

Nonparametric Identification and Estimation of Independent Factor Models

Stéphane Bonhomme
CEMFI, Madrid

Jean-Marc Robin
Université de Paris I, Pantheon-Sorbonne,
CREST-INSEE, Malakoff, and
CEPR, London

June 2005

STILL PRELIMINARY – PLEASE DO NOT QUOTE

Abstract

JEL codes:

Keywords:

1 Introduction

In this paper, we shall consider linear factor models of the form:

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

where Y is a vector of L observed quantities, X is a vector of K unobserved factors, U is a vector of L unobserved errors and Λ is a $L \times K$ matrix of factor loadings. Without loss of generality, it is assumed that all random variables have zero mean.

Linear factor models are routinely used in the social sciences, at least since the introduction of the “g” factor by Spearman (1904) in psychology. Important variants include Principal Component Analysis (PCA) as a reduction dimensionality technique, and “exploratory” and “confirmatory” factor analysis.

In econometrics, several commonly used techniques can be embedded into the framework of Model (1). In cross-section, latent variable models are used to separate the “true” value of a regressor of interest from measurement error (see the survey by Aigner *et al.*, 1984). In the Panel data literature, the so-called “random” individual effects can be seen as factors. Moreover, linear factor models can yield the same covariance structure as many dynamic models used in that literature, provided that a sufficiently large number of factors is allowed for.

A common feature of these techniques is that factor models are used as a way of restricting the distribution of unobserved heterogeneity to warrant identification. Recently, this idea was used explicitly in the context of a Roy model of education with several earnings measures by Carneiro, Hansen and Heckman (2003).

However, the applicability of factor models is thought to be limited due to somewhat arbitrary identification. Traditionally, the identifying assumption of orthogonality of factors and errors (both within and between) is made in (1). We shall refer to this benchmark case as Orthogonal Factor Analysis (OFA), of which PCA can be seen as a particular case. Under the assumption of orthogonality, identification can be analyzed from the covariance equation: $\Sigma_Y = \Lambda \Sigma_X \Lambda^T + \Sigma_U$, where

M^T is the transpose of matrix M , and $\Sigma_Z = \mathbb{E}(ZZ^T)$ is the covariance matrix of Z . A classical restriction is to assume that factors have variance one. In this case:

$$\Sigma_Y = \Lambda\Lambda^T + \Sigma_U, \quad (2)$$

where Σ_U is diagonal. Equation (2) features $L(L - 1)/2$ covariances for $L \times K$ factor loadings. It follows that at most $K = (L - 1)/2$ factors can be identified from the data. This is the first limitation: to identify complex factor structures, one needs a large amount of data, a requirement that is not necessarily met in practice.

The second limitation stems from the fact that, as factors and errors are unobserved, pointwise identification is meaningless. To illustrate this characteristic of OFA models, note that if (Λ, Σ_U) satisfies equation (2), then also does $(\Lambda\Omega, \Sigma_U)$, for all Ω orthogonal, *i.e.* such that $\Omega\Omega^T = I_K$. Identification is thus defined up to a rotation matrix.

In this paper, we propose a solution to overcome these limitations. Our starting point is to note that, if (2) uses all the information of factor model (1) under *normality*, this is not necessarily true if either factors or errors are not normal, and *independence* of factors and errors is assumed instead of uncorrelatedness.

Relaxing the implicit normality assumption in factor models allows us to extend the literature in two different ways. First, by using identification from higher-order moments of the data, we show that we are able to mitigate the two limitations mentioned above. We show that, provided that there is enough skewness (respectively kurtosis) in the data, up to $L - 1$ (resp. L) factors are identified in Model (1). We also show that, in this case, identification holds up to trivial sign and permutation restrictions. In other words: up to this irreducible ambiguity, higher moments allow one to choose a particular rotation matrix.

Second, as factors and errors are not assumed normally distributed *a priori*, their distributions become of interest. We show how we can recover in a second step the densities of factors and errors, provided that the matrix of factor loadings Λ is known.

There are relatively few papers related to non-normal Factor Analysis. Nevertheless, we are

aware of three important approaches from which we shall borrow a lot, and that we now review. Each of these approaches can be seen as a special case of equation (1), together with the central assumption that factors, errors, and factors and errors are all independent.

The seminal paper of Reiersol (1950) considers the estimation of matrix Λ in (1), where factors and errors are assumed independent. He focuses on the particular case of two measurements and one factor ($L = 2$, $K = 1$). After noting that three covariances are not enough to estimate Λ and the variances of the errors, Reiersol shows that using third-order moments adds four equations and three unknown: the third-order moments of X , U_1 and U_2 respectively. He shows that this system is indeed just-identified, provided that X is non-symmetric.

Reiersol's result shows that non-symmetry, and more generally non-normality, can help identification of simple factor structures. This idea has been used several times in the econometrics literature by Madanski (1959), Pal (1980), Cragg (1998), Lewbel (1997), and Erickson and Whitted (2002) among others, to estimate one-factor models. Different estimators, using information from the whole characteristic function of the data instead of moments, were introduced by Spiegelman (1979) and Van Montfort *et al.* (1989). A closely related idea is the use of heteroskedasticity as a source of identification, as *e.g.* in Lewbel (2004).

Non-normality is also at the heart of Independent Component Analysis (ICA), although to our knowledge the connection has not been made so far. ICA is currently used in the Signal Processing literature, to separate sources (*i.e.* factors) from a given set of sensors (measurements). The basic ICA model assumes $U = 0$ in model (1). Especially related to our approach, several algorithms, such as the JADE algorithm of Cardoso and Souloumiac (1993), are based on second and fourth-order moments of the data; that is: identification comes from non-normal kurtosis. JADE allows explicit and computationally easy estimation of matrix Λ , through the joint diagonalization of matrices of fourth-order cumulants. Since its introduction, JADE has become increasingly popular in the Signal Processing community (Cardoso, 1999).

Recently, two extensions of the basic ICA model have been studied in this literature. The first

one, the so-called “noisy” ICA model, assumes normally distributed errors in (1). As, in social sciences, errors are fundamental to the extent that they both embody true measurement errors as well as factors specific to a given measurement, we think important to relax this restriction, by allowing general– not necessarily normal– errors. Moreover, noisy ICA models are usually estimated provided that the covariance structure of the errors is known. We extend this framework by devising algorithms which jointly estimate the error variances and the matrix of factor loadings.

The second extension of ICA models is provided by “overcomplete” ICA, where there can be more factors than measurements. Arguably, if errors are assumed non-normal, Model (1) can be seen as a special case of overcomplete ICA models. However, as we are not willing to assume normality of the errors, we do not want either to rule out normality on *a priori* grounds. Moreover, current methods to solve the “overcomplete” problem assume particular functional forms for the factors, such as mixtures of normals. A distribution-free estimation method has recently been proposed by de Lathauwer (2003) in the case where measurements are complex. In the real case, arguably the interesting one in econometrics, no similar algorithm has been proposed so far.

The third approach we shall build on concerns the identification and estimation of factor and error distributions. The seminal result we shall use is due to Kotlarski (1967). Assuming that $\Lambda = (1, 1)^T$ is known, Kotlarski shows that under general conditions the distributions of the single factor and the errors are identified. Kotlarski’s insight has been used in the deconvolution literature to build convergent estimators of the three densities. Horowitz and Markatou (1996) construct such an estimator in a particular case. Recently, Hall and Yao (2003) and Li and Vuong (1998) introduced two new estimators, each of them based on a different proof of Kotlarski’s result.

The outline of the paper is as follows. In Section 2, we show identification of both matrix Λ and factor and error densities in Model (1) under general conditions. We also prove necessary conditions, a result that seems new in the literature. The end of the section is devoted to the proof of parametric identification results, based on second-, third- and fourth-order cumulants of the data. Namely, we show that if there is sufficient skewness/ kurtosis in the data, then factor

loadings are identified.

In Section 3, we focus on the estimation of factor loadings. As the parametric identification proofs of Section 2 are constructive, we are in position to construct analog estimators based on these results. The algorithms we propose can be seen as extensions of Cardoso and Souloumiac’s (1993) JADE to deal with independent errors of an unknown form. As the original JADE algorithm, they do not require any complex maximization scheme. Moreover, we show that choosing the “right” number of factors can be done by using simple tests of ranks of some matrices of cumulants.

Given a root- N consistent estimate of Λ , we explain in Section 4 how to generalize Kotlarski’s result to Model (1). We devise uniformly convergent estimators of the factor and error densities, and explain how to obtain explicit convergence rates. In the particular case where $\Lambda = (1, 1)^T$, our estimator is close to the one introduced in Li and Vuong (1998).

In Section 5, we investigate the finite-sample properties of our estimators through Monte-Carlo simulations. We find that both the extent of non-normality in the data and the sample size are likely to be critical in the application of our method.

In Section 6, we apply our methodology to estimate returns to education. Using data from the French Labor Force Survey, we model the joint relationship between education and wages using a factor structure. We dispose of two measures of educational attainment: the number of years of schooling, and a second measure based on credentials or qualifications. We identify two factors in this relationship. The first factor is essentially the one that can be identified by classical IV methods, using the second measure of education as an instrument for the first one. This factor corresponds to a higher return than the one calculated by OLS. In contrast, the second factor that we identify is positively correlated to the number of schooling and negatively to the wage. Estimating factor and error distributions, we then are able to compute the correlation between the two factors and other labor market outcomes. The results confirm the existence of two very different factors in educational attainment, suggesting that classical measures of educational attainment such as the number of years of schooling might be highly heterogeneous.

Lastly, Section 7 concludes.

2 Identification of linear independent factor models

We consider linear factor models with the following structure:

Definition 1 (*Linear independent factor models*) A linear independent factor model for a vector $Y = (Y_1, \dots, Y_L)^T$ of $L \geq 2$ real-valued random variables (measurements)¹ is a DGP:

$$Y = \mu + \Lambda X + U,$$

where μ is a L -vector of non stochastic scalars and the triple (Λ, X, U) , that we call representation, is such that

- (i) $X = (X_1, \dots, X_K)^T$ is a random vector of $K \geq 1$ real valued, mutually independent and non degenerate random variables (factors) with zero means and unit variances,
- (ii) $\Lambda = (\lambda_1, \dots, \lambda_K)$ is a (L, K) matrix of scalar parameters (factor loadings),
- (iii) any column $\lambda_k \in \mathbb{R}^{L \times 1}$ of Λ has at least two non-zero entries,
- (iv) $U = (U_1, \dots, U_L)^T$ is a vector of L real-valued random variables (errors) with zero means and finite variances, which are mutually independent and independent of factors.

We start by noting that one can always center the measurements and errors. For this reason, we shall assume $\mu = 0$ is the rest of the paper.

In Condition (i) of this definition, we assume that factors have unit variances. Clearly, some normalization is necessary for the factor distributions *and* the factor loadings to be simultaneously identified, as one can change ΛX into $(\Lambda D) (D^{-1} X)$ for any diagonal matrix with non zero entries in the diagonal. We could alternatively assume that one of the rows of Λ is composed of factor loadings equal to one.

Condition (iii) is necessary to distinguish factors from errors. If there is one factor, say X_k , which associated vector of factor loadings, λ_k , has only one non zero entry, say $\lambda_{\ell k} \neq 0$ for measurement

¹We denote as A^T the transpose of a matrix A to avoid confusion with the notation for derivatives of univariate scalar functions.

Y_ℓ , then such a factor model is equivalent to a model where there is no factor X_k and the error on the ℓ th measurement is $U_\ell + \lambda_{\ell k} X_k$. Without specifying the distribution of U_ℓ one obviously cannot identify the distribution of X_k from the distribution of U_ℓ . Condition (iii) is thus a necessary condition for the identification of any “noisy” independent factor model.

Conditions (i) and (iv) assume that factor and error variables have finite variances. It follows that the characteristic functions of factors and errors are well defined, continuous and bounded and are of class C^2 on \mathbb{R} .

Moreover, it will be useful to work with cumulant generating functions (c.g.f.),² so we also assume that factor and error variables’ characteristic functions are nowhere equal to zero.

Assumption 2 (*Existence of cumulant generating functions*) *The characteristic functions of factors and errors do not vanish anywhere.*

Note that, as far as identification is concerned, it would be sufficient to assume that the characteristic functions of factors and errors are *almost everywhere* non zero, so that cumulant generating functions are a.e. well defined and two times differentiable (because second order moments exist).

Finally, if factors are normal then ICA boils down to Principal Component Analysis (PCA) which is now well understood. In the rest of the paper, we shall assume that all factors are non normal:

Assumption 3 (*Non-gaussianity*) *Factor variables X_k , $k = 1, \dots, K$, are non normal.*

Admittedly, intermediate cases where several factors are non normal and others are can be of great interest. The approach that we introduce in this paper does not directly generalize to this case. We leave this interesting question for future work.

We now turn to the definition of identifiability. For all K , let us define the set of sign-permutation matrices as the set \mathcal{S}_K composed of the products DP , where D is a diagonal matrix

²Let X be a vector of K random variables. Its characteristic function is the Fourier transform of its probability measure μ : $\varphi_X(t) = \mathbb{E} \exp(it^T X) = \int \exp(it^T X) \mu(dt)$, and its cumulant generating function is: $\kappa_X(t) = \ln \varphi_X(t)$, where \ln denotes the principal branch of the logarithm.

with coefficients equal to 1 or -1, and P is a permutation matrix.

For given values of L and K , let us consider (Λ, X, U) as in Definition 1 under Assumptions 2 and 3. Clearly:

$$\Lambda X = (\Lambda S) (S^T X),$$

and:

$$\mathbb{E}(S X X^T S^T) = S \mathbb{E}(X X^T) S^T = I_K,$$

for all $S \in \mathcal{S}_K$. Therefore, (Λ, X, U) is defined up to a multiplication by a sign-permutation matrix. A more remarkable result is that the representation (Λ, X, U) is generally *unique* up to permutation. The purpose of this section is to give a precise meaning to this statement (Theorems 6 and 9 below).

Let us start with a definition of identifiability:

Definition 4 (*Semiparametric Identification*) *A representation (Λ, X, U) as in Definition 1 is said identifiable if for every other representation $(\tilde{\Lambda}, \tilde{X}, \tilde{U})$ there exists a matrix S in \mathcal{S}_K such that: $\tilde{\Lambda} = \Lambda S$, $\tilde{X} \stackrel{d}{=} S^T X$, and $\tilde{U} \stackrel{d}{=} U$, where $\stackrel{d}{=}$ means “equal in distribution.”*

Note that the group \mathcal{S}_K is a *finite* subgroup of the infinite orthogonal group \mathcal{O}_K , up to which identification is defined in classical (Orthogonal) Factor Analysis. The quotient group $\mathcal{O}_K/\mathcal{S}_K$ is thus also infinite. As illustrated in Reiersol’s (1950) paper and in Independent Component Analysis (ICA), the independence and non-normality of the factor and error distributions allows one to choose one of the equivalence classes of $\mathcal{O}_K/\mathcal{S}_K$ as *the* relevant rotation. Independence and non-normality therefore considerably reduce the model’s indeterminacy.

The previous definition of identification draws information from the whole distribution of observed measurements. For practical reasons, it is useful to understand how much of the model’s structure can be identified from a finite set of parameter of that distribution. We thus also define parametric identification as follows.

Definition 5 (Parametric identification of order p) Let p be an integer not smaller than two. A representation (Λ, X, U) as in Definition 1 is said identified to the p th order if for any other factor representation $(\tilde{\Lambda}, \tilde{X}, \tilde{U})$, the equality of all moments of order less than p of $\Lambda X + U$ and $\tilde{\Lambda}\tilde{X} + \tilde{U}$ implies that there exists a sign-permutation matrix $S \in \mathcal{S}_K$ such that $\tilde{\Lambda} = \Lambda S$, and \tilde{X} and $S^T X$ (resp. \tilde{U} and U) have all moments of order less than p equal.

2.1 Identifying restrictions

In this subsection and the next, we shall develop some implications of the factor structure introduced in Definition 1 in terms of cumulant generating functions and their derivatives. These results will prove especially useful in the identification proofs of 2.3 and 2.4, and in the construction of our estimators in the next Section.

Denote the cumulant generating functions (the log of characteristic functions) of Y, X_k and U_ℓ as κ_Y, κ_{X_k} and κ_{U_ℓ} . The independence assumptions of the linear factor model of Definition 1 imply that, for $t = (t_1, \dots, t_L) \in \mathbb{R}^L$,

$$\kappa_Y(t) = \sum_{k=1}^K \kappa_{X_k}(\lambda_k^T t) + \sum_{\ell=1}^L \kappa_{U_\ell}(t_\ell). \quad (3)$$

Let $\alpha = (\alpha_1, \dots, \alpha_L)$ be a multi-index with length p ($\alpha_\ell \in \{0, \dots, L\}$ and $|\alpha| \equiv \alpha_1 + \dots + \alpha_L = p$). For any vector $x = (x_1, \dots, x_L) \in \mathbb{R}^L$, define the monomial $x^\alpha = x_1^{\alpha_1} \dots x_L^{\alpha_L}$. Then, assuming that derivatives exist, we have

$$\kappa_Y^{(\alpha)}(t) \equiv \partial_\alpha \kappa_Y(t) \equiv \frac{\partial^{|\alpha|} \kappa_Y(t)}{\partial t_1^{\alpha_1} \dots \partial t_L^{\alpha_L}} = \sum_{k=1}^K \lambda_k^\alpha \kappa_{X_k}^{(p)}(\lambda_k^T t) + \sum_{\ell=1}^L \mathbf{1}\{\alpha_\ell = p\} \kappa_{U_\ell}^{(p)}(t_\ell), \quad (4)$$

where $\kappa_{X_k}^{(p)}$ and $\kappa_{U_\ell}^{(p)}$ are the p th derivative of κ_{X_k} and κ_{U_ℓ} .

Let $\bar{\Delta}_p$ be the set of multi-indices of length p . Let Δ_p be the set of multi-indices of length p except the L ones of the form $(\dots, 0, p, 0, \dots)$:

$$\begin{aligned} \bar{\Delta}_p &= \left\{ \alpha = (\alpha_1, \dots, \alpha_L) \in \{0, \dots, p\}^L : |\alpha| = \alpha_1 + \dots + \alpha_L = p \right\}, \\ \Delta_p &= \left\{ \alpha = (\alpha_1, \dots, \alpha_L) \in \{0, \dots, p-1\}^L : |\alpha| = \alpha_1 + \dots + \alpha_L = p \right\}, \end{aligned}$$

and let $\#\overline{\Delta}_p$ (resp. $\#\Delta_p$) be the number of elements in $\overline{\Delta}_p$ (resp. Δ_p). For instance, for $p = 2$: $\#\overline{\Delta}_2 = L(L + 1)/2$ and $\#\Delta_2 = L(L - 1)/2$.

Let $\overline{\kappa}_Y^{(p)}(t) = \left(\kappa_Y^{(\alpha)}(t), \alpha \in \overline{\Delta}_p \right)$, for $t \in \mathbb{R}^L$, be the $(\#\overline{\Delta}_p, 1)$ -vector of p th-order partial derivatives, and let $\kappa_Y^{(p)}(t) = \left(\kappa_Y^{(\alpha)}(t), \alpha \in \Delta_p \right)$, for $t \in \mathbb{R}^L$, be the $(\#\Delta_p, 1)$ -vector of p th-order partial *cross*-derivatives of κ_Y . Let also

$$\kappa_X^{(p)}(t) = \left(\kappa_{X_1}^{(p)}(t_1), \dots, \kappa_{X_K}^{(p)}(t_K) \right)^T, \quad t = (t_1, \dots, t_K) \in \mathbb{R}^K,$$

be the $(K, 1)$ -vector of p th-order derivatives of the cumulant generating functions of the K factors.

Lastly, let \overline{A}_p and A_p be defined as the operators on a matrix Λ that change its columns into column-wise cross-products:

$$\begin{aligned} \overline{A}_p(\Lambda) &= [\lambda_1^\alpha, \dots, \lambda_K^\alpha; \alpha \in \overline{\Delta}_p], \\ A_p(\Lambda) &= [\lambda_1^\alpha, \dots, \lambda_K^\alpha; \alpha \in \Delta_p]. \end{aligned}$$

Equation (4) implies the following restrictions on factor cumulant generating functions:

$$\overline{\kappa}_Y^{(p)}(t) = \overline{A}_p(\Lambda) \kappa_X^{(p)}(\Lambda^T t) + \overline{\kappa}_U^{(p)}(t), \quad (5)$$

and hence:

$$\kappa_Y^{(p)}(t) = A_p(\Lambda) \kappa_X^{(p)}(\Lambda^T t). \quad (6)$$

2.2 Moment restrictions

If one lets $t = 0$ in equation (6) one obtains a set of restrictions on the cumulants (moments) of factors and measurements. For a given order p , let $\alpha \in \Delta_p$ be a multi-index of length p . Write α as $\alpha = \iota_{\ell_1} + \dots + \iota_{\ell_p}$, where ι_ℓ denotes the ℓ th column of the identity matrix of dimension L . Then, the second-order cumulants of zero-mean random variables are equal to their covariances:

$$\kappa_Y^{\iota_{\ell_1} + \iota_{\ell_2}}(0) \equiv \text{Cum}(Y_{\ell_1}, Y_{\ell_2}) = \mathbb{E}(Y_{\ell_1} Y_{\ell_2}), \quad (7)$$

Third-order cumulants are:

$$\begin{aligned} \kappa_Y^{\iota_{\ell_1} + \iota_{\ell_2} + \iota_{\ell_3}}(0) &\equiv \text{Cum}(Y_{\ell_1}, Y_{\ell_2}, Y_{\ell_3}) \\ &= \mathbb{E}(Y_{\ell_1} Y_{\ell_2} Y_{\ell_3}). \end{aligned} \quad (8)$$

And fourth-order cumulants:

$$\begin{aligned}
\kappa_Y^{\iota_{\ell_1} + \iota_{\ell_2} + \iota_{\ell_3} + \iota_{\ell_4}}(0) &\equiv \text{Cum}(Y_{\ell_1}, Y_{\ell_2}, Y_{\ell_3}, Y_{\ell_4}) \\
&= \mathbb{E}(Y_{\ell_1} Y_{\ell_2} Y_{\ell_3} Y_{\ell_4}) - \mathbb{E}(Y_{\ell_1} Y_{\ell_2}) \mathbb{E}(Y_{\ell_3} Y_{\ell_4}) \\
&\quad - \mathbb{E}(Y_{\ell_1} Y_{\ell_3}) \mathbb{E}(Y_{\ell_2} Y_{\ell_4}) - \mathbb{E}(Y_{\ell_2} Y_{\ell_3}) \mathbb{E}(Y_{\ell_1} Y_{\ell_4}).
\end{aligned} \tag{9}$$

We shall use moment conditions up to the fourth order. Restrictions (6) when $t = 0$ take the following form. Covariance restrictions:

$$\begin{aligned}
\text{Cum}(Y_{\ell_1}, Y_{\ell_2}) &= \sum_{k=1}^K \lambda_{\ell_1, k} \lambda_{\ell_2, k}, \\
\ell_1, \ell_2 &= 1, \dots, L, \ell_1 \neq \ell_2,
\end{aligned} \tag{10}$$

as $\text{Var}(X_k) = 1$. Third-order restrictions:

$$\begin{aligned}
\text{Cum}(Y_{\ell_1}, Y_{\ell_2}, Y_{\ell_3}) &= \sum_{k=1}^K \lambda_{\ell_1, k} \lambda_{\ell_2, k} \lambda_{\ell_3, k} \mathbb{E}(X_k^3), \\
\ell_1, \ell_2, \ell_3 &= 1, \dots, L, \ell_1 \neq \ell_2 \text{ or } \ell_1 \neq \ell_3,
\end{aligned} \tag{11}$$

as $\mathbb{E}(X_k) = 0$. Lastly, fourth-order restrictions:

$$\begin{aligned}
\text{Cum}(Y_{\ell_1}, Y_{\ell_2}, Y_{\ell_3}, Y_{\ell_4}) &= \sum_{k=1}^K \lambda_{\ell_1, k} \lambda_{\ell_2, k} \lambda_{\ell_3, k} \lambda_{\ell_4, k} \text{kur}(X_k), \\
\ell_1, \ell_2, \ell_3, \ell_4 &= 1, \dots, L, \ell_1 \neq \ell_2 \text{ or } \ell_1 \neq \ell_3 \text{ or } \ell_1 \neq \ell_4,
\end{aligned} \tag{12}$$

where $\text{kur}(X_k) = \mathbb{E}(X_k^4) - 3\mathbb{E}(X_k^2)^2 = \mathbb{E}(X_k^4) - 3$ is the fourth-order cumulant (kurtosis excess) of X_k .

