Assuming no noise, factor loadings satisfy the following system of matrix equations:

$$\Omega_Y(\ell, m) = \Lambda D_4(\ell, m) \Lambda^T, \quad (\ell, m) \in \overline{\Delta}_{L,2}, \tag{34}$$

$$\Sigma_Y = \Lambda \Lambda^T, \tag{35}$$

for diagonal matrices $D_4(\ell, m)$ (see Section 2.4).

In an influential paper, Cardoso and Souloumiac (1993) propose the following procedure to estimate Λ using this system of restrictions.

1. "Whiten" the data, i.e. compute $\widetilde{Y} = P^- Y$, where P is a $L \times L$ such that $PP^T = \Sigma_Y$ (for example, a Cholesky decomposition) and A^- is a generalized inverse of P, e.g. $P^- = \left[P^T P\right]^{-1} P^T$.

2. Compute $\Omega_{\widetilde{Y}}(\ell, m)$, for all $(\ell, m) \in \overline{\Delta}_{L,2}$. These matrices satisfy the restrictions:

$$V^T \Omega_{\widetilde{Y}}(\ell, m) V = D_4(\ell, m),$$

   where $V = P^- \Lambda$ is an orthonormal matrix of dimensions K.

3. Compute V as an orthonormal matrix minimizing the sum of squares of the off-diagonal elements of matrices $V^T \Omega_{\widetilde{Y}}(\ell, m) V$. Cardoso and Souloumiac (1993) develop a simple and efficient algorithm to perform this optimisation (using Jacobi rotations), that is detailed in Section B of the Appendix.[13]

To apply this algorithm on a sample $\{Y_1, ..., Y_N\}$ of i.i.d. observations, replace expectations by sample means. The theoretical restrictions then only hold approximately but the joint diagonalisation algorithm still delivers an orthonormal matrix $\widehat{V}$ such that all matrices $\widehat{V}^T \widehat{\Omega}_{\widetilde{Y}}(\ell, m) \widehat{V}$ are approximately diagonal. An estimate of Λ is then simply obtained as $\widehat{\Lambda} = \widehat{P}\widehat{V}$. Cardoso

[13] A MATLAB code of the JADE algorithm is available on Cardoso's web page: http://www.tsi.enst.fr/ cardoso/Algo/Jade/jadeR.m.

and Souloumiac (1993) call JADE this empirical procedure (Joint Approximate Diagonalisation of Eigenmatrices).

The JADE algorithm has several attractive properties. As it uses *all* fourth-order cumulants of the data, it is much less sensitive to spectrum degeneracy than single diagonalization algorithms (see Cardoso, 1999). Moreover, the cost to pay for these efficiency gains is reasonable, as Jacobi rotation-based algorithms are fast to converge. Lastly, JADE is *equivariant* in the sense that changing Y into WY, for any invertible matrix W, changes $\widehat{\Lambda}$ into $W\widehat{\Lambda}$.

3.3 Asymptotic theory for JADE

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.

To proceed, let $\widehat{A}_1, ..., \widehat{A}_J$ be root-N consistent and asymptotically normal estimators of J symmetric $K \times K$ matrices $A_1, ..., A_J$. Construct $\widehat{A} = \left[\widehat{A}_1, ..., \widehat{A}_J\right]$ and $A = [A_1, ..., A_J]$ by concatenation. Let $\mathbb{V}_A$ be the asymptotic variance of $N^{1/2}\operatorname{vec}(\widehat{A})$. The JADE estimator is

$$\widehat{V} = \arg\min_{V \in \mathcal{O}_K} \sum_{j=1}^{J} \operatorname{off}(V^T \widehat{A}_j V),$$

where $\operatorname{off}(M) = \sum_{i \neq j} m_{ij}^2$ and $\mathcal{O}_K$ is the set of orthonormal $K \times K$ matrices.

Assume that there exists $V \in \mathcal{O}_K$ such that, for all $j = 1, ..., J$, $V^T A_j V = D_j$, where D_j is the diagonal matrix with diagonal elements $d_{j1}, ..., d_{jK}$. Define the $K \times K$ matrices:

$$R(D_j) = \left[\frac{(d_{jk} - d_{jm})}{\sum_{j'=1}^{J}(d_{j'k} - d_{j'm})^2}; \quad (k,m) \in \{1, ..., K\}^2\right].$$

Lastly, let W be the following $K^2 \times JK^2$ matrix:

$$W = \left[\operatorname{diag}(\operatorname{vec}(R(D_1))), ..., \operatorname{diag}(\operatorname{vec}(R(D_J)))\right].$$

We show the following result in Appendix C.

Theorem 7 *Assume that* $\sum_{j=1}^{J}(d_{jk} - d_{jm})^2 \neq 0$ *for all* $k \neq m$. *Then*

$$N^{1/2}\left(\operatorname{vec}(\widehat{V}) - \operatorname{vec}(V)\right) \underset{N \to \infty}{\xrightarrow{L}} \mathcal{N}\left(0, \mathbb{V}_V\right),$$

where:

$$\mathbb{V}_V = (I_K \otimes V)W(I_J \otimes V^T \otimes V^T)\mathbb{V}_A(I_J \otimes V \otimes V)W^T(I_K \otimes V^T). \tag{36}$$

Let us consider the particular case of $J = 1$. In this case, (36) 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 W are equal to $1/(d_{1k} - d_{1m})$, for $k \neq m$. The variance of eigenvectors thus increases when two eigenvalues of A_1 get close to each other.

In the general case of more than one matrix ($J > 1$), a precise estimation requires $\sum_j (d_{jk} - d_{jm})^2$ not to be close to zero, for all indices (k, m). Now, the larger J and the less likely it is that $d_{jk} = d_{jm}$ for all j. Cardoso (1999) already noted that joint diagonalisation algorithms seemed less sensitive to the presence of multiple roots than usual diagonalisation techniques.[14] Theorem 7 allows to better understand the conditions granting a good precision.

Basing identification on fourth-order moments, indices are $j = (\ell, m)$, and matrices A_j and D_j are of the form: $\Omega_Y(\ell, m)$ and $D_4 \operatorname{diag}(\Lambda_\ell \odot \Lambda_m)$, respectively. If there exist k, k' such that $d_{jk} = d_{jk'}$ for all j, it must be that

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

for all ℓ, m. This cannot happen if at most one factor has zero kurtosis excess and the columns of Λ 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 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 kurtosis excess. We shall produce Monte-Carlo simulations to illustrate this point.

Practical remark. In practice, we do *not* recommend to use formula (36) to compute standard errors. Instead, we suggest to compute standard errors or coverage intervals by standard bootstrap (with appropriate recentring). The reason is that the expression in (36) involves second moments of third and/or fourth-order moments of the data, which are difficult to esti-

[14] See also the asymptotic distribution of estimators of *common* principal components derived by Flury (1984, 1986).

mate precisely. Our simulations show extremely imprecise estimates of matrix $\mathbb{V}_A$, even with very large samples (more than $10,000$ observations). In contrast, bootstrap provides good approximations of the true variance-covariance matrix of the JADE estimator.

3.4 The quasi-JADE algorithm

The JADE algorithm is only valid if there is no noise ($U = 0$). However, Lemmas 1 and 2 show that the first four moments of error variables can be estimated independently of factor loadings. Given error moments, one can then apply JADE.

We call quasi-JADE the following procedure.

1. Estimate matrices C and/or $\overline{C}$ of Lemmas 1 and 2. These matrices are easily obtained by Singular Value Decomposition of matrices Ω_Y and Ξ_Y.

2. Estimate error variances $\mathrm{Var}\,(U_\ell)$, third-order cumulants $\kappa_3\,(U_\ell)$ and/or fourth-order cumulants $\kappa_4\,(U_\ell)$ using the restrictions in Lemmas 1 and 2. One should impose the non negativity of error variances, as well as the positive semi-definiteness of matrix $\Sigma_Y - \Sigma_U$.

3. Proceed to the joint diagonalisation (i.e. steps 2 and 3 of the JADE algorithm) of matrices $P^- \left[\Gamma_Y\,(\ell) - \kappa_3\,(U_\ell)\,\mathrm{Sp}_{L,\ell}\right] P^{-T}$ and/or $P^- \left[\Omega_Y\,(\ell,m) - \delta_{\ell m}\kappa_4\,(U_\ell)\,\mathrm{Sp}_{L,\ell}\right] P^{-T}$, where P is a full column rank $L \times K$ matrix such that $\Sigma_Y - \Sigma_U = PP^T$. We suggest to compute P as the first K columns of the Cholesky decomposition of matrix $\Sigma_Y - \Sigma_U$. Let V be the orthonormal matrix of joint eigenvectors. Then $\Lambda = PV$.

4. Estimate factor cumulants $\kappa_3\,(X_k)$ and $\kappa_4\,(X)$ by OLS from restrictions:

$$\begin{aligned}\left[V^T P^- \left[\Gamma_Y\,(\ell) - \kappa_3\,(U_\ell)\,\mathrm{Sp}_{L,\ell}\right] P^{-T} V\right]_{k,k} &= \lambda_{\ell k}\kappa_3\,(X_k)\,,\\ \left[V^T P^- \left[\Omega_Y\,(\ell,m) - \delta_{\ell m}\kappa_4\,(U_\ell)\,\mathrm{Sp}_{L,\ell}\right] P^{-T} V\right]_{k,k} &= \lambda_{\ell k}\lambda_{mk}\kappa_4\,(X_k)\,,\end{aligned}$$

   where $[A]_{i,j}$ denotes the (i,j) entry of matrix A.

Quasi-JADE is only marginally more complicated to implement than JADE,[15] and is almost as fast to converge. However, allowing for errors has a cost. Whereas JADE is equivariant,

[15] GAUSS codes for quasi-JADE can be downloadable from the first author's web-page: http://www.cemfi.es/bonhomme/.

quasi-JADE is only *scale-invariant* in the sense that changing measurement units – that is, Y into DY, for any invertible diagonal matrix D – changes the estimate $\widehat{\Lambda}$ into $D\widehat{\Lambda}$ exactly (as it changes theoretically Λ into $D\Lambda$).

Efficiency improvements. As the original JADE algorithm, quasi-JADE is obviously not efficient. First, it operates a sequence of minimum distance estimations instead of estimating all parameters jointly. Second, it does not use the optimal metric in these minimum distance problems. Third, it does not use all the structural moment restrictions. For example, the diagonal matrices $D_4(\ell, m)$ in (12) are related to Λ but we do not use this restriction.

A natural alternative to our approach would be to use all cumulant restrictions (8)-(10)-(12) in estimation. However, these restrictions are highly nonlinear polynomial equations, which are difficult to solve using standard gradient algorithms or any other general-purpose solving technique. We shall make this point more precise in the simulation section. Second, there is considerable evidence that the optimal metric does not outperform the identity metric in finite samples (see Altonji and Segal, 1994, 1996).

