Generalized Nonparametric Deconvolution with an Application to Earnings Dynamics[1]

Stéphane Bonhomme
CEMFI, Madrid

Jean-Marc Robin
Université de Paris 1 - Panthéon - Sorbonne
University College London
and Institute for Fiscal Studies

June 2006

[1] We thank Laura Hospido for providing the data. We also thank Manuel Arellano, Jean-Pierre Florens and Eric Renault for comments and suggestions. The usual disclaimer applies.

Abstract

In this paper, we study nonparametric methods for estimating the distributions of factor and error variables in linear independent multi-factor models, under the assumption that factor loadings are known. We show that one can nonparametrically identify and estimate the distributions of up to $L\left(L+1\right)/2$ factors and errors given L measurements. Nonparametric deconvolution estimators of these distributions are constructed and their asymptotic properties are studied. Monte-Carlo simulations show good finite-sample performance, less so if distributions are highly skewed or leptokurtic. In addition, we provide an application to data from the PSID, analyzing a simple permanent/transitory decomposition of individual wages.

JEL codes: C13, C14.

Keywords: Factor models, nonparametric estimation, deconvolution, Fourier transformation, earnings dynamics.

1 Introduction

In this paper, we study nonparametric methods for identifying and estimating the distributions of factor and error variables in linear multi-factor models of the form: $Y = \Lambda X + U$, where Y is a vector of L observed outcomes, X is a vector of K unobservable common factors, U is a vector of L independent errors and Λ is a $L \times K$ matrix of parameters (factor loadings). The critical assumption is that all components of X and U are mutually independent. Throughout the paper, we assume that a root-N consistent estimator of the matrix of factor loadings Λ is available and that the number of factors is known.

Examples of applications of factor models are numerous in econometrics. In financial econometrics, Principal Component Analysis (PCA) is widely used to uncover fundamental factors driving the dynamics of stock returns. Macroeconometric structural VAR models aim at identifying the effects of various structural shocks on macro-variables. Microeconometric applications flourish.[1]

In this paper we consider the following particular application. The log of earnings at time t of some individual i, y_{it}, is modelled as the sum of a fixed effect α_i, a random walk component, p_{it}, and a transitory, say i.i.d., component, r_{it}:

$$\begin{aligned} y_{it} &= \alpha_i + p_{it} + r_{it} \\ &= \alpha_i + p_{i0} + u_{i1} + ... + u_{it} + v_{it}, \quad t = 1, ..., L, \end{aligned} \tag{1}$$

where u_{it} and v_{it} are independent permanent and transitory shocks. This is a linear factor model with $2L + 1$ factors/errors. The generalized nonparametric deconvolution method of this paper allows to recover nonparametrically the distributions of the transitory and permanent shocks.

In many applications of factor models, it is important to estimate both factor loadings and a prediction of the underlying factors given observables. If the matrix Λ is full column rank and there are no errors, then simple matrix inversion works for this purpose, as in PCA. However, if either errors are present or there are more factors than measurements, then more sophisticated methods have to be used. This is the case of model (1).

One possible solution uses flexible parametric families of distributions, such as the family of normal mixtures. There are two main difficulties with this approach: First, it is difficult to estimate the degree of mixing—the number of components—of a mix-

[1] Factor models are also useful in nonlinear models. For example, in a series of recent papers, Heckman and various coauthors estimate Roy models of education choice, assuming outcome and treatment variables independent given individual heterogeneity. They model unobserved heterogeneity variables as independent factors. See Carneiro, Hansen and Heckman (2003), Cunha, Heckman and Navarro (2005) and Heckman and Navarro (2005).

ture model. Second, whether statistical inference is frequentist or Bayesian, computer-intensive techniques such as the EM algorithm and MCMC algorithms are requested.

The alternative approach that we adopt in this paper uses nonparametric deconvolution methods. The most classical deconvolution problem has one factor and one error, and the error distribution is known. With repeated observations, one can also solve the problem of one-factor models with unknown error distributions. We briefly review this literature in the next section. To our best knowledge no general solution to the linear multi-factor case has been proposed so far.[2]

Our contribution is twofold: First, we show that $\frac{L(L+1)}{2}$ factor and error distributions are nonparametrically identified under relatively mild conditions on second-order partial derivatives of the logarithm of the characteristic function of the data. Second, we propose a nonparametric estimation procedure for factor and error distributions based on these second-order functional restrictions. Factor and error densities then straightforwardly follow by integration and inverse Fourier transformation. This procedure can be seen as a generalization of Li and Vuong's (1998) estimator to the multi-factor case.

We prove that our estimator converges uniformly to the true density when the sample size tends to infinity. Moreover, we provide upper bounds for the convergence rates in special cases. Following Hu and Ridder (2005), we do not assume bounded support for factor and error variables. It is indeed generally not possible for the cumulant generating function of a bounded random variable to be everywhere nonvanishing. Allowing for unbounded support makes asymptotic results more difficult to obtain but gives them greater generality.

Our findings are consistent with those in the deconvolution literature. Asymptotic convergence rates are slower than root-N, and can be very slow in special cases. We illustrate our approach by means of Monte Carlo simulations. We find that the shape of factor distributions strongly influences the finite-sample performance of the estimator. In particular, the estimation of factor distributions is more difficult when these distributions are skewed or leptokurtic.

Finally, we apply our methodology to individual earnings data from the PSID. We estimate the distributions of the permanent and transitory shocks in model (1). This model generalizes the model of Horowitz and Markatou (1996) by allowing permanent shocks and non symmetric distributions. Our results show that both shocks exhibit more kurtosis than the normal distribution. We then compute a prediction of both factors for

[2] The case of nonlinear models with one variable measured with errors has also received some attention. See Li (2002), Schennach (2004) and Hu and Ridder (2005).

each individual in the sample given observables. Correlating these imputed values to job change propensity, we find that frequent job changers face more permanent and more transitory earnings shocks than job stayers.

The outline of the paper is as follows. First, we provide a short review of the literature. In Section 3, we discuss identification. In Section 4, we propose our nonparametric estimator and we study its asymptotic properties in Section 5. Sections 6 and 7 present some simulations and the application. Lastly, Section 8 concludes.

2 Review of the literature

The deconvolution problem. The problem is to estimate the distribution of X (with density f_X) given the density of U (f_U) and a sample of i.i.d. observations of a random variable Y such that $Y = X + U$. It is assumed that X and U are independent and absolutely continuous. Let φ_Y, φ_Y and φ_U denote the characteristic functions (c.f.) of Y, X and U. Assume that φ_U is nonvanishing everywhere. Then,

$$\varphi_X(t) = \frac{\varphi_Y(t)}{\varphi_U(t)},$$

and the probability density function (p.d.f.) of X, say f_X, follows as the inverse Fourier transform of φ_X:

$$\begin{aligned} f_X(x) &= \frac{1}{2\pi}\int e^{-itx}\varphi_X(t)dt, \\ &= \frac{1}{2\pi}\int e^{-itx}\frac{\varphi_Y(t)}{\varphi_U(t)}dt. \end{aligned} \tag{2}$$

As noted by several authors (e.g. Horowitz, 1998), this argument proves identification, but cannot be directly used for estimation since the above integral does not necessarily converge when the characteristic functions are replaced by their empirical analogs. Consistent estimation of f_X therefore requires some regularization, resulting in low convergence rates.

Generalizing the results of Carroll and Hall (1988), Fan (1991) shows that the convergence rate of the deconvolution estimator crucially depends on the *smoothness* of the distributions of X and U, as characterized by the speed of convergence of characteristic functions to 0. A normal distribution has a c.f. which converges to 0 at an exponential rate. The worst case is obtained when the error is much smoother than the factor. This happens for example when the factor's c.f. converges at a polynomial rate and the error at an exponential rate, as in the normal case.

An alternative way of solving the deconvolution problem uses projection estimators. See Pensky and Vidakovic (1999), Fan and Koo (2002), Carrasco and Florens (2005), Comte, Rozenholc and Taupin (2006), and references therein. The deconvolution equation in the density space,

$$f_Y(y) = \int f_U(y-x)f_X(x)dx,$$

is an inverse problem of the type: $Tf_X = f_Y$, where T is a linear operator. By restricting the domain of T, one obtains a compact operator. One can then, for example, compute the singular value decomposition of T coupled with Tikhonov regularization to invert the integral equation. It may be possible to generalize this approach to the multi-factor deconvolution problem. However, the operator linking factor densities to the density of the data is no more linear in that case. One solution to this problem could be to use a local linearization of the operator around the true values (Carrasco *et al.*, 2006). Still, this approach is not straightforward in the present case and we leave it for future research.

Repeated measurements. Now, let there be two measurements of X:

$$\begin{cases} Y_1 = X + U_1, \\ Y_2 = X + U_2, \end{cases}$$

with X, U_1 and U_2 independent. Let us also assume that measurements are centered, and X, U_1 and U_2 have zero mean. Kotlarski (1967) shows that the distributions of all three variables X, U_1 and U_2 are identified by the distribution of $Y = (Y_1, Y_2)$ (see e.g. Rao, 1992, p. 21).