For easy manipulation, we shall use these moment conditions in matrix form. Let

$$\Sigma_Y = (\text{Cum}(Y_{\ell_1}, Y_{\ell_2}))_{(\ell_1, \ell_2) \in \{1 \dots L\}^2},$$

be the (L, L) covariance matrix of Y . Similarly, let

$$\begin{aligned}
\Gamma_Y &= (\text{Cum}(Y_{\ell_1}, Y_{\ell_2}, Y_{\ell_3}))_{(\ell_1, \ell_2, \ell_3) \in \{1 \dots L\} \times \#\Delta_2}, \\
\Gamma_Y &= (\text{Cum}(Y_{\ell_1}, Y_{\ell_2}, Y_{\ell_3}, Y_{\ell_4}))_{(\ell_1, \ell_2, \ell_3, \ell_4) \in \#\bar{\Delta}_2 \times \#\Delta_2},
\end{aligned}$$

be the $(L, \#\Delta_2)$ and $(\#\bar{\Delta}_2, \#\Delta_2)$ matrices of third- and fourth- order cumulants of Y , respectively.

Lastly, let Σ_U be the (L, L) diagonal matrix of variances of U , and let D_3 and D_4 be the (K, K) matrices of third- and fourth- order cumulants of X , respectively.

Then the factor structure given by Definition 1 implies:

$$\begin{aligned}\Sigma_Y &= \Lambda\Lambda^T + \Sigma_U, \\ \Gamma_Y &= \Lambda D_3 A_2(\Lambda)^T, \\ \Omega_Y &= \bar{A}_2(\Lambda) D_4 A_2(\Lambda)^T.\end{aligned}$$

2.3 Semiparametric identification

The two previous subsections clarified the structure of linear independent factor models, and revealed an essentially algebraic, or more precisely polynomial, structure. We here use these restrictions to give necessary and sufficient conditions for identification of independent factor models. In 2.4, we shall study two parametric particular cases of these theorems. We first proceed by giving sufficient conditions for identification.

Theorem 6 (*Sufficient conditions for identification*) *Let (Λ, X, U) be a representation as in Definition 1 under Assumptions 2 and 3. Let $(\tilde{\Lambda}, \tilde{X}, \tilde{U})$ be an alternative representation. The following two propositions are true:*

- (i) *Every column of Λ is a scalar multiple of a column of $\tilde{\Lambda}$.*
- (ii) *If $A_2(\Lambda)$ is full-column-rank, then (Λ, X, U) identified.*

Theorem 6 says that, if factor variables are not normally distributed, then the matrix of factor loadings is identified whatever the number of factors. This result, which can be seen as an application of a Theorem by Darmois (1953), is well-known in the ICA community, at least since Comon (1994). However, it is likely to seem quite extraordinary to most econometricians accustomed to second-order statistics.

Moreover, it suffices that $rank(A_2(\Lambda)) = K$ for the distributions of factors and errors to be identified. One thus measures how informative the independence and non-normality properties are,

compared to PCA which instead assumes the absence of correlation between factors and errors. A model with $\#\Delta_2 = L(L - 1)/2$ factors is potentially identifiable when factors and errors are assumed independent instead of uncorrelated. This is considerably more than the maximal number of L factors in PCA (where it is to be noted that errors are not allowed for). Moreover, the matrix of factor loadings is identifiable up to a diagonal matrix and a permutation instead of a rotation matrix (an orthogonal matrix), which induces only K constraints on the parameters (if the factor variances are left free) instead of $K(K + 1)/2$.

Theorem 6 is essentially the same as sufficient condition (iii) for uniqueness of ICA derived by Eriksson and Koivunen (2003). The first proposition of Theorem 6 immediately follows from the non-gaussianity of factors and errors by a straightforward application of a result due to Kagan, Linnik and Rao (1973) that we state below. Proposition (ii) of Theorem 6 easily follows from proposition (i).

Theorem 7 (Theorem 10.3.1, Kagan, Linnik and Rao, 1973) *Let A and B be two non-stochastic matrices and let $S = (s_1, \dots, s_m)^T$ and $R = (r_1, \dots, r_n)^T$ 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 i th column of A is the multiple of a column of B .*

We now show that the rank condition in proposition (ii) is generically necessary, that is: for a class of distribution functions topologically dense in the set of continuous distribution functions. As far as we know, this is a new result in the literature on ICA. Let us define first the class of d.f.'s divisible by a normal distribution.

Definition 8 (Distribution divisible by a normal) *Let X be a continuous random variable with density f and characteristic function φ . The distribution of X is divisible by a normal distribution if there exists $\sigma^2 > 0$ such that $\tilde{\varphi}(t) = \varphi(t) \exp\left(\frac{\sigma^2 t^2}{2}\right)$ is the characteristic function of a random variable \tilde{X} .*

The distribution of a variable X is divisible by a normal if and only if $X \stackrel{d}{=} \tilde{X} + N(0, \sigma^2)$, where $N(0, \sigma^2)$ is a normal r.v. with mean 0 and variance σ^2 . Note that this rules out the case of random variables with bounded support. Moreover, the c.f. of X must tend to zero when t tends to infinity faster than the c.f. of a normal distribution. Lastly, notice that the set of distributions divisible by a normal is dense in the set of continuous distribution functions.³

For a representation (Λ, X, U) to be identifiable, the next theorem shows that either $A_2(\Lambda)$ is full-column-rank or it is not, but then at least some of the factor and error variables must not be divisible by a normal distribution.

Theorem 9 (Necessary condition for identification) *Let (Λ, X, U) be a representation as in Definition 1 under Assumptions 2 and 3. Suppose that $A_2(\Lambda)$ is not full-column-rank ($\ker(A_2(\Lambda)) \neq \{0\}$) and that the distributions of factors and errors are divisible by normal distributions. Then (Λ, X, U) is not identifiable.*

We refer the reader to the Appendix for a proof of Theorem 9. We show that under the assumptions of Theorem (9), for any representation (Λ, X, U) such that $A_2(\Lambda)$ is not full-column-rank and the distributions of factors X and errors U are divisible by normal distributions, then one can construct another representation $(\tilde{\Lambda}, \tilde{X}, \tilde{U})$ that is not equal to (Λ, X, U) up to a sign-permutation matrix and that still verifies the equality $\Lambda X + U \stackrel{d}{=} \tilde{\Lambda} \tilde{X} + \tilde{U}$.

2.4 Parametric identification

We shall now give two sufficient conditions for parametric identification of order 3 and 4. The idea of the proofs is first to show that factor loadings can be recovered from a joint diagonalization problem and then apply the following lemma:

Lemma 10 *Let K and L be any integers. Let A_1, \dots, A_L be matrices of $\mathbb{R}^{K \times K}$. Suppose that there exist $K + 1$ couples of vectors: $x^k = (x_1^k, \dots, x_L^k)^T \in \mathbb{R}^L$ and $v^k \in \mathbb{R}^K$, $v^k \neq 0$, $k = 1, \dots, K + 1$,*

³Let X be a continuous random variable. Let $X_n = X + N(0, \sigma^2)$. Then $X_n \xrightarrow{d} X$ when $\sigma^2 \rightarrow 0$.

solutions to a system of L eigenvalue problems:

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

If the set $\{v^1, \dots, v^K\}$ is linearly independent, then x^{K+1} must be equal to some x^k , $k = 1, \dots, K$.

For $L = 1$, this is of course trivially true. Now, take $L = K = 2$. Matrix A_1 has at most two different eigenvalues and so does matrix A_2 . There are potentially four combinations of two eigenvalues that can be solutions to the joint eigenvalue problem. Lemma 10 tells us that there are at most two linearly independent combinations that can be a solution.

The first theorem shows that the matrix of factor loadings is identified from second and third-order moments, for all parameter values but a set of Lebesgue-measure zero that we characterize, provided that $K \leq L - 1$, and none of the distributions X_k is symmetric. This result can be seen as a generalization of Reiersol's (1950) identification result for his measurement error model B.

Theorem 11 (*Parametric identification of order three*) *Assume that*

- (i) $K \leq L - 1$,
- (ii) every submatrix of Λ made of a selection of $L - 1$ rows has rank K ,
- (iii) the first row of Λ has all its components different from zero,
- (iv) the third-order moments of factor variables, X_k , are finite and non zero.

Then the factor loadings in model (1) are identified from second and third-order moments.

The second theorem shows that, again for all parameter values but a set of Lebesgue-measure zero that we characterize, the matrix of factor loadings is identified from second and fourth-order moments for $K \leq \min\{L, L(L - 1)/2\}$,⁴ provided that factor distributions show excess kurtosis (non zero fourth-order cumulant).

Theorem 12 (*Parametric identification of order four*) *Assume*

- (i) $K \leq \min\{L, L(L - 1)/2\}$,

⁴Note that $L \leq L(L - 1)/2$ if $L \geq 3$.

- (ii) there exists a subset of K linearly independent rows among the last $L - 1$ rows of Λ ,
- (iii) $A_2(\Lambda)$ has rank K ,
- (iv) the first row of Λ has all its components different from zero and
- (v) factor variables, X_k , have finite and non zero kurtosis excess.

Then, the factor loadings in model (1) are identified from second and fourth-order moments.

Note that the restrictions on the matrix of factor loadings are weaker in Theorem 12 than in Theorem 11. For example, matrix

$$\Lambda = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ \lambda_{31} & \lambda_{32} \end{pmatrix},$$

with $\lambda_{31} \neq \lambda_{32}$, satisfies conditions (ii)-(iv) of Theorem 12 but not condition (ii) of Theorem 12.

The proofs of these two Theorems, that we detail in the Appendix, are constructive. In both cases, we show that the factor structure can be recovered from fourth-order cumulants by exploiting their algebraic structure and rewriting the set of moment restrictions as a joint diagonalisation problem.

More precisely, under either of the two Theorems' sets of assumptions, we show how to construct a set of $L \times L$ matrices $B(\ell)$, $\ell \leq L$, such that Λ satisfies, for all ℓ :

$$B(\ell)\Lambda = \Lambda D(\ell),$$

where $D(\ell)$ is a $K \times K$ diagonal matrix.

The expression of $B(\ell)$ is given in the Appendix. When using second- and third-order moments only (Theorem 11), the expression of $B(\ell)$ involves third-order cumulants of the data. When using second- and fourth-order moments (Theorem 12), $B(\ell)$ can be directly constructed from fourth-order cumulants of the data.

In the next Section, we shall present an estimation method that explicitly solves this joint diagonalization problem. Such an idea is the basis of many popular algorithms in the ICA literature (see Cardoso, 1999, for a survey of applications restricted to the case $K = L$, and the Appendix for a presentation of the principles of the JADE algorithm). As emphasized in the Introduction, this

literature essentially deals with factor models without errors. As in economics the error variance can be large, we here develop techniques which apply to the ICA model with error structure exactly. We also provide exact conditions for identification and do not rely on genericity arguments. This can prove useful in econometrics, as many relevant economic models are often non generic cases of the general model.

3 Estimation of independent factor models

In this section, we consider the estimation of the factor loadings and factor and error distributions. Our method to recover factor loadings is based on matrices of third- and/ or fourth- order cumulants. We start by showing that these matrices contain information on the number of factors K in the model.

3.1 Estimating the number of factors K

We here show how to estimate the number of factors from the two matrices of third- and fourth-order cumulants: Γ_Y and Ω_Y , respectively.

Following the notations of 2.2, the factor structure implies that:

$$\begin{aligned}\Gamma_Y &= \Lambda D_3 A_2 (\Lambda)^T, \text{ and} \\ \Omega_Y &= \bar{A}_2 (\Lambda) D_4 A_2 (\Lambda)^T.\end{aligned}$$

Let K be the number of factors. Under the assumption that all factors are non-symmetric, $rank(\Gamma_Y) = K$. Under the assumption that all factors have non zero kurtosis excess, $rank(\Omega_Y) = K$. Thus, estimating the number of factors from third- and fourth- order cumulants can be achieved through tests of ranks of matrices of cumulants.

To determine the number of factors, we propose to proceed as follows. First, we estimate Ω_Y .⁵ To estimate the cumulants of Y from the data, we simply replace the expectations in equations (8) and (9) by sample means. This yields a root- N consistent estimator for Ω_Y , where N is the

⁵We here focus on the case of fourth-order moments only. Estimating the rank of Γ_Y is done identically.

sample size. In practice, trimming a small amount of extreme observations is important for the test to have reasonably good size and power performance. In 4.2 we perform a Monte-Carlo study that illustrates this fact.

Then, to simplify the notations, let $p = L(L - 1)/2$ and $q = L(L + 1)/2$. There exists $\widehat{\Omega}_Y$ such that

$$N^{1/2}vec\left(\widehat{\Omega}_Y - \Omega_Y\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{\Omega}), \quad (13)$$

where $\mathbf{\Omega}$ is finite and $rank(\mathbf{\Omega}) = s$, $0 < s \leq pq$. Let $\widehat{\mathbf{\Omega}}$ be a consistent estimate of $\mathbf{\Omega}$.

We propose to estimate K using the sequential procedure developed in Robin and Smith (2000). Let $\widehat{\Omega}_Y = \widehat{B}\widehat{D}\widehat{C}^T$ be the singular value decomposition of $\widehat{\Omega}_Y$, where \widehat{B} and \widehat{C} are (q, q) and (p, p) orthogonal matrices, and \widehat{D} is a (q, p) diagonal matrix. Let $\widehat{d}_1 \geq \dots \geq \widehat{d}_K$ denote the diagonal entries of \widehat{D}^2 (eigenvalues of $\widehat{\Omega}_Y\widehat{\Omega}_Y^T$). For a given null hypothesis: $H_0^r : K = r$, the statistics

$$\mathcal{CRT}_r \equiv N \sum_{i=r+1}^q \widehat{d}_i \quad (14)$$

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

$$(\widehat{C}_{q-r} \otimes \widehat{B}_{p-r})^T \widehat{\mathbf{\Omega}} (\widehat{C}_{q-r} \otimes \widehat{B}_{p-r}), \quad (15)$$

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

To estimate K , we apply the following method: 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 $-N^{-1} \ln \alpha_N^r = o(1)$.

Before turning to the estimation of factor loadings, note that K as estimated in this section is an upper bound of the number of factors that can be estimated from second-, third- and fourth-order

moments of Y . The reason is that the above tests do not use the covariance structure of Y . We shall come back to this point in the next subsection.

3.2 Estimation of factor loadings

The proofs of parametric identification are constructive proofs explaining how to recover factor loadings from data cumulants using simple eigenvalue decomposition techniques. This is certainly useful, if only to provide a set of starting values for a full-information GMM procedure of estimation.

3.2.1 Estimation using fourth-order moments when $K \leq L$

We start by explaining how to construct an analog estimator based on the proof of Theorem 12, in the case where $K \leq L$.

Factor loadings satisfy the following system of identifying restrictions (see 2.2):

$$\begin{aligned}\Sigma_Y &= \Lambda\Lambda^T + \Sigma_U, \\ \Gamma_Y &= \Lambda D_3 A_2(\Lambda)^T, \text{ and} \\ \Omega_Y(\ell) &= \Lambda D_4 \text{diag}(\Lambda_\ell) A_2(\Lambda)^T,\end{aligned}\tag{16}$$

for all $\ell = 1 \dots L$. Note that $\Omega_Y(\ell)$ are composed of L rows of Ω_Y . Thus, $\Omega_4(\ell)$ is a matrix of fourth-order cumulants of Y . For notational convenience, let us also define $\Omega_Y(L+1) \equiv \Gamma_Y$.

Now, let \mathcal{M} be the set of $(K, \#\Delta_2)$ projection matrices selecting K independent rows of $A_2(\Lambda)$. It follows from the assumptions of Theorem 12 that \mathcal{M} is not empty. Let also $(\ell, M) \in \{2 \dots L\} \times \mathcal{M}$.

Then:

$$\Omega_Y(\ell) M^T [\Omega_Y(1) M^T]^- = \Lambda \text{diag}(\Lambda_\ell) \text{diag}(\Lambda_1)^{-1} \Lambda^-,$$

and:

$$\Omega_Y(L+1) M^T [\Omega_Y(1) M^T]^- = \Lambda D_3 D_4^{-1} \text{diag}(\Lambda_1)^{-1} \Lambda^-,$$

where $Z^- = [ZZ^T]^{-1} Z^T$ is the (Moore-Penrose) generalized inverse of matrix Z .

Therefore, an estimation method for Λ , based on one single pair $(\ell, M) \in \{2 \dots L+1\} \times \mathcal{M}$, could be the following:

1. Estimate $\Omega_Y(\ell)$, $\Omega_Y(1)$, and $\widehat{B}(\ell, M) \equiv \widehat{\Omega}_Y(\ell)M^T [\widehat{\Omega}_Y(1)M^T]^-$.
2. Diagonalize $\widehat{B}(\ell, M)$, and recover matrix \widehat{W} as the matrix of eigenvectors.
3. Estimate $\widehat{\Sigma}_Y$.
4. Using (16), estimate Λ by:

$$\widehat{\Lambda}(\ell, M) = \widehat{W} \text{diag} \left(\sqrt{[A_2(\widehat{W})]^- \text{vecs}(\widehat{\Sigma}_Y)} \right),$$

where $\text{vecs}(\widehat{\Sigma}_Y)$ is the $L(L-1)/2$ vector of covariances of Y , ordered as in $A_2(\widehat{W})$.

However, there are at least three practical problems with this approach. First, the spectrum of matrix $\widehat{B}(\ell, M)$ is not necessarily real. Second, matrix $[A_2(\widehat{W})]^- \text{vecs}(\widehat{\Sigma}_Y)$ is not necessarily positive. Lastly, $\widehat{B}(\ell, M)$ can have multiple eigenvalues.

A similar problem was encountered in the ICA literature when Cardoso introduced his FOBI algorithm (Cardoso, 1989). In the absence of errors, a convenient way to solve the problem is to “pre-whiten” the cumulant matrices, rendering them symmetric and thus diagonalizable. Factor loadings are then recovered as the (conveniently normalized) joint orthonormal eigenvectors of these symmetric matrices. In the Appendix, we concisely present the approach used in Cardoso and Souloumiac (1993).

In our context, where error variances are *a priori* unknown, such a “pre-whitening” method is not directly applicable. However, we here show that a similar insight can be used in this case also. We proceed in two stages.

Stage 1: estimation of error variances The idea of our approach is the following: as $B(\ell, M) = \Omega_Y(\ell)M^T [\Omega_Y(1)M^T]^-$ satisfies $B(\ell, M)\Lambda = \Lambda D(\ell, M)$, where $D(\ell, M)$ is diagonal, it follows that

$$\begin{aligned} B(\ell, M) (\Sigma_Y - \Sigma_U) &= B(\ell, M)\Lambda\Lambda^T \\ &= \Lambda D(\ell, M)\Lambda^T, \end{aligned}$$

is symmetric. Thus, for all index ℓ and all matrix M :

$$vec(B(\ell, M)\Sigma_Y - \Sigma_Y B(\ell, M)^T) = [I_L \otimes B(\ell, M) - B(\ell, M) \otimes I_L] vec(\Sigma_U). \quad (17)$$

This is a set of linear restrictions identifying the variance of errors. Standard Least Squares estimates provide an easy way of estimating $\Sigma_U = Var(U)$ from second-, third- and fourth-order cumulants of Y . Note that by replacing $B(\ell, M)$ and Σ_Y by empirical analogs, one introduces statistical errors which can somewhat be reduced by using GLS instead of OLS. We use bootstrap to provide an estimate of the system's variance. Finally, in practice it can be important to impose the second-order constraints: the diagonal matrix Σ_U has positive coefficients, and matrix $\Sigma_Y - \Sigma_U$ is positive definite.

Stage 2: estimation of factor loadings At this point, Stage 1 has provided us with the noise variances. In the second stage, we separate the factors. The insight is that, when Σ_U is consistently estimated, it is possible to “whiten” the cumulant matrices $B(\ell, M)$ and to diagonalize:

$$\begin{aligned} (\Sigma_Y - \Sigma_U)^{-1/2} B(\ell, M) (\Sigma_Y - \Sigma_U)^{1/2} &= (\Sigma_Y - \Sigma_U)^{-1/2} \Lambda D(\ell, M) \left[(\Sigma_Y - \Sigma_U)^{-1/2} \Lambda \right]^T, \quad (18) \\ &\equiv C(\ell, M), \end{aligned}$$

where $(\Sigma_Y - \Sigma_U)^{-1/2} \Lambda$ is unitary by construction, and thus the LHS in (18) is symmetric.

Thus, this second step writes as a joint diagonalization problem of matrices $C(\ell, M)$, for all $(\ell, M) \in \{2 \dots L + 1\} \times \mathcal{M}$. As a particular case, if Σ_U equals zero one finds a similar expression as the one used in JADE.