Nevertheless, there is scope for efficiency improvements. For instance, one can use Generalised Least Squares instead of OLS to estimate error cumulants in Step 2 of the algorithm. Likewise, one can weight the matrices to diagonalise in Step 3. Weights can be some measure of estimation precision, as outlined in 3.1. In simulations, we found that this method yielded little efficiency gains. On real data, however, we found slightly different results that we shall present in section 5. Note that this weighting procedure is *ad hoc*. Issues regarding the optimal weighting of cumulant matrices, based on asymptotic results such as (36), are left for future research.

4 Monte-Carlo simulations

In this section, we study the finite-sample properties of our estimators by numerical simulations. We first consider the estimation of Λ given the true value of K, the number of factors. Then, we present simulations for estimating K.

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)
λ_{12}	.98 (.23)	.98 (.15)	1.00 (.06)	1.00 (.05)
λ_{22}	2.05 (.27)	2.03 (.19)	2.01 (.08)	2.01 (.07)
λ_{32}	.97 (.23)	.98 (.17)	1.00 (.06)	1.00 (.05)
λ_{13}	.97 (.23)	.98 (.15)	.99 (.06)	1.00 (.05)
λ_{23}	.97 (.23)	.98 (.16)	1.00 (.06)	1.00 (.05)
λ_{33}	2.06 (.27)	2.02 (.19)	2.01 (.09)	2.00 (.05)
$\mathrm{Var}(U_1)$	.77 (.59)	.87 (.43)	.96 (.20)	.98 (.16)
$\mathrm{Var}(U_2)$	.76 (.57)	.87 (.43)	.98 (.20)	.98 (.17)
$\mathrm{Var}(U_3)$	.74 (.56)	.86 (.42)	.96 (.20)	.98 (.16)

Table 1: Quasi-JADE based on the 2nd, 3rd and 4th moment restrictions of Lemma 1 (log-normal factors, standard normal errors, $\Lambda = \Lambda_1$)

4.1 Estimation of factor loadings

Table 1 presents the results of 1000 simulations of the model with centred and standardised log-normal factors, standard normal errors and Λ equal to

$$\Lambda_1 \equiv \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

Results are given for various sample sizes N. Monte Carlo standard deviations of estimates are given between brackets. Estimation is based on all second, third and fourth-order moments of the data and uses the restrictions of Lemma 1.

Table 1 shows some evidence of finite sample bias. However, in general the bias is small and rapidly decreases as N increases. By comparison, small sample biases are much larger and convergence is much slower for empirical cumulants. Table 2 shows the means and standard deviations of the empirical skewness and kurtosis of a standardised log-normal variate, for various sample sizes.[16] The striking contrast between Tables 1 and 2 suggests that our algorithm does a good job at extracting the relevant information from higher-order moments of the data, while being relatively immune to the imprecision of their estimation in finite samples.

We then study the robustness of our algorithm to the magnitude of the noise. In Table 3, we compare the performance of quasi-JADE to standard JADE. We run the simulations with normal errors, log-normal factors, a sample size of $N = 1000$ and $\Lambda = \Lambda_1$. The standard

[16] Means and variances were computed from 1000 independent drawings, for each sample size N.

N	500	1000	5000	10000	∞
κ_3	4.51 (1.98)	5.01 (2.36)	5.73 (2.65)	5.89 (2.02)	6.18
κ_4	36.1 (38.4)	48.6 (62.4)	77.0 (132.3)	83.3 (104.7)	110.9

Table 2: Empirical skewness and excess kurtosis of a log-normal random variable

deviation of errors can take four values: 0.1, 0.5, 1 and 2. We see that the performances of quasi-JADE deteriorate as the signal-to-noise ratio decreases. However, the bias remains limited even for rather large error variances. By comparison, JADE bears much larger biases. In particular, while our quasi-JADE algorithm yields consistent estimates of factor loadings, the inconsistency of ordinary (noise-free) ICA is severe, even when the magnitude of the error variances is not especially large (for example for a variance of one; which here implies that $\mathrm{Var}(U_\ell)/\mathrm{Var}(Y_\ell) = 20\%$).

Next, we compare these results with Minimum Distance based on the complete set of moment restrictions. The estimation is based on second and fourth-order restrictions:

$$\begin{cases} \Sigma_Y = \Lambda\Lambda^T + \Sigma_U, \\ \Omega_Y = \overline{Q} D_4 Q^T, \end{cases}$$

where Ω_Y is the 6×3 matrix of fourth-order cumulants of Y given by (16) and where Q and $\overline{Q}$ depend on Λ.

In all the simulations that we performed, Full Minimum Distance proved to be highly unstable. Maximisation with respect to the whole set of parameters (Λ, Σ_U, D_4) converged (numerically) in none of the cases that we considered. To obtain a more stable algorithm, admittedly at the cost of lower efficiency, we treated the coefficients of D_4 as nuisance parameters. Precisely, we minimised the Minimum Distance norm, evaluated at $(\Lambda, \Sigma_U, D_4(\Lambda))$, with respect to (Λ, Σ_U) alone and where $D_4(\Lambda)$ is such that

$$\mathrm{vec}\left[D_4(\Lambda)\right] = \left(Q \otimes \overline{Q}\right)^{-} \mathrm{vec}(\Omega_Y).$$

Note that using the optimal metric to estimate $D_4(\Lambda)$ from restriction $\Omega_Y = Q D_4 \overline{Q}^T$ given Λ yielded even greater instability. Incorporating third-order moment restrictions into the algorithm had the same effect.

Table 4 presents simulation results with log-normal factors, normal errors and $\Lambda = \Lambda_1$.

JADE

$\mathrm{Var}(U_\ell)$	.01	.25	1	4
$\widehat{\lambda}_{11}$	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)
λ_{12}	1.00 (.11)	1.00 (.12)	.97 (.24)	.78 (.86)
λ_{22}	2.00 (.07)	2.11 (.07)	2.37 (.12)	2.86 (.32)
λ_{32}	1.00 (.12)	1.03 (.13)	1.08 (.22)	1.08 (.76)
λ_{13}	1.00 (.11)	.87 (.13)	.61 (.20)	.16 (.69)
λ_{23}	1.00 (.11)	.87 (.12)	.62 (.20)	.15 (.67)
λ_{33}	2.00 (.08)	2.02 (.09)	2.13 (.16)	2.52 (.43)

quasi-JADE

$\mathrm{Var}(U_\ell)$	.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)
λ_{12}	1.00 (.16)	.99 (.13)	.98 (.15)	.97 (.33)
λ_{22}	1.97 (.11)	2.02 (.11)	2.03 (.19)	2.02 (.41)
λ_{32}	.99 (.16)	.99 (.13)	.98 (.17)	.97 (.32)
λ_{13}	1.00 (.16)	1.00 (.14)	.98 (.15)	.96 (.32)
λ_{23}	1.00 (.16)	1.00 (.13)	.98 (.16)	.96 (.32)
λ_{33}	1.98 (.11)	2.02 (.11)	2.02 (.19)	2.01 (.42)
$\mathrm{Var}(U_1)$	.04 (.11)	.18 (.22)	.87 (.43)	3.77 (.98)
$\mathrm{Var}(U_2)$	.04 (.11)	.17 (.23)	.87 (.43)	3.77 (.94)
$\mathrm{Var}(U_3)$	.04 (.11)	.17 (.22)	.86 (.42)	3.77 (.97)

Table 3: Robustness to noise of JADE and quasi-JADE (log-normal factors, standard normal errors, $N = 1000$, $\Lambda = \Lambda_1$)

$\text{Var}(U_\ell)$	.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)
λ_{12}	.99 (.10)	.99 (.11)	.99 (.13)	.96 (.26)
λ_{22}	2.04 (.13)	2.04 (.12)	2.03 (.17)	2.04 (.44)
λ_{32}	.99 (.10)	.99 (.11)	.99 (.13)	.96 (.27)
λ_{13}	.99 (.11)	.98 (.11)	.98 (.13)	.96 (.28)
λ_{23}	.98 (.10)	.99 (.10)	.99 (.13)	.95 (.27)
λ_{33}	2.04 (.13)	2.04 (.13)	2.03 (.18)	2.00 (.42)
$\text{Var}(U_1)$	-.09 (.32)	.11 (.37)	.86 (.44)	3.75 (1.28)
$\text{Var}(U_2)$	-.11 (.33)	.11 (.34)	.87 (.42)	3.63 (2.27)
$\text{Var}(U_3)$	-.12 (.34)	.12 (.35)	.88 (.45)	3.78 (1.09)
% convergence	99.9%	100.0%	99.8%	84.3%

Table 4: Minimum Distance estimator based on 2nd and 4th order moments ($K = 3$, log-normal factors, normal errors, $V(U) = .25$)

Results are presented conditional on numerical convergence.[17] Starting values were chosen equal to the true parameters. First, we find that, conditional on numerical convergence, Full Minimum Distance is slightly more efficient than our algorithm in finite sample. This result was to be expected, as our algorithm uses only a subset of the moment conditions implied by the factor model. However, the difference in variances is not large, especially when looking at factor loadings. Second, this efficiency gain is obtained at two costs. The first one is numerical instability, which is illustrated by the final row of Table 4. When error variances are larger ($\text{Var}(U_\ell) = 4$), maximisation failed to converge in 157 cases out of 1000. The second cost is computing time, which increases rapidly with the number of factors.

Next, we investigate the sensitivity of our algorithm to the amount of factor kurtosis. The sample size is $N = 1000$. Errors are standard normal variables. To vary the kurtosis, we generate factors as mixtures of two independent normals. Let $W_1 \sim N(0, 1/2)$, and let $\rho \in]0, 1[$. Define $W_2 \sim N(0, (2-\rho)/(2-2\rho))$, independent of W_1. Then, it is straightforward to see that X, defined as W_1 with probability ρ and W_2 with probability $1-\rho$, has variance one and its kurtosis excess is $\kappa_4(\rho) = 3\rho/(4(1-\rho))$. Table 5 displays Monte Carlo simulation results for values of ρ yielding kurtosis equal to $\frac{1}{2}$, 2, 5, 10 and 100. In the first column of Table 5, we report

[17] We declared numerical convergence achieved when the gradient of the GMM criterion was inferior to 10^{-3} in absolute value after 5000 GMM iterations.