Various nonparametric estimators of the distributions of X and U have been proposed in the literature. Horowitz and Markatou (1996) focus on the case where the distribution functions (d.f.) of U_1 and U_2 are identical and symmetric. Then, the information contained in the three univariate distributions of Y_1, Y_2 and $Y_2 - Y_1$ is enough to construct consistent estimators of f_X and $f_{U_1} = f_{U_2}$. Li and Vuong (1998) consider the general case of non symmetric distributions for U_1 and U_2, possibly different.

We now describe Li and Vuong's estimator in some details as our paper extends their approach to the multi-factor case. The c.f. of $Y = (Y_1, Y_2)$ is

$$\varphi_Y(t_1, t_2) = \varphi_X(t_1 + t_2)\varphi_{U_1}(t_1)\varphi_{U_2}(t_2). \tag{3}$$

Hence,

$$\frac{\partial \ln \varphi_Y(0,t)}{\partial t_1} = \left(\ln \varphi_X\right)'(t),$$

as $\left(\ln \varphi_{U_1}\right)'(0) = i\mathbb{E}U_1 = 0$. Li and Vuong estimate $\varphi_X(t)$ as

$$\widehat{\varphi}_X(t) = \exp \int_0^t \frac{\partial \ln \widehat{\varphi}_Y(0,u)}{\partial t_1} du, \tag{4}$$

where $\widehat{\varphi}_Y(t) = \frac{1}{N}\Sigma_{i=1}^N e^{ity_i}$ is a consistent estimate of $\varphi_Y(t)$ from an i.i.d. sample $(y_1, .., y_N)$. The p.d.f. of X is recovered by inverse Fourier transformation, as in the deconvolution problem. Li and Vuong derive the convergence rates of their estimator in several cases, depending on the smoothness of the distributions. Li (2002) uses this estimator in the context of nonlinear errors-in-variables models.

In a recent paper, Hall and Yao (2003) build an alternative uniformly consistent estimator of f_X from the distributions of Y_1, Y_2 and Y_1+Y_2. Condition (3) indeed implies the following restriction:[3]

$$\frac{\varphi_Y(t,t)}{\varphi_Y(t,0)\varphi_Y(0,t)} = \frac{\varphi_{Y_1+Y_2}(t)}{\varphi_{Y_1}(t)\varphi_{Y_2}(t)} = \frac{\varphi_X(2t)}{\varphi_X(t)^2}. \tag{5}$$

Function $h\left(t\right) = \frac{\varphi_{Y_1+Y_2}(t)}{\varphi_{Y_1}(t)\varphi_{Y_2}(t)}$ has an immediate empirical counterpart and f_X can be estimated either as a discrete approximation verifying restriction (5) or by inverse Fourier transformation of the analytical solution to equation (5), that is:

$$\varphi_X\left(t\right) = \prod_{j=0}^{\infty} h\left(\frac{t}{2^{j+1}}\right)^{2j}. \tag{6}$$

Extending this approach to the multi-factor case is not straightforward. We provide a generalization in a technical appendix to this paper.[4]

3 Identification

In this section, we study the identification of factor densities. We shall most of the time eliminate the distinction between factors and errors, because, as far the identification and estimation of their distribution is concerned, the distinction is not essential. So, we now consider the case of DGPs of the form: $Y = AX$, where

1. $Y = (Y_1, ..., Y_L)^T$ is a vector of $L \geq 2$ zero-mean real-valued random variables (where ^{T} denotes the matrix transpose operator).

2. $X = (X_1, ..., X_K)^T$ is a random vector of K real valued, mutually independent and non degenerate random variables, with zero mean and finite variances.

3. $A = [a_{ij}]$ is a known $L \times K$ matrix of scalar parameters.

[3] This restriction is not exactly identical to the one exploited by Hall and Yao. The difference is not essential however.

[4] "Generalizing the Hall and Yao Estimator", available at http://www.cemfi.es/∼bonhomme/

In this paper, we assume that factor loadings are known to the researcher. Alternatively, we may assume that a root-N consistent estimator of A is available. The asymptotic results derived in this paper remain unchanged, as we will find convergence rates of density estimators that are slower than root-N. If we are interested in functionals of the density estimators such as means and variances, then the randomness of the estimator of matrix A should be taken into account.

3.1 The identification theorem

The following two assumptions will be maintained throughout the paper.

Assumption 1 *Any two columns of A are linearly independent.*

Assumption 2 *The characteristic functions of factor variables X_1,...,X_K are everywhere non vanishing and twice differentiable.*

In the factor analysis literature, variables Y_ℓ are called measurements, and parameters a_{ij} are called factor loadings. If the kth column of A (say, $A_{[\cdot,k]}$) contains only one nonzero element, variable X_k is called an error and otherwise, it is called a factor. In the various examples that we shall consider, we use $U_1, ..., U_L$ to denote error variables, reserving the notation X for common factors if A of the form $A = (\Lambda, I_L)$.

Assumption 1 is clearly necessary for identification. If two columns of A are proportional, say $A_{[\cdot,k]} = \alpha A_{[\cdot,j]}$, then $A_{[\cdot,k]}X_k + A_{[\cdot,j]}X_j = A_{[\cdot,j]}(\alpha X_k + X_j)$ and there is obviously no way to separately identify the distribution of X_k from the distribution of X_j.

However, Assumptions 1 and 2 are not sufficient for factor distributions to be identified. For identification to hold, we need an additional rank condition on matrix A. Let $Q(A)$ be the matrix operator that changes $A = [a_{ij}] \in \mathbb{R}^{L\times K}$ into the $\frac{L(L+1)}{2} \times K$ matrix which generic element is $a_{\ell k}a_{mk}$, when the row index is $(\ell, m) \in \{1, ..., L\}^2$, $\ell \leq m$, and the column index is $k \in \{1, ..., K\}$. In the sequel, we write Q for $Q(A)$ to simplify the notations. Let us assume that matrix Q has rank K. Then the following theorem, due to Székely and Rao (2000, p. 85), shows that the distribution of X_k is identified.

Theorem 3 *If assumptions 1-2 hold, and if Q has rank K, then the distribution of $(X_1, ..., X_K)$ is uniquely determined.*

Theorem 3 gives a sufficient condition for nonparametric identification. In addition, Székely and Rao (2000) prove that this condition is necessary. Precisely they prove that, if rank$(Q) < K$ then one can always find a distribution for Y for which factor distributions are not uniquely determined. However, Q being full-column rank is not necessary for identification if factor variances are known.

In this paper we shall assume that Assumption 2 is satisfied. The existence of first derivatives in a neighborhood of zero (i.e. finite expectations) is sufficient for this purpose. We now explain how Assumption 2 yields convenient identifying restrictions to construct estimators of factor distributions.

3.2 Moment restrictions

Notations. Under assumption 2, cumulant generating functions (c.g.f.) are well defined and everywhere two times differentiable. Let us denote the characteristic functions of Y and X_k as φ_Y and φ_{X_k}, and their c.g.f.'s as $\kappa_Y = \ln \varphi_Y$ and $\kappa_{X_k} = \ln \varphi_{X_k}$. The independence assumptions and the linear factor structure imply that, for all $t = (t_1, ..., t_L) \in \mathbb{R}^L$,

$$\kappa_Y(t) \equiv \ln \left[\mathbb{E} \exp \left(it^T Y\right)\right] = \sum_{k=1}^{K} \kappa_{X_k} \left(t^T A_{[\cdot,k]}\right). \tag{7}$$

We denote as $\partial_\ell \kappa_Y(t)$ the ℓth partial derivative of $\kappa_Y(t)$ and as $\partial^2_{\ell m} \kappa_Y(t)$ the second-order partial derivative of $\kappa_Y(t)$ with respect to t_ℓ and t_m:

$$\partial_\ell \kappa_Y(t) = i \frac{\mathbb{E}\left[Y_\ell e^{it^T Y}\right]}{\mathbb{E}\left[e^{it^T Y}\right]}, \tag{8}$$

$$\partial^2_{\ell m} \kappa_Y(t) = -\frac{\mathbb{E}\left[Y_\ell Y_m e^{it^T Y}\right]}{\mathbb{E}\left[e^{it^T Y}\right]} + \frac{\mathbb{E}\left[Y_\ell e^{it^T Y}\right]}{\mathbb{E}\left[e^{it^T Y}\right]} \frac{\mathbb{E}\left[Y_m e^{it^T Y}\right]}{\mathbb{E}\left[e^{it^T Y}\right]}. \tag{9}$$

Let $\Delta_{L,2} = \left\{(\ell, m) \in \{1, ..., L\}^2, \ell \leq m\right\}$ be a set of $L(L+1)/2$ bidimensional indices. We denote as $\nabla \kappa_Y(t)$ the L-dimensional gradient vector and as $\nabla^2 \kappa_Y(t)$ the vector of all $\frac{L(L+1)}{2}$ non redundant second-order partial derivatives arranged in lexicographic order of (ℓ, m) in $\Delta_{L,2}$. Lastly, for any $\tau = (\tau_1, ..., \tau_K) \in \mathbb{R}^K$, we denote as

$$\begin{aligned} \boldsymbol{\kappa}_X(\tau) &= \left(\kappa_{X_1}(\tau_1), ..., \kappa_{X_K}(\tau_K)\right)^T, \\ \boldsymbol{\kappa}'_X(\tau) &= \left(\kappa'_{X_1}(\tau_1), ..., \kappa'_{X_K}(\tau_K)\right)^T, \\ \boldsymbol{\kappa}''_X(\tau) &= \left(\kappa''_{X_1}(\tau_1), ..., \kappa''_{X_K}(\tau_K)\right)^T, \end{aligned}$$

the K-dimensional vectors of factor cumulant generating functions and their first and second derivatives.

