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

March 2006

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. High-order moments are shown to yield full identification of the matrix of factor loadings if factor distributions are sufficiently skewed or kurtotic. 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 microdata on wages and education, and to financial data on stock returns.

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. 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. Linear factor models have also been used in nonlinear empirical microeconomic models. For example, Carneiro, Hansen and Heckman's (2003) Roy model of educational choice is a successful application of factor models 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 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 is 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. In the same spirit, Carneiro, Hansen and Heckman (2003) assume that there is at least two specific measurements for each factor.

In this paper, we show that these 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, which implies that factors are *not Gaussian*. If $K \leq L$, we show that the matrix of factor loadings Λ

[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.

is generically identified from second, third and fourth-order moments of the data. If $K < L$, we show that Λ is identified from second and third-order moments only. In both cases, identification is unambiguously defined up to multiplication of each column by ± 1 and column permutations.

The importance of the assumptions of independence and non normality for the identification of one-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 generalization 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 for *blind signal separation*.[3] In this vast literature, one of the most popular methods is Cardoso and Souloumiac's (1993) JADE algorithm. This is a joint diagonalization 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 an important topic in the neural networks literature and Hyvärinen's (1999) FastICA algorithm has become another very popular algorithm.[4]

One serious drawback of ICA, at least for econometric applications, is that it rules out measurement errors. The estimated model is $Y = \Lambda X$, with $K = L$, not $Y = \Lambda X + U$. Neglecting noise can be a source of severe biases, as we shall show. All existing extensions of ICA allowing for noise make parametric assumptions on the distributions of errors (usually Gaussian) and factors (usually Gaussian mixtures).[5] As far as we know, our paper is the

[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 than 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 designation "Independent Component Analysis" was first proposed par Comon (1994). See Hyvärinen *et al.* (2001) and Cardoso (1999) for surveys.

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

[5]For example, Moulines *et al.* (1997) and Attias (1999) use 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 method which combines PCA and JADE to reduce the size of the noise. As such, the estimator they propose is still inconsistent in the presence of noise.

first, out of a long list of contributions, to propose a semiparametric statistical procedure for consistently estimating a $K \times K$ matrix of factor loadings from data moments in a linear factor model with error distributions of unknown form.

We develop an algebraic procedure that builds on Cardoso and Souloumiac's JADE algorithm. Our *quasi-JADE* algorithm proceeds in two stages: First, we estimate the second, third and fourth-order error moments, which 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 diagonalization algorithm to the "whitened" data. Notice that we therefore do not need to assume full independence between factor and error components, only that they are orthogonal up to third or fourth order.

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 first discuss the estimation of the number of common factors using Robin and Smith's (2000) rank test. We then present Cardoso and Souloumiac's (1993) JADE algorithm and study its asymptotic properties. Lastly, we introduce 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 two rather different empirical setups.

We first estimate the returns to schooling in France, using microdata from the French Labor Force Survey. Our method allows to identify two factors of individual wages and education. Interestingly, while the first factor has a positive effect on wages, the second factor is positively related to education, yet negatively to wages. This is evidence that there exist individual characteristics which are valued by the education institution but not by the labor market. Moreover, the exhibited factor structure is consistent with the standard model of education returns if one allows measurement errors on the education measure and unobserved heterogeneity. We then reconsider Fama and French's (1993) factor analysis of US stock returns.[6] Fixing the number of factors to three, as Fama and French do, we apply our blind signal extraction procedure and estimate independent (up to third or fourth order) factor components which turn out to be

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

strongly correlated with those derived by Fama and French.

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).[7] 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). Factors and errors are unobserved.

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

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. Moreover, if two columns of Λ are proportional, then one can aggregate the corresponding factors into a single one.

Assumption A2 *Any column of Λ contains at least two non zero entries and any two columns of Λ are not proportional.*

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 normalized to the identity matrix I_K. Thirdly, even if $\Sigma_X = I_K$ and Σ_U is diagonal, Λ is

[7] In the rest of the paper, M^T denotes the transpose of matrix M.

identified only up to an orthogonal matrix; that is, if Λ satisfies the covariance restrictions, then so does ΛP, for any orthogonal matrix P. This indeterminacy characterizes both ordinary Factor Analysis and PCA (e.g. Lawley and Maxwell, 1971).

In this paper, we maintain the assumptions that $\Sigma_X = I_K$ and Σ_U is diagonal, and we strengthen the absence of correlations between factors and errors by making them independent.

Assumption A3 *Factors have unit variances.*

Assumption A4 *All factor and error variables are mutually independent.*

A triple (Λ, X, U), satisfying Assumptions A1, A2, A3 and A4 and such that $\Lambda X + U$ has the same distribution as Y is called a *representation* of Y.

The aim of this section is to determine conditions under which Λ, the matrix of factor loadings, is identified. We shall start by a general semiparametric identification result stating conditions under which the family of equivalent representations (Λ, X, U) (i.e. such that $\Lambda X + U$ has the same distribution as Y) share the same Λ. Then, we shall investigate the conditions under which one can uniquely reconstruct Λ from a small number of moments of Y instead of its entire distribution.

2.1 A first identification result

For any value of 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 representation. Clearly, for all $S \in \mathcal{S}_K$, $\left(\Lambda S, S^T X, U\right)$ is another representation. Hence, identification has to be defined modulo the set $\mathcal{S}_K$.

Notice 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 modulo $\mathcal{S}_K$, instead of modulo $\mathcal{O}_K$, will result in a considerable reduction of model indeterminacy.

We first state the following general identification theorem.

Theorem 1 ***(Sufficient conditions for semiparametric identification of*** Λ***)*** *Let* (Λ, X, U) *be a representation of a vector of measurements* Y*, such that the components of* X *are not normal. Let* $(\widetilde{\Lambda}, \widetilde{X}, \widetilde{U})$ *be an equivalent representation. The following two assertions are true:*

(i) *Every column of* Λ *is a scalar multiple of a column of* $\widetilde{\Lambda}$*.*

(ii) *Assume, in addition, that the* $\frac{L(L-1)}{2} \times K$ *matrix*

$$Q(\Lambda) = [\lambda_{\ell 1}\lambda_{m1}, ..., \lambda_{\ell K}\lambda_{mK}; \ell, m = 1, ..., L, \ell < m]$$

is full column rank. Then matrix $\widetilde{\Lambda}$ *differs from matrix* Λ *by a sign-permutation matrix.*

Theorem 1 follows by a straightforward application of a result due to Kagan, Linnik and Rao (1973) that is stated in Section A.1 of the Mathematical Appendix (see also Eriksson and Koivunen, 2003). This theorem shows that, if factors are non normal, then Λ is identified up to a multiplicative diagonal matrix and column permutations (i.e. if Λ works, then so does ΛDP with D diagonal and P a permutation matrix). Recovering the right scale shift requires an additional restriction ($Q(\Lambda)$ full column rank) which implies that $K \leq \frac{L(L-1)}{2}$. The rest of this section aims at demonstrating that, under these two assumptions, only a few moments of Y suffice to identify Λ up to column-sign and column-permutation.

2.2 Moment restrictions

Define multivariate cumulants as follows. The second-order cumulant of a couple of zero-mean random variables (Z_1, Z_2) is equal to their covariance:

$$\text{Cum}(Z_1, Z_2) = \mathbb{E}(Z_1 Z_2), \tag{2}$$

The third-order cumulant of (Z_1, Z_2, Z_3) is the third-order moment:

$$\text{Cum}(Z_1, Z_2, Z_3) = \mathbb{E}(Z_1 Z_2 Z_3). \tag{3}$$

The fourth-order, multivariate cumulant of (Z_1, Z_2, Z_3, Z_4) is

$$\begin{aligned} \text{Cum}(Z_1, Z_2, Z_3, Z_4) &= \mathbb{E}(Z_1 Z_2 Z_3 Z_4) - \mathbb{E}(Z_1 Z_2)\mathbb{E}(Z_3 Z_4) \\ &\quad -\mathbb{E}(Z_1 Z_3)\mathbb{E}(Z_2 Z_4) - \mathbb{E}(Z_1 Z_4)\mathbb{E}(Z_2 Z_3). \end{aligned} \tag{4}$$

Straightforward algebra shows that the linear factor structure and the assumption of orthogonality between factors and errors up to fourth order, that follows from independence, implies the following set of multilinear restrictions. Let $\ell_1, ..., \ell_p$ be p indices in $\{1, ..., L\}$, for $p = 2, 3, 4$. Then,

$$\begin{aligned} \text{Cum}\left(Y_{\ell_1}, ..., Y_{\ell_p}\right) &= \sum_{k=1}^{K} \left(\prod_{i=1}^{p} \lambda_{\ell_i,k} \right) \kappa_p\left(X_k\right) + \text{Cum}\left(U_{\ell_1}, ..., U_{\ell_p}\right) \\ &= \begin{cases} \sum_{k=1}^{K} \left(\lambda_{\ell k}\right)^p \kappa_p\left(X_k\right) + \kappa_p\left(U_\ell\right), & \text{if } \ell_1 = ... = \ell_p \equiv \ell, \\ \sum_{k=1}^{K} \left(\prod_{i=1}^{p} \lambda_{\ell_i,k}\right) \kappa_p\left(X_k\right), & \text{otherwise,} \end{cases} \end{aligned} \tag{5}$$

where $\kappa_p(Z) = \text{Cum}(Z, Z, ..., Z)$ (repeat Z p times) denotes the pth cumulant of a univariate random variable Z.[8]

2.3 Algebraic structure of moment restrictions

Moment restrictions of all orders have a common multilinear structure which can be conveniently expressed in matrix form, as in ordinary Factor Analysis.