A convenient way to estimate this system is to use the Jacobi algorithm developed by Cardoso and Souloumiac (1993). As JADE is a joint diagonalization algorithm, it is able to deal with cases where several of the above matrices have multiple roots. Moreover, the algorithm is very efficient computationally: in most cases, numerical convergence is achieved in less than ten iterations. See Appendix for a presentation.

In the case of the JADE algorithm, all the cumulant matrices are treated symmetrically. In econometric applications, however, we do not think that equi-weighting is appropriate. We propose

the following weighting scheme: first, we bootstrap the estimation of $\Lambda(\ell, M)$, for all (ℓ, M) , and compute:

$$\omega_{\ell, M} = \left\| \text{vec} \left(\widehat{\Lambda}(\ell, M) \right) \right\|_2^2, \quad (19)$$

where $\|v\|_2$ is the euclidian norm of vector v . Then, instead of diagonalizing matrices $C(\ell, M)$, we diagonalize $\omega_{\ell, M}^{-1} C(\ell, M)$ in a joint orthogonal basis.

To summarize, we propose to estimate Λ by the following procedure:

1. Estimate Σ_Y , $B(\ell, M)$ for all $(\ell, M) \in \{2 \dots L+1\} \times \mathcal{M}$, and estimate their covariance matrices by a first bootstrap procedure.
2. Estimate Σ_U by GLS in (17), replacing Σ_Y , $B(\ell, M)$ for all (ℓ, M) and their covariance matrices by the estimates obtained in the previous step.
3. For all (ℓ, M) , compute $\widehat{C}(\ell, M)$ and impose symmetry by $\widetilde{C}(\ell, M) = \frac{1}{2}(\widehat{C}(\ell, M) + \widehat{C}(\ell, M)^T)$. Obtain $\widehat{V}(\ell, M)$ as the orthogonal matrix of eigenvectors of $\widetilde{C}(\ell, M)$ and compute

$$\widehat{\Lambda}(\ell, M) = (\widehat{\Sigma}_Y - \widehat{\Sigma}_U)^{1/2} \widehat{V}(\ell, M),$$

where $\widehat{\Sigma}_Y$ and $\widehat{\Sigma}_U$ were obtained in the two first steps.

4. Bootstrap Step 3. This yields the covariance matrix of $\text{vec}(\widehat{\Lambda}(\ell, M))$, for all (ℓ, M) .
5. Compute the weights $\omega_{\ell, M}$ by (19).
6. Estimate \widehat{W} as the orthogonal matrix of joint eigenvectors of $\omega_{\ell, M}^{-1} \widetilde{C}(\ell, M)$, for all (ℓ, M) , and compute

$$\widehat{\Lambda} = (\widehat{\Sigma}_Y - \widehat{\Sigma}_U)^{1/2} \widehat{W},$$

where $\widehat{\Sigma}_Y$ and $\widehat{\Sigma}_U$ were obtained in the two first steps.

7. Obtain the covariance matrix of $\text{vec}(\widehat{\Lambda})$ by bootstrapping Step 6.

In the rest of the text we shall refer to this algorithm as 2S-AD (2-Stage Approximate Diagonalization). Note that the choice of $\Omega_Y(1)$ as the baseline matrix is arbitrary. In practice, we shall consider all pairs of matrices and compute:

$$B(\ell, \ell') = \Omega_Y(\ell')M(\ell)^T [\Omega_Y(\ell)M(\ell)^T]^{-1},$$

for $\ell < \ell'$, where we choose matrix $M(\ell)$ so as to guarantee that $\Omega_Y(\ell)M(\ell)^T$ is full-column rank. A convenient procedure is to compute its Singular Value Decomposition, as $\Omega_Y(\ell)M(\ell)^T = VDW^T$, and to set $M(\ell) = W_{1:K}^T$, where $W_{1:K}$ is composed of the first K columns of matrix W .

This algorithm can be modified in several ways. For instance, one can use only several moments, *i.e.* several matrices $B(\ell, M)$, to estimate Λ . In particular, if factors are symmetric (that is: if the hypothesis that the rank of Γ_Y is zero is not rejected by the test presented in 3.1) then, for all matrix M , $B(0, M)$ is not informative and can be dropped from the set of matrices $B(\ell, M)$.

3.2.2 Estimation using third-order moments when $K \leq L - 1$

We here treat the case considered in Theorem 11; that is: $K < L$ and only third-order moments are available. In econometrics, unlike in ICA, there are at least two reasons to consider this case. First, there are no *a priori* reasons to assume that factors are symmetric. In contrast, the signal processing literature is mainly motivated by the application to symmetric signals. Second, it is often the case in econometrics that sample sizes are large enough to estimate third-order moments precisely, yet too small to rely on fourth-order moments. These two reasons might explain in part why the econometrics literature reviewed in the Introduction has focused mainly on algorithms that use third-order moments.

The algorithm we propose differs from 2S-AD to the only extent as matrices $B(\ell, M)$ are con-

structured. In the proof of Theorem 11 we show the identity:

$$\begin{aligned}\Gamma_Y(\ell) &\equiv \Lambda D_3 \text{diag}(\Lambda_\ell) \Lambda^T, \\ &= \Gamma_Y(-\ell) \times \begin{pmatrix} 1 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & \dots & 0 \\ -\frac{v_{1,L}}{v_{\ell,L}} & \dots & -\frac{v_{\ell-1,L}}{v_{\ell,L}} & -\frac{v_{\ell+1,L}}{v_{\ell,L}} & \dots & -\frac{v_{L,L}}{v_{\ell,L}} \\ 0 & \dots & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & \dots & 1 \end{pmatrix}^T,\end{aligned}$$

where

$$\Gamma_Y(-\ell) = \Lambda D_3 \text{diag}(\Lambda_\ell) (\Lambda_{-\ell})^T$$

is composed of L-1 columns of Γ_Y except the ℓ th one, and v_{ij} are the elements of matrix V such that VDW^T is the Singular Value Decomposition of Γ_Y .

Let M be a permutation matrix selecting K linearly independent rows of Λ . Under the Theorem's assumptions, $\Gamma_Y(1)M^T$ is full-column rank. We then define $B(\ell, M) = \Gamma_Y(\ell)M^T [\Gamma_Y(1)M^T]^-$ for all $\ell \in \{2\dots L\}$, where Z^- is the left generalized inverse of matrix Z .

To estimate $B(\ell, M)$ we propose to proceed as follows:

1. Estimate Γ_Y as $\widehat{\Gamma}_Y$.
2. Compute the Singular Value Decomposition of $\widehat{\Gamma}_Y$, say $\widehat{V}\widehat{D}\widehat{W}^T = \widehat{\Gamma}_Y$.
3. Use the previous estimates to compute $\widehat{\Gamma}_Y(\ell)$ for all ℓ . Note that $Cum(Y_\ell, Y_{\ell'}, Y_m)$ is a submatrix of Γ_Y .
4. Compute $\widehat{B}(\ell, M) = \widehat{\Gamma}_Y(\ell)M^T [\widehat{\Gamma}_Y(1)M^T]^-$.

The rest of the algorithm is identical to 2S-AD. For easy reference, we shall refer to this algorithm as 2S-AD3 (third-order cumulants only).

3.2.3 Remark on the estimation of the number of factors

Stage 1 in 2S-AD or 2S-AD3 yields an estimate of Σ_U . Now, under the model's assumptions:

$$\text{rank}(\Sigma_Y - \Sigma_U) = \text{rank}(\Lambda\Lambda^T) = K.$$

Therefore, from Stage 1 it is possible to infer the number of factors. In practice, we propose to proceed as follows:

1. Estimate the rank of Γ_Y and Ω_Y as in 3.1. Let K_1 be the maximum of the two estimates.
2. Fix K_1 . Estimate Σ_U by Stage 1 of 2S-AD or 2S-AD3, depending on the test statistics.
3. Estimate the rank of $\Sigma_Y - \Sigma_U$. Let K_2 be the estimate.
4. Fix $K = \min(K_1, K_2)$. Estimate Λ by Stage 2 of 2S-AD or 2S-AD3.

3.3 Estimation of the distributions of factor and error variables

The purpose of this subsection is to construct uniformly convergent estimators of factor and error densities. We shall assume that Λ is known. One can think of Λ as a first-step estimate obtained by the algorithm 2S-AD (or 2S-AD3).

Estimation of characteristic functions of factors: Under the assumption that $A_2(\Lambda)$ is full-column-rank, the generalized inverse

$$A_2^-(\Lambda) = [A_2(\Lambda)^T A_2(\Lambda)]^{-1} A_2(\Lambda)^T$$

is well defined and one can write:

$$\kappa_X''(\Lambda^T t) = A_2^-(\Lambda) \kappa_Y^{(2)}(t).$$

Evaluating this equality at $t = [\Lambda^-]^T u$ yields, for all $u \in \mathbb{R}^K$:

$$\kappa_X''(u) = A_2^-(\Lambda) \kappa_Y^{(2)}([\Lambda^-]^T u).$$

Let a_k^T denote the k th row of $A_2^-(\Lambda)$ and b_k the k th column of $[\Lambda^-]^T$. Then:

$$\kappa_{X_k}''(\tau) = a_k^T \kappa_Y^{(2)}(\tau b_k). \tag{20}$$

Integrating over τ , we obtain:

$$\varphi_{X_k}(\tau) = \exp\left(\int_0^\tau \left[\int_0^v a_k^T \kappa_Y^{(2)}(ub_k) du\right] dv\right). \tag{21}$$

Connexion with Li and Vuong's (1998) estimator: We propose to estimate φ_{X_k} by (21).

Hence, estimation requires a double integration. In the more restrictive case where $K < L$ and the conditions of Theorem 11 are satisfied, we now show that one can consistently estimate φ_{X_k} by relying on a simple integration only. Doing so, the resulting estimator appears as a direct generalization of Li and Vuong's (1998).

To see why this is so, let first $\ell \in \{1 \dots L\}$. Let us assume that $\lambda_{\ell k} \neq 0$ for all $k = 1, \dots, K$, and $\Lambda_{-\ell}$, which is matrix Λ without its ℓ th row, is full-column-rank. Then, remark that, for all (ℓ, m) components of $\kappa_Y^{(2)}(t)$, $m \in \{1, \dots, L\}$, $\ell < m$:

$$\int_0^\tau \frac{\partial^2 \kappa_Y}{\partial t_m \partial t_\ell}(u \iota_\ell) du = \frac{\partial \kappa_Y}{\partial t_m}(\tau \iota_\ell), \quad (22)$$

where ι_ℓ denotes the ℓ th column of the identity matrix of size L . Moreover, if λ_k denotes the k th column of matrix Λ then:

$$\left[\frac{\partial^2 \kappa_Y(t)}{\partial t_m \partial t_\ell}; m = 1, \dots, L, m \neq \ell \right] = \Lambda_{-\ell} \text{diag}(\Lambda_\ell) \kappa_X''(t^T \Lambda) = \Lambda_{-\ell} [\lambda_{\ell k} \kappa_{X_k}''(\lambda_k^T t); k = 1, \dots, K].$$

Let $c_k^T = (c_{km}, m = 1, \dots, L, m \neq \ell)$ denote the k th row of the generalized inverse of $\Lambda_{-\ell}$: $\Lambda_{-\ell}^- = (\Lambda_{-\ell}^T \Lambda_{-\ell})^{-1} \Lambda_{-\ell}^T$. We have

$$\lambda_{\ell k} \kappa_{X_k}''(\lambda_k^T t) = c_k^T \left(\frac{\partial^2 \kappa_Y(t)}{\partial t_m \partial t_\ell}; m \neq \ell \right) = \sum_{\substack{m=1 \\ m \neq \ell}}^L c_{km} \frac{\partial^2 \kappa_Y}{\partial t_m \partial t_\ell}(t).$$

Evaluating at $t = \tau \iota_\ell$ yields:

$$\lambda_{\ell k} \kappa_{X_k}''(\lambda_{\ell k} \tau) = \sum_{\substack{m=1 \\ m \neq \ell}}^L c_{km} \frac{\partial^2 \kappa_Y}{\partial t_m \partial t_\ell}(\tau \iota_\ell).$$

Integrating with respect to τ yields, using (22):

$$\kappa_{X_k}'(\tau) = \sum_{\substack{m=1 \\ m \neq \ell}}^L c_{km} \frac{\partial \kappa_Y}{\partial t_m} \left(\frac{\tau}{\lambda_{\ell k}} \iota_\ell \right).$$

We thus obtain the expression:

$$\begin{aligned} \kappa_{X_k}(\tau) &= \sum_{\substack{m=1 \\ m \neq \ell}}^L c_{km} \int_0^\tau \frac{\partial \kappa_Y}{\partial t_m} \left(\frac{u}{\lambda_{\ell k}} \iota_\ell \right) du, \\ &= \int_0^\tau c_k^T \kappa_Y^{(1)} \left(\frac{u}{\lambda_{\ell k}} \iota_\ell \right) du, \end{aligned}$$

or:

$$\varphi_{X_k}(\tau) = \exp\left(\int_0^\tau c_k^T \kappa_Y^{(1)}\left(\frac{u}{\lambda_{\ell k}}\iota_\ell\right)du\right). \quad (23)$$

This is the formula used by Li and Vuong (1998) for the special case of $K = 1, L = 2$ and $\Lambda = (1, 1)^T$. Note that one can symmetrize the estimator by averaging over all ℓ such that $\lambda_{\ell k} \neq 0$:

$$\varphi_{X_k}(\tau) = \exp\left(\int_0^\tau \frac{\sum_{\ell/\lambda_{\ell k} \neq 0} \omega_\ell c_k^T \kappa_Y^{(1)}\left(\frac{u}{\lambda_{\ell k}}\iota_\ell\right)}{\sum_{\ell/\lambda_{\ell k} \neq 0} \omega_\ell} du\right).$$

Estimation of factor densities: In either of the two cases considered in Theorems 11 and 12, we propose to estimate factor characteristic functions by equations (23) and (21), respectively.

To do so, we estimate

$$\begin{aligned} \kappa_Y(t) &= \exp\left(\mathbb{E}\left[e^{it^T Y}\right]\right), \\ \frac{\partial \kappa_Y(t)}{\partial t_\ell} &= i \frac{\mathbb{E}\left[Y_\ell e^{it^T Y}\right]}{\mathbb{E}\left[e^{it^T Y}\right]} \end{aligned}$$

and

$$\frac{\partial^2 \kappa_Y(t)}{\partial t_\ell \partial t_m} = -\frac{\mathbb{E}\left[Y_\ell Y_m e^{it^T Y}\right]}{\mathbb{E}\left[e^{it^T Y}\right]} + \frac{\mathbb{E}\left[Y_\ell e^{it^T Y}\right]}{\mathbb{E}\left[e^{it^T Y}\right]} \frac{\mathbb{E}\left[Y_m e^{it^T Y}\right]}{\mathbb{E}\left[e^{it^T Y}\right]} \quad (24)$$

by their empirical analog, replacing mathematical expectations by arithmetic means.

We then estimate the factor distribution functions by inverse Fourier transform:

$$\begin{aligned} \widehat{f}_{X_k}(x) &= \frac{1}{2\pi} \int_{-T_N}^{T_N} \widehat{\varphi}_{X_k}(\tau) \exp(-i\tau x) d\tau \\ &= \frac{1}{2\pi} \int_{-T_N}^{T_N} \exp(-i\tau x + \widehat{\kappa}_{X_k}(\tau)) d\tau, \end{aligned} \quad (25)$$

where T_N tends to infinity at a rate to be specified in 3.4.

Estimation of error densities: To estimate error distributions, we propose to proceed similarly, taking advantage of the fact that, for all $\ell \in \{1 \dots L\}$:

$$\begin{aligned} \kappa_{U_\ell}''(\tau) &= \kappa_{Y_\ell}''(\tau) - \sum_{k=1}^K \lambda_{\ell k}^2 \kappa_{X_k}''(\lambda_{\ell k} \tau), \\ &= \kappa_{Y_\ell}''(\tau) - (\Lambda_\ell * \Lambda_\ell) A_2(\Lambda)^- \kappa_Y^{(2)}(\tau \iota_\ell), \end{aligned}$$

and hence:

$$\varphi_{U_\ell}(\tau) = \varphi_{Y_\ell}(\tau) \exp \left(\int_0^\tau \left[\int_0^v -(\Lambda_\ell * \Lambda_\ell) A_2(\Lambda)^{-1} \kappa_{Y'}^{(2)}(u_\ell) du \right] dv \right). \quad (26)$$

We propose to replace $\varphi_{Y_\ell}(\tau)$ and $\kappa_{Y'}^{(2)}(u_\ell)$ in (26) by their empirical analog, and to estimate:

$$\widehat{f}_{U_\ell}(x) = \frac{1}{2\pi} \int_{-T_N}^{T_N} \widehat{\varphi}_{U_\ell}(\tau) \exp(-i\tau x) d\tau, \quad (27)$$

where the ‘‘trimming’’ parameter T_N can be chosen different from its value in (21) (see 3.5 for practical estimation details).

3.4 Asymptotic properties of the estimators

Empirical moments are root- N consistent estimators of the true moments. It then follows that the empirical cumulants are also root- N consistent. Therefore, the factor loadings’ estimates, $\widehat{\Lambda}$, are also root- N consistent and asymptotically normally distributed under the usual GMM regularity conditions.

Therefore, without loss of generality, we can from now on assume that Λ is known, as one can estimate factor loadings at a rate that is higher than the rate of convergence that can be achieved for factor and error distributions.

The rest of this subsection is devoted to the study of the estimators of characteristic functions (c.f.’s) and distribution functions (d.f.’s) introduced above (in 3.3). More precisely, we shall consider the estimator of the c.f. of factors given by the empirical counterpart of either (21) or (23). Then, the estimator of factor d.f. is given by (25).

We now proceed to show that \widehat{f}_{X_k} is a uniformly convergent estimator of f_{X_k} , for $k = 1 \dots K$, provided that the characteristic functions of factors and errors do not vanish anywhere. The convergence of \widehat{f}_{U_ℓ} , $\ell = 1 \dots L$, can be proved similarly under the same assumptions.

Assume that the model (1) is identified and that Assumption 2 is satisfied. Li and Vuong (1998) and Hall and Yao (2003) in addition to assuming that characteristic functions are nowhere zero also assume support boundedness.⁶ As emphasized by Hu and Ridder (2003) these two assumptions are

⁶Horowitz and Markatou (1996) do not assume boundedness. However, as pointed out by Hu and Ridder (2003), their proofs implicitly require this hypothesis.

indeed incompatible. Hu and Ridder develop an argument that is supposed to work in the case of factor and error distributions with unbounded support and with c.f. vanishing a countable number of times. As economic outcome variables often have large supports (think of wages or income variables) we shall follow Hu and Ridder and develop a proof of consistency of the estimator that works with factor and error distributions with unbounded supports. However, we refer the reader to Hu and Ridder for insights on how to allow the characteristic functions to attain zero for values of the argument that is not a dense subset of \mathbb{R} .

In Appendix 7.6 and 7.7, we prove the following Theorem that shows that \widehat{f}_{X_k} converges uniformly to f_{X_k} :

Theorem 13 *Assume that the model of Definition 1 is identified and that Assumption 2 is satisfied. Suppose that there exists a non increasing and positive function $\underline{h}(t)$ and a non-negative function $\overline{h}(t)$, both being defined over \mathbb{R}^+ , such that \overline{h}^{K+1} is integrable, and such that, for all $W = X_1, \dots, X_K, U_1, \dots, U_L$:*

$$\overline{h}(|t|) \geq |\varphi_W(t)| \geq \underline{h}(|t|), \text{ for } |t| \text{ large enough.}$$

Then, \widehat{f}_{X_k} is a uniformly convergent estimator of the d.f. f_{X_k} of X_k :

$$\sup_{x \in \text{Supp}(X_k)} \left| \widehat{f}_{X_k}(x) - f_{X_k}(x) \right| \rightarrow 0, \quad a.s.$$

Since Fan (1991), the asymptotic properties of deconvolution estimators are known to depend on the smoothness of the distributions; that is: on the tails of their characteristic functions. Fan (1991), and subsequently Li and Vuong (1998) distinguish *smooth* and *supersmooth* distributions.

Using the notations of Theorem 13, smooth factors and errors correspond to polynomial \underline{h} and \overline{h} . Examples are uniform or gamma distributions. Supersmooth factors and errors, *e.g.* normally distributed, correspond to exponential \underline{h} and \overline{h} .

Li and Vuong (1998) prove the convergence of their estimator in the four cases where the factor and the errors are either smooth or supersmooth. They also provide expressions of the asymptotic

convergence rates in these four cases. Note that it is possible, at the cost of more complex notations, to adapt the proof of Theorem 13 to deal with different functions \underline{h} and \overline{h} for factors X_k and errors U_ℓ . Following the steps of the proof, the interested reader can also obtain explicit convergence rates. The expression of the asymptotic rates in the two case where both factors and errors are smooth or supersmooth are available from the authors upon request.

3.5 Practical estimation of factor and error distributions

We here present the method that we use to implement the estimators introduced in 3.3. The convergence properties proved in 3.4 are based on Assumption 2, which requires that the characteristic functions of factors and errors (and, as a consequence, the c.f.'s of measurements as well) are everywhere non-vanishing.

In practice, however, empirical c.f.'s and their derivatives are rarely well estimated in the tails (Gibbs oscillations). In the present case, this problem has two consequences: first, second derivatives of κ_Y are badly estimated far from zero; for such indices τ , formulas such as (21) are likely to render the estimation very noisy. Second, empirical c.f.'s can vanish even if the true c.f. does not, causing problems in the estimation of cumulant generating functions (as c.f.'s appear at the denominator of (24)).

In an important paper, Diggle and Hall (1993) propose an intuitive way to choose T_N . We follow their insight and proceed as follows. First, we regress $\ln |\psi_Y(\tau b_k)|$ on different powers of $|\tau|$ and $\ln |\tau|$, on a region where this relationship is approximately linear. As argued by Diggle and Hall, this step relies on the subjective judgement of the researcher.

Let us for instance assume that $\ln |\psi_Y(\tau b_k)| \approx \alpha \ln |\tau|$, with $\alpha < 0$. Then $|\psi_Y(\tau b_k)| \approx |\tau|^\alpha$. This corresponds to the smooth case mentioned above. Diggle and Hall (1994) then propose to choose T_N as the solution of:

$$|T_N|^\alpha = N^{-1/2}.$$

We proceed similarly for error distributions, replacing $|\psi_Y(\tau b_k)|$ by $|\psi_Y(\tau \iota_\ell)|$ in the argument.

More rigorous strategies could be devised. However, we observed that this method yielded convenient choices for T_N . With this value of T_N , we performed the integration by multiplying the c.f in the inverse Fourier transforms by a “damping factor” as in Diggle and Hall (1993). The factor we choose is given by:

$$d(\tau) = \mathbf{1}\{|\tau| < (1 - \mu)T_N\} + \mathbf{1}\{(1 - \mu)T_N \leq |\tau| \leq T_N\} \cdot \frac{1}{\mu} [1 - |\tau|/T_N],$$

where we set $\mu = .05$.