ρ	(Uniform)	2/5	4/7	20/23	40/43	400/403	(Lognormal)
κ_4	-6/5	1/2	1	5	10	100	≈110
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)
λ_{12}	.97 (.49)	1.00 (.71)	.98 (.65)	.96 (.30)	.98 (.21)	.99 (.19)	.98 (.16)
λ_{22}	1.98 (.44)	1.71 (.69)	1.83 (.64)	2.02 (.35)	2.02 (.26)	2.01 (.18)	2.03 (.18)
λ_{32}	.98 (.49)	1.00 (.72)	.95 (.66)	.97 (.30)	.98 (.20)	.99 (.18)	.98 (.16)
λ_{13}	.96 (.49)	1.12 (.74)	1.05 (.70)	.97 (.29)	.99 (.20)	.99 (.17)	.98 (.15)
λ_{23}	.94 (.49)	1.12 (.75)	1.05 (.69)	.97 (.29)	.98 (.19)	.99 (.18)	.98 (.15)
λ_{33}	1.97 (.43)	1.83 (.57)	1.89 (.56)	2.03 (.32)	2.03 (.25)	2.02 (.18)	2.03 (.20)
$\mathrm{Var}(U_1)$	.71 (.65)	.92 (.84)	.76 (.79)	.77 (.63)	.88 (.53)	.92 (.40)	.86 (.44)
$\mathrm{Var}(U_2)$	.75 (.65)	.89 (.83)	.69 (.78)	.75 (.64)	.83 (.55)	.93 (.40)	.87 (.43)
$\mathrm{Var}(U_3)$	.74 (.66)	.93 (.82)	.76 (.80)	.77 (.64)	.84 (.53)	.91 (.40)	.86 (.44)

Table 5: Quasi-Jade with factors of increasing kurtosis (factors are normal mixtures, standard normal errors, $N = 1000$, $\Lambda = \Lambda_1$)

the results corresponding to factors following a (standardised) uniform distribution over $[-1, 1]$. The uniform distribution is platykurtic, with $\kappa_4 = -6/5$. The last column shows the results for (standardised) log-normal factors, the kurtosis excess of which is 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 kurtosis excess is to zero, the greater the estimator's bias and the lower its precision.

We now set $K < L$ and compare the quasi-JADE procedures based on the restrictions of Lemma 1 and 2. The estimator based on the restrictions of Lemma 1 uses all second, third and fourth-order moment restrictions while the estimator based on the restrictions of Lemma 2 only uses second and third-order moments and assumes that all factors are skewed. Table 6 reports simulations with log-normal factors, standard normal errors with variance 1, and matrix Λ is equal to:

$$\Lambda_2 \equiv \begin{pmatrix} 2 & 1 \\ 1 & 2 \\ 1 & 1 \end{pmatrix}. \tag{37}$$

Table 6 shows, quite surprisingly, that fourth-order moments yield rather small additional efficiency gains. This illustrative table suggests that an algorithm based on third-order moments only could do well in practice, provided that there is enough skewness in the data. On the other hand, adding moment restrictions (and there can be a lot of fourth-order moment restrictions)

N	500	500	1000	1000	5000	5000
Cumulants	2,3,4	2,3	2,3,4	2,3	2,3,4	2,3
λ_{11}	1.95 (.28)	1.93 (.32)	1.98 (.19)	1.97 (.24)	2.00 (.08)	2.00 (.08)
λ_{21}	1.96 (.30)	1.91 (.37)	1.99 (.16)	1.96 (.23)	1.00 (.09)	2.00 (.05)
λ_{31}	.97 (.23)	.98 (.25)	.98 (.17)	.98 (.20)	1.00 (.08)	1.00 (.08)
λ_{12}	2.02 (.24)	2.03 (.27)	2.01 (.17)	2.01 (.20)	1.00 (.08)	2.00 (.08)
λ_{22}	1.02 (.28)	1.05 (.32)	1.00 (.18)	1.02 (.22)	2.00 (.09)	1.00 (.08)
λ_{32}	2.01 (.12)	1.99 (.14)	2.01 (.10)	2.00 (.11)	1.00 (.05)	2.00 (.05)
$\text{Var}(U_1)$	.98 (.21)	1.01 (.16)	.98 (.15)	1.00 (.13)	.97 (.09)	1.00 (.06)
$\text{Var}(U_2)$	.94 (.21)	.99 (.20)	.96 (.15)	1.00 (.15)	.97 (.08)	1.00 (.07)
$\text{Var}(U_3)$	.94 (.22)	1.00 (.20)	.96 (.15)	1.00 (.15)	.98 (.09)	1.00 (.07)

Table 6: Comparing the two quasi-JADE algorithms based on Lemma 1 and 2 (log-normal factors, standard normal errors, $\Lambda = \Lambda_2$)

does not seem to increase the bias, which is reassuring.

Lastly, we investigate the finite-sample performance of our algorithm when the number of measurements and factors increases. Table 7 illustrates the cases $L = K = 5$ and $L = K = 10$, respectively. In both cases, Λ has entries equal to 2 everywhere on the diagonal, and equal to one everywhere else.We only report the estimates of the first column of Λ and the variance of the first error, the other estimates being qualitatively similar. These simulations show that the performances of our algorithm are 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. In comparison, the Minimum Distance algorithm discussed above turned out to be infeasible in practice for L as low as five, the computing time becoming prohibitive.

4.2 Estimation of the number of factors

We here report a Monte-Carlo study of the rank tests detailed in 3.1.

We first compute the empirical size of the test based on matrix Ω_Y for various values of factor kurtosis. The simulation scheme is the same as for the results reported in Table 5. The true value of Λ is Λ_2 (as in (37)) and we test $K = 2$ against $K = 3$.

Table 8 shows substantial size distortion. This especially happens when the kurtosis excess is low (in absolute value) – that is, when fourth-order cumulants contain very little information on the factor structure – or large – that is, when fourth-order moments are imprecisely estimated.

	$L = K = 5$			$L = K = 10$		
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)
$\mathrm{Var}(U_1)$	.58 (.56)	.81 (.44)	.95 (.20)	.40 (.55)	.49 (.53)	.88 (.28)

Table 7: Increasing the number of factors and measurements (log-normal factors, standard normal errors)

ρ	-	2/5	4/7	20/23	40/43	400/403
$\kappa_4(\rho)$	-6/5	1/2	1	5	10	100
$\alpha = .10$	.90	.73	.82	.87	.85	.62
$\alpha = .20$	.79	.57	.67	.74	.69	.43
$\alpha = .30$	.67	.44	.54	.61	.57	.29
$\alpha = .40$	.58	.33	.42	.50	.45	.19
$\alpha = .50$	.47	.24	.32	.40	.35	.11
$\alpha = .60$	.37	.16	.22	.32	.26	.05
$\alpha = .70$	.27	.10	.13	.24	.19	.02
$\alpha = .80$	.20	.05	.08	.15	.11	.01
$\alpha = .90$	.10	.02	.04	.06	.04	.00

Table 8: Size of the rank test based on Ω_Y for increasing kurtosis (factors of normal mixtures, errors are Gaussian, $N = 1000$, $\Lambda = \Lambda_2$)

However, for reasonable values of kurtosis excess (less than a few tens),[18] the risk of unvalueing the number of factors exists but remains limited.

Fourth-order moments of unbounded distributions are imprecisely estimated because there is a non negligible probability of drawing values that are much higher than the mode of the distribution, around which most of the distribution is concentrated (peakedness). For the lognormal distribution, for example, drawing one very large value displaces the fourth-order moment to the right of its theoretical value. However, the lognormal distribution is positively skewed and there is thus a bigger probability of drawing small values, so that most of the time the fourth-order moment is underestimated. Since an excessively long tail yields imprecise estimates of higher-order moments, one may be willing to trade a bit of bias against increased precision. We thus experimented with trimming and effectively found that a certain amount of trimming (of measurement variables) improved the size of the test, but too much trimming deteriorated it. As it is impossible to say what is the "optimal" amount of trimming without knowing the model, data trimming is hardly advisable in practice.

In Section 3.1, we proposed to improve the size properties of the rank test by considering a weighted average of cumulant matrices $\Omega_Y(\ell, m)$ – i.e. $\Omega_{Y,w}$ in equation (32) – instead of Ω_Y. Table 9 provides a comparison of rank tests based on different cumulant matrices. We focus on the most difficult case of log-normal factors, normal errors and a sample size of 1000. The first column reports the size of the rank test based on Ω_Y, the second column corresponds to matrix $\Omega_{Y,w}$, and the third and last column refers to matrix Γ_Y (third-order cumulants). The weighting scheme definitely improves the size of the test of $K = 2$ against $K = 3$. However, the rank test still underrejects noticeably, in particular when the theoretical probability of rejection is low. Finally, third-order moments are more precisely estimated and, consequently, the empirical size of the rank test based on Γ_Y is close to the nominal size (third column).

This confirms that applying the characteristic root test to matrices of higher-order cumulants

[18] Financial data are certainly the most kurtotic economic data. The S&P 500 daily returns for 1986 to 1996 have an extremely high kurtosis of about 111. This can be ascribed to the October 1987 stock market crash (Duffie and Pan, 1997). However, between January 1969-December 2004, Lin and Hung (2005), report, for daily 1-, 30-, 100- and 300-day return data on the S&P 500 index, kurtosis values of 36.02, 5.80, 3.77 and 2.99.

Matrix	Ω_Y	$\sum_{\ell<m} w_{\ell m}\Omega_Y(\ell,m)$	Γ_Y
α =.10	.56	.87	.90
α =.20	.34	.71	.79
α =.30	.20	.56	.69
α =.40	.12	.44	.58
α =.50	.08	.32	.48
α =.60	.05	.21	.38
α =.70	.02	.13	.29
α =.80	.01	.06	.16
α =.90	.00	.01	.07

Table 9: Size of the rank test applied to various matrices: Ω_Y, $\sum_{\ell<m} w_{\ell m}\Omega_Y(\ell,m)$ and Γ_Y (log-normal factors, standard normal errors, $N = 1000$, $\Lambda = \Lambda_2$)

ρ	-	2/5	4/7	20/23	40/43	400/403
$\kappa_4(\rho)$	-6/5	1/2	1	5	10	100
α =.10	.99	.81	.81	1.00	1.00	.89
α =.20	.99	.63	.66	1.00	1.00	.80
α =.30	.98	.68	.51	.99	1.00	.72
α =.40	.97	.36	.39	.99	1.00	.64
α =.50	.96	.26	.29	.98	.99	.56
α =.60	.94	.18	.22	.96	.98	.47
α =.70	.93	.11	.16	.92	.96	.35
α =.80	.89	.06	.10	.86	.90	.22
α =.90	.83	.02	.04	.72	.77	.12

Table 10: Power of the improved rank test, Ω_Y, Factors with increasing Kurtosis (standard normal errors, $N = 1000$, $\Lambda = \Lambda_2$)

should be done with a certain amount of caution when they are too imprecisely estimated. However, the results in Tables 8 and 9 show that, for reasonable magnitudes of skewness and kurtosis excess (say less than a few tens). The size properties of the CR test based on third and fourth-order cumulant matrices are satisfactory.