Using (5) with $p = 2$, second-order restrictions take the usual matrix form:

$$\Sigma_Y = \Lambda\Lambda^T + \Sigma_U, \tag{6}$$

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

Now, define the following $L \times L$ matrices of third-order cumulants

$$\Gamma_Y(\ell) = \left[\text{Cum}\left(Y_i, Y_\ell, Y_j\right)\right]_{(i,j)\in\{1,...,L\}^2}, \quad \ell \in \{1...L\}. \tag{7}$$

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

$$\Gamma_Y(\ell) = \Lambda D_3 \,\text{diag}\left(\Lambda_\ell\right) \Lambda^T + \kappa_3\left(U_\ell\right) \text{Sp}_{L,\ell}, \tag{8}$$

where $\Lambda_\ell^T \in \mathbb{R}^{K\times 1}$ is the ℓth row of Λ, D_3 is the diagonal matrix with $\kappa_3(X_k)$ 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).

[8] For a zero mean variable Z, the first four univariate cumulants are thus defined as:

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

Lastly, let $\overline{\Delta}_L = \left\{(i,j) \in \{1,...,L\}^2, i \leq j\right\}$ and $\Delta_L = \left\{(i,j) \in \{1,...,L\}^2, i < j\right\}$. One can also define the following $L \times L$ matrices of fourth-order cumulants:

$$\Omega_Y(\ell, m) = \left[\text{Cum}(Y_i, Y_\ell, Y_m, Y_j)\right]_{(i,j)\in\{1,...,L\}^2}, \quad (\ell, m) \in \overline{\Delta}_L. \tag{9}$$

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

$$\Omega_Y(\ell, m) = \Lambda D_4 \,\text{diag}(\Lambda_\ell \odot \Lambda_m)\,\Lambda^T + \delta_{\ell m}\kappa_4(U_\ell)\,\text{Sp}_{L,\ell}, \tag{10}$$

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.

Restrictions (8) and (10) clearly show the same algebraic structure as restriction (6). This property will be the source of identifying restrictions yielding exact identification of the factor structure.

2.4 Moment-based 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.

We first consider the case of factor models without errors. In this case, second, third and fourth-order restrictions (6), (8), (10) imply that matrix Λ satisfies simultaneously

$$\Sigma_Y = \Lambda\Lambda^T, \tag{11}$$

$$\Gamma_Y(\ell) = \Lambda D_3 \,\text{diag}(\Lambda_\ell)\,\Lambda^T, \quad \ell \in \{1...L\}, \tag{12}$$

$$\Omega_Y(\ell, m) = \Lambda D_4 \,\text{diag}(\Lambda_\ell \odot \Lambda_m)\,\Lambda^T, \quad (\ell, m) \in \overline{\Delta}_L. \tag{13}$$

Left and right-multiplying (11), (12) and (13) 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:

$$\Sigma_Y^{-1/2}\Gamma_Y(\ell)\,\Sigma_Y^{-T/2} = VD_3 \,\text{diag}(\Lambda_\ell)\,V^T, \quad \ell \in \{1...L\},$$

$$\Sigma_Y^{-1/2}\Omega_Y(\ell, m)\,\Sigma_Y^{-T/2} = VD_4 \,\text{diag}(\Lambda_\ell \odot \Lambda_m)\,V^T, \quad (\ell, m) \in \overline{\Delta}_L,$$

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

Theorem 2 ***(Moment-based identification of*** Λ ***in the noise-free case)*** *Assume that (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 (11) and (13).

If (iv') at most one factor variable has zero skewness, then factor loadings are identified from second and third-order moment restrictions (11) and (12).

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 (11), (12) and (13).*

The proof is in the mathematical appendix. Theorem 2 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 (11) and (13) as the basis of their JADE algorithm. In usual applications of ICA methods, factors are thought to be symmetric. For this reason, third-order information is *a priori* neglected in that literature. However, there is no strong argument for 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 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 first application in Section 5 will provide an illustration of this remark.

2.5 Moment-based 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 some fourth-order cumulants of factors are zero or not.

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

$$\Omega_Y = \left[\text{Cum}\left(Y_i, Y_j, Y_\ell, Y_m\right); (i,j) \in \overline{\Delta}_L, (\ell,m) \in \Delta_L\right] \in \mathbb{R}^{\frac{L(L+1)}{2} \times \frac{L(L-1)}{2}}. \tag{14}$$

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

$$\Omega_Y = \overline{Q} D_4 Q^T, \tag{15}$$

where

$$Q \equiv Q(\Lambda) = \left[\lambda_{\ell k}\lambda_{mk}; (\ell,m) \in \Delta_L, k \in \{1,...,K\}\right] \in \mathbb{R}^{\frac{L(L-1)}{2} \times K}, \tag{16}$$

$$\overline{Q} \equiv \overline{Q}(\Lambda) = \left[\lambda_{\ell k}\lambda_{mk}; (\ell,m) \in \overline{\Delta}_L, k \in \{1,...,K\}\right] \in \mathbb{R}^{\frac{L(L+1)}{2} \times K}. \tag{17}$$

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:*

$$\overline{C}^T \text{vech}\left(\Sigma_Y\right) = \sum_{\ell=1}^{L} \text{Var}\left(U_\ell\right) \overline{C}_{(\ell,\ell)}, \tag{18}$$

$$\overline{C}^T \text{vech}\left(\Gamma_Y(\ell)\right) = \kappa_3\left(U_\ell\right) \overline{C}_{(\ell,\ell)}, \tag{19}$$

$$\overline{C}^T \text{vech}\left(\Omega_Y(\ell,\ell)\right) = \kappa_4\left(U_\ell\right) \overline{C}_{(\ell,\ell)}, \tag{20}$$

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.[9]

3. *Matrix* $\left[\overline{C}_{(1,1)},...,\overline{C}_{(L,L)}\right]$ *is full rank and* $\mathrm{Var}\left(U_\ell\right)$, $\kappa_3\left(U_\ell\right)$ *and* $\kappa_4\left(U_\ell\right)$ *are uniquely defined by identification restrictions (18), (19) and (20).*

The proof is in Section A.3 of the mathematical appendix. The following theorem then follows straightforwardly.

Theorem 3 ***(Sufficient conditions for moment-based identification of*** Λ ***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.*

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

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

$$\Omega_Y\left(j\right) = \left[\mathrm{Cum}\left(Y_i, Y_j, Y_\ell, Y_m\right); i \in \{1,...,L\}, (\ell,m) \in \Delta_L\right] \in \mathbb{R}^{L \times \frac{L(L-1)}{2}}. \tag{21}$$

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

$$\Omega_Y\left(j\right) = \Lambda\,\mathrm{diag}\left(\Lambda_j\right) D_4 Q^T. \tag{22}$$

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

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

The factor structure implies that

$$\Gamma_Y = \Lambda D_3 Q^T. \tag{24}$$

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

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, $\Omega_Y(1)$,..., $\Omega_Y(L)$ horizontally:

$$\Xi_Y = [\Gamma_Y, \Omega_Y(1), ..., \Omega_Y(L)]. \tag{25}$$

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* ℓ*th row of* C*. The second, third and fourth-order moments of* U_ℓ*, for all* $\ell \in \{1, ..., L\}$*, satisfy the linear restrictions:*

$$C^T \begin{pmatrix} \mathbb{E}(Y_1 Y_\ell) \\ \vdots \\ \mathbb{E}(Y_L Y_\ell) \end{pmatrix} = \mathrm{Var}(U_\ell) C_\ell, \tag{26}$$

$$C^T \begin{pmatrix} \mathbb{E}(Y_1 Y_\ell^2) \\ \vdots \\ \mathbb{E}(Y_L Y_\ell^2) \end{pmatrix} = \kappa_3(U_\ell) C_\ell. \tag{27}$$

and

$$C^T \begin{pmatrix} \mathbb{E}(Y_1 Y_\ell^3) - 3\mathbb{E}(Y_1 Y_\ell)\mathbb{E}(Y_\ell^2) \\ \vdots \\ \mathbb{E}(Y_L Y_\ell^3) - 3\mathbb{E}(Y_L Y_\ell)\mathbb{E}(Y_\ell^2) \end{pmatrix} = \kappa_4(U_\ell) C_\ell. \tag{28}$$

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.4 of the mathematical appendix. The following theorem then follows immediately.