In practice, the oscillations were significantly reduced by this smoothing procedure, although we were not able to eliminate them completely.

4 Monte-Carlo simulations

In this section, we study the finite-sample properties of our estimators by numerical simulations. In the first subsection, we present the results of the estimation of factor loadings (3.2). Then, we present simulation results for the rank tests introduced Robin and Smith (2000). Lastly, in the last subsection, we turn to the finite-sample behavior of our density estimators (3.3).

4.1 Estimation of factor loadings

The DGP in this subsection and the next corresponds to the Model:

$$Y = \Lambda X + U,$$

of Definition 1, where $\Lambda(L) = I_L + J_L$, I_L is the identity matrix and J_L is the $L \times L$ matrix of ones. All variables are centered, factors have unit variances, and errors have the same variance, not necessarily equal to one.

Table 1 presents the results of 1000 simulations of the model with $L = K = 3$, for (centered and standardized) log-normal factors, and normal errors. The standard deviation of errors is 1. Results are given for various sample sizes N . Standard errors of factor loadings are given between brackets. The chosen algorithm is 2S-AD (using second-, third- and fourth- order moments of the data).

N	500	1000	5000	10000
$\widehat{\lambda}_{11}$	1.99 (.20)	1.99 (.13)	2.00 (.05)	2.00 (.04)
$\widehat{\lambda}_{21}$.97 (.21)	.99 (.14)	.99 (.06)	1.00 (.04)
$\widehat{\lambda}_{31}$.97 (.21)	.99 (.14)	.99 (.06)	1.00 (.04)
$\widehat{\lambda}_{12}$	1.01 (.22)	1.01 (.13)	1.00 (.06)	1.00 (.04)
$\widehat{\lambda}_{22}$	2.00 (.20)	2.01 (.12)	2.00 (.06)	2.00 (.04)
$\widehat{\lambda}_{32}$	1.01 (.22)	1.00 (.13)	1.00 (.06)	1.00 (.04)
$\widehat{\lambda}_{13}$.98 (.22)	.99 (.13)	1.00 (.06)	1.00 (.04)
$\widehat{\lambda}_{23}$.98 (.23)	.99 (.13)	1.00 (.06)	1.00 (.04)
$\widehat{\lambda}_{33}$	1.99 (.22)	2.00 (.13)	2.00 (.05)	2.00 (.04)
$\widehat{V}(U_1)$.97 (.39)	.96 (.26)	.99 (.12)	1.00 (.09)
$\widehat{V}(U_2)$.96 (.37)	.97 (.26)	1.00 (.12)	1.00 (.09)
$\widehat{V}(U_3)$.95 (.38)	.96 (.24)	.99 (.12)	1.00 (.10)

Table 1: L=K=3, Log-normal factors, normal errors, $V(U) = 1$

Table 1 illustrates the well-known instability of higher-order moments. For $N = 500$, there is evidence of bias on the factor loadings. However, this bias is not severe and rapidly vanishes as N increases.⁷ For this reason, our method is likely to be well-suited for large cross-sections (see also Lewbel, 1997, for a similar point).

To illustrate the performance of 2S-AD, we report in Table 2 the mean and standard error of κ_3 and κ_4 for a standardized lognormal variate, and increasing sample sizes.⁸ Table 2 shows that the third-order cumulant of the lognormal suffers from a substantial finite-sample bias. Moreover, standard errors are large. The situation is still much worse as far as the fourth-order cumulant is concerned, as none of the estimates is significantly different from zero at conventional levels, even for a sample size of 10000.

The striking contrast between Tables 1 and 2 suggests that our algorithm does a good job at extracting the relevant information from higher-order moments of the data, while being relatively immune to the imprecision of their estimation in finite samples.

⁷When estimating the model with different–non symmetric– matrices of factor loadings, the bias for $N = 500$ was somewhat stronger. However, it vanished rapidly when N increased in all the cases that we considered.

⁸Means and Variances were computed from 1000 independent drawings, for each sample size N .

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

Table 2: Skewness and excess kurtosis of a lognormal variate

N	500	1000	5000	10000
$\widehat{\lambda}_{11}$	1.00 (.19)	.99 (.17)	.99 (.14)	1.00 (.09)
$\widehat{\lambda}_{21}$.99 (.17)	1.00 (.16)	.99 (.14)	1.00 (.08)
$\widehat{\lambda}_{31}$	-.02 (.17)	-.03 (.13)	-.01 (.09)	-.01 (.07)
$\widehat{\lambda}_{12}$	1.01 (.19)	1.00 (.15)	1.01 (.10)	1.00 (.08)
$\widehat{\lambda}_{22}$	-.00 (.17)	-.01 (.14)	.01 (.11)	.00 (.08)
$\widehat{\lambda}_{32}$	1.00 (.17)	1.01 (.14)	1.00 (.08)	1.00 (.06)
$\widehat{\lambda}_{13}$.00 (.21)	.01 (.18)	.00 (.10)	.00 (.05)
$\widehat{\lambda}_{23}$	1.02 (.15)	1.02 (.11)	1.00 (.06)	1.00 (.04)
$\widehat{\lambda}_{33}$	1.00 (.14)	1.01 (.11)	1.00 (.06)	1.01 (.04)
$\widehat{V}(U_1)$.87 (.44)	.94 (.26)	.97 (.19)	.97 (.13)
$\widehat{V}(U_2)$.90 (.44)	.90 (.26)	.98 (.17)	.98 (.12)
$\widehat{V}(U_3)$.93 (.44)	.89 (.24)	.96 (.15)	.96 (.11)

Table 3: L=K=3, Log-normal factors, normal errors, $V(U) = 1$, Λ_0

In Table 3 we keep the same design, instead that Λ is now:

$$\Lambda_0 = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

Matrix Λ_0 violates one of the identification conditions in Theorem 12, as all the rows of Λ_0 contain a zero element. However, as illustrated by Table 3, the algorithm still works well in that case.

Coming back to matrix Λ , we then compare our algorithm with standard GMM applied to the cumulant equations:

$$\Sigma_Y = \Lambda \Lambda^T + \Sigma_U,$$

$$\Gamma_Y = \Lambda D_3 A_2(\Lambda)^T,$$

$$\Omega_Y = \overline{A}_2(\Lambda) D_4 A_2(\Lambda)^T.$$

In all the simulations that we performed, GMM procedures proved to be highly unstable. For instance, maximization with respect to the whole set of parameters (Λ, D_U, D_3, D_4) converged (numerically) in none of the cases that we considered.

N	500	1000	5000	10000
$\widehat{\lambda}_{11}$	2.00 (.08)	2.00 (.06)	2.00 (.03)	2.00 (.02)
$\widehat{\lambda}_{21}$.99 (.12)	1.00 (.09)	1.00 (.04)	1.00 (.03)
$\widehat{\lambda}_{31}$.99 (.13)	.99 (.09)	1.00 (.04)	1.00 (.03)
$\widehat{\lambda}_{12}$.99 (.12)	1.00 (.09)	1.00 (.04)	1.00 (.03)
$\widehat{\lambda}_{22}$	1.99 (.08)	2.00 (.06)	2.00 (.03)	2.00 (.02)
$\widehat{\lambda}_{32}$.99 (.13)	1.00 (.09)	1.00 (.04)	1.00 (.03)
$\widehat{\lambda}_{13}$.99 (.13)	1.00 (.09)	1.00 (.04)	1.00 (.03)
$\widehat{\lambda}_{23}$	1.00 (.13)	1.00 (.09)	1.00 (.04)	1.00 (.03)
$\widehat{\lambda}_{33}$.99 (.09)	2.00 (.06)	2.00 (.03)	2.00 (.02)
$\widehat{V}(U_1)$.24 (.14)	.24 (.10)	.25 (.05)	.25 (.04)
$\widehat{V}(U_2)$.24 (.13)	.24 (.10)	.24 (.05)	.25 (.04)
$\widehat{V}(U_3)$.24 (.13)	.23 (.11)	.24 (.05)	.25 (.04)
% convergence	.539	.516	.530	.505

Table 4: K=3, Log-normal factors, normal errors, $V(U) = .25$, GMM

To obtain a more stable algorithm, admittedly at the cost of lower efficiency, we treat D_3 and D_4 as nuisance parameters. Precisely, we minimize the GMM norm, evaluated at $(\Lambda, D_U, \widehat{D}_3(\Lambda), \widehat{D}_4(\Lambda))$, with respect to (Λ, D_U) alone, where $\widehat{D}_3(\Lambda)$ and $\widehat{D}_4(\Lambda)$ are obtained as the OLS estimates of:⁹

$$(A_2(\Lambda) \otimes \Lambda) \text{vec}(D_3) = \text{vec}(\Gamma_Y),$$

$$(A_2(\Lambda) \otimes \overline{A}_2(\Lambda)) \text{vec}(D_4) = \text{vec}(\Omega_Y).$$

Table 4 presents the results corresponding to this GMM method. Results are presented conditional on numerical convergence.¹⁰ Starting conditions were chosen equal to the true parameters.

We draw two conclusions from the comparison of Table 4 and Table 1. First, conditional on convergence, full GMM is more efficient than our algorithm in finite sample. This result was to be expected, as our algorithm uses only part of the moment (2^d , 3^d and 4^{rth}) conditions implied by the factor model. However, the difference in variances is not large, especially when looking at the factor loadings.

Second, this efficiency gain is obtained at the cost of high numerical instability. With a prob-

⁹In the cases we considered, GLS estimation of $\widehat{D}_3(\Lambda)$ and $\widehat{D}_4(\Lambda)$ yielded a very unstable, non-convergent, algorithm.

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

$\mathbb{V}(U)$.01	.25	1	4
$\widehat{\lambda}_{11}$	2.00 (.07)	2.11 (.08)	2.36 (.12)	2.81 (.46)
$\widehat{\lambda}_{21}$	1.00 (.11)	1.00 (.12)	.95 (.24)	.72 (.86)
$\widehat{\lambda}_{31}$	1.00 (.11)	1.03 (.14)	1.08 (.22)	1.05 (.77)
$\widehat{\lambda}_{12}$	1.00 (.11)	1.00 (.12)	.97 (.24)	.78 (.86)
$\widehat{\lambda}_{22}$	2.00 (.07)	2.11 (.07)	2.37 (.12)	2.86 (.32)
$\widehat{\lambda}_{32}$	1.00 (.12)	1.03 (.13)	1.08 (.22)	1.08 (.76)
$\widehat{\lambda}_{13}$	1.00 (.11)	.87 (.13)	.61 (.20)	.16 (.69)
$\widehat{\lambda}_{23}$	1.00 (.11)	.87 (.12)	.62 (.20)	.15 (.67)
$\widehat{\lambda}_{33}$	2.00 (.08)	2.02 (.09)	2.13 (.16)	2.52 (.43)

Table 5: Robustness to noise, JADE, log-normal factors, normal errors, $N = 1000$

ability of non-convergence higher than 40%, Full GMM appears largely unreliable, at least in our experiments. In contrast, 2S-AD gives very reasonable results.

We then study the robustness of 2S-AD to the presence of noise. To measure how “noisy” a model is, we shall use the Signal-to-Noise Ratio (SNR), defined as $\mathbb{V}(\Lambda_\ell X)/\mathbb{V}(Y_\ell)$, for given ℓ . In the case of matrix $\Lambda(L)$, this ratio is independent of ℓ , and is analogous to a R-squared in standard regression analysis.

In Tables 5 and 6, we compare the performance of 2S-AD to that of Cardoso and Souloumiac’s (1993) JADE (see the Appendix for a description of the JADE algorithm). In Tables 5 and 6, errors are still assumed normal, and factors log-normal. The standard deviation of errors takes four values: .1, .5, 1 and 2.

We see from Table 6 that the performance of our algorithm decreases as the Signal-to-Noise Ratio decreases. However, for large sample sizes the bias and efficiency properties remain acceptable, even in the case of rather large error variances. On the other hand, Table 5 shows that the performance of the JADE algorithm worsens very rapidly with the size of the noise. In particular, while our algorithm yields consistent estimates of the factor loadings, the inconsistency of standard ICA methods is severe, even when the magnitude of the error is not especially large (variance 1; that is a SNR of 80%).

In Tables 7, 8 and 9, we replicate the results of Tables 1, 5 and 6, respectively, in the case

$V(U)$.01	.25	1	4
$\widehat{\lambda}_{11}$	1.98 (.06)	2.00(.09)	1.99 (.13)	1.93 (.38)
$\widehat{\lambda}_{21}$	1.00 (.09)	.99 (.11)	.99 (.14)	.96 (.32)
$\widehat{\lambda}_{31}$	1.00 (.09)	1.00 (.11)	.99 (.14)	.97 (.37)
$\widehat{\lambda}_{12}$	1.00 (.09)	1.00 (.11)	1.01 (.13)	.98 (.31)
$\widehat{\lambda}_{22}$	1.98 (.06)	2.00 (.10)	2.01 (.12)	1.94 (.42)
$\widehat{\lambda}_{32}$	1.00 (.09)	1.00 (.11)	1.00 (.13)	.97 (.35)
$\widehat{\lambda}_{13}$	1.00 (.09)	.99 (.10)	.99 (.13)	.98 (.36)
$\widehat{\lambda}_{23}$	1.00 (.09)	.99 (.11)	.99 (.13)	.97 (.36)
$\widehat{\lambda}_{33}$	1.99 (.06)	2.00 (.08)	2.00 (.13)	1.92 (.42)
$\widehat{V}(U_1)$.04 (.09)	.22 (.16)	.96 (.26)	3.99 (.80)
$\widehat{V}(U_2)$.04 (.08)	.23 (.16)	.97 (.26)	3.98 (.75)
$\widehat{V}(U_3)$.04 (.09)	.23 (.17)	.96 (.24)	3.98 (.84)

Table 6: Robustness to noise, 2S-AD, log-normal factors, normal errors, $N = 1000$

N	500	1000	5000	10000
$\widehat{\lambda}_{11}$	1.94 (.23)	1.96 (.12)	1.97 (.06)	1.97 (.05)
$\widehat{\lambda}_{21}$.99 (.23)	1.01 (.13)	1.00 (.06)	1.00 (.05)
$\widehat{\lambda}_{31}$.97 (.23)	.99 (.13)	.99 (.06)	.99 (.05)
$\widehat{\lambda}_{12}$	1.00 (.22)	1.00 (.14)	1.00 (.06)	1.01 (.04)
$\widehat{\lambda}_{22}$	2.01 (.23)	2.02 (.14)	2.01 (.06)	2.01 (.04)
$\widehat{\lambda}_{32}$.97 (.25)	.97 (.15)	.98 (.06)	.98 (.04)
$\widehat{\lambda}_{13}$	1.00 (.23)	1.00 (.12)	1.01 (.06)	1.01 (.04)
$\widehat{\lambda}_{23}$.97 (.24)	.98 (.15)	1.00 (.06)	.99 (.04)
$\widehat{\lambda}_{33}$	2.04 (.24)	2.04 (.14)	2.04 (.06)	2.03 (.05)
$\widehat{V}(U_1)$	1.10 (.36)	1.10 (.24)	1.09 (.13)	1.10 (.10)
$\widehat{V}(U_2)$.87 (.36)	.90 (.25)	.95 (.13)	.95 (.10)
$\widehat{V}(U_3)$.79 (.36)	.83 (.28)	.88 (.14)	.89 (.11)

Table 7: L=K=3, Log-normal factors, dependent chi-square errors, $V(U) = 1$

of non-normal noise in the case where the model is misspecified. Namely, errors are uncorrelated chi-square random variables that are not independent. Specifically, we construct U as a $\mathcal{N}(0, \Lambda)^2$ variable that we center and standardize, and finally multiply by $\sqrt{\mathbb{V}(U)}$.

Comparing Table 7 with Table 1 shows that dependence in the errors creates a bias in the estimates of error variances and factor loadings. Moreover, unlike in Table 1, the bias does not vanish as the sample size increases. However, for errors of moderate size (variance 1) the bias appears to be small, at least for the dependence introduced in this design.

The situation is very different when the variance of the errors increases, as illustrated in the last

$\mathbb{V}(U)$.01	.25	1	4
$\widehat{\lambda}_{11}$	2.00 (.06)	2.11 (.07)	2.34 (.11)	2.89 (.29)
$\widehat{\lambda}_{21}$	1.00 (.10)	.99 (.12)	.91 (.21)	.56 (.46)
$\widehat{\lambda}_{31}$	1.00 (.10)	1.02 (.12)	1.02 (.21)	.67 (.53)
$\widehat{\lambda}_{12}$	1.00 (.11)	1.01 (.13)	.99 (.21)	.91 (.45)
$\widehat{\lambda}_{22}$	2.00 (.07)	2.12 (.08)	2.37 (.11)	2.96 (.24)
$\widehat{\lambda}_{32}$	1.00 (.11)	1.04 (.13)	1.06 (.20)	.81 (.55)
$\widehat{\lambda}_{13}$.99 (.09)	.88 (.10)	.69 (.19)	.69 (.49)
$\widehat{\lambda}_{23}$.99 (.10)	.87 (.11)	.69 (.17)	.65 (.51)
$\widehat{\lambda}_{33}$	2.00 (.07)	2.02 (.07)	2.18 (.12)	2.87 (.20)

Table 8: Robustness to noise, JADE, log-normal factors, dependent chi-square errors, $N = 1000$

$\mathbb{V}(U)$.01	.25	1	4
$\widehat{\lambda}_{11}$	2.00 (.07)	2.01(.08)	1.96 (.12)	1.74 (.57)
$\widehat{\lambda}_{21}$.97 (.09)	.99 (.10)	1.01 (.13)	1.08 (.59)
$\widehat{\lambda}_{31}$.95 (.10)	.98 (.11)	.99 (.13)	1.11 (.63)
$\widehat{\lambda}_{12}$	1.00 (.09)	.99 (.10)	1.00 (.14)	.98 (.50)
$\widehat{\lambda}_{22}$	1.97 (.06)	2.00 (.09)	2.02 (.14)	1.72 (.65)
$\widehat{\lambda}_{32}$	1.05 (.09)	1.00 (.10)	.97 (.15)	.85 (.61)
$\widehat{\lambda}_{13}$.97 (.09)	.99 (.11)	1.00 (.12)	.92 (.51)
$\widehat{\lambda}_{23}$	1.05 (.09)	1.01 (.10)	.98 (.15)	.93 (.59)
$\widehat{\lambda}_{33}$	1.98 (.07)	2.01 (.10)	2.04 (.14)	1.79 (.70)
$\widehat{V}(U_1)$.04 (.08)	.24 (.16)	1.10 (.24)	4.37 (1.14)
$\widehat{V}(U_2)$.04 (.08)	.22 (.17)	.90 (.25)	3.88 (.92)
$\widehat{V}(U_3)$.04 (.09)	.21 (.16)	.83 (.28)	3.66 (.93)

Table 9: Robustness to noise, 2S-AD, log-normal factors, dependent chi-square errors, $N = 1000$

column of Table 6. There, even if the bias is still much smaller than the one that JADE yields for the same DGP (see Table 5), the performance of 2S-AD is especially bad when the error variance is larger ($\mathbb{V}(U) = 4$).

To conclude this series of experiments, the size of the errors appears to be critical. Large error variances are a strong source of deterioration of the algorithm, and more so if the true errors are not independent.

Next, in Table 10, we investigate the sensitivity of our algorithm to the extent of non-normality of the measurements. We consider 2S-AD, use second and fourth-order cumulants of the data only, and we vary the kurtosis of the factors. Sample size is $N = 1000$.

ρ	-	2/5	4/7	20/23	40/43	-
$\kappa_4(\rho)$	-6/5	1/2	1	5	10	≈ 110
$\widehat{\lambda}_{11}$	2.00 (.52)	1.85 (.74)	1.93 (.67)	2.06 (.35)	2.03 (.26)	2.02 (.17)
$\widehat{\lambda}_{21}$.93 (.56)	.91 (.86)	.91 (.85)	.92 (.32)	.95 (.20)	.97 (.14)
$\widehat{\lambda}_{31}$.93 (.54)	.90 (.89)	.91 (.82)	.93 (.31)	.97 (.19)	.98 (.14)
$\widehat{\lambda}_{12}$.86 (.55)	.73 (.83)	.84 (.82)	.97 (.31)	.99 (.19)	.99 (.13)
$\widehat{\lambda}_{22}$	1.96 (.52)	1.85 (.89)	1.81 (.81)	2.09 (.37)	2.06 (.19)	2.04 (.17)
$\widehat{\lambda}_{32}$.86 (.58)	.85 (.86)	.83 (.85)	.96 (.32)	.97 (.25)	.99 (.13)
$\widehat{\lambda}_{13}$.91 (.56)	.83 (.82)	.79 (.77)	.93(.31)	.97 (.20)	.99 (.14)
$\widehat{\lambda}_{23}$.92 (.56)	.82 (.85)	.80 (.82)	.93 (.32)	.98 (.19)	.99 (.14)
$\widehat{\lambda}_{33}$	1.97 (.56)	1.65 (.97)	1.74 (.88)	2.08 (.35)	2.04 (.26)	2.03 (.17)
$\widehat{V}(U_1)$.49 (.70)	.18 (.71)	.18 (.85)	.70 (.94)	.79 (.57)	.90 (.43)
$\widehat{V}(U_2)$.46 (.68)	.15 (.56)	.15 (.61)	.67 (.81)	.78 (.60)	.86 (.44)
$\widehat{V}(U_3)$.51 (.72)	.20 (.76)	.17 (.61)	.65 (.74)	.80 (.66)	.87 (.44)

Table 10: Factors with increasing Kurtosis, $\mathbb{V}(U) = 1$, $N = 1000$

A simple way to build distributions with arbitrary positive kurtosis excess is to construct factors as mixtures of two independent normals. Let $W_1 \sim N(0, 1/2)$, and let $\rho \in]0, 1[$. Define $W_2 \sim N(0, (2 - \rho)/(2 - 2\rho))$, independent of W_1 . Then it is straightforward to see that X define as the mixture of (W_1, ρ) and $(W_2, 1 - \rho)$ has variance one, and kurtosis excess $\kappa_4(\rho) = 3\rho/(4(1 - \rho))$.

In Table 10, errors are $N(0, 1)$. In the first column of Table 10, we report the results corresponding to factors following a (standardized) uniform distribution over $[-1, 1]$. Such a distribution is platykurtic, with $\kappa_4 = -6/5$. The last column shows the results for (standardized) log-normal factors, the kurtosis excess of which is equal to $e^4 + 2e^3 + 3e^2 - 6 \approx 110$.

Table 10 shows that the impact of non-normal kurtosis (remember that, except in the log-normal case, factors are assumed symmetric) on the performance of the algorithm is radical. The closer the kurtosis excess to zero, the higher the bias and the higher the variance estimates.