We end this section by a study of the power of the rank test based on $\Omega_{Y,w}$. Table 10 display empirical power computations for various levels of kurtosis. The true value of Λ is Λ_1 and we test $K = 2$ against $K = 3$. For low significance values (α less than 10%) the power of the test is good even if factors are excessively leptokurtic. For intermediate values of the kurtosis excess, the power is good whatever the α level.

5 Application to the returns to schooling

In this section, we apply our methodology to the estimation of the returns to schooling. We consider the relationship between wage and education. Chamberlain and Grilliches (1975, 1977) provide insightful examples of the use of factor models in this context. We first construct a second measure of educational attainment, and we estimate a one-factor model to correct for measurement error in the first education measure. We then apply the methods of this paper and estimate a second factor.

5.1 The data

We use data from the French Labor Force Survey of 1995. This is a large and representative cross-section of the French labor force which provides detailed information on individual education. We exclude women, out-of-employment individuals, and workers with missing data for either (monthly) wages, hours worked or education. We trim the sample of the first and last percentiles of the wage, hour and education data. We finally obtain a sample of 21,794 workers.

We divide monthly wages by hours worked to obtain wage rates. We define Y as the residual of the regression of wage rates on a set of regressors, including a quartic in age. We construct two education variables. The first one is the "age at the end of school", which broadly corresponds to the number of years of schooling (minus 6) in France. This variable, denoted as D, is the usual regression variable in most studies of the returns to schooling. The second one (say "diploma") codes the highest diploma obtained by the individual into 16 categories (no diploma, elementary level, middle school, high school, college, plus various declinations of these different levels into vocational and non vocational). To make this variable continuous and comparable to D, we construct the variable D^* equal to the median value of D by diploma.

Table 11 shows the moments of the three variables of interest. The correlation between D and D^* is only 0.76, indicating that both measures of education are correlated, yet not perfectly. The OLS coefficients of the separate regressions of Y on D and on D^* are 0.044 and 0.060, respectively. The second measure yields a slightly higher return. This difference is quasi entirely explained by the difference in the variances of education measures. (The variance of

	Wage Y	Years of Schooling D	Diploma D^*
Mean	0	17.7	17.6
Standard error	.29	2.64	2.17
Skewness	.29	.61	.61
Kurtosis	.079	-.015	.18
		Covariances	
Y	0.086	0.304	0.284
D	0.304	6.95	4.33
D^*	0.284	4.33	4.71

Table 11: Moments of the variables

D^* is less than that of D, as it should be because D^* is a conditional median of D.) It is not due – as one could maybe have expected – to a difference in the means of D and D^* (class repetition, for example, making D^* less than D on average).

The two education variables are only slightly negatively skewed and exhibit little kurtosis excess. Yet, the joint distribution of (Y, D, D^*) displays a statistically significant amount of skewness and kurtosis. To check that, we estimate the three characteristic roots of matrices Γ_Y and Ω_Y, as well as their bootstrap standard errors.[19] These estimates are: 1.17 (.02), .073 (.0072) and .0072 (.0037) for the three CRs of Γ_Y, and .38 (.060), .146 (.017) and .042 (.0051) for those of Ω_Y. The third eigenvalue of Γ_Y is small and insignificant. These results are confirmed by the CR test applied to matrices Γ_Y and Ω_Y and reported in Table 12. The null hypothesis that Γ_Y has rank 2 is not rejected by the data at the 5% level. The test rejects the hypothesis that the rank of Ω_Y is less than 3 at the 1% level. However, the third eigenvalue of Ω_Y is much smaller that the first two. In any case, there is definite evidence that the joint distribution of (Y, D, D^*) is not normal. One can thus safely apply the methods introduced in this paper.

5.2 Estimation results

We start by estimating the matrix of factor loadings under the assumption that $K = 1$. Factor loadings can then be estimated from covariance calculations only. We report the PCA estimates in the first column of Table 13 (PCA). The implied return to education, as measured by $\frac{\lambda_{11}}{\lambda_{21}}$ is .066, higher than the return estimated by OLS but comparable to the OLS estimate of the

[19] As in the rest of this section, standard errors are computed by 500 bootstrap replications. Standard errors are given in brackets.

	Γ_Y	Ω_Y
Rank	0	0
Statistic	29994	20.2
Critical value .05	57.40	2.20
p-value	.00	.00
Rank	1	1
Statistic	114.0	2.34
Critical value .05	7.74	.12
p-value	.00	.00
Rank	2	2
Statistic	1.10	.185
Critical value .05	1.32	.0091
p-value	.072	.00

Table 12: Rank tests

	$K=1$ PCA	$K=1$ quasi-JADE	$K=1$ quasi-JADE*	$K=2$ quasi-JADE	$K=2$ quasi-JADE*
$\widehat{\lambda}_{11}$	.141 (.0021)	.140 (.0025)	.143 (.0021)	.154 (.0058)	.151 (.0063)
$\widehat{\lambda}_{21}$	2.15 (.019)	2.09 (.030)	2.16 (.018)	2.11 (.029)	2.21 (.029)
$\widehat{\lambda}_{31}$	2.01 (.019)	2.04 (.030)	2.00 (.014)	2.02 (.029)	1.93 (.031)
$\frac{\widehat{\lambda}_{11}}{\widehat{\lambda}_{21}}$	6.6%	6.7%	6.6%	7.3%	6.8%
$\widehat{\lambda}_{12}$	-	-	-	-.085 (.019)	-.057 (.0087)
$\widehat{\lambda}_{22}$	-	-	-	.241 (.098)	.507 (.164)
$\widehat{\lambda}_{32}$	-	-	-	.323 (.082)	.129 (.113)
$\widehat{V}(U_1)$	.066 (.00086)	.068 (.0018)	.065 (.0009)	.055 (.0047)	.060 (.0028)
$\widehat{V}(U_2)$	2.31 (.062)	2.69 (.177)	2.25 (.065)	2.46 (.128)	1.82 (.197)
$\widehat{V}(U_3)$	.672 (.053)	.545 (.130)	.72 (.046)	.53 (.132)	.97 (.100)

Table 13: Factor loadings and error variances (quasi-JADE*: weighted quasi-JADE)

regression of Y on D^*. We find that X_1 accounts for 23% of the variance of wages, 67% of the variance of D but 86% of the variance of D^*. These results are consistent with D^* being a "better" measure of educational attainment than D.[20]

We then estimate the one-factor model using higher-order moments of the data. Columns 2 and 3 of Table 13 present the estimates of the vector of factor loadings using the quasi-JADE algorithm, using second, third and fourth-order cumulants and the restrictions of Lemma 2. In column 3, we report the results for the version of the algorithm weighting the matrices of cumulants according to their precision (quasi-JADE*). The results of all three columns are

[20]Note that PCA yields the same estimate of $\frac{\lambda_{11}}{\lambda_{21}}$ as instrumenting D by D^* in the 2SLS regression of Y on D.

	$K=1$ quasi-JADE	$K=1$ quasi-JADE*	$K=2$ quasi-JADE	$K=2$ quasi-JADE*
$\kappa_3(X_1)$	1.36 [1.32, 1.40]	1.27 [1.22, 1.29]	1.28 [1.23, 1.35]	1.66 [1.61, 1.74]
$\kappa_3(X_2)$	-	-	-.29 [-.93, .03]	-.88 [-11.1, .51]
$\kappa_4(X_1)$	.31 [.25, .38]	.34 [.26, .41]	.29 [.23, .35]	.34 [.26, .40]
$\kappa_4(X_2)$	-	-	34.4 [14.0,206]	97.2 [22.0, 525]

Table 14: Factor cumulants (quasi-JADE*: weighted quasi-JADE)

remarkably similar.

Next, we turn to the estimation of the two-factor model, reported in the last two columns of Table 13. The estimates of factor loadings associated to the first factor are very close to the values estimated using the one-factor model. The second factor is positively correlated with the number of years of schooling D and is negatively correlated with the wage Y. This pattern is consistent with the interpretation that a overspecialisation in education is negatively valued on the labor market (the returns to obtaining a Ph. D. are slight; see Ashenfelter and Mooney, 1968, Hamermesh and Donald, 2004).

Notice that, using third-order moments only, we obtained very imprecise estimates (not reported). This is because the second factor is found to have a nearly symmetric distribution. We report in Table 14 the estimates of factor cumulants. We give between brackets the bootstrap 90% coverage intervals. The results show that the first factor is skewed to the left, with rather small kurtosis. Moreover, the second factor shows little skewness but displays much kurtosis excess. This implies that the second factor is essentially identified from fourth-order moments of the data.

Finally, we tried to investigate the existence of a third factor without success. The estimates were far too imprecise. In any case, if a third factor exists, it has very little explanatory power on individual earnings.

6 Conclusion

It is well known that non normality is an important source of identification in linear measurement error models. In this paper, we extend this insight to general linear independent factor models. We prove that $L(L-1)/2$ factors can be generically identified from a set of L mea-

surement. Contrary to ordinary Factor Analysis, identification is unambiguously defined up to sign and permutation normalisations.

We also prove that second, third and/or fourth-order moments of the data provide sufficient information to identify and estimate the first four moments of at most L factors. We then extend and adapt a well-known technique of Independent Component Analysis (ICA), Cardoso and Souloumiac's (1993) JADE algorithm, to construct estimators of factor loadings in the case where errors are not negligible. We propose a multi-step procedure (quasi-JADE) in which we estimate error moments in a first stage, and then apply Cardoso and Souloumiac's approximate joint diagonalisation algorithm.

The independent factor structure generates many overidentifying restrictions on higher-order moments. This may explain the encouraging Monte Carlo simulation results that we obtained. In contrast with previous evidence on the use of higher-order moments for estimation,[21] we find, for sufficiently non symmetric and/or kurtotic data, small biases and precise estimates, even in relatively small samples. The estimation methodology is then applied to earnings and education data. Besides the common factor that IV and PCA estimates already reveal (explaining the bias toward zero of the OLS estimate of the returns to the number of years of education on individual earnings), quasi-ICA reveals an interesting second factor that is negatively correlated with earnings and positively correlation with education. This is evidence that there exist individual characteristics which are valued by the education institution but not by the labour market.

In the future, we plan to pursue two directions of research. First, this paper leaves many methodological questions unanswered. In particular, efficiency issues concerning the quasi-JADE estimators, as well as the properties of the tests of the number of factors, seem worth investigating further. Moreover, it would be interesting to extend existing algorithms to deal with more factors than measurements ($K > L$). In the ICA literature, this case is referred to as *overcomplete* ICA. De Lathauwer (2003) presents an algorithm comparable to JADE that works for $K > L$ in the case of complex measurements. In the real case, the one of interest in

[21] See the results reported in Madansky (1959), and the survey by Aigner *et al.* (1984).

econometrics, we are not aware of similar semi-parametric methods.

The second direction of research concerns the extension of the methods of this paper to dynamic settings, where observations can be correlated over time (*e.g.* Bai and Ng, 2002, and Bai, 2003).