Theorem 4 ***(Sufficient conditions for moment-based identification of*** Λ ***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 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 5 ***(Sufficient conditions for moment-based identification of*** Λ ***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 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:

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

The ratio of the two factor loadings is then

$$\frac{\lambda_{21}}{\lambda_{11}} = \frac{\mathbb{E}(Y_1 Y_2 Y_2)}{\mathbb{E}(Y_1 Y_1 Y_2)}. \tag{29}$$

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 proposed by Geary (1942). Interestingly, the quasi-JADE estimator that we shall develop in the next section also satisfies equation (29) in the case $(L, K) = (2, 1)$. 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.[10]

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 $(\ell, m) \in \Delta_L$ (i.e. $\ell < m$). 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,

$$\Omega_{Y,w} \equiv \sum_{(\ell,m)\in\Delta_L} w_{\ell,m} \Omega_Y(\ell, m) = \Lambda D_4 \operatorname{diag}\left(Q^T w\right) \Lambda^T. \tag{30}$$

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 simple average of the asymptotic variances of the components of the empirical analog $\widehat{\Omega}_Y(\ell, m)$ 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 $\widehat{\Gamma}_Y$.

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

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 $\widehat{\Xi}_Y$.

Alternatively, in the same spirit as in the previous paragraph, remark that, under the assumption that all factors are skewed or kurtotic, all matrices

$$\Xi_{Y,w} = \Gamma_Y + \sum_{j=1}^{L} w_j \Omega_Y(j) = \Lambda \left[D_3 + D_4 \operatorname{diag}\left(\Lambda^T w\right) \right] Q^T \tag{31}$$

have rank K, for almost all weights $w = (w_1, ..., w_L)^T \in \mathbb{R}^L$.[11] Matrices $\Xi_{Y,w}$ can therefore be used to estimate the number of factors K, if $K \leq L$. We suggest to set w_j equal to the 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:

$$\Omega_Y(\ell, m) = \Lambda D_4(\ell, m) \Lambda^T, \quad (\ell, m) \in \overline{\Delta}_L, \tag{32}$$

$$\Sigma_Y = \Lambda \Lambda^T, \tag{33}$$

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 $L \times K$ matrix 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 $(\ell, m) \in \overline{\Delta}_L$. These matrices satisfy the restrictions:

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

[11] 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 $k = 1...K$.

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

3. Compute V as an orthogonal 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 optimization (using Jacobi rotations), that is detailed in Section B of the Appendix.[12]

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 diagonalization algorithm still delivers an orthogonal 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 and Souloumiac (1993) call JADE this empirical procedure (Joint Approximate Diagonalization 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 algorithms based on Jacobi rotations are fast to converge.

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 $N^{\frac{1}{2}} \operatorname{vec}(\widehat{A})$. 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 $\operatorname{off}(M) = \sum_{i \neq j} m_{ij}^2$ for a matrix $M = [m_{ij}]$, and $\mathcal{O}_K$ is the set of orthogonal $K \times K$ matrices.

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

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:

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

We show the following result in Appendix C.

Theorem 6 *Assume that* $\sum_{j=1}^{J} (d_{jk} - d_{jm})^2 \neq 0$ *for all* $k \neq m$. *Then*

$$N^{\frac{1}{2}} \Big(\operatorname{vec}(\widehat{V}) - \operatorname{vec}(V) \Big) \underset{d}{\to} \mathcal{N}\left(0, \mathbb{V}_V\right),$$

where:

$$\mathbb{V}_V = (I_K \otimes V) W (I_J \otimes V^T \otimes V^T) \mathbb{V}_A (I_J \otimes V \otimes V) W^T (I_K \otimes V^T). \tag{34}$$

Let us consider the particular case of $J = 1$. In this case, (34) 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 ($J > 1$), precise estimation requires $\sum_j (d_{jk} - d_{jm})^2$ to be away from zero, for all indices (k, m). Cardoso (1999) already noted that joint diagonalization algorithms seemed less sensitive to the presence of multiple roots than usual diagonalization techniques.[13] Theorem 6 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

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

[13] See also the asymptotic distribution of estimators of Common Principal Components derived by Flury (1984, 1986).

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 (34) to compute standard errors. Instead, we suggest to compute standard errors or coverage intervals by standard bootstrap (with appropriate recentering). The reason is that the expression in (34) involves variances of third and/or fourth-order moments of the data, which are difficult to estimate precisely. Our simulations show extremely imprecise estimates of matrix $\mathbb{V}_A$, even with very large samples (more than $10,000$ observations). In contrast, the bootstrap provides good approximations of the true variance-covariance matrix of the JADE estimator.

3.4 The quasi-JADE algorithm

When the error components are not negligible, Lemmas 1 and 2 deliver moment restrictions which identify the first four moments of error variables independently of factor loadings. We call quasi-JADE the following estimation 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 $\mathrm{Var}\left(U_\ell\right)$, third-order cumulants $\kappa_3\left(U_\ell\right)$ and/or fourth-order cumulants $\kappa_4\left(U_\ell\right)$ 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 diagonalization (i.e. steps 2 and 3 of the JADE algorithm) of matrices $P^- \left[\Gamma_Y\left(\ell\right) - \kappa_3\left(U_\ell\right) \mathrm{Sp}_{L,\ell}\right] P^{-T}$ and/or $P^- \left[\Omega_Y\left(\ell, m\right) - \delta_{\ell m}\kappa_4\left(U_\ell\right) \mathrm{Sp}_{L,\ell}\right] P^{-T}$, 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 orthogonal matrix of joint eigenvectors. Then $\Lambda = PV$.

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

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

for $\ell, m = 1, ..., L$, $\ell \leq m$, and where $[A]_{i,j}$ denotes the (i,j) entry of matrix A.

Quasi-JADE is only marginally more complicated to implement than JADE,[14] and is almost as fast to converge.

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 moment restrictions optimally (the optimal metric). Third, it does not use all the structural moment restrictions. For example, the diagonal matrices $D_4(\ell, m)$ in (10) are related to Λ but we do not use this restriction.

A natural alternative to our approach would be to use all cumulant restrictions (6), (8) and (10) 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 diagonalize 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

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

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)
$\text{Var}(U_1)$	.77 (.59)	.87 (.43)	.96 (.20)	.98 (.16)
$\text{Var}(U_2)$	.76 (.57)	.87 (.43)	.98 (.20)	.98 (.17)
$\text{Var}(U_3)$	.74 (.56)	.86 (.42)	.96 (.20)	.98 (.16)

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

efficiency gains. Note that this weighting procedure is *ad hoc*. Issues regarding the optimal weighting of cumulant matrices, based on asymptotic results such as (34), are left for future research.

4 Monte-Carlo simulations

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

4.1 Estimation of factor loadings

Table 1 displays means and standard deviations of the Monte Carlo distributions of factor loadings estimates obtained from 1000 simulations of samples of various sizes generated by standardized log-normal factors, standard normal errors and Λ equal to

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

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 that finite sample biases are small and rapidly decrease as N increases. By

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

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 log-normal variate, for various sample sizes.[15] The striking contrast between Tables 1 and 2 suggests that our algorithm does a good job at extracting the relevant information from high-order moments of the data, while being relatively immune to the imprecision of their estimation in finite samples.

We then study the robustness of the JADE and quasi-JADE algorithms to noise (see Table 3). We run the simulations with normal errors, log-normal factors, a sample size of $N = 1000$ and $\Lambda = \Lambda_1$. The standard deviation of errors can take four values: 0.1, 0.5, 1 and 2. The performance of quasi-JADE deteriorates as the signal-to-noise ratio decreases. However, biases remain limited even for rather large error variances. By comparison, JADE (ordinary noise-free ICA) produces large finite sample biases (although $N = 1000$ is not such a small sample size). Notice that these biases are 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 quasi-JADE to Minimum Distance based on the complete set of moment restrictions (GMM). 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 (14) and where Q and $\overline{Q}$ depend on Λ.

In all the simulations that we performed, GMM proved to be highly unstable. Minimization 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

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

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, $V(U) = .25$)

efficiency, we treated the coefficients of D_4 as nuisance parameters. Precisely, we minimized the GMM norm, evaluated at $(\Lambda, \Sigma_U, D_4(\Lambda))$, with respect to (Λ, Σ_U) alone and where $D_4(\Lambda)$ is such that

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

Note that using the optimal metric to estimate $D_4(\Lambda)$ from restriction $\Omega_Y = QD_4\overline{Q}^T$ 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 $\Lambda = \Lambda_1$. Conditional on numerical convergence,[16] GMM yields only slightly more precise estimates of factor loadings than quasi-JADE. However, as error variances get larger, the GMM algorithm fails to reach convergence more frequently (less than 1% of the time when $\text{Var}(U_\ell) \leq 1$ but 15% of the time when $\text{Var}(U_\ell) = 4$). Note also that, for GMM, computing time 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.[17] Table 5 summarizes Monte Carlo

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

[17] 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 the

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)
$\text{Var}(U_1)$	.71 (.65)	.92 (.84)	.76 (.79)	.77 (.63)	.88 (.53)	.92 (.40)	.86 (.44)
$\text{Var}(U_2)$	.75 (.65)	.89 (.83)	.69 (.78)	.75 (.64)	.83 (.55)	.93 (.40)	.87 (.43)
$\text{Var}(U_3)$	.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, $N = 1000$, $\Lambda = \Lambda_1$)

distributions for kurtosis values in $\frac{1}{2}$, 2, 5, 10 and 100. In the first column of Table 5, we also report results for the case of uniformly distributed factors. The uniform distribution is platykurtic, with $\kappa_4 = -6/5$. The last column shows results for log-normal factors, with kurtosis excess 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 quasi-JADE based on second, third and fourth-order moments (using the restrictions of Lemma 1) to quasi-JADE based on second and third-order moments only (using the restrictions of Lemma 2), which yields consistent estimates when all factors are skewed. Table 6 reports simulations with log-normal factors, standard normal errors with variance 1, and matrix Λ is equal to

$$\Lambda_2 \equiv \begin{pmatrix} 2 & 2 \\ 2 & 1 \\ 1 & 2 \end{pmatrix}. \tag{35}$$

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, and relying on orthogonality up to the third order, is likely to do well in practice, provided that there is enough skewness in factors. On the other hand, adding moment restrictions does

random variable X defined as W_1 with probability ρ and W_2 with probability $1-\rho$ has variance one and kurtosis excess equal to $\kappa_4(\rho) = 3\rho/(4(1-\rho))$.