Restrictions. First-differentiating equation (7) yields

$$\nabla \kappa_Y(t) = A\boldsymbol{\kappa}'_X\left(A^T t\right) = \sum_{k=1}^{K} \kappa'_{X_k}\left(t^T A_{[\cdot,k]}\right) A_{[\cdot,k]}. \tag{10}$$

In general $K > L$ as there are L errors and at least one common factor. So there are more function κ'_{X_k} than $\partial_\ell \kappa_Y$. To obtain an invertible system, we differentiate one more time:

$$\nabla^2 \kappa_Y(t) = Q\boldsymbol{\kappa}''_X\left(A^T t\right), \tag{11}$$

where Q is the matrix $L(L+1)/2 \times K$ matrix defined above. Equation (11) is a generalization to the entire spectrum of the variance equality used in factor analysis (e.g. Anderson and Rubin, 1956). It is the basis of our identification and estimation strategy.

Assume Q full column rank. This requires in particular that $K \leq \frac{L(L+1)}{2}$. One can invert equation (11) as

$$\boldsymbol{\kappa}''_X\left(A^T t\right) = Q^{-}\nabla^2 \kappa_Y(t), \tag{12}$$

where Q^- is a generalized inverse of Q.[5] This equation provides a set of overidentifying restrictions which can be exploited to yield an expression for κ''_{X_k}.

Let $\mathcal{T}_k = \left\{t \in \mathbb{R}^L | t^T A_{[\cdot,k]} = 1\right\}$. $\mathcal{T}_k$ is not empty as there is at least one non zero element in $A_{[\cdot,k]}$. Let $\left(Q^-\right)_{[k,\cdot]}$ denote the kth row of Q^-. Then, for all $t \in \mathcal{T}_k$ and $\tau_k \in \mathbb{R}$,

$$\kappa''_{X_k}(\tau_k) = \left(Q^-\right)_{[k,\cdot]} \nabla^2 \kappa_Y(\tau_k t).$$

Integrating with respect to τ_k, using the constants of integration: $\kappa'_{X_k}(0) = i\mathbb{E}X_k = 0$ and $\kappa_{X_k}(0) = 0$, yields

$$\kappa_{X_k}(\tau_k) = \int_0^{\tau_k}\int_0^{u} \left(Q^-\right)_{[k,\cdot]} \nabla^2 \kappa_Y(vt)\, dv du. \tag{13}$$

Equation (13) can directly be used for estimation of factor characteristic functions and densities, as we shall explain in the next section.

Example: the measurement error model. Let

$$\begin{cases} Y_1 = X + U_1, \\ Y_2 = aX + U_2, \end{cases}$$

with $a \neq 0$ and $X \in \mathbb{R}$. We here assume that a is known.[6]

[5] That is: $Q^- = \left(Q^T W Q\right)^{-1} Q^T W$, for a symmetric, positive definitive matrix W.

[6] A root-N consistent estimator of a could be obtained using third-order moments of Y, by regressing Y_2 on Y_1 using $Y_1 Y_2$ as instrument, provided that the distribution of X is skewed (Geary, 1942, Reiersol, 1950).

Then,

$$\kappa_Y(t_1,t_2) = \kappa_X(t_1+at_2) + \kappa_{U_1}(t_1) + \kappa_{U_2}(t_2),$$

and

$$\underbrace{\begin{pmatrix} \partial^2_{11}\kappa_Y(t_1,t_2) \\ \partial^2_{12}\kappa_Y(t_1,t_2) \\ \partial^2_{22}\kappa_Y(t_1,t_2) \end{pmatrix}}_{\nabla^2\kappa_Y(t)} = \underbrace{\begin{pmatrix} 1 & 1 & 0 \\ a & 0 & 0 \\ a^2 & 0 & 1 \end{pmatrix}}_{Q} \underbrace{\begin{pmatrix} \kappa''_X(t_1+at_2) \\ \kappa''_{U_1}(t_1) \\ \kappa''_{U_2}(t_2) \end{pmatrix}}_{\boldsymbol{\kappa}''_X(A^Tt)}.$$

The restriction:

$$\kappa''_X(t_1+at_2) = \frac{1}{a}\partial^2_{12}\kappa_Y(t_1,t_2), \quad \forall(t_1,t_2),$$

implies that, for all τ and t_1,

$$\kappa''_X(\tau) = \frac{1}{a}\partial^2_{12}\kappa_Y\left(t_1, \frac{1}{a}\tau - \frac{1}{a}t_1\right),$$

and

$$\begin{aligned} \kappa_X(\tau) &= \frac{1}{a}\int_0^\tau\int_0^u \partial^2_{12}\kappa_Y\left(t_1, \frac{1}{a}v - \frac{1}{a}t_1\right)dvdu, \\ &= \int_0^\tau\left[\partial_1\kappa_Y\left(t_1, \frac{1}{a}u - \frac{1}{a}t_1\right) - \partial_1\kappa_Y\left(t_1, -\frac{1}{a}t_1\right)\right]du. \end{aligned}$$

Taking $t_1 = 0$ yields Li and Vuong's (1998) solution:

$$\kappa_X(\tau) = \int_0^\tau \partial_1\kappa_Y\left(0, \frac{1}{a}u\right)du.$$

This discussion shows that Li and Vuong's estimator is particular for two reasons: First, in general, the double integrals of second-order derivatives of κ_Y in (13) will not simplify into a simple integral of first derivatives. Second, the choice of which component of t to fix a priori is arbitrary. It does not matter for identification. Yet, intuitively, overidentifying restrictions could be used to obtain more efficient estimates.

4 Estimation

We here introduce our estimator of factor densities. In the next section, we shall discuss its asymptotic properties.

4.1 The estimator

The estimation procedure goes through the following steps.

First step: Following most of the literature on deconvolution, given an i.i.d. sample of size N, we first estimate κ_Y and its derivatives by empirical analogs, replacing the mathematical expectations in (7), (8) and (9) by arithmetic means:

$$\widehat{\kappa}_Y(t) = \ln\left(\mathbb{E}_N\left[e^{it^TY}\right]\right), \tag{14}$$

$$\widehat{\partial_\ell \kappa_Y}(t) = i\frac{\mathbb{E}_N\left[Y_\ell e^{it^TY}\right]}{\mathbb{E}_N\left[e^{it^TY}\right]} = \partial_\ell \widehat{\kappa}_Y(t), \tag{15}$$

and

$$\widehat{\partial^2_{\ell m}\kappa_Y}(t) = -\frac{\mathbb{E}_N\left[Y_\ell Y_m e^{it^TY}\right]}{\mathbb{E}_N\left[e^{it^TY}\right]} + \frac{\mathbb{E}_N\left[Y_\ell e^{it^TY}\right]}{\mathbb{E}_N\left[e^{it^TY}\right]}\frac{\mathbb{E}_N\left[Y_m e^{it^TY}\right]}{\mathbb{E}_N\left[e^{it^TY}\right]} = \partial^2_{\ell m}\widehat{\kappa}_Y(t), \tag{16}$$

where $\mathbb{E}_N$ denotes the empirical expectation operator.

Second step: As the choice of t in $\mathcal{T}_k = \left\{t \in \mathbb{R}^L | t^T A_{[\cdot,k]} = 1\right\}$, along which to perform the integration yielding $\kappa_{X_k}(\tau_k)$, is arbitrary, one can estimate κ_{X_k} by averaging solution (13) over a distribution of points in $\mathcal{T}_k$, that is,

$$\begin{aligned}\widehat{\kappa}_{X_k}(\tau) &= \int_0^\tau \int_0^u \left(Q^-\right)_{[k,\cdot]} \left(\int \nabla^2 \widehat{\kappa}_Y(vt)\, dW(t)\right) dvdu \\ &= \int_0^\tau \int_0^u \left(Q^-\right)_{[k,\cdot]} \left(\sum_{j=1}^p w_j \nabla^2 \widehat{\kappa}_Y(vt_j)\right) dvdu, \end{aligned}\tag{17}$$

where $W = \sum_{j=1}^p w_j \delta_{t_j}$ is a discrete probability distribution on $\mathcal{T}_k$.

Third step: We then estimate the factor distribution functions by inverse Fourier transformation:

$$\begin{aligned}\widehat{f}_{X_k}(x) &= \frac{1}{2\pi}\int_{-T_N}^{T_N} \widehat{\varphi}_{X_k}(\tau)e^{-i\tau x}d\tau \\ &= \frac{1}{2\pi}\int_{-T_N}^{T_N} \exp\left[-i\tau x + \widehat{\kappa}_{X_k}(\tau)\right]d\tau,\end{aligned}\tag{18}$$

where the smoothing parameter T_N tends to infinity at a rate to be specified.