These results confirm the intuition that “strict” non-normality is not sufficient for the factor loadings to be estimable in practice. For the model parameters to be well estimated, factor distributions have to be “sufficiently far” from the normal in a probabilistic (and somewhat informal) sense.

Moreover, comparing the last column in Table 10 to column 2 in Table 1 suggests that the use

N	500	500	1000	1000	5000	5000
	2S-AD	2S-AD3	2S-AD	2S-AD3	2S-AD	2S-AD3
$\widehat{\lambda}_{11}$	2.01 (.11)	1.99 (.11)	2.00 (.08)	2.00 (.07)	2.00 (.04)	2.00 (.03)
$\widehat{\lambda}_{21}$	1.00 (.17)	1.00 (.15)	.99 (.11)	1.00 (.10)	1.00 (.05)	1.00 (.04)
$\widehat{\lambda}_{31}$	1.00 (.10)	.99 (.09)	.99 (.07)	1.00 (.06)	1.00 (.03)	1.00 (.03)
$\widehat{\lambda}_{12}$.99 (.14)	1.00 (.14)	1.00 (.10)	1.00 (.10)	1.00 (.05)	1.00 (.04)
$\widehat{\lambda}_{22}$	1.99 (.18)	2.00 (.13)	2.00 (.10)	2.00 (.07)	2.00 (.04)	2.00 (.03)
$\widehat{\lambda}_{32}$.98 (.12)	1.00 (.10)	1.00 (.07)	1.00 (.06)	1.00 (.03)	1.00 (.03)
$\widehat{V}(U_1)$.96 (.28)	.98 (.24)	.98 (.18)	.99 (.17)	1.00 (.09)	1.00 (.08)
$\widehat{V}(U_2)$.95 (.26)	.98 (.25)	.97 (.19)	.99 (.17)	.99 (.09)	1.00 (.08)
$\widehat{V}(U_3)$	1.01 (.30)	1.01 (.24)	1.01 (.22)	1.00 (.18)	1.00 (.11)	1.00 (.09)

Table 11: Comparison of 2S-AD and 2S-AD3 in the case $L = 3$, $K = 2$, $\mathbb{V}(U) = 1$

of third-order moments together with fourth-order moments, can reduce the bias and increase the efficiency. This appears to be the case especially for error variances. To study the identifying power of third-order moments only, we then consider both the 2S-AD and 2S-AD3 algorithms in the case $L = 3$, $K = 2$. The first algorithm uses the information from second to fourth-order moments. The second algorithm, building on Reiersol (1950), uses only second and third-order moments of the data to identify the factor loadings (see 3.2).

Table 11 gives the results of the estimation of the model with (standardized) log-normal factors and normal errors with variance 1. Matrix Λ is given by:

$$\Lambda = \begin{pmatrix} 2 & 1 \\ 1 & 2 \\ 1 & 1 \end{pmatrix}.$$

For log-normal factors, Table 11 shows remarkably close results for both algorithms. This illustrative table suggests that an algorithm based on third-order moments only could do well in practice, provided that there is enough skewness in the data.

We conclude from these experiments that the behavior of our algorithm seems satisfactory in the case of a low number of measurements. Good performance of 2S-AD seems to depend critically on three conditions, namely: (1) that sample sizes are sufficiently large, (2) that error variances are not too large, and (3) that factor distributions are sufficiently far from normal.

In the remainder of this subsection, we investigate the finite-sample performance of our algo-

N	500	1000	5000
$\hat{\lambda}_{11}$	2.00 (.27)	2.00 (.15)	2.00 (.07)
$\hat{\lambda}_{21}$	1.01 (.24)	.99 (.15)	1.00 (.07)
$\hat{\lambda}_{31}$	1.01 (.26)	1.00 (.15)	1.00 (.06)
$\hat{\lambda}_{41}$	1.02 (.25)	.99 (.15)	.99 (.06)
$\hat{\lambda}_{51}$	1.02 (.25)	.99 (.15)	1.00 (.07)
$\hat{V}(U_1)$.91 (.38)	.96 (.26)	.99 (.12)

Table 12: L=K=5, Log-normal factors, normal errors, $V(U) = 1$

N	500	1000	5000
$\hat{\lambda}_{11}$	1.70 (.79)	1.97 (.46)	2.01 (.16)
$\hat{\lambda}_{21}$.93 (.56)	.99 (.40)	1.00 (.15)
$\hat{\lambda}_{31}$.88 (.55)	.95 (.39)	1.00 (.15)
$\hat{\lambda}_{41}$.93 (.55)	.98 (.38)	1.00 (.15)
$\hat{\lambda}_{51}$.97 (.56)	1.01 (.37)	1.00 (.15)
$\hat{\lambda}_{61}$.92 (.55)	.98 (.36)	1.00 (.15)
$\hat{\lambda}_{71}$.97 (.54)	1.02 (.37)	1.00 (.15)
$\hat{\lambda}_{81}$.82 (.53)	.99 (.35)	1.00 (.15)
$\hat{\lambda}_{91}$.93 (.54)	.99 (.35)	1.00 (.15)
$\hat{\lambda}_{10,1}$.89 (.56)	.97 (.36)	1.00 (.15)
$\hat{V}(U_1)$	1.14 (.48)	.88 (.40)	.94 (.18)

Table 13: L=K=10, Log-normal factors, normal errors, $V(U) = 1$

rithm when the numbers of measurements and factors increases. Tables 12 and 13 illustrate the cases $L = K = 5$ and $L = K = 10$, respectively. In these two tables, we report only the estimates of the factor loadings corresponding to the first factor, and the variance of the first error, the other estimates being qualitatively similar.

Focusing first on Table 12 shows similar results as in Table 1, for $L = K = 5$. Even for a sample size of 500, there is little or no bias on the factor loadings, and the bias on the error variance is moderate. This result reinforces the above evidence that, when errors are not too large the algorithm shows satisfying finite-sample properties.

Turning next to Table 13 shows somewhat different results, at least in the case of moderate sample sizes. Indeed, for $N = 500$ the bias on the first of the ten factor loadings corresponding to the first factor is substantial. Note however that the bias decreases rapidly when N increases. This quite poorer performance of the algorithm when L and K increase simultaneously was to be

expected. It is intuitively more difficult to “separate out” ten factors, than three or five.

Is this bias coming from the weighting procedures that we use ? Focusing on GMM estimation of covariance structures, Altonji and Segall (1996) argue that a positive correlation between the error in the estimation of the weighting matrix and the error in the estimation of the covariance matrix of interest is likely to arise. In return, this correlation generates a downward bias on the covariance estimates.

In their Monte-Carlo experiments, Altonji and Segal (1996) find that the bias on covariance estimates can be very large for moderate sample sizes. As higher-order cumulants are much more sensitive to extreme observations than covariances are, the Altonji-Segal bias is expected to be especially severe in the case of the estimators considered in this paper.

To assess the relevance of this point in our case, we performed a Monte-Carlo experiment using the design of Table 13, estimating the model by equal weighting; that is: we used OLS instead of GLS to estimate Σ_U , and we joint diagonalized the un-weighted cumulant matrices in the second stage to estimate Λ . The results we obtained were close to the ones reported in Table 13. In particular, for $N = 500$ the downward bias on factor loadings was still substantial.

This result suggests that the bias might, to first order, be due to the difficulty to estimate higher-order cumulants in finite samples. In this case, the correlation between estimates and the weighting GMM matrix is likely to be a second-order concern.

4.2 Estimation of the number of factors

We here perform a Monte-Carlo study of the rank test detailed in 3.1 when applied to matrices of higher-order cumulants. In Tables 14 and 15, we report the size of the test for various specifications of the DGP:

$$Y = \Lambda X + U,$$

where:

$$\Lambda = \begin{pmatrix} 2 & 1 \\ 1 & 2 \\ 1 & 1 \end{pmatrix}.$$

In Tables 14 and 15, each factor follows a standardized lognormal distribution. Moreover, errors are normal with variance $V(U)$ set to .25 and 4.

Table 14 reports the size of the test for Γ_Y ; that is, the null hypothesis is $rank(\Gamma_Y) = 2$. Nominal sizes $(1 - \alpha)$ vary between .90 and .10. Columns 1 and 4 in that Table show little evidence of size distortion. The results seem not to depend on the magnitude of the noise. Moreover, columns 7 and 10 of the same Table show the results for a larger sample size ($N=5000$). Real and nominal sizes are very close in this case.

These results, well in line with the asymptotic behavior of the test (see Robin and Smith, 2000) contrast strongly with the sizes reported in Table 15. Columns 1, 3, 7 and 10 in that Table show the size of the test for Ω_Y . There is very strong evidence of distortion, as the probability of rejecting the null hypothesis is almost always close to zero, irrespective of the true nominal size. Moreover, increasing the sample size (columns 7 and 10) does little to limitate this distortion. Only for virtually infinite sample size ($N = 50000$, column 14) are the real sizes somewhat closer to the nominal ones. However, in this case also size distortion is severe.

We interpret these results as reflecting the high sensibility of fourth-order cumulants to extreme observations, or outliers. Consistently with this interpretation, we modify the test procedure by trimming the data. Our trimming procedure consists in dropping the observations which correspond to the τ 's upper percentile of $\sum_{\ell=1}^L |Y_\ell|$, where τ is a parameter.

In Tables 14 and 15, we also report the size calculations in the presence of such trimming. The results seem consistent with our hypothesis, as is illustrated in columns 2 and 5 of Table 15. Note that if some trimming can reduce size distortion, too much trimming can have perverse effects. In column 9 of Table 15 for instance, corresponding to a 1% trimming on a sample of 5000 observations, real sizes are clearly *overestimated*. A similar conclusion is drawn from column 3 in Table 14.

We conclude from this study that dropping some extreme values from the sample might be necessary in order to apply the test of ranks of 3.1 to matrices of higher-order cumulants. Moreover,

N	1000	1000	1000	1000	1000	1000	5000	5000	5000	5000	5000	5000	50000
$V(U)$.25	.25	.25	4	4	4	.25	.25	.25	4	4	4	.25
Trimming	.000	.005	.010	.000	.005	.010	.000	.001	.005	.000	.001	.005	.000
$\alpha = .10$.85	.93	.97	.84	.91	.93	.90	.91	.99	.89	.92	.98	.92
$\alpha = .20$.71	.86	.93	.69	.84	.84	.78	.83	.97	.79	.83	.94	.81
$\alpha = .30$.55	.79	.90	.55	.74	.78	.65	.73	.94	.69	.74	.92	.70
$\alpha = .40$.43	.72	.85	.44	.65	.68	.57	.64	.91	.60	.66	.88	.61
$\alpha = .50$.33	.65	.81	.33	.56	.59	.45	.55	.88	.49	.55	.84	.52
$\alpha = .60$.23	.56	.75	.24	.47	.49	.33	.46	.83	.39	.44	.77	.41
$\alpha = .70$.15	.47	.65	.16	.39	.39	.24	.36	.78	.29	.35	.71	.29
$\alpha = .80$.08	.36	.54	.09	.27	.28	.14	.26	.68	.17	.25	.63	.19
$\alpha = .90$.03	.22	.39	.04	.06	.15	.05	.15	.53	.08	.14	.49	.10

Table 14: Size, Γ_Y

N	1000	1000	1000	1000	1000	1000	5000	5000	5000	5000	5000	5000	50000
$V(U)$.25	.25	.25	4	4	4	.25	.25	.25	4	4	4	.25
Trimming	.000	.005	.010	.000	.005	.010	.000	.001	.005	.000	.001	.005	.000
$\alpha = .10$.12	.96	.99	.10	.81	.79	.30	.93	1.00	.42	.87	.89	.78
$\alpha = .20$.04	.92	.98	.06	.68	.65	.06	.87	1.00	.28	.78	.78	.61
$\alpha = .30$.02	.87	.96	.04	.55	.52	.02	.82	1.00	.20	.67	.68	.40
$\alpha = .40$.01	.82	.93	.03	.44	.39	.01	.75	.99	.15	.58	.58	.21
$\alpha = .50$.00	.74	.90	.02	.35	.31	.01	.68	.99	.11	.47	.48	.11
$\alpha = .60$.00	.65	.84	.02	.26	.25	.00	.57	.98	.07	.37	.38	.06
$\alpha = .70$.00	.50	.73	.01	.19	.16	.00	.45	.97	.04	.27	.29	.03
$\alpha = .80$.00	.34	.59	.01	.11	.09	.00	.31	.94	.03	.17	.19	.02
$\alpha = .90$.00	.17	.38	.00	.06	.04	.00	.14	.88	.01	.06	.09	.01

Table 15: Size, Ω_Y

Tables 14 and 15 suggest that a small amount of trimming might substantially improve the size properties of the test in this context.

Tables 16 and 17 show the effect of trimming extreme values of the data on the power performance of the test. The chosen alternative is:

$$\Lambda = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix},$$

that is: $rank(\Gamma_Y) = 3$ (resp. $rank(\Omega_Y) = 3$). Tables 16 and 17 show unsurprisingly that reducing the volatility of cumulant matrices through trimming increases the probability that H_0 is rejected.

In this case as well, the improvement is especially strong in the case of fourth-order cumulants, and seems to be taking place for a small amount of trimming.

N	1000	1000	1000	1000	1000	1000	5000	5000	5000	5000	5000	5000
$V(U)$.25	.25	.25	4	4	4	.25	.25	.25	4	4	4
Trimming	.000	.005	.010	.000	.005	.010	.000	.001	.005	.000	.001	.005
$\alpha = .10$.89	1.00	1.00	.33	.97	.91	1.00	1.00	1.00	1.00	1.00	1.00
$\alpha = .20$.44	1.00	1.00	.08	.91	.82	1.00	1.00	1.00	.93	1.00	1.00
$\alpha = .30$.17	1.00	1.00	.02	.87	.72	.99	1.00	1.00	.77	1.00	1.00
$\alpha = .40$.05	1.00	1.00	.00	.82	.63	.95	1.00	1.00	.62	1.00	1.00
$\alpha = .50$.02	1.00	1.00	.00	.75	.52	.83	1.00	1.00	.50	1.00	1.00
$\alpha = .60$.01	1.00	1.00	.00	.66	.42	.61	1.00	1.00	.33	1.00	1.00
$\alpha = .70$.00	1.00	1.00	.00	.55	.32	.38	1.00	1.00	.17	1.00	1.00
$\alpha = .80$.00	.99	1.00	.00	.39	.22	.16	1.00	1.00	.06	1.00	1.00
$\alpha = .90$.00	.96	1.00	.00	.22	.12	.03	1.00	1.00	.01	1.00	1.00

Table 16: Power, Γ_Y

N	1000	1000	1000	1000	1000	1000	5000	5000	5000	5000	5000	5000
$V(U)$.25	.25	.25	4	4	4	.25	.25	.25	4	4	4
Trimming	.000	.005	.010	.000	.005	.010	.000	.001	.005	.000	.001	.005
$\alpha = .10$.01	1.00	1.00	.00	.92	.87	.39	1.00	1.00	.14	1.00	1.00
$\alpha = .20$.00	.99	1.00	.00	.81	.78	.07	1.00	1.00	.02	1.00	1.00
$\alpha = .30$.00	.97	.99	.00	.68	.70	.01	1.00	1.00	.01	1.00	1.00
$\alpha = .40$.00	.93	.97	.00	.57	.60	.00	1.00	1.00	.00	1.00	1.00
$\alpha = .50$.00	.83	.93	.00	.47	.50	.00	.00	1.00	.00	1.00	1.00
$\alpha = .60$.00	.73	.89	.00	.34	.40	.00	.99	1.00	.00	.99	1.00
$\alpha = .70$.00	.58	.79	.00	.23	.29	.00	.99	1.00	.00	.98	1.00
$\alpha = .80$.00	.40	.64	.00	.13	.18	.00	.97	1.00	.00	.92	1.00
$\alpha = .90$.00	.21	.40	.00	.06	.10	.00	.87	1.00	.00	.76	1.00

Table 17: Power, Ω_Y

4.3 Estimation of factor densities

In this section, we investigate the finite-sample properties of the estimator introduced in 3.3. We set L to 3, and set matrix Λ to

$$\Lambda = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

We also assume that both factors and errors follow the same distribution. For reasons of symmetry, we shall present the estimation results for the first factor and first error component only.

In Figures 1 to 10, the upper panel corresponds to the first factor X_1 , and the lower panel to the first error U_1 . For each simulation, we draw 100 independent realizations of Y . We plot in thick line the true factor (resp. error) distribution and in dashed line the pointwise median of the of curve estimates. Pointwise 90% upper and lower percentiles are given in dotted lines.

Figures 1 to 5 show the estimation results for normally distributed factors and errors. Sample size is $N = 1000$. Let us first focus on Figure 1, for which the error variance is set to .25. In this case, the Signal-to-Noise Ratio is equal to .96;¹¹ that is: errors account for a very small part of total variance. In this case, factors seem very well estimated. The upper Panel in Figure 1 shows little or no bias, and very precise estimates. It is to be noted that the estimator seems to be fitting the mode and the general shape of the distributions better than the tails. On the other hand, the lower Panel in Figure 1 shows that error densities are not well reproduced for such small (relative) error variances.

The situation is pretty similar in Figure 2, which presents the results for the same specification, and error variances set to 1. In this case, the SNR is .85, and factors are still very well approximated, the precision being somewhat worse in the tails. The size of the error does not seem to allow precise estimation.

When increasing the error variance to 4, and lowering the SNR to .60, the estimation of the error densities improves substantially, while the one of the factor densities worsens. As illustrated

¹¹As $(4 + 1 + 1)/(4 + 1 + 1 + .25) = .96$.

by Figure 3, for both factors and errors there seems to be evidence of a downward bias on the estimated densities. However, factor densities are still relatively well fitted in this case, despite large error variances. This result suggests that our estimator might be robust to a reasonable amount of noise, at least when factor distributions are sufficiently “well-behaved”.

In Figure 4, we replicate the specification of Figure 2 for a larger sample size of 5000. There is little improvement compared to Figure 4. The results, both for factor and errors, suggest that the bias vanishes slowly when T increases.

Lastly, in Figure 5, the sample size is 1000, and errors are normal with variances 16, corresponding to a SNR of .27. This design is the symmetric of the one used in Figure 1. Factor densities are extremely poorly estimated, while the precision of the estimation of error densities is high.

Together, Figures 1 to 5 suggest that the finite-sample performances of our estimator depend firstly on the relative variance of the distribution to be estimated, and secondly on the sample size. In the rest of this subsection, we investigate to which extent the performance depends on the shape of the distributions.

In the design of Figure 6, errors are still normally distributed and factors are mixtures of independent normals, as introduced in the previous subsection. Factors are associated to parameter $\rho = 40/43$; that is: their kurtosis excess is 10. Lastly, errors have variance 1. The upper panel in Figure 6 shows that factor densities are still globally well estimated in this case. In particular, the estimated densities also present significant “peakedness” around their mode. However, the tails of the estimated curves do not reproduce well the true ones. We observe large oscillations of the Gibbs type, a phenomenon much noticed in the deconvolution literature.¹² This result suggests that the performance of our estimator might be worse, the fatter the tails of the distribution to be fitted. Comparing Figure 6 with Figure 2 suggests that it might be more difficult to estimate a factor density when its tails are fatter than the tails of the error distributions.

In Figure 7 we modify the design of Figure 6 and assume that factors are normal and errors

¹²These oscillations may largely be a result of the way we construct our estimator. See below.

follow a mixture of independent normals, as above. Then the factor density seems very precisely estimated, and there is little evidence of bias. This is consistent with the intuition that the thinner the tails of the factors relative to the errors', the better the estimator's properties. When both factors and errors have fatter tails, estimation is not as good. This point is illustrated by Figure 8, where both factors and errors follow a normal mixture with excess kurtosis 10.

In Figures 9 and 10 we consider the case where factor distributions have "very fat" tails. Namely, in both figures, factors are distributed as lognormals, and errors are normal. In Figure 9, errors have variance .25. The upper Panel shows that the true and estimated densities are significantly different in this case. First, salient characteristics of the lognormal, such as the high mode and the large skewness, are only partially reproduced by the estimator. Second, the tails seem badly estimated. This negative conclusion is even stronger in the case of Figure 10, where error variances are larger (equal to 4).

We conclude from this exercise that the finite-sample performance of our estimator critically depends on the tails of the distributions to be estimated. When tails are fat (*e.g.* normal distribution) or "not too" fat (*e.g.* mixtures) the estimator behaves well. This result is to be noted, as convergence rates of deconvolution estimators are famous for being extremely low (*e.g.* Fan, 1991 and Horowitz and Markatou, 1996). In this case, factor densities are precisely estimated, provided that error variances are not too large.

However, the study of the lognormal case shows that our estimator can behave badly when tails are fat. In the lognormal case, fat tails are associated to a high mode. Admittedly, this conjunction makes the lognormal an especially hard distribution to be fitted.

Better finite-sample properties for fat-tailed distributions could be achieved at the cost of complicating the estimator presented in 3.3. For instance, splines or wavelets could be used to smoothen the estimator. We leave the comparison of these various improvements to future research.

5 Application: a two-factor modeling of returns to schooling

In this section, we apply the above methodology to the estimation of returns to schooling. We deliberately abstract from participation and employment biases, and consider the simplest framework with one measure of schooling, D , and one outcome, Y . In the following, D will be the number of years of schooling and Y will be the hourly wage.

It is well-known that if D is measured with error then OLS of Y on D yields a biased estimate of the relationship between wages and “true” education. A popular solution to this problem is to use a second measure of education. We start from this benchmark, and construct an alternative measure of education based on credentials. Using IV or Orthogonal Factor Analysis, we then estimate the coefficients of the “true”, and unobserved, education.

With three measurements, estimating one factor is how far one can go by using second-order information only. We then try to make a step forward, and apply the algorithms of this paper to identify at most three factors. We find evidence that there is one other factor in the data, that has a natural interpretation.

5.1 The data

We use data from the French Labor Force Survey for 1995. This is a large and representative cross-section of the French labor force. Education data is especially precise.

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. This is relevant in our case, as using higher-order moments is well-known to be sensitive to outliers. In this way, we obtain a sample of 21794 workers.

We divide monthly wages by hours worked by months to obtain wage rates. We define Y as a residual of wage rate on a set of regressors, including a quartic in age.

Our education data consists of two variables. The first one is the “age at the end of school”, which broadly corresponds to the number of years of schooling (minus 6) in France. This variable,

denoted as D , is the usual regression variable in most studies of returns to schooling.

To obtain a second measure of schooling, we use the particularity of school systems in continental Europe. In France, as in the rest of Europe, credentials (or qualifications) are thought to be at least as relevant as number of years of schooling on the labor market.