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, $\Lambda = \Lambda_2$)

not increase the bias in this case.

Lastly, we investigate the finite-sample performance of our algorithm when the number of measurements and the number of factors increase. Table 7 illustrates the cases of $L = K = 5$ 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 GMM algorithm discussed above turned out to be impractical for L as low as five, 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 design is the same as for the results reported in Table 5. The true value of Λ is Λ_2 given by (35), and we test $K = 2$ against $K = 3$.

Table 8 shows substantial size distortion. This especially happens when 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
α =.10	.90	.73	.82	.87	.85	.62
α =.20	.79	.57	.67	.74	.69	.43
α =.30	.67	.44	.54	.61	.57	.29
α =.40	.58	.33	.42	.50	.45	.19
α =.50	.47	.24	.32	.40	.35	.11
α =.60	.37	.16	.22	.32	.26	.05
α =.70	.27	.10	.13	.24	.19	.02
α =.80	.20	.05	.08	.15	.11	.01
α =.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, $N = 1000$, $\Lambda = \Lambda_2$)

Matrix	Ω_Y	$\Omega_{Y,w}$	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, $\Omega_{Y,w}$ and Γ_Y (log-normal factors, standard normal errors, $N = 1000$, $\Lambda = \Lambda_2$)

However, for reasonable values of kurtosis excess,[18] the risk of underestimating the number of factors exists but remains limited.

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 (30) – instead of Ω_Y. Table 9 provides a comparison of rank tests based on different cumulant matrices. We focus on the 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 high-order cumulants should be done with some 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 (see footnote 18), the size properties of the rank test based on third and fourth-order cumulant matrices are satisfactory.

[18] Stock returns are well-known for presenting high kurtosis. 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. See also our analysis of Fama and French's (1993) data on US stock returns (section 5.2).

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, $\Omega_{Y,w}$, Factors with increasing Kurtosis (standard normal errors, $N = 1000$, $\Lambda = \Lambda_2$)

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 Applications

We consider two applications: a factor model for individual data on wages and education (returns to schooling) and a factor model for stock returns.

5.1 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.

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

	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

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 regressor 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 a new 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.

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

	Γ_Y	Ω_Y	$\Omega_{Y,w}$
Rank	0	0	0
Statistic	29994	3646	20.2
Critical value .05	57.40	386.1	2.20
p-value	.00	.00	.00
Rank	1	1	1
Statistic	114.0	491.0	2.34
Critical value .05	7.74	45.4	.12
p-value	.00	.00	.00
Rank	2	2	2
Statistic	1.10	36.0	.185
Critical value .05	1.32	7.62	.0091
p-value	.072	.00	.00

Table 12: Rank tests

and Ω_Y, as well as their bootstrap standard errors.[19] These estimates are: 1.17 (1.14,1.20), .07 (.06,.08) and .007 (.001,.014) for the three CRs of Γ_Y, and .38 (.28,.49), .15 (.12,.18) and .04 (.03,.05) for those of Ω_Y. 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. There is thus evidence that the joint distribution of (Y, D, D^*) is not normal.

Estimation results. We started 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 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 estimated the one-factor model using high-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

[19] As in the rest of this section, 5%-95% confidence intervals are computed by 500 bootstrap replications with appropriate recentering. Confidence intervals are given between brackets.

[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.

	$K=1$ PCA	$K=1$ quasi-JADE(4)	$K=1$ quasi-JADE(3,4)	$K=2$ quasi-JADE(4)	$K=2$ quasi-JADE(3,4)
$\widehat{\lambda}_{11}$	.141 (.138,.145)	.154 (.136,166)	.142 (.137,.148)	.172 (.146,.200)	.166 (.145,.182)
$\widehat{\lambda}_{21}$	2.15 (2.12,2.19)	2.09 (2.02,2.18)	2.13 (2.09,2.20)	2.05 (1.96,2.16)	2.09 (2.02,2.19)
$\widehat{\lambda}_{31}$	2.01 (1.98,2.03)	2.05 (1.95,2.14)	2.03 (1.96,2.11)	2.02 (1.93,2.12)	2.02 (1.93,2.10)
$\frac{\widehat{\lambda}_{11}}{\widehat{\lambda}_{21}}$	6.6%	7.4%	6.7%	8.5%	7.9%
$\widehat{\lambda}_{12}$	-	-	-	-.138 (-.212,-.067)	-.136 (-.209,-.040)
$\widehat{\lambda}_{22}$	-	-	-	.360 (.009,.561)	.316 (.091,.459)
$\widehat{\lambda}_{32}$	-	-	-	.475 (.310,.660)	.381 (.131,.484)
$\widehat{V}(U_1)$	.066 (.065,.067)	.052 (.041,.070)	.066 (.060,.069)	.038 (.000,.060)	.040 (.010,.063)
$\widehat{V}(U_2)$	2.31 (2.22,2.40)	2.56 (2.06,2.90)	2.43 (2.04,2.65)	2.61 (1.85,3.04)	2.50 (1.92,2.84)
$\widehat{V}(U_3)$	.672 (.604,.745)	.426 (.000,.850)	.586 (.177,.867)	.385 (.000,.766)	.500 (.089,.889)

Table 13: Factor loadings and error variances (quasi-JADE(4): uses second and fourth-order moments; quasi-JADE(3,4): uses second, third and fourth-order moments)

algorithm. In column 2, we report the results for the version of the algorithm using second and fourth-order cumulants and the restrictions of Lemma 1. In column 3, second, third and fourth-order cumulants are used and the restrictions of Lemmas 1 and 2 are combined. The results of all three columns are remarkably similar.

Next, we turned 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.

We then performed a test of overidentifying restrictions, based on the JADE criterion (sum of squares of the off-diagonal elements of the jointly diagonalized matrices). We bootstrapped the test statistic 500 times to compute p-values. We found p-values of 26% and 27% for the two versions of quasi-JADE, with $K = 2$. Thus, according to this criterion, at all conventional levels, the data do not reject the validity of the overidentifying restrictions imposed in quasi-JADE.

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. 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

	$K=1$ quasi-JADE(4)	$K=1$ quasi-JADE(3,4)	$K=2$ quasi-JADE(4)	$K=2$ quasi-JADE(3,4)
$\kappa_3(X_1)$	-	1.34 (1.29,1.39)	-	1.17 (1.08,1.30)
$\kappa_3(X_2)$	-	-	-	.087 (-.709,6.10)
$\kappa_4(X_1)$	.612 (.391,.854)	.741 (.354,1.02)	.627 (.439,.768)	.665 (.445,.841)
$\kappa_4(X_2)$	-	-	13.6 (3.58,196)	15.5 (4.28,580)

Table 14: Factor cumulants (quasi-JADE(4): uses second and fourth-order moments; quasi-JADE(3,4): uses second, third and fourth-order moments)

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.

Interpretation. We thus obtain the following factor structure:

$$\begin{cases} Y = .17X_1 - .14X_2 + U_1 \\ D = 2X_1 + .4X_2 + U_2 \\ D^* = 2X_1 + .4X_2 + U_3 \end{cases} \tag{36}$$

This structure is consistent with the interpretation of $E = 2X_1 + .4X_2$ as "true" education, being measured with error by D and D^*. The number of years of education, D, faces large measurement errors ($\text{Var}(U_2) = 2.6$ and $\text{Var}(E) = 5.6$) in comparison to the other education measure based on the highest diploma obtained ($\text{Var}(U_3) = .4$).

Education results from two independent factors X_1 and X_2, the interesting factor being X_2, which negatively correlates wages and education. It could be that some specific taste for education favors career choices with low labor market value (wage) but with high amenity content (compare Ph.D. versus MBA degrees, and academic versus business careers). Alternatively, Card (2001) shows that a negative correlation between wage and education can be obtained if individual-specific marginal costs of education increase faster, across individuals, than individual-specific returns to education. The negative correlation component between education and wages could thus reflect the fact that individuals with high innate "ability" avoid paying their relatively higher cost of education.

Interestingly, model (36) is consistent with a classical Mincer equation, whereby the wage

depends on education via the linear relationship:

$$Y = \alpha E + V. \tag{37}$$

Effectively, given α, model (36) implies (37) if we specify V as

$$V = (.17 - 2\alpha)X_1 - (.14 + .4\alpha)X_2 + U_1. \tag{38}$$

In model (37)-(38), E is orthogonal to V if and only if $\alpha = 6.8\%$. This is the value that was estimated by PCA (i.e. regress Y on D, instrumenting D by D^*) in the first column of Table 13. For any other value of α, E and V are correlated. In this case, one should not only worry about measurement error but also about endogeneity biases generated by unobserved heterogeneity, here captured by factors X_1, X_2. If α is larger than 6.8%, then the correlation between education (E) and error (V) is negative. Whenever α is less than the PCA estimate, the correlation is positive.