APPENDIX

A Mathematical proofs

A.1 Proof of Theorem 1

The proof of proposition (i) is a straightforward consequence of Theorem 10.3.1 in Kagan, Linnik and Rao (1973).

Theorem 8 ***(Theorem 10.3.1, Kagan, Linnik and Rao, 1973)*** *Let A and B be two non-stochastic matrices and let $S = (s_1, ..., s_m)^T$ and $R = (r_1, ..., r_n)^T$ be two random vectors with independent components. Assume that AS and BR have the same distribution. If s_i, for some $i \leq m$, is not normal, then the ith column of A is the multiple of a column of B.*

Assume that $\Lambda X + U$ and $\widetilde{\Lambda}\widetilde{X} + \widetilde{U}$ have the same distribution. The components of vectors (X^T, U^T) and $(\widetilde{X}^T, \widetilde{U}^T)$, respectively, are independent. Let $k \leq K$. Since X_k is not normal, Kagan *et al.*'s result applies to show that the k's column of Λ, say Λ_k, is the multiple of a column of the $L \times (K + L)$ matrix $(\widetilde{\Lambda}, I_L)$, where I_L is the $L \times L$ identity matrix. Since every column of matrices Λ and $\widetilde{\Lambda}$ has at least two non-zero coefficients, it must be that Λ_k is the multiple of a column of $\widetilde{\Lambda}$. This shows proposition (i).

To show proposition (ii) let $\boldsymbol{\kappa}''_Y(t) = \left(\kappa_Y^{(\alpha)}(t), \alpha \in \Delta_{L,2}\right)$, for $t \in \mathbb{R}^L$, be the $\#\Delta_{L,2} \times 1$ vector of second-order partial cross-derivatives of $\kappa_Y(t)$. Let also

$$\begin{aligned}
\boldsymbol{\kappa}_X(t) &= (\kappa_{X_1}(t_1), ..., \kappa_{X_K}(t_K))^T \\
\boldsymbol{\kappa}'_X(t) &= \left(\kappa'_{X_1}(t_1), ..., \kappa'_{X_K}(t_K)\right)^T, \\
\boldsymbol{\kappa}''_X(t) &= \left(\kappa''_{X_1}(t_1), ..., \kappa''_{X_K}(t_K)\right)^T, \quad t = (t_1, ..., t_K) \in \mathbb{R}^K.
\end{aligned}$$

Equation (3) implies the following restrictions on factor cumulant generating functions:

$$\boldsymbol{\kappa}''_Y(t) = Q(\Lambda)\boldsymbol{\kappa}''_X\left(\Lambda^T t\right). \tag{A1}$$

To show the second proposition, remark that it must be that

$$Q(\Lambda)\boldsymbol{\kappa}''_X\left(t^T\Lambda\right) = Q(\widetilde{\Lambda})\boldsymbol{\kappa}''_{\widetilde{X}}\left(t^T\widetilde{\Lambda}\right). \tag{A2}$$

By proposition (i), every column of Λ is a scalar multiple of a column of $\widetilde{\Lambda}$. Since $\text{rank}(Q(\Lambda)) = K$, it follows that there exists no couple of columns of Λ which are proportional. Therefore, there exist a permutation matrix P and a diagonal matrix D with non zero entries in the diagonal such that $\widetilde{\Lambda} = \Lambda DP$. Now, since $\ker(Q(\Lambda)) = 0$, (A2) implies:

$$\boldsymbol{\kappa}''_X\left(t^T\Lambda\right) = \boldsymbol{\kappa}''_{DP\widetilde{X}}\left(t^T\Lambda\right). \tag{A3}$$

Taking this equation at $t = 0$ and using the normalization assumption (i) in Definition **??** yields:

$$D^2 = \text{Var}(DP\widetilde{X}) = \text{Var}(X) = I_K.$$

Moreover, integrating the differential equation (A3) shows that $\boldsymbol{\kappa}_X$ and $\boldsymbol{\kappa}_{DP\widetilde{X}}$ differ by an affine function. By definition, $\boldsymbol{\kappa}_X(0) = \boldsymbol{\kappa}_{DP\widetilde{X}}(0) = 0$. By assumption, since the factor distributions have zero mean: $\boldsymbol{\kappa}'_X(0) = \boldsymbol{\kappa}'_{DP\widetilde{X}}(0) = 0$. This shows that the d.f. of X and $DP\widetilde{X}$ are equal. Lastly, the d.f. of U and $\widetilde{U}$ are equal by deconvolution, since the characteristic functions of the factors are nonvanishing everywhere.

This ends the proof.

A.2 Proof of Theorem 2

Let $x = (x_1, ..., x_K)^T \in \ker(Q(\Lambda))$ such that $x \neq 0$. For all $k = 1, ..., K$, and all $\ell = 1, ..., L$, define

$$\begin{aligned} \psi_k(\tau_k) &= \kappa_{X_k}(\tau_k) - x_k \frac{\tau_k^2}{2}, \ \forall \tau_k \in \mathbb{R}, \\ \zeta_\ell(t_\ell) &= \kappa_{U_\ell}(t_\ell) + (\Lambda_\ell \otimes \Lambda_\ell)\, x \frac{t_\ell^2}{2}, \ \forall t_\ell \in \mathbb{R}, \end{aligned}$$

where $\Lambda_\ell = (\lambda_{\ell 1}, ..., \lambda_{\ell K})$ is the ℓth row of Λ and $\otimes$ is the Kronecker product ($(\Lambda_\ell \otimes \Lambda_\ell)\, x = \sum_{k=1}^K x_k \lambda_{\ell k}^2$). If $x_k \geq 0$, $\psi_k$is the the cumulant generating function (c.g.f.) of the convolution of the distribution of X_k and the normal distribution $\mathcal{N}\left(0, \sqrt{x_k}\right)$. Now, suppose that $x_k < 0$. The distribution of X_k is divisible by a normal distribution, say $\mathcal{N}\left(0, \sigma_k^2\right)$. If $\sigma_k^2 + x_k > 0$, then ψ_k is the c.g.f. of some random variable that is the sum of the random variable with c.g.f. $\kappa_{X_k}(\tau_k) + \frac{\sigma_k^2 \tau_k^2}{2}$ and of the normal variable $\mathcal{N}\left(0, \sigma_k^2 + x_k\right)$. The same argument applies to ζ_ℓ. If $(\Lambda_\ell \otimes \Lambda_\ell)\, x \leq 0$, then ζ_ℓ is the c.g.f. of $U_\ell + \mathcal{N}\left(0, -(\Lambda_\ell \otimes \Lambda_\ell)\, x\right)$. Otherwise, U_ℓ is divisible by a normal distribution, say $\mathcal{N}\left(0, \omega_\ell^2\right)$. If $\omega_\ell^2 - (\Lambda_\ell \otimes \Lambda_\ell)\, x > 0$, then ζ_ℓ is the c.g.f. of some random variable that is the sum of the random variable whose c.g.f. is $\kappa_{U_\ell}(t_\ell) + \frac{\omega_\ell^2 t_\ell^2}{2}$ and the normal variable $N\left(0, \omega_\ell^2 - (\Lambda_\ell \otimes \Lambda_\ell)\, x\right)$. Rescale x if necessary so that $x_k > -\sigma_k^2$, for all $k = 1, ..., K$, and $\omega_\ell^2 > (\Lambda_\ell \otimes \Lambda_\ell)\, x$, for all $\ell = 1, ..., L$. One can thus construct $K + L$ non degenerate, independent random variables with zero mean and finite variance: $Z_1, ..., Z_K, \widetilde{U}_1, ..., \widetilde{U}_L$, with given c.g.f.'s $\psi_1, ..., \psi_K, \zeta_1, ..., \zeta_L$.

Next, for all $t = (t_1, ..., t_L)^T$, define

$$\begin{aligned} \kappa(t) &\equiv \sum_{k=1}^K \psi_k(\boldsymbol{\lambda}_k^T t) + \sum_{\ell=1}^L \zeta_\ell(t_\ell) \\ &= \sum_{k=1}^K \kappa_{X_k}(\boldsymbol{\lambda}_k^T t) - \sum_{k=1}^K x_k \frac{\left(\boldsymbol{\lambda}_k^T t\right)^2}{2} + \sum_{\ell=1}^L \kappa_{U_\ell}(t_\ell) + \sum_{\ell=1}^L (\Lambda_\ell \otimes \Lambda_\ell)\, x \frac{t_\ell^2}{2}. \end{aligned}$$

As $Q(\Lambda)x = 0$, $\sum_{k=1}^K x_k \lambda_{\ell k} \lambda_{mk} = 0$ for all $\ell \neq m$ in $\{1, ..., L\}$. Hence,

$$\begin{aligned} \sum_{k=1}^K x_k \left(\boldsymbol{\lambda}_k^T t\right)^2 &= \sum_{k=1}^K x_k \sum_{\ell=1}^L \lambda_{\ell k}^2 t_\ell^2 = \sum_{\ell=1}^L \sum_{k=1}^K x_k \lambda_{\ell k}^2 t_\ell^2 \\ &= \sum_{\ell=1}^L (\Lambda_\ell \otimes \Lambda_\ell)\, x t_\ell^2; \end{aligned}$$

and, therefore,

$$\kappa(t) = \sum_{k=1}^{K} \kappa_{X_k}(\boldsymbol{\lambda}_k^T t) + \sum_{\ell=1}^{L} \kappa_{U_\ell}(t_\ell) = \kappa_Y(t).$$

Now, define D as the diagonal of order K with diagonal entries: $d_k = \sqrt{1 - x_k}$. Rescale x if necessary such that D is invertible. Then:

$$\text{Var}(D^{-1}Z) = D^{-1}\,\text{diag}(1 - x_k)D^{-1} = I_K.$$

It follows that $\left(\Lambda D, D^{-1}Z, \widetilde{U}\right)$ is an alternative representation to (Λ, X, U).

Lastly, we have to show that these two representations are different. Note that, by the above construction, one can find an *infinity* of alternative representations $\left(\widetilde{\Lambda}, \widetilde{X}, \widetilde{U}\right)$ by appropriately rescaling x. Since the cardinal of $\mathcal{S}_K$ is *finite*, it follows that (Λ, X, U) is not identified. This ends the proof.

A.3 Proof of Theorem 3

To prove Theorem 3, 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).