The “dipl1” variable is the FLFS divides diplomas into 16 categories, including “less than high school”, but also different categories for college graduates, special categories for vocational studies... We transform this discrete variable into a continuous one, that we call D^* , by assigning to every individual in a given “dipl1” category the median of D in the sample conditional on belonging to this category. Doing so, we are interested in mitigating the error coming from fact that many students repeat the same class several times, artificially increasing their number of years of schooling.

Interestingly, the two measures of educational levels obtained in this way are correlated, yet not perfectly, as illustrated by the correlation matrix of the data (Y, D, D^*) :

$$\widehat{\Sigma}_Y = \begin{pmatrix} .086 & .304 & .284 \\ .304 & 6.95 & 4.33 \\ .284 & 4.33 & 4.71 \end{pmatrix}.$$

We then compute the OLS coefficients in the regression of Y on D and D^* , and find .04 and .06, respectively. Thus, the two education measures yield similar OLS coefficients in wage regressions. According to both results, an additional year of schooling is on average associated with a wage increase of between 4.4% and 6.0%. Interestingly, the second measure is associated with a slightly higher return, consistently with the view that the measure based on credentials is a second indicator of schooling, for which measurement error is (slightly) lower.

5.2 One indicator

Let us start with the simplest factor structures, using the wage Y , and only the first education indicator D . The above theory of identification shows that if we allow for measurement error in both variables, then (at most) one single factor is identifiable.

We proceed first by determining if one factor is indeed identifiable from the data we consider. For this we test the rank of Γ_Y and Ω_Y , as explained in 3.1.

	(Y, D) Γ_Y	(Y, D) Ω_Y	(Y, D^*) Γ_Y	(Y, D^*) Ω_Y	(D, D^*) Γ_Y	(D, D^*) Ω_Y
Rank	0	0	0	0	0	0
Statistic	6454	148	7198	2502	2925982	1169670
Critical value .05	43.7	799	20.1	260	5066	139100
p-value	1.00	.60	1.00	1.00	1.00	1.00

Table 18: Rank tests, one indicator

Table 18, columns one and two, presents the estimation results. According to the test's p-values, the null hypothesis that Γ_Y is zero is rejected at any (reasonable) level. However, the hypothesis that Ω_Y is zero is not rejected at conventional levels.

To strengthen the conclusions that might be drawn from Table 18, we performed the same test for the two (1,1) submatrices of Γ_Y and the two (2,1) matrices $\Omega_Y(1)$ and $\Omega_Y(2)$, respectively. We found that the p-values corresponding to the two elements of Γ_Y are 1.00, and that the ones corresponding to $\Omega_Y(1)$ and $\Omega_Y(2)$ are 1.00 and .46, respectively. This last result contrasts with the p-value of the rank test for Ω_Y , as the hypothesis that $\Omega_Y(1)$ is zero is strongly rejected by the data. According to these results, both 2S-AD3 (based on submatrices of Γ_Y) and 2S-AD (based on Γ_Y , $\Omega_Y(1)$ and $\Omega_Y(2)$) might be applicable to estimate factor loadings. In contrast, the test based on Ω_Y only would have led to the conclusion that 2S-AD cannot be applied because of lack of non-normal kurtosis. Hence, multiple tests on different submatrices of cumulants can be an important check when the number of factors K is to be chosen.

In columns three and four of Table 18, we repeat this exercise for D^* as the indicator of education. The hypotheses that either Γ_Y or Ω_Y are zero are strongly rejected. We obtained the same conclusion after testing the ranks of the above submatrices, for which the p-values were all equal to 1.00. The results, reported in columns five and six of the Table, are similar.

In columns one and two of Table 19, we report the results of the estimation of a one-factor model for Y and D . In column 1, the algorithm based on second to fourth-order moments was used (2S-AD), while in column 2 we report the results of the algorithm based on second and third-order

	(Y, D) 2S-AD	(Y, D) 2S-AD3	(Y, D*) 2S-AD	(Y, D*) 2S-AD3	(D, D*) 2S-AD	(D, D*) 2S-AD3
$\hat{\lambda}_{11}$.200 (.008)	.157 (.004)	.204 (.011)	.156 (.003)	-	-
$\hat{\lambda}_{21}$	1.53 (.053)	1.94 (.055)	-	-	2.37 (.084)	2.28 (.018)
$\hat{\lambda}_{31}$	-	-	1.41 (.074)	1.82 (.032)	1.79 (.062)	1.90 (.013)
$\hat{V}(U_1)$.045 (.003)	.061 (.001)	.047 (.005)	.061 (.001)	-	-
$\hat{V}(U_2)$	4.73 (.15)	3.20 (.21)	-	-	1.17 (.40)	1.75 (.064)
$\hat{V}(U_3)$	-	-	2.88 (.23)	1.39 (.11)	1.38 (.23)	1.09 (.036)

Table 19: Factor loadings and error variances, one indicator

moments only (2S-AD3).¹³

Focusing on the relationships between Y and D , and Y and D^* , the factor seems very well estimated in all cases. Moreover, the return to the factor X_1 is always higher than the OLS coefficient, and ranges between .081 and .145.

5.3 Two indicators: factor loadings

Let us first model the data by a one-factor model; that is: $L = 3$ and $K = 1$. Then it is well-known that the factor loadings can be estimated from covariance calculations only. We report the result of OFA estimation in the first column of Table 21. The estimated factor loadings share a lot of resemblance with the ones reported in Table 19. In particular, the implicated return to X_1 is .07, again higher than the return estimated by OLS. Second, the R^2 in the wage equation, or the SNR of the wage, is .23 for the wage.¹⁴ Moreover, measurement error in the education data seems rather large, as the SNR for D is .67. Interestingly, the share of the factor variance in total variance is higher for D^* (.86), consistently with the intuition that D^* is a “better” measure of educational attainment than D .

Alternatively, one can estimate the return to X_1 by IV, using D^* as an instrument for D in the wage regression. The result is virtually identical to the one obtained by OFA.

As emphasized in the Introduction, one is the maximal number of factors identifiable from

¹³Note that in the case of one factor, recovering the variances of the errors is sufficient to recover Λ . Writing $\Sigma - \Sigma_U = PP^T$, where P is a vector of size L , it follows that $\Lambda = P$. Indeed, the one-factor case is the only one where the identification up to a rotation matrix is not ambiguous.

¹⁴As $1 - .066/.086 \approx .23$

	Γ_Y	Ω_Y
Rank	0	0
Statistic	2959457	1186756
Critical value .05	4766	139380
p-value	1.00	1.00
Rank	1	1
Statistic	296	969
Critical value .05	5.82	326
p-value	1.00	1.00
Rank	2	2
Statistic	.095	282
Critical value .05	.104	41.0
p-value	.939	1.00

Table 20: Rank tests, two indicators

covariances only. We then investigate the possibility of estimating more factors in the relationship between the wage and the two measures of education.

As a first step, we report in Table 20 the results of the test of rank for matrices Γ_Y and Ω_Y . According to the results in Table 20, the null hypothesis that Γ_Y has rank 2 is not rejected by the data, at the 5% level. On the other hand, the test rejects the hypothesis that the rank of Ω_Y is less than 3. We draw two conclusions from these calculations. First, the data that we use seems significantly non-normal. This suggests that one can apply the methods introduced in this paper. Second, tests of ranks of higher-order cumulant matrices imply that at most $K = 3$ factors can be identified from second, third and fourth-order moments.

Following the suggestions in 3.2, we first set $K = 3$ and estimated Σ_U by 2S-AD. Then, we computed the estimate of $\Sigma_Y - \Sigma_U$ and found a rank of 2. We then conclude that, because of the covariance structure of the data, at most 2 factors can be identified. Moreover, higher-order information suggests that there is enough skewness/ kurtosis in the data to estimate precisely the factor loadings of two factors.

We thus estimate two factor models for two specifications: $K = 1$ and $K = 2$. Columns 2 and 3 of Table 21 present the estimates for $K = 1$, using 2S-AD and 2S-AD3 for estimation. Lastly, column 4 features the estimation results corresponding to the combination of 2S-AD and 2S-AD3,

	K=1		K=1		K=1		K=2		K=2	
	OFA	2S-AD	2S-AD3	2S-AD+2S-AD3	2S-AD	2S-AD3	2S-AD+2S-AD3	2S-AD	2S-AD3	2S-AD+2S-AD3
$\hat{\lambda}_{11}$.141 (.0021)	.147 (.0024)	.147 (.0027)	.150 (.0024)	.149 (.0039)	.149 (.0074)	.149 (.0031)	.149 (.0039)	.149 (.0074)	.149 (.0031)
$\hat{\lambda}_{21}$	2.15 (.019)	2.07 (.026)	2.07 (.033)	2.03 (.031)	2.21 (.023)	1.62 (.122)	2.26 (.030)	2.21 (.023)	1.62 (.122)	2.26 (.030)
$\hat{\lambda}_{31}$	2.01 (.019)	1.93 (.019)	1.93 (.026)	1.89 (.026)	1.92 (.017)	1.71 (.081)	1.92 (.022)	1.92 (.017)	1.71 (.081)	1.92 (.022)
$\hat{\lambda}_{12}$	-	-	-	-	-.065 (.014)	.033 (.013)	-.053 (.011)	-.065 (.014)	.033 (.013)	-.053 (.011)
$\hat{\lambda}_{22}$	-	-	-	-	.370 (.075)	1.88 (.115)	.588 (.111)	.370 (.075)	1.88 (.115)	.588 (.111)
$\hat{\lambda}_{32}$	-	-	-	-	.030 (.099)	.84 (.165)	.023 (.077)	.030 (.099)	.84 (.165)	.023 (.077)
$\hat{V}(U_1)$.066 (.00086)	.064 (.0014)	.064 (.0016)	.064 (.0013)	.060 (.0030)	.062 (.0022)	.061 (.0022)	.060 (.0030)	.062 (.0022)	.061 (.0022)
$\hat{V}(U_2)$	2.31 (.062)	2.88 (.135)	2.68 (.162)	3.00 (.144)	1.97 (.088)	.80 (.447)	1.50 (.184)	1.97 (.088)	.80 (.447)	1.50 (.184)
$\hat{V}(U_3)$.672 (.053)	1.07 (.072)	1.15 (.113)	1.28 (.093)	1.03 (.075)	1.03 (.155)	1.02 (.089)	1.03 (.075)	1.03 (.155)	1.02 (.089)

Table 21: Factor loadings and error variances, two indicators

using the two systems of matrix equations in the two estimation stages.

The results of all three columns are remarkably similar, and show that factor loadings are little affected by the introduction of higher moments. This result is in line with the intuition that factors obtained through covariance calculations are highly robust.

Next, we turn to the estimation results for $K = 2$, reported in the three last columns of Table 21. The results somewhat differ whether 2S-AD or 2S-AD3 is used for estimation. In the case of 2S-AD, the first factor seems very similar to the one identified from second-order information only. This factor is associated with a return of around .07, and is precisely estimated. Then, a second factor seems also well estimated. Quite surprisingly, this factor enters positively the number of years of schooling D and is negatively correlated with the wage Y . Moreover, this second factor is not correlated with the “corrected” measure of education D^* .

This second factor has a natural interpretation in the case of the French schooling system. In France, the diploma or qualification that one gets is more valued on the labor market than the raw number of years of schooling. Moreover, it is common to repeat the same class twice. For the same D^* , one could expect two opposite effects of higher D : (1) a selection effect, the less able staying longer at school, or (2) a learning effect, the ones staying longer at school accumulating more human capital. The results of 2S-AD suggest that the first effect dominates the second.

Importantly, covariance calculations are not able to highlight this effect, as illustrated by the comparison of columns 1 and 5 in Table 21. By allowing for one factor only, OFA takes X_2 for measurement error.

Turning to the estimates of 2S-AD3 shows quite different results. The first factor is still well estimated, and is associated to a return of .09. The second factor is also different from the one found by 2S-AD, as it is associated with a positive (albeit very small) return, and is correlated with D^* . One could interpret this result as evidence that the measurement errors in D and D^* are correlated.

The two descriptions of the data yielded by 2S-AD and 2S-AD3 both suggest that there might

be significant heterogeneity in the education measures themselves.¹⁵ To combine these different estimates, we report in the last column of Table 21 the results obtained by the algorithm putting together the moment equations of 2S-AD and 2S-AD3. The results are somewhat closer to the one obtained by using 2S-AD, suggesting that on our data excess kurtosis is more informative than skewness.

5.4 Two indicators: filtering

In the rest of this section, we propose a way to predict the realizations of the two factors at the individual level. This allows us to correlate the unobserved factors with other (observed) labor market outcomes than the wage. To do so, we estimate:

$$\mathbb{E}(X_1|Y, D, D^*), \quad \text{and} \quad \mathbb{E}(X_2|Y, D, D^*), \quad (28)$$

and then correlate these quantities with three covariates:

1. A dummy variable indicating if the individual works or not in the public sector (1), that we call PUB.
2. A variable indicating the economic status of the son (CS), either manual worker (1), intermediate profession (2) or “cadre” (the French upper class, 3).
3. A variable indicating the economic status of the father (CF), classified similarly as CS.

Conditional expectations in (28) are functions of the whole distributions of factors and errors, not only of their first two moments. Figures 11 and 12 present the densities of factors and errors, respectively. To estimate these d.f.’s, we fixed Λ to the estimate reported in the last column of Table 21.¹⁶

The upper panel in Figure 11 shows the estimated density of the first factor X_1 . The standard normal density is plotted in thick line. Bootstrapped 10% – 90% confidence band are in dotted line.

¹⁵The literature on returns to schooling has recently put a strong emphasis on the presence of heterogeneous returns (*e.g.* Carneiro *et al.*, 2003). In contrast, our results are consistent with education measures being heterogeneous. Clearly, the two interpretations differ strongly in terms of policy implications.

¹⁶To choose T_N , we adopted the method outlined in 3.5. For all five distributions, say Z , we chose T_N such that $\exp(\alpha T_N) = N^{-1/2}$, where α is the coefficient of t^2 in the OLS regression of $\ln |\psi_Z(t)|$.

	N	Y	D	D^*	X_1	X_2
$PUB = 0$	16270	-.03	17.6	17.5	-.06	.05
$PUB = 1$	5524	.09	18.0	17.9	.17	-.14

Table 22: Public sectors: some means

The density of X_1 presents both significant skewness and kurtosis. This result is consistent with the results in the previous subsection where the data was found “non-normal enough” to permit the application of 2S-AD.

The density of X_2 is reported in the lower panel of Figure 2. Unlike the previous case, estimation is imprecise. The very large oscillation in the left tail is of particular concern. This bad estimation is in line with the above Monte-Carlo evidence, as the second factor represents less than 4% of the factor variance. On the contrary, the first factor represents 77% of total variance. However, despite the lack of precision, the estimation result suggests that X_2 is also significantly skewed, and has non-zero kurtosis excess.

Then, in the three panels of Figure 12 we plot the three error densities. There is no “obvious” departure from normality in the first and third errors. However, the estimation of the error corresponding to the D variable pictures a non-symmetric density. Except on the first case, error densities do not seem very precisely estimated.

With the factor and error densities at hand, we then compute the expectations in (28) by using that, for all integrable bivariate function g :

$$\mathbb{E}(g(X_1, X_2)|y, d, d^*) = \int g(x_1, x_2) \left[\frac{f_U(y, d, d^*, x_1, x_2)}{\int f_U(y, d, d^*, x'_1, x'_2) dx'_1 dx'_2} \right] f_{X_1}(x_1) f_{X_2}(x_2) dx_1 dx_2 \quad (29)$$

where:

$$f_U(y, d, d^*, x_1, x_2) \equiv f_{U_1}(y - \lambda_{11}x_1 - \lambda_{12}x_2) f_{U_2}(d - \lambda_{21}x_1 - \lambda_{22}x_2) f_{U_3}(d^* - \lambda_{31}x_1 - \lambda_{32}x_2).$$

To estimate (28), we replace the factor and error densities by their nonparametric estimates in (29), and set $g(x, y) = x$ and $g(x, y) = y$ for the two conditional expectations, respectively.

Having computed the expectations in (28), we correlate them with the variables CS, CF and PUB. Tables 22, 23 and 24 give the results. Overall, working in the public sector seems associated

	N	Y	D	D^*	X_1	X_2
$CS = 1$	14253	-.11	16.9	16.9	-.36	.13
$CS = 2$	5271	.14	18.6	18.5	.43	-.15
$CS = 3$	2070	.39	20.5	20.5	1.34	-.50

Table 23: Economic activity of the son: some means

	N	Y	D	D^*	X_1	X_2
$CF = 1$	13024	-.02	17.2	17.3	-.16	-.02
$CF = 2$	4090	.05	18.5	18.3	.29	.04
$CF = 3$.05	1551	.18	20.3	19.9	1.02	.04

Table 24: Economic activity of the father: some means

with higher X_1 and lower X_2 . Moreover, public employees have on average higher wages and higher education in the two dimensions (D and D^*). The same remark applies to “better” economic activities relative to “worse” ones, *i.e.* French “cadres” relative to French manual workers. These results are consistent with X_1 positively affecting all labor market outcomes, not only the wage. Similarly, higher X_2 seems to be associated with lower overall returns on the labor market.

To identify further the factors X_1 and X_2 , we would need more data. For instance, precise data on whether an individual has repeated one or several classes could be very useful. We end this section by looking at the correlation between the economic activity of the father and the factors.¹⁷ The results are not surprising as far as X_1 s concerned. As expected, children of “cadres” have on average higher X_1 than children of manual workers for instance. They also have higher wages and higher D and D^* education. The result concerning X_2 is more challenging, as X_2 appears not to be correlated to CF.

6 conclusion

7 Proofs

7.1 Proof of Theorem 6

The proof of proposition (i) is a straightforward consequence of Theorem 10.3.1 in Kagan, Linnik and Rao (1973), which states the following proposition. Let A and B be two non-stochastic matrices

¹⁷The father’s wage would have been another possible choice. However, it is not available in the French Labor Force Survey.

of dimensions (p, m) and (p, n) , respectively, and let $S = (s_1, \dots, s_m)^T$ and $R = (r_1, \dots, r_n)^T$ 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 i th column of A is the multiple of a column of B .

Assume that $\Lambda X + U$ and $\tilde{\Lambda}\tilde{X} + \tilde{U}$ have the same distribution. The components of vectors (X^T, U^T) and $(\tilde{X}^T, \tilde{U}^T)$, 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 $(L, K + L)$ matrix $(\tilde{\Lambda}, I_L)$, where I_L is the (L, L) -identity matrix. Since every column of matrices Λ and $\tilde{\Lambda}$ has at least two non-zero coefficients, it must be that Λ_k is the multiple of a column of $\tilde{\Lambda}$. This shows proposition (i).

To show the second proposition, consider equation (6) for $p = 2$. One must have:

$$A_2(\Lambda)\kappa_X^{(2)}(t^T\Lambda) = A_2(\tilde{\Lambda})\kappa_{\tilde{X}}^{(2)}(t^T\tilde{\Lambda}). \quad (30)$$

By proposition (i), every column of Λ is a scalar multiple of a column of $\tilde{\Lambda}$. Since $\text{rank}(A_2(\Lambda)) = K$, it follows that no two columns of Λ are proportional. Therefore, there exist a permutation matrix P and a diagonal matrix D with non zero entries in the diagonal such that $\tilde{\Lambda} = \Lambda DP$. Now, since $\ker(A_2(\Lambda)) = 0$, (30) implies:

$$\kappa_X^{(2)}(t^T\Lambda) = \kappa_{DP\tilde{X}}^{(2)}(t^T\Lambda). \quad (31)$$

Taking this equation at $t = 0$ and using the normalization assumption (i) in Definition 1 yields:

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

Moreover, integrating the differential equation (31) shows that κ_X and $\kappa_{DP\tilde{X}}$ differ by an affine function. By definition, $\kappa_X(0) = \kappa_{DP\tilde{X}}(0) = 0$. By assumption, since the factor distributions have zero mean: $\kappa_X^{(1)}(0) = \kappa_{DP\tilde{X}}^{(1)}(0) = 0$. This shows that the d.f. of X and $DP\tilde{X}$ are equal. Lastly, the d.f. of U and \tilde{U} are equal by deconvolution, since the characteristic functions of the factors are non-vanishing everywhere.

7.2 Proof of Theorem 9

Let $x = (x_1, \dots, x_K)^T \in \ker(A_2(\Lambda)) \setminus \{0\}$. For all $k = 1, \dots, K$, and all $\ell = 1, \dots, L$, define

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

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

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

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

As $A_2(\Lambda)x = 0$, $\sum_{k=1}^K x_k \lambda_{\ell k} \lambda_{mk} = 0$ for all $\ell \neq m$ in $\{1, \dots, L\}$. Hence,

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

and, therefore,

$$\begin{aligned}\kappa(t) &= \sum_{k=1}^K \kappa_{X_k}(\lambda_k^T t) + \sum_{\ell=1}^L \kappa_{U_\ell}(t_\ell) \\ &= \kappa_Y(t).\end{aligned}$$

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

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

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

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

7.3 Proof of Lemma 10

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

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

As (v^1, \dots, 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 $v^{K+1} \neq 0$, there exists k such that $c^k \neq 0$. For this k : $x^k = x^{K+1}$.

7.4 Proof of Theorem 11

Let us first compute the Singular Value Decomposition of matrix Γ_Y :

$$\Gamma_Y = VDW^T,$$

where V and W are unitary, of order L and $L(L-1)/2$, respectively, and $D = \begin{pmatrix} D_1 & 0 \\ 0 & 0 \end{pmatrix}$ is of order $L(L-1)/2$, with D_1 diagonal, of order K , with diagonal elements in decreasing order.

Let V_L be the L th column vector of V . Then, clearly:

$$V_L^T \Lambda = 0.$$

At this stage, we need the following Lemma:

Lemma 14 *Let A and B be $n \times m$ and $m \times p$ matrices, respectively, where n, m, p are non-zero integers. Let us assume that every submatrix of $L-1$ rows of A is full-column rank. Let us write $AB = VDW^T$ the Singular Value Decomposition of AB , and let us denote as V_n the last column vector of V . Then none of the elements of V_n is zero.*

Proof. Let us assume that $v_{jn} = 0$, where v_{jn} denotes the j th element of V_n . As we shall not use any restrictions on matrices D and W , we can assume without loss of generality that $j = 1$. Let us write $V = \begin{pmatrix} a^T & 0 \\ C & b \end{pmatrix}$, where a and b are vectors of order $n-1$, and C is a square matrix of order $n-1$.