Next, suppose then that the true model is (37)-(38) and that E and V are correlated. Unfortunately, labor force survey data do not provide a natural instrument for education, that is a variable Z correlated with E but not with V.[21] Because it is possible to produce many linear combinations of X_1 and X_2 which satisfy these conditions, it follows that α is not identified.

However, one case is of particular interest. This is when we constrain Z to be independent of–and not only orthogonal to–the error V. Then, α can take only two values: $\alpha = 8.5\%$ or $\alpha = -35\%$. The first case corresponds for example to the instrument $Z = X_1$, the second case to $Z = X_2$. Obviously, $\alpha = 8.5\%$ is the only plausible case. By comparison, the OLS estimate of the return to education, using the number of years of education as education proxy, is 4.4%. The 4.1% difference between both estimates can be tentatively decomposed as follows: 2.4% is due to measurement error and 1.7% reflects unobserved heterogeneity.

5.2 Stock returns

In an influential paper, Fama and French (1993) identify three factors explaining a large proportion of the variance of time-series of U.S. excess stock returns, $Y_\ell(t) = R_\ell(t) - R_F(t)$,

[21] We experimented with the month of birth without success, the R^2 of the first stage regression being too low.

$\ell = 1, ..., L$. In addition to the market return $(R_M(t) - R_F(t)$, where $R_F(t)$ is the risk-free return), which is the unique factor of the CAPM model, they identify two additional factors:

- $SMB(t)$, or "small minus big", is the difference between the average of the returns on two stock portfolios: one containing firms with market value (price time number of shares) less than the median, and one containing firms with size above the median.

- $HML(t)$, or "high minus low", is the difference between the average of the returns on two stock portfolios: one gathering firms with book-to-market ratio (book value of capital divided by market value, denoted B/M) less that the 30th percentile and another one containing all firms with B/M ratio above the 70th percentile.

Fama and French show that these three factors explain monthly data on 25 portfolios formed by intersecting size and book-to-market quintiles remarkably well. Other relevant contributions by the same authors include Fama and French (1995, 1996) and Davis, Fama and French (2000). Fama and French's factors are now widely used in applied finance to summarize the correlations between bond or stock returns.

In this section we apply quasi-JADE to estimate a linear independent factor model with three factors. Unlike Fama and French, who construct factors on the basis of economic intuition, we shall estimate factors blindly.

The data. We use daily US observations between 01/07/1963 and 31/08/2005 of the returns to 25 stock portfolios formed on size and book-to-market.[22] With monthly data, we obtained similar results, though less precisely estimated. The size and book-to-market breakpoints are NYSE quintiles. There are 10,616 observations. Table 15 shows the mean, standard error, skewness and kurtosis of the returns on the 25 portfolios. Returns are net of the risk-free rate R_F, which varies between .003 and .061 over the period. All returns appear strongly leptokurtic, and somewhat skewed to the left.

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

Size	B/M ratio	Mean	Standard error	Skewness	Kurtosis
Small	Low	.0008	1.10	-.86	13.8
	2	.0293	.91	-.87	13.3
	3	.0341	.77	-.98	15.1
	4	.0428	.71	-1.00	15.4
	High	.0479	.71	-.97	14.2
2	Low	.0123	1.17	-.50	10.0
	2	.0251	.91	-.68	12.4
	3	.0383	.81	-.67	11.6
	4	.0404	.77	-.69	12.9
	High	.0447	.87	-.55	10.2
3	Low	.0148	1.15	-.38	9..8
	2	.0302	.88	-.59	12.8
	3	.0309	.77	-.62	12.9
	4	.0372	.77	-.48	11.9
	High	.0447	.88	-.66	16.5
4	Low	.0220	1.10	-.21	12.0
	2	.0209	.87	-.75	18.5
	3	.0320	.81	-.92	20.4
	4	.0383	.80	-.63	15.7
	High	.0394	.92	-.57	15.3
Big	Low	.0198	1.06	-.37	15.8
	2	.0214	.95	-.96	27.6
	3	.0231	.91	-.90	27.3
	4	.0260	.88	-.92	32.2
	High	.0281	.99	-.57	17.9

Table 15: Moments of the variables

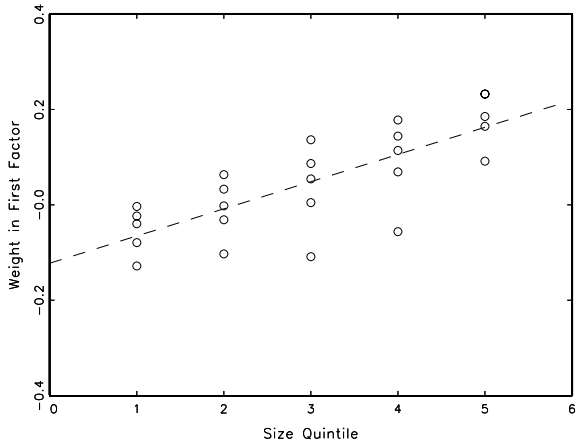
Size	B/M ratio	Factor 1	Factor 2	Factor 3	Error variance	R^2
Small	Low	.57	.79	-.37	.055	.95
	2	.50	.67	-.23	.054	.93
	3	.45	.56	-.14	.035	.93
	4	.41	.52	-.09	.029	.94
	High	.43	.52	-.05	.028	.94
2	Low	.70	.72	-.50	.102	.92
	2	.59	.58	-.26	.049	.94
	3	.54	.51	-.17	.066	.89
	4	.53	.47	-.10	.079	.84
	High	.61	.51	-.08	.118	.82
3	Low	.71	.61	-.58	.122	.90
	2	.62	.47	-.30	.050	.93
	3	.58	.40	-.15	.061	.88
	4	.60	.37	-.10	.080	.84
	High	.70	.40	-.06	.095	.85
4	Low	.74	.46	-.61	.094	.92
	2	.70	.33	-.28	.041	.94
	3	.69	.29	-.16	.038	.94
	4	.68	.28	-.08	.086	.84
	High	.77	.30	-.06	.146	.79
Big	Low	.83	.14	-.52	.125	.87
	2	.84	.08	-.33	.034	.96
	3	.81	.08	-.23	.077	.89
	4	.80	.07	-.12	.061	.91
	High	.83	.13	-.12	.256	.64
% Variance		.544	.258	.099	.099	
Skewness		-1.21	-.76	-.56	-	
Kurtosis		30	28	77	-	

Table 16: Factor loadings, factor moments and error variances

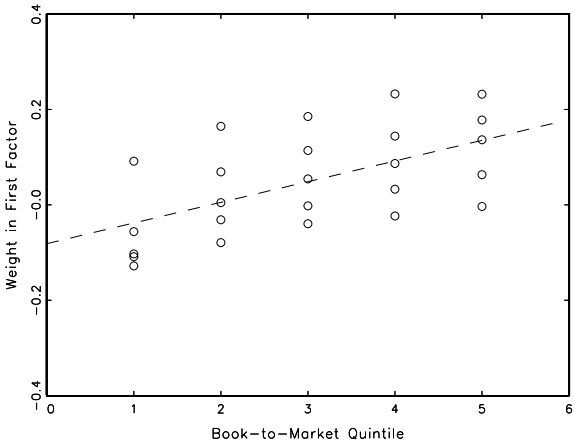
Estimation results. Table 16 presents the estimates of factor loadings and error variances, under the assumption that $K = 3$. Quasi-JADE was applied using second, third and fourth-order cumulants. Interestingly, using only third or only fourth-order moments made very little difference in estimation. Moreover, bootstrapped confidence intervals (not reported) show that most factor loadings are rather precisely estimated. Overall, the three factors account for 90% of total variance. The first factor explains 54%, the second factor 26% and the third factor 10%, about the same as the error term. Lastly, one notices that all three factors are skewed to the left and strongly leptokurtic.