Note that one could use alternative approaches instead of this third deconvolution step. For instance, the "histogram-based" estimator proposed by Hall and Yao (2003) could be used for the purpose of estimating factor densities, given that their c.f.'s have been previously estimated by steps 1 and 2.

4.2 Choice of the weighting distribution W

Empirical characteristic functions are typically well estimated around the origin and badly estimated in the tails. When c.f.'s approach zero, the estimation of c.g.f.'s and their derivatives worsens rapidly. It thus makes sense to choose t such that $\nabla^2 \kappa_Y(\tau_k t)$ is well estimated on a maximal interval. A natural choice is to minimize the Euclidian norm of $\frac{t}{t^T A_{[.,k]}}$, which yields, by Cauchy-Schwartz inequality:

$$t^* = \left(A_{[\cdot,k]}\right)^{-T} = \frac{A_{[\cdot,k]}}{A_{[\cdot,k]}^T A_{[\cdot,k]}}. \tag{19}$$

The simulation section will provide evidence that choosing $W = \delta_{t^*}$ works well in practice.

5 Asymptotic properties

In this section, we study the asymptotic properties of our estimator. All mathematical proofs are in the appendix.

5.1 Consistency

We now proceed to show that $\widehat{f}_{X_k}$ is a uniformly consistent estimator of f_{X_k}, for all $k = 1, ..., K$, provided that the characteristic functions of factors and errors are everywhere nonvanishing.

Li and Vuong (1998) and Hall and Yao (2003) assume bounded supports.[7] However, as recently emphasized by Hu and Ridder (2005), the two assumptions of bounded support and nonvanishing characteristic functions are mutually exclusive. As the c.f.'s of most standard distributions are never zero, and as economic variables often have unbounded support, we relax the assumption of bounded support.

To prove the consistency of our estimators in the case of unbounded support, we first extend Hu and Ridder's Lemma 1 proving a uniform consistency result for empirical characteristic functions. To proceed, write $|t| = \max_\ell |t_\ell|$, for any vector $t \in \mathbb{R}^L$. We have the following lemma.

Lemma 4 *Let X be a scalar random variable and let Y be a vector of L scalar random variables. Let $Z = \left(X, Y^T\right)^T$. 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. (resp. mean) corresponding to a sample $\mathbf{Z}_N \equiv (Z_1, ..., Z_N)$ of N i.i.d. draws from F. Assume that*

[7] Horowitz and Markatou (1996) do not assume support boundedness. However, as pointed out by Hu and Ridder (2005), their proofs implicitely require this hypothesis.

$\mathbb{E}X^2 \leq M_1 < \infty$ and that $\mathbb{E}|Y|^i < \infty$ for all $i \in \{1, ..., L\}$. Define $f_t(x, y) = x\exp(it^T y)$ for $t \in \mathbb{R}^L$. Lastly, let $K_{|X|}(\varepsilon)$ be the positive and nonincreasing function implicitly defined by the equality:

$$\mathbb{E}\left[|X|\,\mathbf{1}\{|X| > K\}\right] = \int_K^\infty u f_{|X|}(u)\,du = \varepsilon.$$

Then, $\sup_{|t|\leq T_N} |\mathbb{E}_N f_t - \mathbb{E} f_t| = O(\varepsilon_N)$, a.s., for all ε_N, T_N such that $\ln T_N = O(\ln N)$ and $\frac{K_{|X|}(\varepsilon_N)}{\varepsilon_N} = o\left[\left(\frac{N}{\ln N}\right)^{\frac{1}{2}}\right]$.

Lemma 4 shows that the rate of convergence of the empirical mean of f_t depends on the tails of the distribution of X: the thicker the right tail of the distribution of $|X|$, the slower the rate of convergence. For example, if X is Gaussian:

$$\frac{1}{\sqrt{2\pi}\sigma}\int_K^\infty x e^{-\frac{x^2}{2\sigma^2}}dx = \frac{1}{\sqrt{2\pi}}\sigma e^{-\frac{K^2}{2\sigma^2}},$$

and $K_{|X|}(\varepsilon)$ tends to infinity when $\varepsilon \downarrow 0$ as $\sqrt{\ln(1/\varepsilon)}$. If 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, $K_{|X|}(\varepsilon)$ diverges as $(1/\varepsilon)^{\frac{1}{a-1}}$.

We now apply Lemma 4 to the first two derivatives of the c.f. of a vector of random variables Y: $\mathbb{E}\left(Y_\ell \exp(it^T Y)\right)$ and $\mathbb{E}\left(Y_\ell Y_m \exp(it^T Y)\right)$, for $\ell, m = 1, ..., L$. This allows us to prove the following uniform consistency result for the characteristic functions of factors.

Theorem 5 *Suppose that there exists an integrable, decreasing function $g_Y : \mathbb{R}^+ \to [0, 1]$, such that $|\varphi_Y(t)| \geq g_Y(|t|)$ as $|t| \to \infty$. Then, there exists $\varepsilon_N \downarrow 0$ and $T_N \to \infty$ such that*

$$\sup_{|\tau|\leq T_N} \left|\widehat{\varphi}_{X_k}(\tau) - \varphi_{X_k}(\tau)\right| = \frac{T_N^2}{g_Y(T_N)^3} O(\varepsilon_N) \quad a.s., \tag{20}$$

where ε_N is the minimum convergence rate satisfying the conditions of Lemma 4 for all functions f_t of the form $\exp\left(it^T Y\right)$, $Y_\ell \exp\left(it^T Y\right)$ and $Y_\ell Y_m \exp\left(it^T Y\right)$, $\ell, m \in \{1, ..., L\}$, and T_N satisfies two constraints: $\ln T_N = O(\ln N)$, and $\frac{T_N^2}{g_Y(T_N)^3}\varepsilon_N = o(1)$.

Theorem 5 shows that the estimator of factor characteristic functions converges to the true c.f's at a polynomial rate. The following theorem states that $\widehat{f}_{X_k}$ converges uniformly to f_{X_k} when the sample size tends to infinity.

Theorem 6 *Suppose that there exists an integrable, decreasing function $g_X : \mathbb{R}^+ \to [0, 1]$ such that $|\varphi_X(\tau)| \geq g_X(|\tau|)$ as $|\tau| \to \infty$. Suppose also that there exist K integrable* *functions $h_{X_k} : \mathbb{R}^+ \to [0,1]$ such that $h_{X_k}(|\tau|) \geq \left|\varphi_{X_k}(\tau)\right|$ as $|\tau| \to \infty$. Then, $\widehat{f}_{X_k}$ is a uniformly consistent estimator of the p.d.f. f_{X_k} of X_k, i.e.*

$$\sup_x \left|\widehat{f}_{X_k}(x) - f_{X_k}(x)\right| = \frac{T_N^3}{g_X(T_N)^3} O(\varepsilon_N) + O\left(\int_{T_N}^{+\infty} h_{X_k}(v)dv\right) = o(1) \quad a.s., \tag{21}$$

where ε_N and T_N are given by Theorem 5 applied to $g_Y(|t|) = g_X(L|A||t|)$, with $|A| = \max_{i,j}(|a_{ij}|)$.

Theorem 6 shows that the convergence rate of the factor density estimator depends on the shape of factor and measurement distributions in two different ways. Firstly, as emphasized by Lemma 4, the rate of convergence of derivatives of factor c.f.'s depends on the tails of factor distributions: the thicker the tails, the slower the rate. Secondly, the rate of convergence of factor densities depends on how fast the tails of factor c.f.'s decay to zero—which characterizes the smoothness of factor distributions and is controlled by functions g_X and h_{X_k}.

Interestingly, smoothness has an ambiguous effect on the convergence rate. This is apparent from the fact that one needs to bound factor c.f.'s from both sides. On the one hand, smooth distributions require less trimming. This effect is reflected in the second term on the right-hand side of equation (21): the same value of $\int_{T_N}^{+\infty} h_{X_k}(v)dv$ can be obtained with less trimming (higher T_N) if $h_{X_k}(v)$ decays faster to zero. On the other hand, it is more difficult to separate the different sources of information if factors are smooth. Looking at equation (20) in Theorem 5, one sees that a given convergence rate of factor characteristic functions can be achieved uniformly over a wider interval $[-T_N, T_N]$ if g_Y (and thus g_X) decays more slowly to zero. Lastly, the smoothness of the other factor distributions (the thickness of the tails of φ_{X_m}, for $m \neq k$) has an unambiguous effect on the convergence rate: the thinner the tails of other factors' c.f.'s, the lower the convergence rate.

5.2 Optimal convergence rates

We now examine how the preceding results can be used in practice to calculate optimal convergence rates. We consider one particular example which illustrates how the convergence rate varies with the thickness of the tails of factor distributions and their degree of smoothness.

The optimal rate of convergence is obtained by choosing an optimal trimming parameter T_N and an optimal rate ε_N so as to minimize:

$$\frac{T_N^3}{g_X(T_N)^3} O(\varepsilon_N) + \int_{T_N}^{+\infty} h_{X_k}(v)dv, \tag{22}$$

under the constraints: $\frac{K_{|X|}(\varepsilon_N)}{\varepsilon_N} = o\left[\left(\frac{N}{\ln N}\right)^{\frac{1}{2}}\right]$, $\ln T_N = O(\ln N)$ and $\frac{T_N^2}{g_X(L|A|T_N)^3}\varepsilon_N = o(1)$.