Then, V being unitary implies $VV^T = I_n$, and thus $\begin{pmatrix} C & b \end{pmatrix} \begin{pmatrix} a & 0 \end{pmatrix}^T = Ca = 0$. As $a \neq 0$, this implies that matrix $\begin{pmatrix} C & b \end{pmatrix}$ is not full-column rank. Consequently, the matrix formed of the last $L-1$ columns of A is not full-column rank, contradicting the assumption. ■

Hence, for all ℓ , denoting $V_{-\ell,L} = (v_{1,L}, \dots, v_{\ell-1,L}, v_{\ell+1,L}, \dots, v_{L,L})^T$:

$$\Lambda_\ell = -\frac{1}{v_{\ell,L}} V_{-\ell,L}^T \Lambda_{-\ell}.$$

Hence:

$$\Lambda = \Psi_\ell \Lambda_{-\ell},$$

where:

$$\Psi_\ell = \begin{pmatrix} 1 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & \dots & 0 \\ -\frac{v_{1,L}}{v_{\ell,L}} & \dots & -\frac{v_{\ell-1,L}}{v_{\ell,L}} & -\frac{v_{\ell+1,L}}{v_{\ell,L}} & \dots & -\frac{v_{L,L}}{v_{\ell,L}} \\ 0 & \dots & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & \dots & 1 \end{pmatrix},$$

is computable from the data.

Let us now define:

$$\Gamma_Y(-\ell) \equiv \Lambda D_3 \text{diag}(\Lambda_\ell) (\Lambda_{-\ell})^T.$$

As $\Gamma_Y(-\ell)$ is composed of $L - 1$ columns of Γ_Y , it is thus identified from the data. Thus, matrices:

$$\begin{aligned} \Gamma_Y(\ell) &\equiv \Gamma_Y(-\ell) \Psi_\ell^T \\ &= \Lambda D_3 \text{diag}(\Lambda_\ell) \Lambda^T, \end{aligned}$$

are also identified from the data.

By assumption, Λ is full-column rank. Let M be a projection matrix selecting K invertible rows of Λ . As both Λ_1 and $\kappa_X^{(3)}(0)$ have only non-zero elements by assumptions, it follows that $M\Gamma_Y(1)M^T$ is invertible. Thus we can express:

$$M\Gamma_Y(\ell)M^T [M\Gamma_Y(1)M^T]^{-1} = M\Lambda \text{diag}(\Lambda_\ell) [\text{diag}(\Lambda_1)]^{-1} [M\Lambda]^{-1}.$$

It then follows from lemma 10 that $x_\ell \equiv \left(\frac{\lambda_{\ell 1}}{\lambda_{11}}, \dots, \frac{\lambda_{\ell K}}{\lambda_{1K}}\right)^T$ is identified (up to permutation) for all ℓ . Hence $\Lambda \text{diag}(\Lambda_1)^{-1}$ is identified up to permutation.

To finish the proof, let $\text{vecs}(\Sigma_Y)$ be the vector of $L(L - 1)/2$ covariances of Y , sorted as the rows of $A_2(\Lambda)$. Under the Theorem's assumptions, $A_2(\Lambda)$ is full-column rank. Moreover, the factor structure implies:

$$A_2(\Lambda)(1, \dots, 1)^T = \text{vecs}(\Sigma_Y).$$

Or, equivalently:

$$A_2(\Lambda \text{diag}(\Lambda_1)^{-1})(\Lambda_1 * \Lambda_1)^T = \text{vecs}(\Sigma_Y),$$

where $*$ is the Hadamard product. As $A_2(\Lambda)$ has rank K , and Λ_1 has no zero element, $A_2(\Lambda \text{diag}(\Lambda_1)^{-1})$ has rank K . It follows that Λ_1 , and hence Λ , are identified up to permutation and sign.

7.5 Proof of Theorem 12

When fourth-order moments are allowed, obtaining a joint eigenvalue system as in Lemma 10 is easier. To see why, let us define:

$$\Omega_Y(\ell) \equiv \Lambda D_4 \text{diag}(\Lambda_\ell) A_2(\Lambda)^T.$$

As $\Omega_Y(\ell)$ is formed of L rows of Ω_Y , it is composed of fourth-order cumulants of the data.

Let M_1 be a projection matrix selecting K linearly independent rows of Λ , and M_2 be a projection matrix selecting K linearly independent rows of $A_2(\Lambda)$. Matrices M_1 and M_2 exist by assumption.

It follows directly from the Theorem's assumptions that $M_1 \Omega_Y(1) M_2^T$ is invertible. Thus we can express:

$$M_1 \Omega_Y(\ell) M_2^T [M_1 \Omega_Y(1) M_2^T]^{-1} \equiv M_1 \Lambda \text{diag}(\Lambda_\ell) \text{diag}(\Lambda_1)^{-1} [M_1 \Lambda]^{-1}.$$

Lemma 10 thus implies that $\Lambda \text{diag}(\Lambda_1)^{-1}$ is identified up to permutation. The end of the proof is as in the proof of Theorem 11.

7.6 Lemma 15

The basic result that we use to show the consistency of our estimators of factor and error distributions is stated in the following lemma that is proved using the theory developed in Chapter II of Pollard (1984).

Lemma 15 *Let $Z = (X, Y)$ be a couple of real random variables. Let F denote the c.d.f. of Z (\mathbb{E} denotes the corresponding expectation operator) and let F_N (resp. \mathbb{E}_N) denote the empirical c.d.f. for N i.i.d. draws $\mathbf{Z}_N \equiv (Z_1, \dots, Z_N)$ from F . Let $t \in \mathbb{R}$. Assume that $\mathbb{E}X^2 \leq M_1 < \infty$ and $\mathbb{E}|XY| \leq M_2 < \infty$. Define $f_t(x, y) = x \exp(ity)$. Then, for any sequence (ε_N) converging to zero*

and any diverging sequence (T_N) ,

$$\sup_{|t| \leq T_N} |\mathbb{E}_N f_t - \mathbb{E} f_t| = O(\varepsilon_N),$$

if $\varepsilon_N^2 \geq \frac{8M_1}{N}$ and $\sum_N \exp \left[\ln \left(\frac{T_N}{\varepsilon_N} \right) - \frac{N\varepsilon_N^2}{128K_N^2} \right] < \infty$, where K_N is taken such that $\mathbb{E} [|X| \{ |X| > K_N \}] <$

ε_N .

The following remarks are in order:

1. If the set of values for t is not \mathbb{R} but a bounded interval and if the support of X is bounded, then Pollard's (1984) Theorem 37 applies (T_N and K_N are fixed constants) and the uniform convergence result holds if $\frac{N\varepsilon_N^2}{\ln N} \rightarrow \infty$ or, equivalently, $\frac{1}{\varepsilon_N} = o\left(\frac{N}{\ln N}\right)^{\frac{1}{2}}$. Then, $\ln\left(\frac{1}{\varepsilon_N}\right) - \frac{1}{2} \ln N \rightarrow -\infty$. As $N\varepsilon_N^2$ increases faster than $\ln N$ and $\ln\left(\frac{1}{\varepsilon_N}\right)$ more slowly, $\exp\left[\ln\left(\frac{T_N}{\varepsilon_N}\right) - \frac{N\varepsilon_N^2}{128K_N^2}\right]$ decreases faster than any power of N and the series $\sum_N \exp\left[\ln\left(\frac{T_N}{\varepsilon_N}\right) - \frac{N\varepsilon_N^2}{128K_N^2}\right]$ converges.
2. Hu and Ridder's Lemma 1 considers the case of $X = 1$, in which case one can take $K_N = 1$. The conditions of Lemma 15 are met if $N\varepsilon_N^2$ increases faster than $\ln N$ and $\ln\left(\frac{T_N}{\varepsilon_N}\right)$ not faster, i.e. $\frac{1}{\varepsilon_N} = o\left(\frac{\ln N}{N}\right)^{\frac{1}{2}}$ and $\ln\left(\frac{T_N}{\varepsilon_N}\right) = O(\ln N)$.
3. In the general case that we consider, it suffices that $\frac{N\varepsilon_N^2}{128K_N^2} - \ln\left(\frac{T_N}{\varepsilon_N}\right) \geq \rho \ln N$, for some $\rho > 1$, for the uniform consistency result in Lemma 15 to hold. In this case, the series $\sum_N \exp\left[\ln\left(\frac{T_N}{\varepsilon_N}\right) - \frac{N\varepsilon_N^2}{128K_N^2}\right]$ is dominated by the Riemann zeta function $\zeta(\rho)$. To obtain a precise definition of the rate of convergence, one needs to know the behavior of the tails of the distribution of X . For example, if X is Gaussian:

$$\int_K^\infty x e^{-\frac{x^2}{2}} dx = e^{-\frac{K^2}{2}}.$$

Consequently, K_N^2 must tend to infinity faster than $\ln(1/\varepsilon_N)$ for $\mathbb{E} [|X| \{ |X| > K_N \}] < \varepsilon_N$.

If however, X is Pareto:

$$\int_K^\infty x \frac{ab^a}{x^{a+1}} dx = \frac{ab^a}{a-1} \frac{1}{K^{a-1}}, \quad a > 1.$$

In which case, K_N^{a-1} must tend to infinity faster than $1/\varepsilon_N$. The fatter the tail of the distribution of X and the slower the rate of convergence. Suppose, for example, that $K_N = 1/\varepsilon_N$ works. One can choose $\varepsilon_N^{\frac{2a}{a-1}} = \rho_1 \frac{\ln N}{N}$ and $\ln T_N = \rho_2 \ln N$, with $\frac{\rho_1}{128} - \frac{a-1}{2a} - \rho_2 > 1$, to meet the conditions of Lemma 15. Clearly, if the tails of the distribution of X decay at polynomial rate, the rate of convergence of the empirical estimator $\mathbb{E}_N f_t$ can be very slow.

Proof. First, remark that

$$\mathbb{E}_N f_t - \mathbb{E} f_t = \mathbb{E}_N \operatorname{Re}(f_t) - \mathbb{E} \operatorname{Re}(f_t) + i [\mathbb{E}_N \operatorname{Im}(f_t) - \mathbb{E} \operatorname{Im}(f_t)]$$

and

$$\sup_{|t| \leq T_N} |\mathbb{E}_N f_t - \mathbb{E} f_t| \leq \sup_{|t| \leq T_N} |\mathbb{E}_N \operatorname{Re}(f_t) - \mathbb{E} \operatorname{Re}(f_t)| + \sup_{|t| \leq T_N} |\mathbb{E}_N \operatorname{Im}(f_t) - \mathbb{E} \operatorname{Im}(f_t)|.$$

It will thus suffice to show that the claimed result is true for the family of functions $\operatorname{Re}(f_t)(x, y) = x \cos(ty)$ for it to be true for functions $\operatorname{Im}(f_t)$ and f_t . So, without loss of generality, we prove the result for real functions $f_t(x, y) = x \cos(ty)$, using the same notation for f_t and its real part.

Let $K_N > 0$. Let us define:

$$g_t = \frac{1}{K_N} f_t \{ |X| \leq K_N \}.$$

Then $\operatorname{Im}(g_t) \subset [-1, 1]$. We thus can apply Theorem 2.3. in Mendelson (2002, p.10):

$$\Pr \left\{ \sup_{|t| \leq T_N} |\mathbb{E}_N g_t - \mathbb{E} g_t| \geq \frac{\varepsilon_N}{K_N} \right\} \leq 8 \mathcal{N}_1(\{g_t\}, \frac{\varepsilon_N}{K_N}) \exp \left[-\frac{N \varepsilon_N^2}{128 K_N^2} \right],$$

for all N such that $N \geq \frac{8K_N^2}{\varepsilon_N^2}$.

Now, the (expected) covering number of $\{g_t\}$ satisfies:

$$\mathcal{N}_1(\{g_t\}, \varepsilon) \leq \frac{2T_N M}{K_N \varepsilon},$$

as: $\mathbb{E}|g_{t_1}(x, y) - g_{t_2}(x, y)| \leq \frac{M}{K_N} |t_1 - t_2|$ for all (t_1, t_2) .

Hence:

$$\Pr \left\{ \sup_{|t| \leq T_N} |\mathbb{E}_N g_t - \mathbb{E} g_t| \geq \frac{\varepsilon_N}{K_N} \right\} \leq 16 \frac{T_N M}{\varepsilon_N} \exp \left[-\frac{N \varepsilon_N^2}{128 K_N^2} \right].$$

Assume also that ε_N tends to zero at infinity. The Borel-Cantelli Lemma then implies that

$$\sup_{|t| \leq T_N} |\mathbb{E}_N [f_t \{ |X| \leq K_N \}] - \mathbb{E} [f_t \{ |X| \leq K_N \}]| = o(\varepsilon_N),$$

as soon as

$$\sum_N \frac{T_N}{\varepsilon_N} \exp\left(-\frac{N\varepsilon_N^2}{128K_N^2}\right) < \infty.$$

We now explain how to choose K_N as in Theorem 24 of Pollard (1984). The integrability of X allows us to choose K_N , for any $\varepsilon_N > 0$, such that $\mathbb{E} [|X| \{ |X| > K_N \}] < \varepsilon_N$. Then, writing $\mathbb{E}_N f$ for the sample mean $\frac{1}{N} \sum_{n=1}^N f_t(X_n, Y_n)$,

$$\begin{aligned} \sup_{|t| \leq T_N} |\mathbb{E}_N f_t - \mathbb{E} f_t| &\leq \sup_{|t| \leq T_N} |\mathbb{E}_N [f_t \{ |X| \leq K_N \}] - \mathbb{E} [f_t \{ |X| \leq K_N \}]| \\ &\quad + \sup_{|t| \leq T_N} \mathbb{E}_N [|f_t| \{ |X| > K_N \}] + \sup_{|t| \leq T_N} \mathbb{E} [|f_t| \{ |X| > K_N \}] \\ &\leq \sup_{|t| \leq T_N} |\mathbb{E}_N [f_t \{ |X| \leq K_N \}] - \mathbb{E} [f_t \{ |X| \leq K_N \}]| + 2\varepsilon_N, \end{aligned}$$

for N large enough, and T_N, ε_N, K_N satisfying the above conditions. Hence:

$$\sup_{|t| \leq T_N} |\mathbb{E}_N f_t - \mathbb{E} f_t| = O(\varepsilon_N).$$

This achieves to prove Lemma 15. ■

7.7 Proof of Theorem 13

We shall need the following technical lemma:

Lemma 16 *Let J be a positive integer. For all $j \in \{1 \dots J\}$ let g_j be a real-valued function, and $f_{j,N}$, $N \in \mathbb{N}$, be a sequence of real-valued functions defined over \mathbb{R} . Assume that g_j is everywhere non-vanishing and that there exists a function h_j defined on \mathbb{R}^+ , strictly decreasing, and converging to zero at infinity such that $|g_j(t)| \geq h_j(|t|)$ for large enough $|t|$. Suppose also that there exists a sequence T_N , $N \in \mathbb{N}$, diverging to infinity, such that:*

$$\sup_{t \in [-T_N, T_N]} |f_{j,N}(t)| \xrightarrow{N \rightarrow \infty} 0, \text{ for all } j \in \{1 \dots J\}.$$

Then there exists a second sequence \tilde{T}_N , $N \in \mathbb{N}$, diverging to infinity, such that:

$$\sup_{t \in [-\tilde{T}_N, \tilde{T}_N]} \left| \frac{f_{j,N}(t)}{g_j(t)} \right| \xrightarrow{N \rightarrow \infty} 0, \text{ for all } j \in \{1 \dots J\}.$$

Proof. Let us fix $j \in \{1 \dots J\}$. Let us define

$$\Delta_{j,N} = \sup_{t \in [-T_N, T_N]} |f_{j,N}(t)|.$$

As h_j is strictly decreasing on \mathbb{R}^+ , it admits an inverse that we denote as h_j^{-1} . Let us then define:

$$\tilde{T}_{j,N} = h_j^{-1} \left([\Delta_{j,N}]^{1/2} \right).$$

As $\Delta_{j,N}$ tends to zero, and as h_j has limit zero at infinity it follows that $\tilde{T}_{j,N}$ diverges to infinity.

Let us lastly define:

$$\tilde{T}_N = \min \left(T_N, \tilde{T}_{1,N}, \dots, \tilde{T}_{J,N} \right),$$

where T_N is given in the Lemma's assumptions. Then \tilde{T}_N diverges to infinity. Let $j \in J$ and $t \in [-T_N, T_N]$. As $\tilde{T}_N \leq \tilde{T}_{j,N}$ and h_j is decreasing on \mathbb{R}^+ , $h_j(|t|) \geq h_j(\tilde{T}_{j,N}) = [\Delta_{j,N}]^{1/2}$. Second, as $\tilde{T}_N \leq T_N$: $|f_{j,N}(t)| \leq \Delta_{j,N}$.

Let T_0 be such that $|g_j(t)| \geq h_j(t)$ for all $|t| \geq T_0$, and let $M_j = \inf_{t \in [-T_0, T_0]} |g_j(t)|$.

It thus follows that:

$$\begin{aligned} \sup_{t \in [-\tilde{T}_N, \tilde{T}_N]} \left| \frac{f_{j,N}(t)}{g_j(t)} \right| &\leq \sup \left(\sup_{t \in [-T_0, T_0]} \left| \frac{f_{j,N}(t)}{g_j(t)} \right|, \sup_{t \in [-\tilde{T}_N, -T_0] \cup [T_0, \tilde{T}_N]} \left| \frac{f_{j,N}(t)}{g_j(t)} \right| \right), \\ &\leq \sup \left(\frac{\sup_{t \in [-T_0, T_0]} |f_{j,N}(t)|}{M_j}, \sup_{t \in [-\tilde{T}_N, -T_0] \cup [T_0, \tilde{T}_N]} \frac{|f_{j,N}(t)|}{h_j(t)} \right), \\ &\leq \sup \left(\frac{\Delta_{j,N}}{M_j}, \frac{\Delta_{j,N}}{[\Delta_{j,N}]^{1/2}} \right), \\ &= o(1). \end{aligned}$$

■

We now can show the following lemma, which states conditions under which $\hat{\varphi}_{X_k}$ is a uniformly convergent estimator of φ_{X_k} .

Lemma 17 *Suppose that there exists a non increasing and positive function $g(t)$, defined for $t \in \mathbb{R}^+$, such that, for all $\ell = 1, \dots, L$, $|\varphi_Y(t\ell)| \geq g(|t|)$, for $|t|$ large enough. Then there exists a sequence $T_N \rightarrow \infty$ such that*

$$\sup_{t \in [-T_N, T_N]} |\widehat{\varphi}_{X_k}(t) - \varphi_{X_k}(t)| = o(1).$$

Proof. (i) We start by noting that one can always redefine g to be even, strictly decreasing on \mathbb{R}^+ and tending to zero at infinity; for instance as: $t \mapsto g(|t|)\exp(-|t|)$.

(ii) Fix $\ell = 1, \dots, L$. Let $\varphi(t) = \mathbb{E}[e^{itY_\ell}]$, $\psi_m(t) = \mathbb{E}[Y_m e^{itY_\ell}]$ and $\xi_{mp}(t) = \mathbb{E}[Y_m Y_p e^{itY_\ell}]$, for any $(m, p) \in \{1, \dots, L\}^2$. Then, Lemma 15 shows that there exists $T_N \rightarrow \infty$ such that

$$\begin{aligned} \sup_{t \in [-T_N, T_N]} |\widehat{\varphi}(t) - \varphi(t)| &= o(1), \\ \sup_{t \in [-T_N, T_N]} \left| \widehat{\psi}_m(t) - \psi_m(t) \right| &= o(1), \\ \sup_{t \in [-T_N, T_N]} \left| \widehat{\xi}_{mp}(t) - \xi_{mp}(t) \right| &= o(1), \end{aligned}$$

hold simultaneously for all $(m, p) \in \{1, \dots, L\}^2$ (take the smallest T_N).

(iii) It thus follows from the lemma's assumptions that the assumptions of Lemma 16 are satisfied for $J = 1$ and one can redefine T_N diverging to infinity such that¹⁸

$$\sup_{t \in [-T_N, T_N]} \left| \frac{\widehat{\varphi}(t) - \varphi(t)}{\varphi(t)} \right| = o(1).$$

(iv) One has

$$\frac{\partial \kappa_Y}{\partial t_m}(t\ell) = i \frac{\psi_m(t)}{\varphi(t)},$$

and

$$\begin{aligned} \frac{\widehat{\psi}_m(t)}{\widehat{\varphi}(t)} - \frac{\psi_m(t)}{\varphi(t)} &= \frac{\widehat{\psi}_m(t)}{\widehat{\varphi}(t)} - \frac{\widehat{\psi}_m(t)}{\varphi(t)} + \frac{\widehat{\psi}_m(t)}{\varphi(t)} - \frac{\psi_m(t)}{\varphi(t)} \\ &= -\frac{\widehat{\psi}_m(t)}{\varphi(t)} \frac{\widehat{\varphi}(t) - \varphi(t)}{\widehat{\varphi}(t) - \varphi(t)} + \frac{1}{\varphi(t)} \left[\widehat{\psi}_m(t) - \psi_m(t) \right]. \end{aligned}$$

¹⁸Note that for a stochastic sequence of functions $f_{j,N}$, the equalities in Lemma 16 have to be understood in the *almost everywhere* sense. This is also the case of all (in)equalities in this section.