Given that the model is nearly noise-free, we predict factor levels as $\widehat{X}(t) = \widehat{\Lambda}^{-}Y(t)$, where

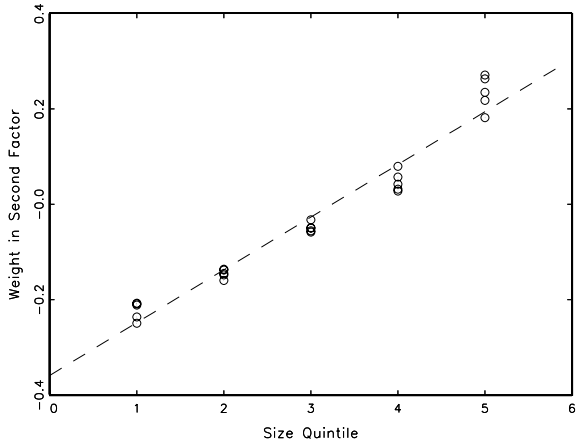
a1) Size/X_1



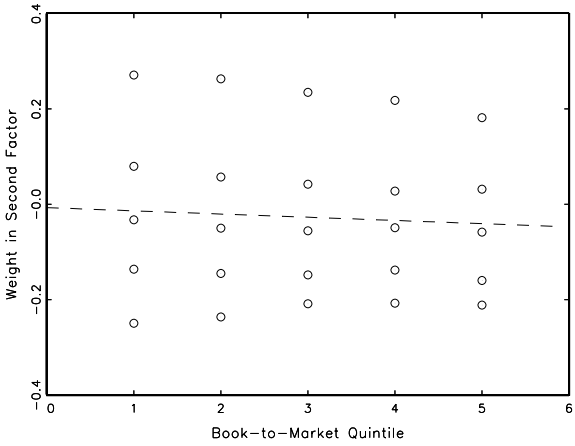
b1) Book-to-Market/X_1



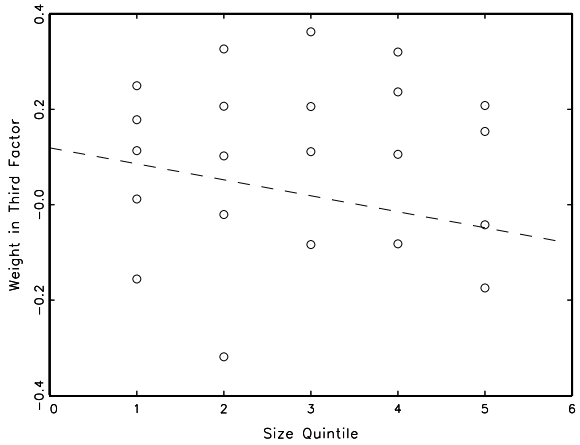
a2) Size/X_2



b2) Book-to-Market/X_2



a3) Size/X_3



b3) Book-to-Market/X_3

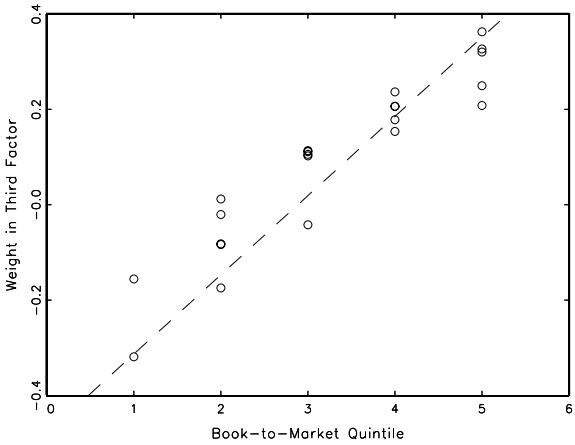
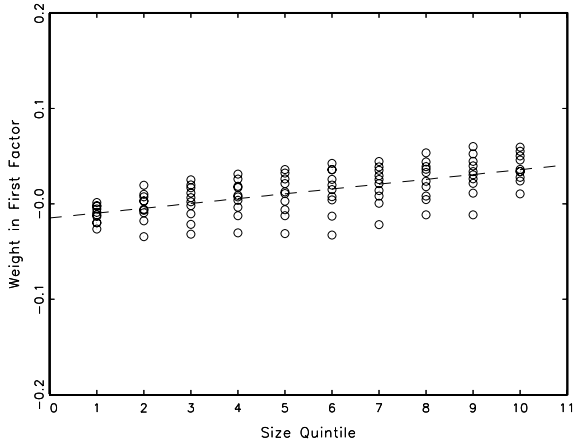
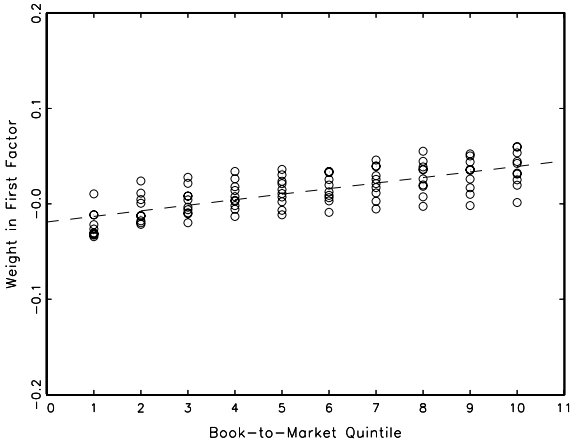


Figure 1: Independent factors against quintiles of size and book-to-market, daily Fama French data, 25 portfolios

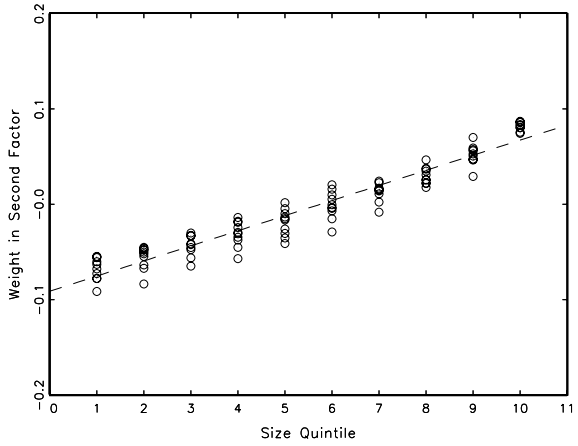
a1) Size/X_1



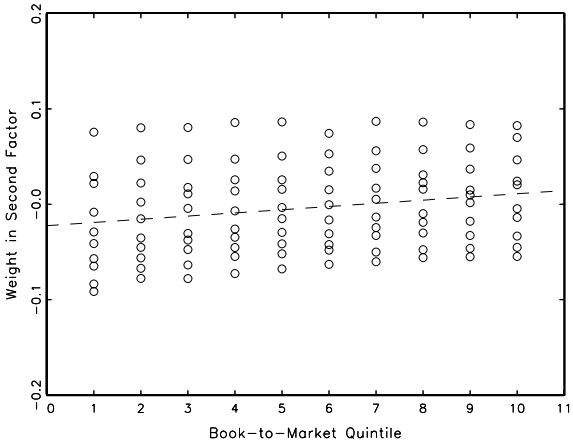
b1) Book-to-Market/X_1



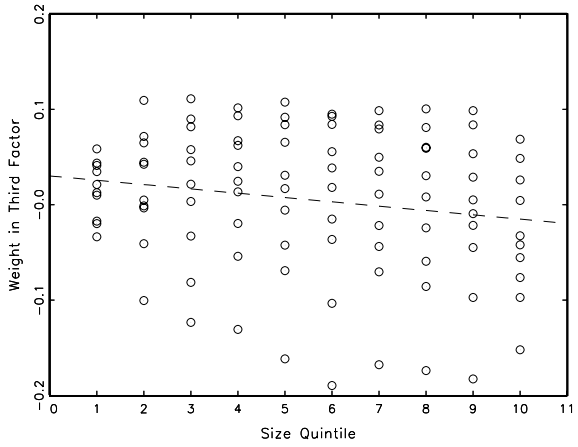
a2) Size/X_2



b2) Book-to-Market/X_2



a3) Size/X_3



b3) Book-to-Market/X_3

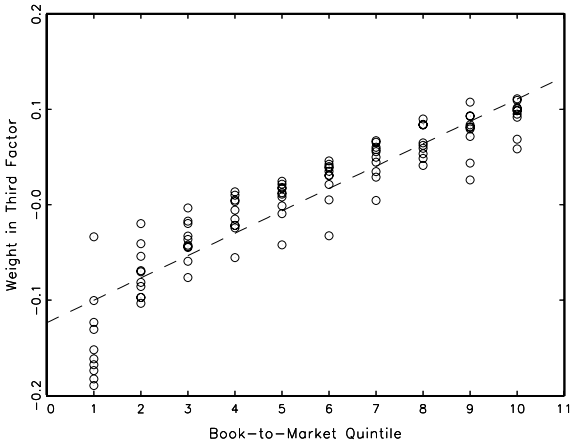


Figure 2: Independent factors against deciles of size and book-to-market, daily Fama French data, 100 portfolios

$\widehat{\Lambda}^{-}$ is the generalized inverse of the estimated matrix of factor loadings $\widehat{\Lambda}$. When the error variance cannot be neglected, predicting factor levels requires a more complicated procedure that is developed in a companion paper (Bonhomme and Robin, 2006).

Figure 1 displays the elements of matrix $\widehat{\Lambda}^{-}$ as a function of size quintiles (panels a)), and B/M quintiles (panels b)). The three columns are reported from top to bottom of the figure. Dashed lines represent OLS fit. Panels a2) and b2) show that the second factor is strongly negatively correlated with size and almost uncorrelated with book-to-market. Conversely, panels a3) and b3) show that the third factor is weakly negatively correlated with size, yet strongly positively correlated with book-to-market. Lastly, the first factor appears positively correlated with size and to book-to-market, although the correlation with the latter is weaker. These results are qualitatively the same if we estimate a three-factor model on 100 portfolios formed as the intersection of size and book-to-market deciles, as shown by Figure 2. Interestingly, unrotated PCA yields similar pictures as Figures 1 and 2. This suggests that unrotated principal components are approximately independent.