To make this procedure operational, one has to specify the tails of factor distributions (determine function $K_{|X|}$) and their degree of smoothness (determine g_X and h_{X_k}). We assume that factor distributions have Pareto tails. Since Fan's (1991) influential paper it has become standard to distinguish between *smooth* distributions (polynomial g_X and h_{X_k} functions) and *supersmooth* distributions (g_X and h_{X_k} decay at a faster, exponential, rate). Examples of smooth distributions are the uniform, gamma or Laplace distributions. The normal distribution is supersmooth. We here focus on the case where all factors are smooth.

The following corollary follows straightforwardly from Theorem 6.

Corollary 7 ***(Smooth, Pareto-tailed factors)*** *Assume that there exist (α_k, β_k), $1 < \beta_k \leq \alpha_k$, such that*

$$|\tau|^{-\alpha_k} \leq \left|\varphi_{X_k}(\tau)\right| \leq |\tau|^{-\beta_k}, \quad |\tau| \to \infty,$$

and $a > 1$ such that $K_{|X_k|}(\varepsilon) \leq (1/\varepsilon)^{\frac{1}{a-1}}$. Then for all $\gamma > 0$ small enough

$$\sup_x \left|\widehat{f}_{X_k}(x) - f_{X_k}(x)\right| = O\left(\left(\frac{\ln N}{N}\right)^{\frac{\beta_k - 1}{2+3\alpha+\beta_k}(1-1/a)(1/2-\gamma)}\right) \quad a.s., \tag{23}$$

for $\alpha = \sum_{k=1}^K \alpha_k$, and for a trimming parameter T_N in $\widehat{f}_{X_k}$ chosen such as

$$T_N = O\left(\left(\frac{N}{\ln N}\right)^{\frac{(1-1/a)(1/2-\gamma)}{2+3\alpha+\beta_k}}\right). \tag{24}$$

Three remarks are in order. First, the convergence rate is polynomial in $\frac{\ln N}{N}$ instead of $\frac{\ln \ln N}{N}$ as in Li and Vuong (1998). This is because we do not require factor distributions to have bounded support. Second, thicker tails (smaller a) require more trimming and yield lower convergence rates. The limit when a tends to infinity gives the convergence rate in the case of factor densities with thin tails (thinner than polynomial), like the normal.

Third, the rate of convergence in (23) increases with β_k and decreases with α. This reflects the ambiguous effect of the degree of smoothness on the convergence rate. To see why, hold α_m, for $m \neq k$, constant. Then the rate increases with β_k and decreases with α_k. To determine the net effect of smoothness in this case, it is necessary to tie α_k to β_k. This can be done by fixing a priori the exact degree of smoothness, i.e. $\alpha_k = \beta_k$. Then, the optimal convergence rate unambiguously increases with β_k and less trimming is necessary to achieve this rate.

Furthermore, the optimal uniform rate of convergence of $\widehat{f}_{X_k}(x)$ unambiguously decreases with $\alpha - \alpha_k = \sum_{m \neq k} \alpha_m$. Hence it is more difficult to identify and estimate the distribution of one factor if the other factors and errors are smoother. When the other factors are not only smoother than a smooth factor X_k, but supersmooth, then, applying the same calculation techniques as in Corollary 7 shows that the rate of convergence of $\widehat{f}_{X_k}(x)$ becomes logarithmic.[8]

5.3 Practical choice of the smoothing parameter T_N

It is tempting to use (24) as a guideline to choose T_N in practice. However, our experiments suggest that doing so one underestimates T_N. The reason could be that this "optimal" T_N maximizes an upper bound for the convergence rate, which can be very conservative. Moreover, in finite samples, the upper bound of the asymptotic rate may be severely inadequate. We obtained much better results by using the following simple method due to Diggle and Hall (1993). Refining the bound further is an interesting issue that we leave for future research.

Diggle and Hall consider the problem of computing T_N in the context of a deconvolution problem, $Y = X + U$, with independent random samples for Y and U. Estimating the density of X as

$$\widehat{f}_X(x) = \frac{1}{2\pi} \int_{-T_N}^{T_N} e^{-itx} \frac{\widehat{\varphi}_Y(t)}{\widehat{\varphi}_U(t)} dt, \tag{25}$$

where $\widehat{\varphi}_Y(t)$ and $\widehat{\varphi}_U(t)$ are the empirical c.f. of Y and U, and maximizing the asymptotic Mean Integrated Squared Error $\int |\widehat{f}_X(x) - f_X(x)|^2 dx$ with respect to T_N, the optimal trimming parameter has to satisfy:

$$\varphi_Y(T_N) = N^{-1/2}. \tag{26}$$

For given t, $|\widehat{\varphi}_Y(t) - \varphi_Y(t)|$ is also of order $N^{-1/2}$. One thus cannot directly replace the unknown c.f. in (26) by its estimated counterpart. Assuming that $|\varphi_Y(t)| = \alpha\, |t|^{-\beta}$ for large enough $|t|$, Diggle and Hall then propose to proceed in two steps. First, estimate α and β by a linear regression of $\ln |\widehat{\varphi}_Y(t)|$ on $\ln |t|$ on an interval where this relationship is approximately linear. Then, substitute $\widehat{\alpha} T_N^{-\widehat{\beta}}$ for $\varphi_Y(T_N)$ into (26) to get T_N.

In addition, Diggle and Hall recommend to perform the integration in (25) by multiplying the argument of the integral by a "damping factor" that aims at reducing the oscillations that often characterize estimated c.f.'s in their tails. That is, estimate the

[8] See Caroll and Hall (1988) and Horowitz and Markatou (1996) for deconvolution estimators with logarithmic rates of convergence.

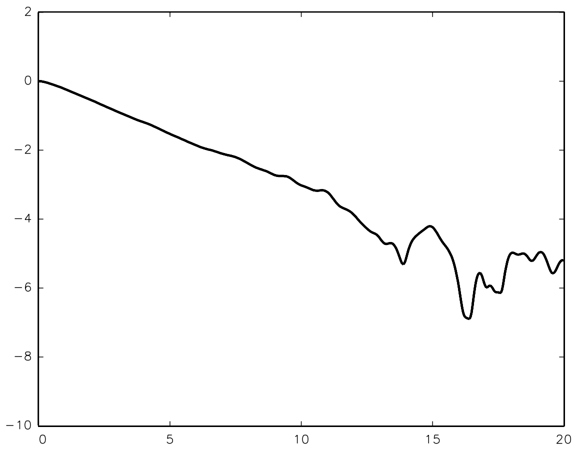


Figure 1: Empirical c.f. of year-to-year wage growth residuals (PSID, 1978-1987) — Plot of $Y = \ln |\widehat{\varphi}_Y(t)|$ against $X = |t|$.

density as

$$\widetilde{f}_X(x) = \frac{1}{2\pi} \int_{-T_N}^{T_N} d(t) e^{-itx} \frac{\widehat{\varphi}_Y(t)}{\widehat{\varphi}_U(t)} dt,$$

where $d(t)$ is the damping function. In practice, we use:

$$d(t) = \mathbf{1}\left\{|t| < (1-\mu)T_N\right\} + \mathbf{1}\left\{(1-\mu)T_N \le |t| \le T_N\right\} \cdot \frac{1}{\mu}\left[1 - |t|/T_N\right].$$

Varying the scalar $\mu \in [0,1]$, we can reduce the magnitude of the oscillations in the tails.

In the case of model $Y = AX$, we suggest to proceed analogously. For $k \in \{1, ..., K\}$, let $t^* = A^{-T}_{[.,k]} = \frac{A_{[.,k]}}{A^T_{[.,k]} A_{[.,k]}}$. Then

$$t^{*T} Y = t^{*T} AX = X_k + \sum_{m \neq k} t^{*T} A_{[.,m]} X_m. \qquad (27)$$

We treat the d.f. of $\sum_{m \neq k} t^{*T} A_{[.,m]} X_m$ in (27) as if it were known. In this case, the problem of estimating the density of factor X_k boils down to a classical deconvolution problem, and the approach in Diggle and Hall (1993) can be applied.

In practice, characteristic functions can have tails decaying at a faster rate than polynomial. An illustration is provided by Figure 1, which plots the (real part of the) empirical c.g.f. (the log-modulus of the c.f.) of the wage data considered in Section 7. The c.g.f. is approximately linear in $|t|$ over a wide range, and becomes more erratic afterwards. It is straightforward to extend Diggle and Hall's method to such cases. In the Monte Carlo simulations that we have done, this simple approach provided a reasonable guide for choosing T_N.

6 Monte-Carlo simulations

In this section, we study the finite-sample behavior of our density estimators.

6.1 The measurement error model

We first consider the estimation of the characteristic function of the unique factor in the measurement error model:

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

where $(X_1, U_1, U_2) \sim \mathcal{N}(0, I_3)$.