Now:

$$\begin{aligned} \sup_{t \in [-T_N, T_N]} \left| \widehat{\psi}_m(t) \right| &\leq \sup_{t \in [-T_N, T_N]} \left| \widehat{\psi}_m(t) - \psi_m(t) \right| + \sup_{t \in [-T_N, T_N]} |\psi_m(t)| \\ &\leq \sup_{t \in [-T_N, T_N]} \left| \widehat{\psi}_m(t) - \psi_m(t) \right| + \mathbb{E} |Y_m| = O(1). \end{aligned}$$

When N and t are large enough $\left| \frac{\widehat{\varphi}(t) - \varphi(t)}{\varphi(t)} \right| = o(1)$, thus $\left[\frac{\widehat{\varphi}(t) - \varphi(t)}{\varphi(t)} + 1 \right]^{-1} = O(1)$. It follows that:

$$\sup_{t \in [-T_N, T_N]} \left| \frac{\widehat{\psi}_m(t)}{\widehat{\varphi}(t)} - \frac{\psi_m(t)}{\varphi(t)} \right| = O(1) \sup_{t \in [-T_N, T_N]} \left| \frac{\widehat{\varphi}(t) - \varphi(t)}{\varphi(t)^2} \right| + \sup_{t \in [-T_N, T_N]} \left| \frac{\widehat{\psi}_m(t) - \psi_m(t)}{\varphi(t)} \right|.$$

As $|\varphi(t)|^2 \geq g(t)^2$ for $|t|$ large enough, and as g^2 satisfies the assumptions of Lemma 16, it follows that Lemma 16 applies (with $J = 2$). Thus there exists a diverging sequence, that we call T_N for simplicity, such that

$$\Delta_N^{(1)} \equiv \sup_{t \in [-T_N, T_N]} \left| \frac{\widehat{\psi}_m(t)}{\widehat{\varphi}(t)} - \frac{\psi_m(t)}{\varphi(t)} \right| = o(1).$$

(v) It is easy to extend these results to second derivatives of cumulant generating functions:

$$\begin{aligned} \zeta_{mp}(t) &\equiv \frac{\partial^2 \kappa_Y}{\partial t_m \partial t_p}(t) \\ &= -\frac{\mathbb{E} [Y_m Y_p \cdot e^{itY_\ell}]}{\mathbb{E} [e^{itY_\ell}]} + \frac{\mathbb{E} [Y_m e^{itY_\ell}]}{\mathbb{E} [e^{itY_\ell}]} \frac{\mathbb{E} [Y_p e^{itY_\ell}]}{\mathbb{E} [e^{itY_\ell}]} \\ &= -\frac{\xi_{mp}(t)}{\varphi(t)} + \frac{\psi_m(t)}{\varphi(t)} \frac{\psi_p(t)}{\varphi(t)}. \end{aligned}$$

The same asymptotic approximations apply to $\widehat{\zeta}_{mp}(t) - \zeta_{mp}(t)$ as

$$\begin{aligned} \widehat{\zeta}_{mp}(t) - \zeta_{mp}(t) &= -\left[\frac{\widehat{\xi}_{mp}(t)}{\widehat{\varphi}(t)} - \frac{\xi_{mp}(t)}{\varphi(t)} \right] \\ &\quad + \left[\frac{\widehat{\psi}_m(t)}{\widehat{\varphi}(t)} - \frac{\psi_m(t)}{\varphi(t)} \right] \frac{\psi_p(t)}{\varphi(t)} + \left[\frac{\widehat{\psi}_p(t)}{\widehat{\varphi}(t)} - \frac{\psi_p(t)}{\varphi(t)} \right] \frac{\psi_m(t)}{\varphi(t)} \\ &\quad + \left[\frac{\widehat{\psi}_m(t)}{\widehat{\varphi}(t)} - \frac{\psi_m(t)}{\varphi(t)} \right] \left[\frac{\widehat{\psi}_p(t)}{\widehat{\varphi}(t)} - \frac{\psi_p(t)}{\varphi(t)} \right], \end{aligned}$$

with $\widehat{\zeta}_{mp}(t) = -\frac{\widehat{\xi}_{mp}(t)}{\widehat{\varphi}(t)} + \frac{\widehat{\psi}_m(t)}{\widehat{\varphi}(t)} \frac{\widehat{\psi}_p(t)}{\widehat{\varphi}(t)}$.

Proceeding as above, we obtain:

$$\sup_{t \in [-T_N, T_N]} \left| \widehat{\xi}_{mp}(t) \right| \leq \sup_{t \in [-T_N, T_N]} \left| \widehat{\xi}_{mp}(t) - \xi_{mp}(t) \right| + \mathbb{E} |Y_m Y_p| = O(1),$$

as Y has finite second-order moments. Then, again as above:

$$\begin{aligned} \sup_{t \in [-T_N, T_N]} \left| \widehat{\zeta}_{mp}(t) - \zeta_{mp}(t) \right| &= O(1) \sup_{t \in [-T_N, T_N]} \left| \frac{\widehat{\varphi}(t) - \varphi(t)}{\varphi(t)^2} \right| + \sup_{t \in [-T_N, T_N]} \left| \frac{\widehat{\xi}_{mp}(t) - \xi_{mp}(t)}{\varphi(t)} \right| \\ &+ O(1) \sup_{t \in [-T_N, T_N]} \left| \frac{\widehat{\varphi}(t) - \varphi(t)}{\varphi(t)^3} \right| + O(1) \sup_{t \in [-T_N, T_N]} \left| \frac{\widehat{\psi}_m(t) - \psi_m(t)}{\varphi(t)^2} \right| \\ &+ O(1) \sup_{t \in [-T_N, T_N]} \left| \frac{\widehat{\varphi}(t) - \varphi(t)}{\varphi(t)^3} \right| + O(1) \sup_{t \in [-T_N, T_N]} \left| \frac{\widehat{\psi}_p(t) - \psi_p(t)}{\varphi(t)^2} \right| \\ &+ \left[O(1) \sup_{t \in [-T_N, T_N]} \left| \frac{\widehat{\varphi}(t) - \varphi(t)}{\varphi(t)^2} \right| + \sup_{t \in [-T_N, T_N]} \left| \frac{\widehat{\psi}_m(t) - \psi_m(t)}{\varphi(t)} \right| \right] \\ &\times \left[O(1) \sup_{t \in [-T_N, T_N]} \left| \frac{\widehat{\varphi}(t) - \varphi(t)}{\varphi(t)^2} \right| + \sup_{t \in [-T_N, T_N]} \left| \frac{\widehat{\psi}_p(t) - \psi_p(t)}{\varphi(t)} \right| \right]. \end{aligned}$$

Again, Lemma 16 applies to show that there exists a diverging sequence, T_N again, such that

$$\Delta_N^{(2)} \equiv \sup_{t \in [-T_N, T_N]} \left| \widehat{\zeta}_{mp}(t) - \zeta_{mp}(t) \right| = o(1).$$

(vi) Finally,

$$\sup_{t \in [-T_N, T_N]} \left| \int_0^t \frac{\widehat{\psi}_m(t')}{\widehat{\varphi}(t')} dt' - \int_0^t \frac{\psi_m(t')}{\varphi(t')} dt' \right| \leq \sup_{t \in [-T_N, T_N]} \left(\left| \frac{\widehat{\psi}_m(t)}{\widehat{\varphi}(t)} - \frac{\psi_m(t)}{\varphi(t)} \right| t \right).$$

Let us then redefine T_N as the minimum of T_N and $[\Delta_N^{(1)}]^{-1/2}$. Then, as in the proof of Lemma

16:

$$\begin{aligned} \sup_{t \in [-T_N, T_N]} \left| \int_0^t \frac{\widehat{\psi}_m(t')}{\widehat{\varphi}(t')} dt' - \int_0^t \frac{\psi_m(t')}{\varphi(t')} dt' \right| &\leq T_N \sup_{t \in [-T_N, T_N]} \left| \frac{\widehat{\psi}_m(t)}{\widehat{\varphi}(t)} - \frac{\psi_m(t)}{\varphi(t)} \right|, \\ &\leq [\Delta_N^{(1)}]^{-1/2} \Delta_N^{(1)}, \\ &= o(1). \end{aligned}$$

Similarly,

$$\begin{aligned} \sup_{t \in [-T_N, T_N]} \left| \int_0^t \int_0^{t'} \widehat{\zeta}_{mp}(t'') dt'' dt' - \int_0^t \int_0^{t'} \zeta_{mp}(t'') dt'' dt' \right| &\leq \sup_{t \in [-T_N, T_N]} \left(\left| \widehat{\zeta}_{mp}(t) - \zeta_{mp}(t) \right| \frac{t^2}{2} \right) \\ &\leq T_N^2 \sup_{t \in [-T_N, T_N]} \left| \widehat{\zeta}_{mp}(t) - \zeta_{mp}(t) \right| \\ &= T_N^2 \Delta_N^{(2)} = o(1), \end{aligned}$$

by appropriately redefining T_N as the minimum of T_N and $[\Delta_N^{(2)}]^{-1/4}$.

To finish the proof, note that, for large enough N and $t \in [-T_N, T_N]$:

$$\begin{aligned} \left| \exp\left(\int_0^t \frac{\widehat{\psi}_m(t')}{\widehat{\varphi}(t')} dt'\right) - \exp\left(\int_0^t \frac{\psi_m(t')}{\varphi(t')} dt'\right) \right| &\leq \sum_{k=1}^{\infty} \left| \int_0^t \frac{\widehat{\psi}_m(t')}{\widehat{\varphi}(t')} dt' - \int_0^t \frac{\psi_m(t')}{\varphi(t')} dt' \right|^k \\ &\leq \frac{\left| \int_0^t \frac{\widehat{\psi}_m(t')}{\widehat{\varphi}(t')} dt' - \int_0^t \frac{\psi_m(t')}{\varphi(t')} dt' \right|}{1 + \left| \int_0^t \frac{\widehat{\psi}_m(t')}{\widehat{\varphi}(t')} dt' - \int_0^t \frac{\psi_m(t')}{\varphi(t')} dt' \right|}, \\ &= O\left(\left| \int_0^t \frac{\widehat{\psi}_m(t')}{\widehat{\varphi}(t')} dt' - \int_0^t \frac{\psi_m(t')}{\varphi(t')} dt' \right|\right), \end{aligned}$$

provided that $\left| \int_0^t \frac{\widehat{\psi}_m(t')}{\widehat{\varphi}(t')} dt' - \int_0^t \frac{\psi_m(t')}{\varphi(t')} dt' \right| < 1$. Hence the result in the case of equation (23). A similar reasoning yields the result for the case of equation (21).

■

The proof of Theorem 13 then immediately follows:

Proof of Theorem 13. For all $\ell \in \{1, \dots, L\}$ and $t \in \mathbb{R}$:

$$\bar{h}_\ell(t) \geq |\varphi_Y(t\iota_\ell)| = \left| \left(\prod_k \varphi_{X_k}(\lambda_{\ell k} t) \right) \varphi_{U_\ell}(t) \right| \geq \underline{h}_\ell(t),$$

for $|t|$ large enough, where

$$\bar{h}_\ell(t) \equiv \left| \left(\prod_k \bar{h}(|\lambda_{\ell k} t|) \right) \bar{h}(t) \right|,$$

and $\underline{h}_\ell(t)$ has a similar expression, for all $t \in \mathbb{R}^+$.

Now, for all x in the support of X_k :

$$\left| \widehat{f}_{X_k}(x) - f_{X_k}(x) \right| \leq \frac{1}{2\pi} \left(\int_{-T_N}^{T_N} |\widehat{\varphi}_{X_k}(v) - \varphi_{X_k}(v)| dv + \int_{-\infty}^{-T_N} |\varphi_{X_k}(v)| dv + \int_{T_N}^{+\infty} |\varphi_{X_k}(v)| dv \right).$$

As \bar{h}_ℓ and \bar{h}^ℓ satisfy the assumptions of Lemma 17, there exists T_N such that the first term on the RHS of this inequality tends to zero. Hence, for N large enough:

$$\left| \widehat{f}_{X_k}(x) - f_{X_k}(x) \right| \leq o(1) + \frac{1}{\pi} \int_{T_N}^{+\infty} \bar{h}_\ell(v) dv = o(1),$$

where the equality comes from the fact that \bar{h}^{K+1} , and hence \bar{h}_ℓ , are integrable and that T_N tends to infinity. ■

8 Appendix

8.1 Principles of the JADE algorithm

The baseline model considered in the ICA literature writes:

$$Y = \Lambda X,$$

with $K = L$, and otherwise the same assumptions as in Section 2. Cardoso and Souloumiac (1993) assume additionally that $L = K$, and propose to consider the following matrices of fourth-order cumulants:

$$\begin{aligned} \Omega(\ell, \ell') &\equiv \Lambda D_4 \text{diag}(\Lambda_\ell * \Lambda_{\ell'}) \Lambda^T, \\ &= (\text{Cum}(Y_m, Y_\ell, Y_{\ell'}, Y_{m'}))_{(m, m') \in \{1 \dots L\}^2}. \end{aligned}$$

Cardoso (*e.g.* 1999) derives from this expression a first way of estimating Λ by using that:

$$\Omega(\ell, \ell') [\Omega(1, 2)]^{-1} \equiv \Lambda \text{diag}(\Lambda_\ell * \Lambda_{\ell'}) \text{diag}(\Lambda_1 * \Lambda_2)^{-1} \Lambda^{-1},$$

provided that $\Omega(1, 2)$ is invertible, for whatever pair $(\ell, \ell') \in \bar{\Delta}_2(L)$.

As pointed out by Cardoso (1999), such an algorithm suffers from two flaws. First, in practice, near non-invertibility of $\Omega(1, 2)$ can cause serious problems. The method is not equivariant; that is: some matrices of factor loadings are well estimated, while others are not. Second, this method uses a small part of the information contained in the fourth-order cumulants: it is not efficient.

Instead, Cardoso and Souloumiac (1993) rely on second-order restrictions to pre-whiten the data. As:

$$\Sigma_Y = \Lambda \Lambda^T,$$

it follows that Λ can be estimated as $\Sigma_Y^{1/2} V$, where V is the orthogonal matrix of eigenvectors of $\Sigma_Y^{-1/2} \Omega(\ell, \ell') \Sigma_Y^{-1/2}$.¹⁹ As the spectrum of a symmetric matrix is real, such an estimation can always be performed. This solves the problem of near non-invertibility.

¹⁹A popular way of “pre-whitening” the data is to compute $\tilde{Y} = \hat{\Sigma}_Y^{-1/2} Y$, where $\hat{\Sigma}_Y^{-1/2}$ is a consistent estimator of $\Sigma_Y^{-1/2}$, and to apply the algorithm to \tilde{Y} .

Then, Cardoso and Souloumiac (1993) improve efficiency and achieve asymptotic equivariance by considering the joint diagonalization of matrices $\Sigma_Y^{-1/2}\Omega(\ell, \ell')\Sigma_Y^{-1/2}$, for all (ℓ, ℓ') , in orthonormal bases. Their algorithm uses simple Jacobi techniques, that we now outline.

8.2 Jacobi methods for joint diagonalization

In this section of the Appendix, we shall follow Cardoso and Souloumiac (1996). A MATLAB listing can be found on the webpage of Cardoso. The GAUSS listings of the main programs that we use in this paper will soon be available on the webpage of the second author of this paper.

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

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

for all $A = (a_{ij})_{(i,j)}$. Then joint diagonalization of \mathcal{A} is achieved by minimizing

$$\sum_{k=1}^K \text{off}(UA_kU^T), \quad (32)$$

with respect to U orthogonal.

Let $\theta \in [-\pi, \pi]$, let $(i, j) \in \{1 \dots L\}^2$ 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:

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

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

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

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

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

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

This closed-form expression for θ_0 allows to minimize (32) 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 $R_{ij}(\theta_0)A_kR_{ij}(\theta_0)^T$ 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 .

8.3 The algorithm that we use in 4.1

To compare this approach to ours, we use in 4.1 an analogue of the algorithm presented in 3.2, adapted to the noise-free case. The algorithm writes as follows:

1. Estimate Σ_Y , $\Omega(\ell, \ell')$ for all $(\ell, \ell') \in \Delta_2(L)$, and their covariance matrices. Covariance matrices are estimated by a first bootstrap procedure.
2. For all ℓ, ℓ' , compute $\widehat{C}(\ell, \ell') = \widehat{\Sigma}_Y^{-1/2}\widehat{\Omega}(\ell, \ell')\widehat{\Sigma}_Y^{-1/2}$. Obtain $\widehat{V}(\ell, \ell')$ as the orthogonal matrix of eigenvectors of $\widehat{C}(\ell, \ell')$ and compute $\widehat{\Lambda}(\ell, \ell') = \widehat{\Sigma}_Y^{1/2}\widehat{V}(\ell, \ell')$.
3. Bootstrap Step 3. This yields the covariance matrix of $vec(\widehat{\Lambda}(\ell, \ell'))$, for all $(\ell, \ell') \in \Delta_2(L)$.
4. Compute weights $\omega_{\ell, \ell'}$ as:

$$\omega_{\ell, \ell'} = \left\| \left\| vec(\widehat{\Lambda}(\ell, \ell')) \right\|_2 \right\|_2^2.$$

5. Joint diagonalize matrices $\omega_{\ell, \ell'}^{-1} \widehat{C}(\ell, \ell')$ by the Jacobi algorithm, obtain \widehat{W} and compute:

$$\widehat{\Lambda} = \widehat{\Sigma}_Y^{1/2} \widehat{W}.$$

6. Bootstrap Step 5 to obtain standard errors.

8.4 Equivariance

Can a similar approach be adopted in the case of the factor model considered in Definition 1? The answer is no, as pre-whitening does not preserve the structure of factor models with errors. In fact:

$$\Sigma^{-1/2} Y = \Sigma^{-1/2} \Lambda X + \Sigma^{-1/2} \Sigma_U,$$

where $\Sigma^{-1/2} \Sigma_U$ is no longer diagonal.

It is to be noted that for this same reason, equivariance does not have the same meaning in our case as in Cardoso and Souloumiac (1993). In the absence of noise, the definition is as follows: an estimator of Y , say $\widehat{\theta}(Y)$, is equivariant if and only if

$$\widehat{\theta}(AY) = A\widehat{\theta}(Y),$$

for all invertible $L \times L$ matrix A .

Allowing for noise, the definition of equivariance becomes: $\widehat{\theta}(Y)$ is equivariant if and only if

$$\widehat{\theta}(DY) = D\widehat{\theta}(Y),$$

for all invertible $L \times L$ *diagonal* matrix D .

For this notion of equivariance, or metric invariance, the algorithm introduced in 3.2 can be shown to be equivariant. To see why, let us denote matrices $\widehat{B}(\ell, M)$ as $\widehat{B}(\ell, M)(Y)$, to emphasize the dependence on the data. Likewise, let us denote all the estimators introduced in 3.2 in the same way.

Let D be an invertible $L \times L$ *diagonal* matrix. Then, it follows from the definition of cumulants that:

$$\widehat{B}(\ell, M)(DY) = \frac{d_\ell}{d_1} D \widehat{B}(\ell, M)(Y) D^{-1},$$

where $D = \text{diag}(d_1, \dots, d_L)$, and $d_0 = 1$ by definition.

It then follows from (17) that:

$$\widehat{\Sigma}_U(DY) = D\widehat{\Sigma}_U(Y)D.$$

Writing the expressions of $\widehat{C}(\ell, M)(Y)$ and $\widehat{C}(\ell, M)(DY)$, ones then sees that this implies:

$$\widehat{V}(\ell, M)(DY) = D^{-1/2} \left(\widehat{\Sigma}_Y - \widehat{\Sigma}_U \right)^{-1/2} D^{1/2} \left(\widehat{\Sigma}_Y - \widehat{\Sigma}_U \right)^{1/2} \widehat{V}(\ell, M)(Y),$$

from which it directly follows that:

$$\widehat{\Lambda}_\ell(DY) = D\widehat{\Lambda}_\ell(Y).$$

This shows equivariance in the above sense.

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] ANDERSON, T. W., AND H. RUBIN (1956): “Statistical Inference in Factor Analysis,” in *Proceedings of the Third Symposium in Mathematical Statistics and Probability*, Vol. 5. University of California press.
- [3] CARDOSO J.-F. (1998): “Blind signal separation : statistical principles,” *Proc. IEEE*, 9(10), 2009-2025.
- [4] CARDOSO J.-F. (1999): “High-order contrasts for independent Component Analysis,” *Neural Computation*, 11, 157-192.
- [5] CARDOSO J.-F., and A. SOULOUMIAC (1993): “Blind Beamforming for Non-Gaussian Signals,” *IEE-Proceedings-F*, 140, 362-370.

- [6] CARDOSO J.-F., and A. SOULOUMIAC (1996): “Jacobi Angles for Simultaneous Diagonalization,” *SIAM J. Mat. An. Appl.*, 17, 161-164.
- [7] 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.
- [8] CARRASCO, M., AND J. P. FLORENS (2000): “Generalization of GMM to a Continuum of Moment Conditions,” *Econometric Theory*, 16, 797-834.
- [9] CARROLL, R. J. , AND P. HALL (1988): “Optimal rates of Convergence for Deconvoluting a Density,” *Journal of the American Statistical Association*, 83, 1184-1186.
- [10] CRAGG, J. G. (1997): “Using Higher Moments to Estimate the Simple Errors-in-Variables Model,” *RAND Journal of Economics*, 28, S71-S91.
- [11] 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.
- [12] 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.
- [13] DIGGLE, P. J., AND P. HALL (1993): “A Fourier Approach to Nonparametric Deconvolution of a Density Estimate,” *Journal of the Royal Statistical Society Series B*, 55, 523-531.
- [14] DUGUË, D. (1951), “Analyticité et convexité des fonctions caractéristiques,” *Annales de l’Institut Henri Poincaré*, Vol. XII, 45-46.
- [15] 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.

- [16] 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.
- [17] FAN, J. Q. (1991): "On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems," *Annals of statistics*, 19, 1257-1272.
- [18] GEWEKE, J., AND ZHOU (1996): "Measuring the Pricing Error of the Arbitrage Pricing Theory," *Review of Financial Studies*, 9, 557-587.
- [19] HALL, P., AND Q. YAO (2003): "Inference in Components of Variance Models with Low Replications," *Annals of Statistics*, 31, 414-441.
- [20] HOROWITZ, J. L. (1998): *Semiparametric Methods in Econometrics*. New-York: Springer-Verlag.
- [21] HOROWITZ, J. L., AND M. MARKATOU (1996): "Semiparametric Estimation of Regression Models for Panel Data," *Review of Economic Studies*, 63, 145-168.
- [22] HYVARINEN A., J. KARHUNEN AND E. OJA (2001), *Independent Component Analysis*, John Wiley & Sons, New York.
- [23] KOTLARSKI, I. (1967): "On Characterizing the Gamma and Normal Distribution," *Pacific Journal of Mathematics*, 20, 69-76.
- [24] 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.
- [25] LI, T. (2002): "Robust and Consistent Estimation of Nonlinear Errors-in-Variables Models," *Journal of Econometrics*, 110, 1-26.
- [26] LI, T., AND Q. VUONG (1998): "Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators," *Journal of Multivariate Analysis*, 65, 139-165.

- [27] MADANSKY, A. (1959): “The Fitting of Straight Lines When Both Variables are Subject to Error,” *Journal of the American Statistical Association*, 54, 173-205.
- [28] MOULINES E., J.F. CARDOSO AND E. GASSIAT (1997): “Maximum likelihood for blind separation and deconvolution of noisy signals,” *Proc. of the IEEE int. conf. on acoustics, speech and signal processing*, 3617-3620.
- [29] PAL, M. (1980): “Consistent Moment Estimators of Regression Coefficients in the Presence of Errors-in-Variables,” *Journal of Econometrics*, 14, 349-364.
- [30] REIERSOL, O. (1950): “Identifiability of a Linear Relation Between Variables which are Subject to Error,” *Econometrica*, 9, 1-24.
- [31] ROBIN, J.M., R.J. SMITH (2000): “Tests of rank,” *Econometric Theory*, vol 16, 151-175.
- [32] SPEARMAN, C. (1904): “General intelligence, objectively determined and measured,” *American Journal of Psychology*, 15, 201-293.
- [33] SPIEGELMAN, C. (1979): “On Estimating the Slope of a Straight Line when Both Variables are Subject to Error,” *Annals of Statistics*, 7, 201-206.
- [34] VAN MONTFORT, K., A. MOOLJAART, AND J. DE LEEUW (1989): “Estimation of Regression Coefficients with the Help of Characteristic Functions,” *Journal of Econometrics*, 41, 267-278.