We then provide a direct comparison of these estimated factors to those used by Fama and French. To do so, we compute Fama and French's factors and correlate them to $\widehat{X}_t$. Panel a) of Table 17 shows these correlations. We see that the three factors estimated by quasi-JADE are strongly correlated with the market, size and book-to-market factors constructed by Fama and French. The correlations are .84, .85 and .90, respectively. Note that Fama and French's factors are correlated. For instance, the market return has correlation $-.24$ and $-.58$ with *SMB* and *HML*, respectively. For this reason, they cannot be equal to the independent factors obtained by quasi-JADE. We then apply JADE to market return, *SMB* and *HML*, and obtain new factors which are by construction independent. Panel b) of table 17 reports the correlations between these new factors and the factors that were initially estimated by quasi-JADE on the 25 portfolios. We find very high correlations (.97, .98 and .93).

It is interesting to evaluate the extent to which these results are driven by the *ex-ante* grouping into size and book-to-market cells. Our experiments on stock data grouped by in-

	X_1	X_2	X_3
$R_M - R_F$	.84	.24	-.41
SMB	-.49	.85	-.09
HML	-.11	.23	.90

a) Fama French

	X_1	X_2	X_3
$R_M - R_F$	.97	-.01	-.09
SMB	-.01	.98	-.06
HML	.09	.07	.93

b) Independent Fama French

Table 17: Fama French factors *versus* Quasi-JADE estimates, daily US data, 25 portfolios

dustries[23] showed that if the first factor remained strongly linked to market return, size and book-to-market were much less strongly correlated with the two other factors. This casts some doubts on the ability of Fama and French's factors to explain very disaggregate data on stock returns with the same success.

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 measurements. Contrary to ordinary Factor Analysis, identification is unambiguously defined up to sign and permutation normalizations.

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 diagonalization algorithm.

The independent factor structure generates many overidentifying restrictions on high-order moments. This may explain the encouraging Monte Carlo simulation results that we obtained. In contrast with previous evidence on the use of high-order moments for estimation,[24] we find,

[23] These data can be found in the section "change in industry portfolios" in French's data library.
[24] See the results reported in Madansky (1959), and the survey by Aigner *et al.* (1984).

for sufficiently non symmetric and/or kurtotic data, small biases and precise estimates, even in relatively small samples.

The estimation methodology is first 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), our method yields an interesting second factor that is negatively correlated with earnings and positively correlated with education. This is evidence that there exist individual characteristics which are valued by the education institution but not by the labor market. Moreover, the exhibited factor structure is consistent with the standard model of education returns if one allows measurement errors on the education measure and unobserved heterogeneity.

In a second application, we consider data on daily US stock returns, grouped into quintiles of size and book-to-market ratio. Fixing the number of factors to three, we estimate independent components which turn out to be strongly correlated with the three factors constructed by Fama and French (1993). In addition to the market factor, we clearly identify a negative "size" effect and a positive "book-to-market" effect. However, our experiments with data grouped by industry suggest that the high correlation between independent factors and Fama and French's heuristic constructs partly results from the fact that firm stocks are *ex-ante* grouped by size and book-to-market ratio.

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 econometrics, we are not aware of similar semi-parametric methods. As far as we know, our paper is the first example of a fast and consistent estimation algorithm for an overcomplete independent factor structure (L factors and L errors). Nevertheless, more work is needed to

deal with more general overcomplete ICA models.

The second direction of research concerns the extension of the method of this paper to the case of a very large number of measurements. Bai and Ng (2002) and Bai (2003) provide extensive analyses of the PCA estimator in this case. Financial and macroeconomic applications motivate the need to extend ICA methods in this direction.

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 7 ***(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 $\widetilde{\Lambda}$.

Let $D = \text{diag}\,(d_k)$ be a $K \times K$ diagonal matrix and P a permutation matrix such that $\widetilde{\Lambda} = \Lambda DP$. Then, by assumption,

$$\text{Var}\,(\Lambda X + U) = \text{Var}\left(\widetilde{\Lambda}\widetilde{X} + \widetilde{U}\right).$$

Hence, for all $\ell < m$,

$$\sum_{k=1}^{K} \lambda_{\ell k}\lambda_{mk} = \sum_{k=1}^{K} \lambda_{\ell k}\lambda_{mk}d_k^2.$$

If matrix

$$Q\,(\Lambda) = [\lambda_{\ell 1}\lambda_{m1}, ..., \lambda_{\ell K}\lambda_{mK}]_{(\ell,m)\in\Delta_L}$$

is full column rank, then it must be that

$$d_k^2 = 1.$$

This ends the proof.

A.2 Proof of Theorem 2

To prove Theorem 2, 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* $(k, k') \in \{1, ..., K\}^2$*,* $k \neq k'$*. 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 2 easily follows.

Fourth-order moments. In the case where $U = 0$, second and fourth-order cumulant restrictions (6)-(10) yield:

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

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 orthogonal eigenvectors of:

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

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 $P^{-}\Omega_Y(\ell,m)P^{-T}$, $(\ell,m)\in\overline{\Delta}_L$, has a unique solution up to column sign and permutation if for all $(k,k')\in\{1...K\}^2$, $k\neq k'$, there exists $(\ell,m)\in\overline{\Delta}_L$ 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 (8) 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 $P^{-}\Omega_Y(\ell,m)P^{-T}$, $(\ell,m)\in\overline{\Delta}_L$, 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 $(\ell,m)\in\overline{\Delta}_L$ 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.3 Proof of Lemma 1

1. Let Ω_Y be defined by (14). As Q has rank K and D_4 is non singular, restrictions (15) imply that

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

has rank K. It follows that there exists $\overline{C} \in \mathbb{R}^{\#\overline{\Delta}_L \times (\#\overline{\Delta}_L - K)}$, full column rank, such that $\overline{C}^T \Omega_Y = 0$. Since $D_4 Q^T$ 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 $(i, j) \in \overline{\Delta}_L$.

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

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

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 $\left[\overline{C}_{(1,1)}, ..., \overline{C}_{(L,L)}\right]^T \in \mathbb{R}^{L \times (\#\overline{\Delta}_L - K)}$ 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 $\overline{C}_{11} \in \mathbb{R}^{\#\Delta_L \times (\#\Delta_{L,2} - K)}$, $\overline{C}_{12} \in \mathbb{R}^{\#\Delta_L \times L}$, $\overline{C}_{21} \in \mathbb{R}^{L \times (\#\Delta_{L,2} - K)}$ 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, $\kappa_3\left(U_\ell\right)$ and $\kappa_4\left(U_\ell\right)$ are identified.

This ends the proof of Lemma 1.

A.4 Proof of Lemma 2

1. The factor structure implies that

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

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

$$\gamma^T\left[D_3Q^T,D_4\,\mathrm{diag}\left(\Lambda_1\right)Q^T,...,D_4\,\mathrm{diag}\left(\Lambda_L\right)Q^T\right]=0.$$

As Q has rank K, it follows that $\gamma^TD_3=0$ and $\gamma^TD_4\,\mathrm{diag}\left(\Lambda_\ell\right)=0$ for all $\ell\in\{1...L\}$. Then, as Λ is full column rank, this implies that $\gamma^TD_4=0$. Lastly, as for all k either $\kappa_3\left(X_k\right)\neq 0$ or $\kappa_4\left(X_k\right)\neq 0$, it follows that $\gamma=0$.

Therefore: $\left[D_3Q^T,D_4\,\mathrm{diag}\left(\Lambda_1\right)Q^T,...,D_4\,\mathrm{diag}\left(\Lambda_L\right)Q^T\right]$ 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 $\left[D_3Q^T,D_4\,\mathrm{diag}\left(\Lambda_1\right)Q^T,...,D_4\,\mathrm{diag}\left(\Lambda_L\right)Q^T\right]$ is full row rank, it must also be that $C^T\Lambda=0$.

2. One thus has

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

or

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

where C_ℓ^T is the ℓth row of C.

Moreover, matrices $\Gamma_Y(\ell)$ defined by (7) 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

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

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 $\operatorname{rank}(\Lambda_{-\ell}) = K$. Hence $\operatorname{Var}(U_\ell)$, $\kappa_3(U_\ell)$ 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:

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

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

$$\sum_{k=1}^{K} \operatorname{off}(U A_k U^T), \tag{B1}$$

with respect to U orthogonal.

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:

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

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

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

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 (B1) 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

First-order conditions. The JADE estimator solves

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

The Lagrangian associated with the minimization problem is:

$$\begin{aligned} \mathcal{L}(V, \gamma) &= \sum_{j=1}^{J} \text{off}(V^T \widehat{A}_j V) + \gamma^T \text{vec}\left(V^T V - I_K\right), \\ &= \sum_{j} \sum_{m \neq k} (v_k^T \widehat{A}_j v_m)^2 + \sum_{k} \gamma_{kk}(v_k^T v_k - 1) + \sum_{m \neq k} \gamma_{mk} v_k^T v_m, \end{aligned}$$

where γ is a vector of K^2 Lagrange multipliers γ_{mk}, and v_k is the kth column of matrix V.