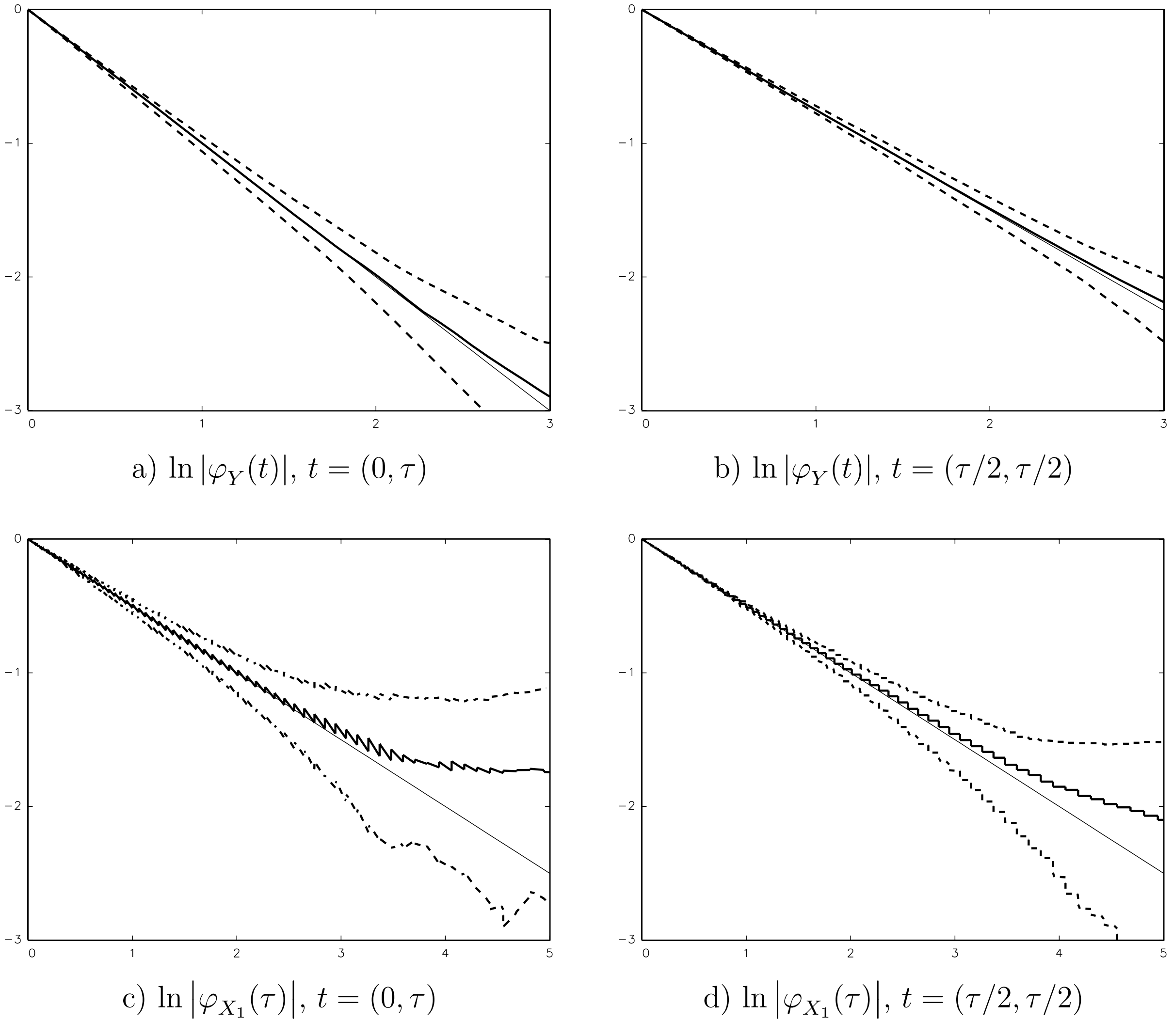
Figure 2 presents Monte Carlo simulations for estimates of the c.f.'s of Y and X_1. For each simulation in this section, we draw 100 independent realizations of Y. The thin line is the true factor characteristic function and the thick line is the pointwise median of the 100 estimates. The dashed lines correspond to the pointwise first and ninth deciles of the Monte Carlo distributions of estimates.

In panels a) and b), we plot the (real part of the) c.g.f. of Y, i.e. $\ln|\widehat{\varphi}_Y(t)| = \ln\left|\mathbb{E}_N\left[e^{it^TY}\right]\right|$, evaluated at $(t_1, t_2) = (0, \tau)$ (panel a) and $(t_1, t_2) = (\tau/2, \tau/2)$ (panel b), where $\tau \in \mathbb{R}^+$. We set the scale on the x-axis equal to τ^2. The c.f. of the standard normal distribution being $\exp(-t^2/2)$, the true value of the c.f. is then a straight line with slope -1 in panel a), and $-3/4$ in panel b).

We note that the c.f. of measurement variables is well estimated over a wide range and that the precision of the estimates is worse at higher frequencies. Note also that the bias is smaller and the precision is higher for $\ln|\widehat{\varphi}_Y(\tau/2, \tau/2)|$ than for $\ln|\widehat{\varphi}_Y(0, \tau)|$.

In panel c), we report estimates of the c.f. of X_1 obtained by Li and Vuong's method; that is: integrating along the direction $(0, \tau)$, for $\tau \in \mathbb{R}^+$. Panel d) shows the results of our preferred method, that is: integrating along the direction of $(\tau/2, \tau/2)$.

We draw two conclusions from this figure. First, the bias is larger and the precision of the estimation of the c.f. of X_1 is much lower than that of the c.f. of Y. A plausible reason is that small errors in the estimation of the tail of a c.f. translate into large errors for its derivative. Second, the c.f. of X_1 is better estimated by the second method, i.e. using as direction of integration the vector of minimal Euclidian norm. This observation is in line with the discussion in 5.3.



a) $\ln|\varphi_Y(t)|$, $t = (0, \tau)$

b) $\ln|\varphi_Y(t)|$, $t = (\tau/2, \tau/2)$

c) $\ln\left|\varphi_{X_1}(\tau)\right|$, $t = (0, \tau)$

d) $\ln\left|\varphi_{X_1}(\tau)\right|$, $t = (\tau/2, \tau/2)$

Figure 2: Monte Carlo simulations for the estimated characteristic functions in the measurement error model

6.2 The 3-measurements, 3-factors, noisy case

Let us now consider the case of a linear factor model with three measurements, three factors and three errors.[9] The DGP is: $Y = \Lambda X + U$, with $Y \in \mathbb{R}^3$ $(L = 3)$, $X \in \mathbb{R}^3$ is the vector of common factors and $U \in \mathbb{R}^3$ is the vector of errors. We set

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

and assume that all factors follow the same distribution, and likewise for all errors. For reasons of symmetry, we shall present the estimation results for the first factor and the first error component only. The sample size is $N = 1000$.

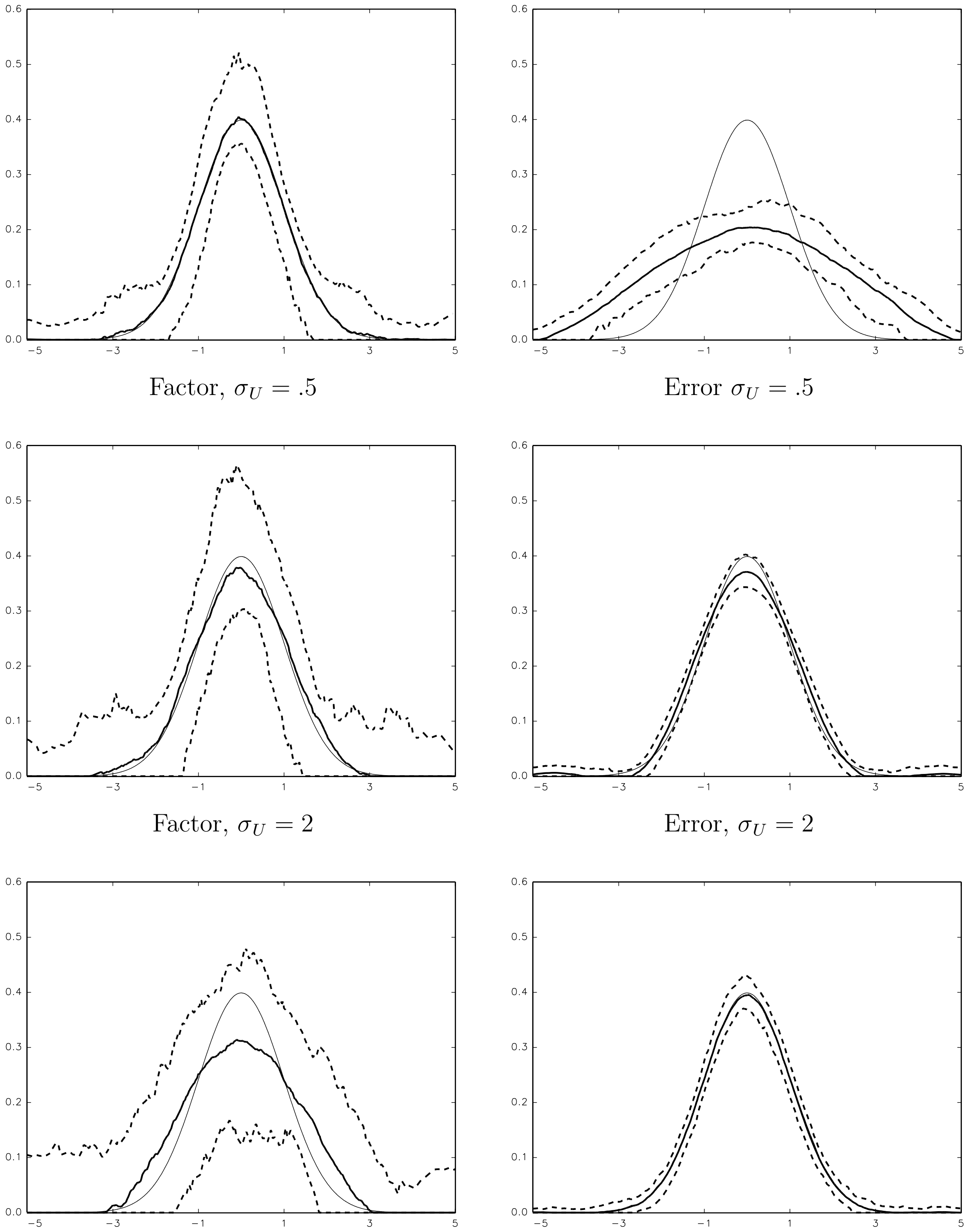
The normal-supersmooth case. We first study the case where factors and errors are normally distributed. Factors have unit variance and we run three different simulations with error standard deviation σ_U equal to .5, 2 and 4. Figure 3 presents the simulation results. In this figure and all figures in the current subsection, the left panels corresponds to the first factor X_1, and the right panels to the first error U_1. The thin line is the true factor (resp. error) distribution and the thick line is the pointwise median of 100 estimates. As before, the dashed lines correspond to the pointwise first and ninth deciles of the Monte Carlo distributions of estimates.