Lemma 4 *Let K and L be any integers. Let $A_1, ..., A_L$ be matrices of $\mathbb{R}^{K\times K}$. Suppose that there exist $x^k = \left(x_1^k, ..., x_L^k\right)^T \in \mathbb{R}^L$ and $v^k \in \mathbb{R}^K$, $v_k \neq 0$, $k = 1, ..., K+1$, solutions to the joint diagonalization problem:*

$$x_\ell^k v^k = A_\ell v^k, \quad \forall \ell = 1, ..., L.$$

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

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

$$\begin{aligned}
\sum_{k=1}^{K} c_k x_\ell^k v^k &= \sum_{\ell=1}^{K} c_k A_\ell v^k \\
&= A_\ell \sum_{\ell=1}^{K} c_k v^k \\
&= A_\ell v^{K+1} \\
&= x_\ell^{K+1} v^{K+1} \\
&= x_\ell^{K+1} \left(\sum_{k=1}^{K} c_k v^k\right).
\end{aligned}$$

As $(v^1, ..., v^K)$ 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:

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

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

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

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

The proof of Theorem 3 easily follows.

Fourth-order moments. In the case where $U = 0$, second and fourth-order cumulant restrictions (8)-(12) yield:

$$\begin{aligned} \Omega_Y(\ell,m) &= \Lambda D_4 \operatorname{diag}(\Lambda_\ell \odot \Lambda_m) \Lambda^T, \quad (\ell,m) \in \overline{\Delta}_{L,2}, \\ \Sigma_Y &= \Lambda\Lambda^T. \end{aligned}$$

To show that Λ is identified from this system, let P be the Cholesky decomposition of Σ_Y, such that $PP^T = \Sigma_Y - \Sigma_U$, and P is a lower triangular $L \times K$ full-column rank matrix.

Then $P^-\Lambda$, where P^- is a generalized inverse of P (*e.g.* $P^- = \left[P^T P\right]^{-1} P^T$), is a matrix of joint orthonormal eigenvectors of:

$$P^-\Omega_Y(\ell,m)P^{-T} = P^-\Lambda D_4 \operatorname{diag}(\Lambda_\ell \odot \Lambda_m)\Lambda^T P^{-T}, \quad (\ell,m) \in \overline{\Delta}_{L,2}.$$

In general, there can be infinitely many joint eigenvectors to a set of matrices if all matrices have multiple roots. Lemma 4 shows that the problem of diagonalizing matrices $P^-\Omega_Y(\ell,m)P^{-T}$, $(\ell,m) \in \overline{\Delta}_{L,2}$, has a unique solution up to column sign and permutation if for all $(k,k') \in \{1...K\}^2$, $k \neq k'$, there exists $(\ell,m) \in \overline{\Delta}_{L,2}$ 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 Λ are linearly independent, this condition is always satisfied. It follows that V, and thus $\Lambda = PV$, are identified (up to column sign and permutation).

Third-order moments. The same argument applies to third-order cumulant matrices $\Gamma_Y(\ell)$. Indeed, in the noise-free case third-order restrictions (10) become

$$\Gamma_Y(\ell) = \Lambda D_3 \operatorname{diag}(\Lambda_\ell)\Lambda^T, \quad \ell \in \{1...L\},$$

where $\Gamma_Y(\ell)$, for all $\ell \in \{1...L\}$, is a $L \times L$ matrix of third-order cumulants of the data, and D_3 is the diagonal matrix of factor cumulants.

In this case, Lemma 4 shows that the problem of diagonalizing matrices $P^{-}\Gamma_Y(\ell)P^{-T}$, $\ell \in \{1...L\}$, has a unique solution up to column sign and permutation if for all $(k,k') \in \{1...K\}^2$, $k \neq k'$, there exists $\ell \in \{1...L\}$ 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. Lemma 4 shows that the problem of diagonalizing matrices $P^{-}\Omega_Y(\ell,m)P^{-T}$, $(\ell,m) \in \overline{\Delta}_{L,2}$, and $P^{-}\Gamma_Y(\ell)P^{-T}$, $\ell \in \{1...L\}$, has a unique solution up to column sign and permutation if for all $(k,k') \in \{1...K\}^2$, $k \neq k'$, there exists $(\ell,m) \in \overline{\Delta}_{L,2}$ 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 Λ that this condition is always satisfied.

A.4 Proof of Lemma 1

1. Let Ω_Y be defined by (16). As Q has rank K and D_4 is non singular, restrictions (17) imply that

$$\Omega_Y = \overline{Q}D_4Q^T,$$

has rank K. It follows that there exists $\overline{C} \in \mathbb{R}^{\#\overline{\Delta}_{L,2}\times(\#\overline{\Delta}_{L,2}-K)}$, full column rank, such that $\overline{C}^T\Omega_Y = 0$. Since D_4Q^T has rank K, it must also be that $\overline{C}^T\overline{Q} = 0$.

2. Let vech be the operator stacking all elements on and below the main diagonal of a $L \times L$ symmetric matrix column by column into a $\frac{L(L+1)}{2}$-vector. Then,

$$\begin{aligned}\operatorname{vech}(\Omega_Y(\ell,m)) &= \operatorname{vech}\left(\Lambda D_4 \operatorname{diag}(\Lambda_\ell \odot \Lambda_m)\Lambda^T + \delta_{\ell m}\kappa_4(U_\ell)\operatorname{Sp}_{L,\ell}\right), \\ &= \overline{Q}D_4(\Lambda_\ell \odot \Lambda_m) + \delta_{\ell m}\kappa_4(U_\ell)\operatorname{vech}\left(\operatorname{Sp}_{L,\ell}\right),\end{aligned}$$

where $\operatorname{Sp}_{L,\ell}$ is the sparse matrix of dimension (L,L) with only one 1 in position (ℓ,ℓ). It follows that

$$\overline{C}^T\operatorname{vech}(\Omega_Y(\ell,m)) = \delta_{\ell m}\kappa_4(U_\ell)\overline{C}_{(\ell,\ell)},$$

where $\overline{C}_{(\ell,\ell)}$ is the (ℓ,ℓ)th column of $\overline{C}^T$, and the columns of $\overline{C}^T$ (the rows of $\overline{C}$) are indexed by $(i,j) \in \overline{\Delta}_{L,2}$.

Moreover, the second-order restrictions are equivalently written as

$$\begin{aligned}\operatorname{vech}(\Sigma_Y) &= \operatorname{vech}\left(\Lambda\Lambda^T + \Sigma_U\right), \\ &= \overline{Q}\mathbf{1}_K + \operatorname{vech}(\Sigma_U),\end{aligned}$$

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

$$\overline{C}^T \operatorname{vech}(\Sigma_Y) = \overline{C}^T \operatorname{vech}(\Sigma_U) = \sum_{\ell=1}^{L} \operatorname{Var}(U_\ell)\,\overline{C}_{(\ell,\ell)}.$$

Lastly, consider

$$\begin{aligned}\operatorname{vech}(\Gamma_Y(\ell)) &= \left[\operatorname{Cum}(Y_\ell, Y_i, Y_j), (i,j) \in \overline{\Delta}_{L,2}\right] \\ &= \operatorname{vech}\left(\Lambda D_3 \operatorname{diag}(\Lambda_\ell)\Lambda^T + \kappa_3(U_\ell)\operatorname{Sp}_{L,\ell}\right).\end{aligned}$$

This vector of third-order moments of Y satisfies the equality

$$\operatorname{vech}(\Gamma_Y(\ell)) = \overline{Q}D_3\Lambda_\ell + \kappa_3(U_\ell)\operatorname{vech}\left(\operatorname{Sp}_{L,\ell}\right).$$

It follows that

$$\overline{C}^T \operatorname{vech}(\Gamma_Y(\ell)) = \kappa_3(U_\ell)\,\overline{C}_{(\ell,\ell)}.$$

3. Lastly, we show that the submatrix $\left[\overline{C}_{(1,1)}, ..., \overline{C}_{(L,L)}\right]^T \in \mathbb{R}^{L\times(\#\overline{\Delta}_{L,2}-K)}$ of $\overline{C}$ is full-row rank. To show this assertion, partition $\overline{C}$ as

$$\overline{C} = \begin{bmatrix} \overline{C}_{11} & \overline{C}_{12} \\ \overline{C}_{21} & \overline{C}_{22} \end{bmatrix},$$

with $\overline{C}_{11} \in \mathbb{R}^{\#\Delta_{L,2}\times(\#\Delta_{L,2}-K)}$, $\overline{C}_{12} \in \mathbb{R}^{\#\Delta_{L,2}\times L}$, $\overline{C}_{21} \in \mathbb{R}^{L\times(\#\Delta_{L,2}-K)}$ and $\overline{C}_{22} \in \mathbb{R}^{L\times L}$. To simplify the notations, suppose that rows $\overline{C}_{(1,1)}^T, ..., \overline{C}_{(L,L)}^T$ are located at the bottom of $\overline{C}$, so that $\left[\overline{C}_{21}, \overline{C}_{22}\right] = \left[\overline{C}_{(1,1)}, ..., \overline{C}_{(L,L)}\right]^T$. Without loss of generality, one can assume that $\overline{C}_{21} = 0$ and that $\overline{C}_{11}$ is a basis of the null space of Q^T. Now, suppose that $\overline{C}_{22}$ is singular. Then there exists a linear combination of the columns of $\overline{C}_{22}$ that is equal to zero. The same linear combination of the columns of $\overline{C}_{12}$ is both linearly independent of $\overline{C}_{11}$, as $\overline{C}$ is full-column rank, and orthogonal to the columns of Q. This contradicts the assumption that Q has rank K. Consequently, $\overline{C}_{22}$ is non singular and $\left[\overline{C}_{21}, \overline{C}_{22}\right]$ is full-row rank.

As matrix $\left[\overline{C}_{(1,1)}, ..., \overline{C}_{(L,L)}\right]^T$ is full-row rank, it follows that error variances are identified. Moreover, it also follows that $\overline{C}_{(\ell,\ell)} \neq 0$. So, $\kappa_3(U_\ell)$ and $\kappa_4(U_\ell)$ are identified.

This ends the proof of Lemma 1.

A.5 Proof of Lemma 2

1. The factor structure implies that

$$\begin{aligned}\Xi_Y &= \left[\Gamma_Y, \Omega_Y(1), ..., \Omega_Y(L)\right], \\ &= \Lambda\left[D_3Q^T, D_4\operatorname{diag}(\Lambda_1)Q^T, ..., D_4\operatorname{diag}(\Lambda_L)Q^T\right].\end{aligned}$$

Let $\gamma \in \mathbb{R}^K$ such that

$$\gamma^T \left[D_3 Q^T, D_4 \operatorname{diag}(\Lambda_1) Q^T, ..., D_4 \operatorname{diag}(\Lambda_L) Q^T \right] = 0.$$

As Q has rank K, it follows that $\gamma^T D_3 = 0$ and $\gamma^T D_4 \operatorname{diag}(\Lambda_\ell) = 0$ for all $\ell \in \{1...L\}$. Then, as Λ is full column rank, this implies that $\gamma^T D_4 = 0$. Lastly, as for all k either $\kappa_3(X_k) \neq 0$ or $\kappa_4(X_k) \neq 0$, it follows that $\gamma = 0$.