Differentiating the Lagrangian with respect to v_ℓ, for $\ell = 1...K$, yields:

$$\frac{\partial \mathcal{L}\left(\widehat{V}, \widehat{\gamma}\right)}{\partial v_\ell} = 2\sum_j \sum_{k\neq\ell} (\widehat{v}_k^T \widehat{A}_j \widehat{v}_\ell)\widehat{A}_j \widehat{v}_k + 2\widehat{\gamma}_{\ell\ell}\widehat{v}_\ell + \sum_{k\neq\ell} \widehat{\gamma}_{k\ell}\widehat{v}_k = 0. \tag{C2}$$

Then, multiplying (C2) by $\widehat{v}_m^T$, for $m \neq \ell$, gives:

$$2\sum_j \sum_{k\neq\ell} (\widehat{v}_k^T \widehat{A}_j \widehat{v}_\ell)\widehat{v}_m^T \widehat{A}_j \widehat{v}_k + \widehat{\gamma}_{m\ell} = 0. \tag{C3}$$

Using that $\widehat{\gamma}_{m\ell} = \widehat{\gamma}_{\ell m}$ by symmetry, it follows from (C3) that

$$\sum_j \sum_{k\neq\ell} (\widehat{v}_k^T \widehat{A}_j \widehat{v}_\ell)\widehat{v}_m^T \widehat{A}_j \widehat{v}_k = \sum_j \sum_{k\neq m} (\widehat{v}_k^T \widehat{A}_j \widehat{v}_m)\widehat{v}_\ell^T \widehat{A}_j \widehat{v}_k,$$

or, equivalently, as $\widehat{A}_j$ is symmetric for all j:

$$\sum_j \widehat{v}_\ell^T \widehat{A}_j \left(\sum_{k\neq\ell} \widehat{v}_k \widehat{v}_k^T\right) \widehat{A}_j \widehat{v}_m = \sum_j \widehat{v}_m^T \widehat{A}_j \left(\sum_{k\neq m} \widehat{v}_k \widehat{v}_k^T\right) \widehat{A}_j \widehat{v}_\ell.$$

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

$$\sum_j \widehat{v}_\ell^T \widehat{A}_j \left(I_K - \widehat{v}_\ell \widehat{v}_\ell^T\right) \widehat{A}_j \widehat{v}_m = \sum_j \widehat{v}_m^T \widehat{A}_j \left(I_K - \widehat{v}_m \widehat{v}_m^T\right) \widehat{A}_j \widehat{v}_\ell,$$

which we write after rearranging:

$$\sum_j \widehat{v}_\ell^T \widehat{A}_j \widehat{v}_m \left(\widehat{v}_m^T \widehat{A}_j \widehat{v}_m - \widehat{v}_\ell^T \widehat{A}_j \widehat{v}_\ell\right) = 0. \tag{C4}$$

Equation (C4) holds for all $\ell < m$. The JADE estimator $\widehat{V}$ solves these $K(K-1)/2$ non redundant equations, together with the $K(K+1)/2$ orthogonality constraints:

$$\widehat{v}_\ell^T \widehat{v}_m = \delta_{\ell m}, \text{ for all } \ell \leq m. \tag{C5}$$

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

$$\widetilde{V} = \arg\min_{V\in\mathcal{O}_K} \sum_{j=1}^J \text{off}(V^T A_j V).$$

Then, as: $\min_{V\in\mathcal{O}_K} \sum_{j=1}^J \text{off}(V^T A_j V) = 0$ at the true value, it follows that $\widetilde{V}^T A_j \widetilde{V} = \widetilde{D}_j$ is diagonal for all j. As for all $k \neq m$ there exists $j \in \{1....J\}$ such that $d_{jk} \neq d_{jm}$, one can apply Lemma 4 to show that $\widetilde{V}$ is equal to the true V, up to column sign and permutation. This shows the identification of V. Consistency follows from classical arguments, as the parameter space $\mathcal{O}_K$ is compact.

Asymptotic distribution. A first-order Taylor expansion of (C4) around the true value V yields:

$$\sum_j^J v_m^T \widehat{A}_j v_k \left(v_k^T \widehat{A}_j v_k - v_m^T \widehat{A}_j v_m\right) + \sum_j^J \left(v_k^T \widehat{A}_j v_k - v_m^T \widehat{A}_j v_m\right) \left(v_m^T \widehat{A}_j (\widehat{v}_k - v_k) + v_k^T \widehat{A}_j (\widehat{v}_m - v_m)\right)$$
$$+ \sum_j^J v_m^T \widehat{A}_j v_k \left(v_k^T \widehat{A}_j (\widehat{v}_k - v_k) - v_m^T \widehat{A}_j (\widehat{v}_m - v_m)\right) = o_p\left(N^{-1/2}\right). \tag{C6}$$

As $\underset{N\to\infty}{\text{plim}}\,\widehat{A}_j = A_j$ for all j, and as $v_k^T A_j v_m = 0$ for all $k \neq m$, (C6) yields:

$$\sum_j^J (d_{jk} - d_{jm})\, v_m^T \left(\widehat{A}_j - A_j\right) v_k + \sum_j^J (d_{jk} - d_{jm}) \left(v_m^T A_j(\widehat{v}_k - v_k) + v_k^T A_j(\widehat{v}_m - v_m)\right) = o_p\left(N^{-1/2}\right),$$

where $d_{jk} = v_k^T A_j v_k$ are the diagonal elements of $V^T A_j V$.

At this stage, it is convenient to define $\widehat{x}_{mk} \equiv v_m^T(\widehat{v}_k - v_k)$. As $v_m^T A_j = d_{jm} v_m^T$, one has:

$$\sum_j^J (d_{jk} - d_{jm})\, v_m^T \left(\widehat{A}_j - A_j\right) v_k + \sum_j^J (d_{jk} - d_{jm}) (d_{jm}\widehat{x}_{mk} + d_{jk}\widehat{x}_{km}) = o_p\left(N^{-1/2}\right). \tag{C7}$$

Now, a Taylor expansion of the orthogonality constraints (C5) yields:

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

Thus, (C7) can be rewritten as:

$$\sum_j^J (d_{jk} - d_{jm})^2\, \widehat{x}_{mk} = -\sum_j^J (d_{jk} - d_{jm})\, v_m^T \left(\widehat{A}_j - A_j\right) v_k + o_p\left(N^{-1/2}\right). \tag{C8}$$

Let $\widehat{X} = V^T\left(\widehat{V} - V\right)$. Then equation (C8) is equivalently written, in matrix form, as:

$$\text{vec}\left(\widehat{X}\right) \;=\; -W\left(I_J \otimes V^T \otimes V^T\right)\left(\text{vec}\left(\widehat{A}\right) - \text{vec}\,(A)\right) + o_p\left(N^{-1/2}\right),$$

where W, A and $\widehat{A}$ have been defined in the text. Note that W is provided that $\sum_j^J (d_{jk} - d_{jm})^2 \neq 0$ for all $k \neq m$.

Then, as:

$$\text{vec}\left(\widehat{X}\right) = \left(I_K \otimes V^T\right)\left(\text{vec}\left(\widehat{V}\right) - \text{vec}\,(V)\right),$$

it follows that

$$N^{\frac{1}{2}}\left(\text{vec}\left(\widehat{V}\right) - \text{vec}\,(V)\right) \;=\; -\left(I_K \otimes V\right) W\left(I_J \otimes V^T \otimes V^T\right) N^{\frac{1}{2}}\left(\text{vec}\left(\widehat{A}\right) - \text{vec}\,(A)\right) + o_p\,(1),$$

from which

$$N^{\frac{1}{2}}\left(\text{vec}(\widehat{V}) - \text{vec}(V)\right) \underset{d}{\to} \mathcal{N}\left(0, \mathbb{V}_V\right),$$

where the expression of $\mathbb{V}_V$ is given by (34).

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 $\text{rank}\left(\Sigma_{\text{vec}\left(\widehat{B}\right)}\right) = s$, $0 < s \leq pq$.[25] 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)

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

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

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

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 $-N^{-1} \ln \alpha_N^r = o(1)$.

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

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

[7] 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.

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

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

[10] BONHOMME, S., and J.M. ROBIN (2006): "Nonparameteric Identification and Estimation of Factor Distributions in Independent Factor Models ," *mimeo.*

[11] CARD, D. (2001): "Estimating the Returns to Schooling: Progress on Some Persistent Econometric Problems ," *Econometrica,* 69, 1127-1160.

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

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

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

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

[16] 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.

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

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

[19] CUNHA, F., J.J. HECKMAN, and S. NAVARRO (2005): "Separating Uncertainty from Heterogeneity in Life Cycle Earnings," *Oxford Economic Papers,* 57, 191-261.

[20] 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.

[21] DAVIS, J., E. FAMA, and K. FRENCH (2000): "Characteristics, Covariances and Average Returns," *Journal of Finance*, 55(1), 389-406.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] FAMA, E., and K. FRENCH (1992): "The Cross-Section of Expected Stock Returns," *Journal of Finance*, 47(2), 427-465.

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

[29] FAMA, E., and K. FRENCH (1995): "Size and Book-to-Market Factors in Earnings and Returns," *Journal of Finance*, 50(1), 131-155.

[30] FAMA, E., and K. FRENCH (1996): "Multifactor Explanations of Asset Pricing Anomalies," *Journal of Finance*, 51(1), 55-84.

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

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

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

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

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

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

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

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

[39] 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.

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

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

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

[43] 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.

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

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

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

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

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

[49] 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.

[50] XU, L. (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.

[51] XU, L. (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.

[52] 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.

[53] 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.