It is clear that the estimation of factor p.d.f.'s deteriorates as the signal gets more noisy. For moderate error variances there is only a small finite sample bias and the precision is fairly good. Note that, as far as each factor is concerned the two other factors are an additional source of noise. So, even with σ_U equal to zero, identifying three source distributions from three linear mixtures with such a precision is already very satisfactory.

When error variances get larger, confidence bands for factor distributions become wider. Simultaneously, it becomes easier to identify the error distributions.

Smooth factors and smooth or supersmooth errors. In the two top-panels of Figure 4, factors and errors follow a centered double exponential or Laplace distribution. The density of the standard Laplace distribution is $f_X(x) = 1/\sqrt{2}\exp(-\sqrt{2}|x|)$ and its characteristic function is $\varphi_X(t) = 1/(1 + \frac{1}{2}t^2)$. It is an example of smooth distribution. It is also symmetric and leptokurtic (the kurtosis of the standard Laplace distribution is

[9] Note that the discussion following Theorem 3 shows that $L(L+1)/2$ is the maximum number of factors/errors identifiable in a linear factor model. So $K = 3$ (one factor and two errors) is the maximal number when $L = 2$, and $K = 6$ (three factors and three errors) is the maximal number when $L = 3$.



Factor, $\sigma_U = .5$

Error $\sigma_U = .5$

Factor, $\sigma_U = 2$

Error, $\sigma_U = 2$

Factor, $\sigma_U = 4$

Error, $\sigma_U = 4$

Figure 3: Monte Carlo simulations for density estimates in the linear factor model with 3 measurements, 3 factors and 3 errors — normal distributions

equal to 6). We set the variance of factors equal to one and the standard deviation of errors equal to 2.

Figure 4 shows that in this case the general shape of the double exponential is well reproduced, albeit not perfectly. The estimator shows some incapacity to capture the right peakedness and the singularity at zero.

In the two bottom-panels of Figure 4, factors are still Laplace, and errors are normally distributed with a standard deviation of 2. In this case, factors are smooth and errors are supersmooth. Therefore, we expect the convergence rate of our estimator to be extremely low (see section 5). Nonetheless, there is no striking difference between the two figures in the top-left and bottom-left panels, although the confidence band is wider in the latter case. Hence, at least in this particular case, the deconvolution problem does not seem to be rendered much more difficult by the presence of a smoother error distribution.

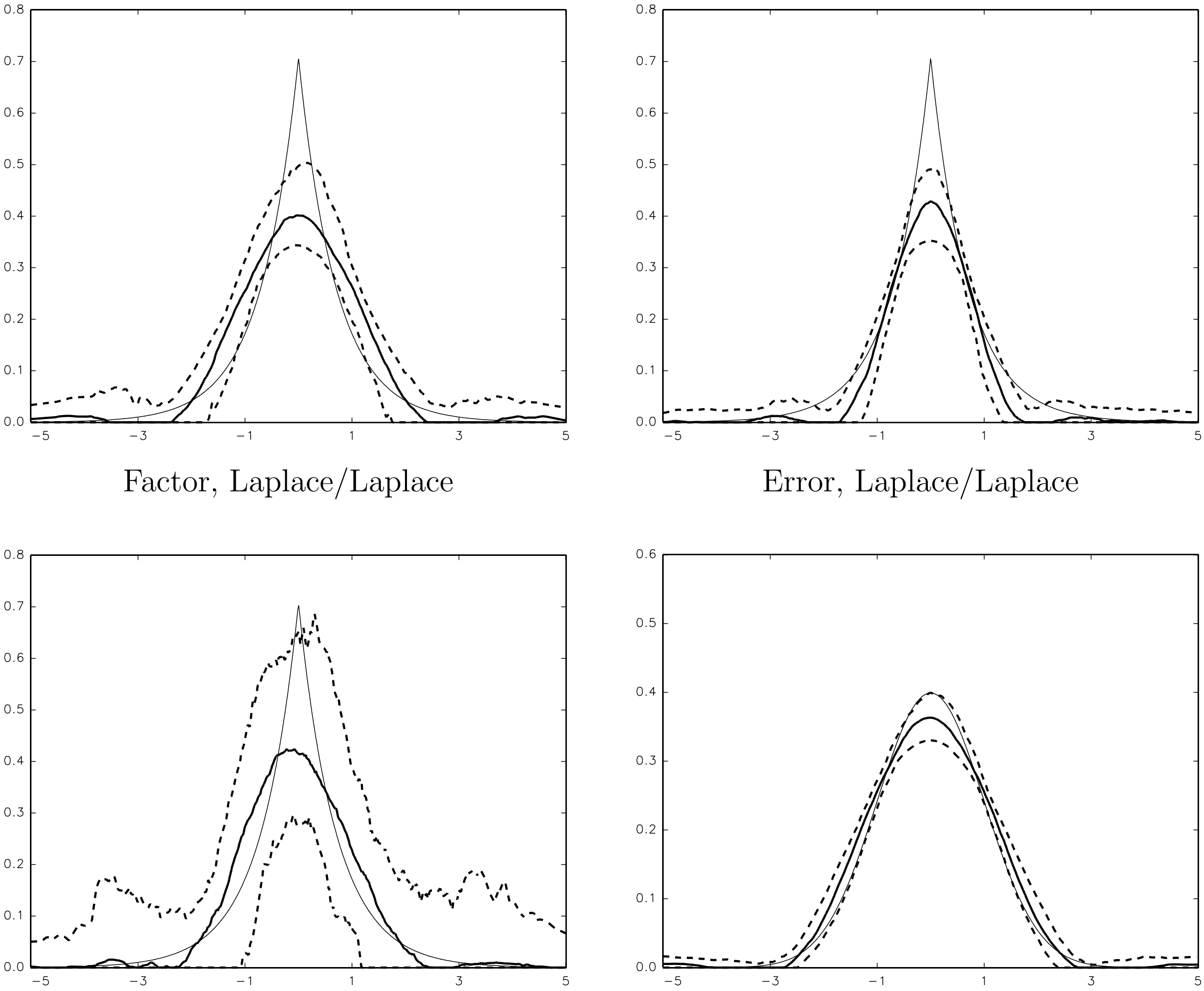
Skewness and kurtosis. In the last part of this section, we study the ability of our estimator to deal with skewed and/or leptokurtic distributions. In all remaining simulations, factors and errors follow the same distribution up to scale, and σ_U is 2.

In the four panels of Figure 5, factors follow a Gamma distribution, with parameters $(5,1)$ and $(2,1)$, respectively. For these values of the parameters, factor skewness is $2/\sqrt{5} \approx .89$ and $\sqrt{2}$, respectively, and kurtosis excess is equal to 1.2 and 3. The results suggest that our estimator captures skewness reasonably well. However, the estimation is less precise if the distribution is less symmetric (the second row of Figure 5).

Next, we turn to kurtosis. In the two upper panel of Figure 6, factors are mixtures of independent normals.[10] The kurtosis excess of factor densities is set equal to 100. As in the case of the Laplace distribution, the estimator does not manage to capture the right amount of peakedness. The density at the mode is underestimated and precision is low.