Therefore: $\left[D_3 Q^T, D_4 \operatorname{diag}(\Lambda_1) Q^T, ..., D_4 \operatorname{diag}(\Lambda_L) Q^T \right]$ as rank K. As Λ has rank K by assumption, Ξ_Y has also rank K.

Then, let $C \in \mathbb{R}^{L \times (L-K)}$ such that

$$C^T \Xi_Y = 0.$$

As $\left[D_3 Q^T, D_4 \operatorname{diag}(\Lambda_1) Q^T, ..., D_4 \operatorname{diag}(\Lambda_L) Q^T \right]$ is full row rank, it must also be that $C^T \Lambda = 0$.

2. One thus has

$$\begin{aligned} C^T \Sigma_Y &= C^T \Lambda \Lambda^T + C^T \Sigma_U, \\ &= C^T \Sigma_U \\ &= [\operatorname{Var}(U_1) C_1, ..., \operatorname{Var}(U_L) C_L] \end{aligned}$$

or

$$C^T \begin{pmatrix} \operatorname{Cov}(Y_1, Y_\ell) \\ \vdots \\ \operatorname{Cov}(Y_L, Y_\ell) \end{pmatrix} = \operatorname{Var}(U_\ell) C_\ell, \quad \ell = 1, ..., L,$$

where C_ℓ^T is the ℓth row of C.

Moreover, matrices $\Gamma_Y(\ell)$ defined by (9) satisfy the equality:

$$\Gamma_Y(\ell) = \Lambda D_3 \operatorname{diag}(\Lambda_\ell) \Lambda^T + \kappa_3(U_\ell) \operatorname{Sp}_{L,\ell}.$$

Hence

$$\begin{aligned} C^T \Gamma_Y(\ell) &= C^T \Lambda D_3 \operatorname{diag}(\Lambda_\ell) \Lambda^T + \kappa_3(U_\ell) C^T \operatorname{Sp}_{L,\ell}, \\ &= \kappa_3(U_\ell) C^T \operatorname{Sp}_{L,\ell}, \end{aligned}$$

or

$$C^T \begin{pmatrix} \operatorname{Cum}(Y_1, Y_\ell, Y_\ell) \\ \vdots \\ \operatorname{Cum}(Y_L, Y_\ell, Y_\ell) \end{pmatrix} = \kappa_3(U_\ell) C_\ell.$$

Lastly,

$$\Omega_Y(\ell, \ell) = \Lambda D_4 \operatorname{diag}(\Lambda_\ell \odot \Lambda_\ell) \Lambda^T + \kappa_4(U_\ell) \operatorname{Sp}_{L,\ell}$$

implies that

$$C^T \Omega_Y(\ell, \ell) = \kappa_4(U_\ell) C^T \operatorname{Sp}_{L,\ell}$$

and

$$C^T \begin{pmatrix} \mathrm{Cum}\,(Y_1, Y_\ell, Y_\ell, Y_\ell) \\ \vdots \\ \mathrm{Cum}\,(Y_L, Y_\ell, Y_\ell, Y_\ell) \end{pmatrix} = \kappa_4\,(U_\ell)\, C_\ell.$$

3. Let $\Lambda_{-\ell}$ be matrix Λ without its ℓth row. As $\Lambda_{-\ell}$ has rank K by assumption, it follows from equality $C^T\Lambda = 0$ that $C_\ell \neq 0$. Otherwise, one would have $C_{-\ell}^T\Lambda_{-\ell} = 0$ for a full $(L-1)\times(L-K)$ matrix $C_{-\ell}$, contradicting the assumption that $\mathrm{rank}(\Lambda_{-\ell}) = K$. Hence $\mathrm{Var}\,(U_\ell)$, $\kappa_3\,(U_\ell)$ and $\kappa_4\,(U_\ell)$ are identified.

This ends the proof of Lemmas 2 and 3.

B The JADE algorithm

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

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

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

$$\sum_{k=1}^{K} \mathrm{off}(U A_k U^T), \tag{B4}$$

with respect to U orthonormal.

Let $\theta \in [-\pi, \pi]$, let $(i,j) \in \{1...L\}^2$ and let $R_{ij}(\theta)$ be the $L \times L$ matrix equal to zero everywhere 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:

$$O_{i,j}(\theta) = \sum_{k=1}^{K} \mathrm{off}\left(R_{ij}(\theta) A_k R_{ij}(\theta)^T\right).$$

Lastly, let $h_{i,j}(A) = (a_{ii} - aij, a_{ij} + a_{ji})$, and let:

$$G_{i,j} = \sum_{k=1}^{K} h_{i,j}^T(A_k) h_{i,j}(A_k) = (g_{ij})_{i,j=1,2}.$$

Then, 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 (B4) by the following algorithm:

1. Start with $U(0) = I_L$.

2. Begin loop on step s.

3. Begin loop on (i, j).
4. Compute $G_{i,j}$.
5. Compute θ_0.
6. If θ_0 is different enough from zero, continue. Else stop.
7. Compute $R_{ij}(\theta_0)A_k R_{ij}(\theta_0)^T$ and modify $\mathcal{A}$ consequently.
8. Update $U(s)$ as $U(s+1) = R_{ij}(\theta_0)U(s)$.
9. End loop on (i, j).
10. End loop on s.

C Asymptotic Theory of the JADE Estimator

C.1 First-Order conditions of the JADE algorithm

Let A_j, $j = 1...J$ be a set of K symmetric matrices. Cardoso and Souloumiac's JADE algorithm finds an orthogonal basis in which the A_j matrices are simultaneously approximately diagonal. The JADE estimator solves:

$$\min_{\Lambda} \sum_{j=1}^{J} \text{off}(\Lambda^T A_j \Lambda),$$

where $\text{off}(M) = \sum_{i \neq j} m_{ij}^2$, and the maximization is taken with respect to $K \times K$ orthogonal matrices $\Lambda = (\lambda_1, ..., \lambda_K)$, where λ_k is the $K \times 1$ column of Λ.

The Lagrangian of this system writes:

$$\mathcal{L} = \sum_{j}^{J} \sum_{m \neq k} (\lambda_k^T A_j \lambda_m)^2 + \sum_k \gamma_k (\lambda_k^T \lambda_k - 1) + \sum_{m \neq k} \gamma_{mk} \lambda_k^T \lambda_m,$$

where scalars γ_k and γ_{mk} are the Lagrange multipliers associated with the orthogonality constraints.

Differentiating the Lagrangian with respect to λ_m, for $m = 1...K$, yields:

$$0 = \frac{\partial \mathcal{L}}{\partial \lambda_m} = 4 \sum_{j}^{J} \lambda_m^T A_j \left(\sum_{k \neq m} \lambda_k \lambda_k^T \right) A_j + 2\gamma_m \lambda_m^T + \sum_{m \neq k} (\gamma_{mk} + \gamma_{km}) \lambda_k^T.$$

At this stage, it is convenient to remark that $\gamma_{mk} + \gamma_{km}$ is symmetric. Hence, for all $k \neq m$, right-multiplication by λ_n yields:

$$\sum_{j}^{J} \lambda_m^T A_j \left(\sum_{k \neq m} \lambda_k \lambda_k^T \right) A_j \lambda_n = \sum_{j}^{J} \lambda_n^T A_j \left(\sum_{k \neq n} \lambda_k \lambda_k^T \right) A_j \lambda_m.$$

As A_j^2 is symmetric, and as $\sum_{k=1}^{K} \lambda_k \lambda_k^T = I_K$ from the orthogonality constraints, this expressions simplifies into:

$$\sum_{j}^{J} \lambda_m^T A_j \left(\lambda_k \lambda_k^T - \lambda_m \lambda_m^T \right) A_j \lambda_k = 0,$$

or:

$$\sum_{j}^{J} \lambda_m^T A_j \lambda_k \left(\lambda_k^T A_j \lambda_k - \lambda_m^T A_j \lambda_m \right) = 0, \tag{C5}$$

for all $m \neq k$. System (C5) comprises $K(K-1)/2$ equations. The JADE algorithm solves this moment condition under the $K(K+1)/2$ orthogonality constraints.

C.2 Asymptotic distribution

Let $\widehat{A}_1, ..., \widehat{A}_J$ be a set of $K \times K$ of root-N consistent and asymptotically normally distributed estimators of matrices $A_1, ..., A_J$. Let $\text{Vec}(A)$ be the JK^2-dimensional vector obtained by stacking vectors $\text{vec}(A_1)$,..., $\text{vec}(A_J)$, and let $\text{Vec}(\widehat{A})$ be similarly defined. It is assumed:

$$N^{1/2} \left(\text{Vec}(\widehat{A}) - \text{Vec}(A) \right) \underset{N \to \infty}{\to} \mathcal{N}(0, \Omega_A),$$

where Ω_A is a $JK^2 \times JK^2$ positive definite matrix. Let Λ be the true matrix of eigenvectors, and let $\widehat{\Lambda}$ be its JADE estimate. It follows from standard GMM arguments (*e.g.* Newey and McFadden, 1994) that $\widehat{\Lambda}$ is root-N consistent and asymptotically normal.

Then from (C5), $\widehat{\Lambda}$ solves:

$$\sum_{j}^{J} \widehat{\lambda}_m^T \widehat{A}_j \widehat{\lambda}_k \left(\widehat{\lambda}_k^T \widehat{A}_j \widehat{\lambda}_k - \widehat{\lambda}_m^T \widehat{A}_j \widehat{\lambda}_m \right) = 0.$$

A first-order expansion of this equality around Λ yields:

$$\begin{aligned} 0 \quad = \quad & \sum_{j}^{J} \lambda_m^T \widehat{A}_j \lambda_k \left(\lambda_k^T \widehat{A}_j \lambda_k - \lambda_m^T \widehat{A}_j \lambda_m \right) + \sum_{j}^{J} \left(\lambda_k^T \widehat{A}_j \lambda_k - \lambda_m^T \widehat{A}_j \lambda_m \right) \left(\lambda_m^T \widehat{A}_j (\widehat{\lambda}_k - \lambda_k) + \lambda_k^T \widehat{A}_j (\widehat{\lambda}_m - \lambda_m) \right) \\ & + \sum_{j}^{J} \lambda_m^T \widehat{A}_j \lambda_k \left(\lambda_k^T \widehat{A}_j (\widehat{\lambda}_k - \lambda_k) - \lambda_m^T \widehat{A}_j (\widehat{\lambda}_m - \lambda_m) \right). \end{aligned} \tag{C6}$$

Note that, as $\Lambda^T A_j \Lambda$ is asymptotically diagonal, the third term in (C6) is asymptotically zero. As for the second term, let us denote as $d_{jk} = \lambda_k^T A_j \lambda_k$ the diagonal terms of $\Lambda^T A_j \Lambda$, and let us define $X_k = (x_{mk})_m \equiv \Lambda^T (\widehat{\lambda}_k - \lambda_k)$. Note that a first-order expansion of the orthogonality condition $\Lambda^T \Lambda = I_K$ yields $x_{mk} + x_{km} = 0$, for all m, k.

Therefore, the second term in (C6) is asymptotically equivalent to:

$$\sum_{j} (d_{jk} - d_{jm})(d_{jk} x_{km} + d_{jm} x_{mk}),$$

which can be conveniently rewritten as:

$$\left(- \sum_{j} (d_{jk} - d_{jm})^2 \right) x_{mk}.$$

Lastly, let us consider the first term in (C6). Let V be the $L \times L$ matrix, of which the (k,k) elements are zero, and the (k,m), $k \neq m$, elements, are:

$$V_{km} = \sum_{j}^{J} \lambda_m^T \widehat{A}_j \lambda_k \left(\lambda_k^T \widehat{A}_j \lambda_k - \lambda_m^T \widehat{A}_j \lambda_m \right).$$

Standard calculations yield the variance-covariance matrix of $\text{vec}(V)$ as

$$\text{Var}\left(\text{vec}(\Lambda)\right) = \left[V(D_1), ..., V(D_J)\right] \Omega_A \left[V(D_1), ..., V(D_J)\right]^T,$$

where for all diagonal matrix D with diagonal coefficients d_k, $V(D) = \text{diag}(\text{vec}(S(D)))$, with $S(D)$ a $K \times K$ matrix with zero elements on the diagonal, and (k, m) element equal to $d_k - d_m$.

For convenience, we shall define matrix $W = [\text{diag}(\text{vec}(R(D_1))), ..., \text{diag}(\text{vec}(R(D_J)))]$, with $R(D_j)$ a $K \times K$ matrix with zero elements on the diagonal, and (k, m) element equal to $d_{jk} - d_{jm} / \sum_{j'} (d_{j'k} - d_{j'm})^2$.

Combining the previous result, and recalling that $(\widehat{\lambda}_k - \lambda_k) = \Lambda X_k$, one obtains the asymptotic variance-covariance matrix of $\text{vec}(\Lambda)$ as:

$$\text{Var}\left(\text{vec}(\Lambda)\right) = (I_K \otimes \Lambda) W (I_J \otimes \Lambda^T \otimes \Lambda^T) \Omega_A (I_J \otimes \Lambda \otimes \Lambda) W^T (I_K \otimes \Lambda^T).$$

This ends the proof.

D Robin and Smith's (2000) rank test

Let $\widehat{B}$ be a root-N consistent estimator of a (p, q), $p \geq q$, matrix B, such that

$$N^{1/2} \,\text{vec}\left(\widehat{B} - B\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_{\text{vec}\left(\widehat{B}\right)}\right),$$

where $\Sigma_{\text{vec}\left(\widehat{B}\right)}$ is definite and $\text{rank}\left(\Sigma_{\text{vec}\left(\widehat{B}\right)}\right) = s$, $0 < s \leq pq$.[22] Let $\widehat{\Sigma}_{\text{vec}\left(\widehat{B}\right)}$ be a consistent estimate of $\Sigma_{\text{vec}\left(\widehat{B}\right)}$. Let $\widehat{B} = \widehat{C}\widehat{D}\widehat{E}^T$ be the singular value decomposition of $\widehat{B}$, where $\widehat{C}$ and $\widehat{E}$ are (p, p) and (q, q) orthogonal matrices and $\widehat{D}$ is a (q, p) diagonal matrix. Let $\widehat{d}_1 \geq ... \geq \widehat{d}_K$ denote the diagonal entries of $\widehat{D}^2$ (the eigenvalues of $\widehat{B}^T\widehat{B}$). For a given null hypothesis: $H_0^r : K = r$, the statistics

$$\mathcal{CRT}_r \equiv N \sum_{i=r+1}^{q} \widehat{d}_i$$

has the same limiting distribution as $\sum_{i=1}^{t} d_i^r Z_i^2$, where $d_1^r \geq ... \geq d_t^r$, $t \leq \min\{s, (p-r)(q-r)\}$, are the non-zero ordered eigenvalues of the matrix

$$(\widehat{E}_{q-r} \otimes \widehat{C}_{p-r})^T \widehat{\Sigma}_{\text{vec}\left(\widehat{B}\right)} (\widehat{E}_{q-r} \otimes \widehat{C}_{p-r}),$$

where $\widehat{E}_{q-r}$ and $\widehat{C}_{p-r}$ are the last $q-r$ and $p-r$ columns of $\widehat{E}$ and $\widehat{C}$, respectively, and $\{Z_i\}_{i=1}^t$ are independent standard normal variates.

To estimate K, we apply the following procedure. Start with $r = 0$. Test H_0^1 against $\widetilde{H}_0^1 : K > 0$. If H_0^1 is rejected, test H_0^2 against $\widetilde{H}_0^2 : K > 1$. And so on until one accepts H_0^r against $\widetilde{H}_0^r : K > r$. The test p-values can be approximated by drawing many independent values of the limiting statistics $\sum_{i=1}^{t} d_i^r Z_i^2$. This procedure delivers a consistent estimate of K if the asymptotic sizes α_N^r used for the sequential tests are such that $\alpha_N^r = o(1)$ and $-N^{-1} \ln \alpha_N^r = o(1)$.

[22] Note that $s < \dim(V)$ because of the symmetry properties of Γ_Y and Ω_Y.

References

[1] AIGNER, D. J., C. HSIAO, A. KAPTEYN, AND T. WANSBEEK (1984): "Latent Variable Models in Econometrics," in *Handbook of Econometrics*, Vol. II, ed. by Z. Griliches and M. D. Intriligator. Amsterdam: North Holland.

[2] ALTONJI, J.G. and L.M. SEGAL (1994): "Small Sample Bias in GMM Estimation of Covariance Structures", NBER Technical Working Paper no. 156, National Bureau of Economic Research, Cambridge, MA.

[3] ALTONJI, J.G. and L.M. SEGAL (1996): "Small Sample Bias in GMM Estimation of Covariance Structures", *Journal of Business and Economic Statistics*, 14, 353-366.

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

[5] ASHENFELTER, O. and J. MOONEY, "Graduate Education, Ability and Earnings," *Review of Economics and Statistics*, 50 (February 1968): 78-86.

[6] ANDERSON, T. W. (1984): "An Introduction to Multivariate Statistical Analysis, New York: Wiley.

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

[8] BACK, A. D. and A. S. WEIGEND (1997), "A First Application of Independent Component Analysis to Extracting Structure from Stock Returns," *International Journal of Neural Systems*, Vol. 8, No.4, 473-484.

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

[10] CARDOSO J.-F. (1998): "Blind signal separation : statistical principles," *Proc. IEEE,* 9(10), 2009-2025.

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

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

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

[14] CARNEIRO, P., K. T. HANSEN, AND J. J. HECKMAN (2002): "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.

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

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

[17] CUNHA, F., J.J. HECKMAN and S. NAVARRO (2005), "Seperating uncertainty from heterogeneity in life cycle earnings," *Oxford Economic Papers,* 57, 191-261.

[18] 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.

[19] DE LATHAUWER, L. (2003): "Simultaneous Matrix Diagonalization: the Overcomplete Case," *Proc. of the 4th International Symposium on ICA and Blind Signal Separation, Nara, Japan*, 812-825.

[20] DOZ, C. and E. RENAULT (2005): "Factor Stochastic Volatility in Mean Models: a GMM approach," mimeo.

[21] DUFFIE, D. and J. PAN (1997): "An Overview of Value at Risk," *Journal of Derivatives*, 7-49, reprinted in *Options Markets*, edited by G. Constantinides and A. G. Malliaris, London: Edward Elgar, 2001.

[22] 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.

[23] 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.

[24] FLURY, B. (1984): "Common Principal Components in K Groups," *J. Am. Stat. Ass.*, 79, 892-898.

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

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

[27] HAMERMESH, D. S. and S. G. DONALD, "The Effects of College Curriculum on Earnings: Accounting for Non-ignrable Non-response Bias," Texas University, *mimeo.*

[28] HECKMAN, J.J. and S. NAVARRO (2005), "Dynamic Discrete Choice and Dynamic Treatment Effects," University of Chicago, *mimeo.*

[29] HYVARINEN A. (1999), "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis," *IEEE Transactions on Neural Networks*, 10(3):626–634.

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

[31] IKEDA, S. and K. TOYAMA (2000), "Independent component analysis for noisy data–MEG data analysis," *Neural Networks*, Vol.13, No.10, 1063-1074.

[32] LAWLEY, D.N, and A.E. MAXWELL (1971): *Factor Analysis as a Statistical Method.* London: Butterworth.

[33] 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.

[34] LEWBEL, A. (2004): "Identification of Heteroskedastic Endogenous or Mismeasured Regressor Models," Boston College, *mimeo.*

[35] LIN, Y.-N. and K. HUNG (2005): "The Volatility Risk Premium Embedded in S&P 500 Index Returns," mimeo

[36] MADANSKY, A. (1959): "The Fitting of Straight Lines When Both Variables are Subject to Error," *Journal of the American Statistical Assiociation*, 54, 173-205.

[37] 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.

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

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

[40] ROBIN, J.M., R.J. SMITH (2000): "Tests of rank," *Econometric Theory*, vol. 16, 151-175

[41] SPEARMAN, C. (1904): "General intelligence, objectively determined and measured," *American Journal of Psychology*, 15, 201-293.

[42] SPIEGELMAN, C. (1979): "On Estimating the Slope of a Straight Line when Both Variables are Subject to Error," *Annals of Statistics*, 7, 201-206.

[43] VAN MONTFORT, K., A. MOOIJAART, AND J. DE LEEUW (1989): "Estimation of Regression Coefficients with the Help of Characteristic Functions," *Journal of Econometrics*, 41, 267-278.

[44] XU, Lei (2000), " Temporal BYY Learning for State Space Approach, Hidden Markov Model and Blind Source Separation ", *IEEE Trans on Signal Processing*, Vol. 48, No. 7, 2132-2144.

[45] XU, Lei (2001), "BYY Harmony Learning, Independent State Space and Generalized APT Financial Analyses", *IEEE Trans. on Neural Networks*, Vol. 12, No.4, 822-849. An Errata to this paper is given on IEEE Trans. on Neural Networks, Vol. 13, No.4, 1023, July, 2002.

[46] XU, L. (2003), "Independent Component Analysis and Extensions with Noise and Time: A Bayesian Ying-Yang Learning Perspective", *Neural Information Processing Letters and Reviews*, Vol.1, No.1, 1-52.

[47] YIP, F. and L. XU (2000), "An Application of Independent Component Analysis in the Arbitrage Pricing Theory," IEEE-INNS-ENNS *International Joint Conference on Neural Networks* (IJCNN'00)-Vol. 5, 5279,