Lastly, we report in the two lower panels of Figure 6 the simulation results for log-normally distributed factors. Their skewness and kurtosis excess are approximately 6.2 and 110, respectively. It is clear from the figure that factor densities are badly estimated in this case. The estimated left tail is particularly out of range. This illustrates the difficulty of identifying support bounds from finite samples.

[10] More precisely, we 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 $\kappa_4(\rho) = 3\rho/(4(1-\rho))$.



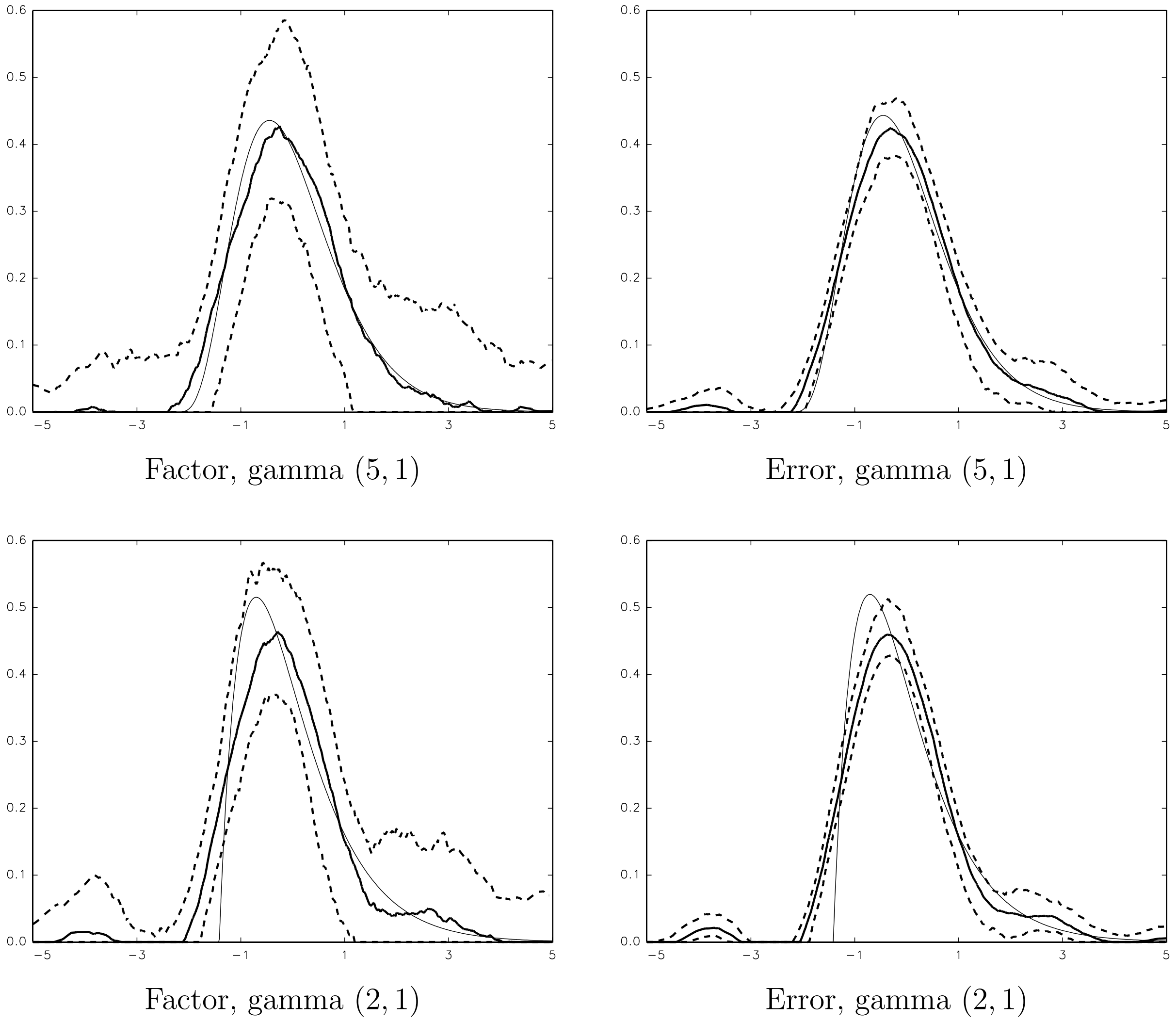
Factor, Laplace/Laplace

Error, Laplace/Laplace

Factor, Laplace/normal

Error, Laplace/normal

Figure 4: Monte Carlo simulations for density estimates in the linear factor model with 3 measurements, 3 factors and 3 errors — Laplace and normal distributions



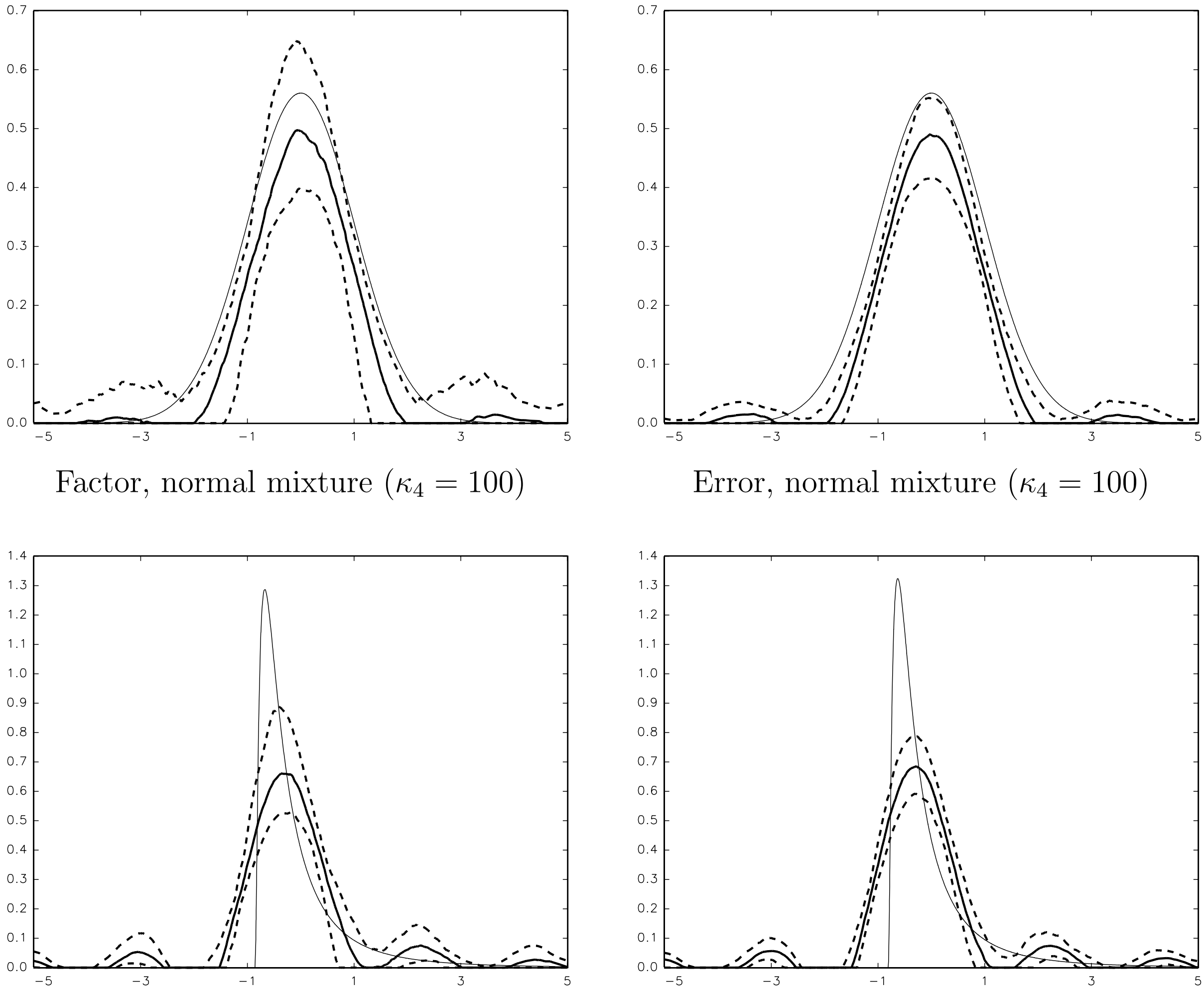
Factor, gamma $(5, 1)$

Error, gamma $(5, 1)$

Factor, gamma $(2, 1)$

Error, gamma $(2, 1)$

Figure 5: Monte Carlo simulations for density estimates in the linear factor model with 3 measurements, 3 factors and 3 errors — Gamma distributions



Factor, normal mixture ($\kappa_4 = 100$)

Error, normal mixture ($\kappa_4 = 100$)

Factor, log-normal

Error, log-normal

Figure 6: Monte Carlo simulations for density estimates in the linear factor model with 3 measurements, 3 factors and 3 errors — normal mixtures and log-normal distributions

Job changes	All	0	1/2	3+
Age	37.4	39.6	37.4	36.2
High school dropout	.21	.23	.23	.17
High school graduate	.54	.59	.52	.53
Hours	2183	2186	2193	2172
Married	.85	.84	.84	.86
White	.69	.63	.69	.74
Children	1.6	1.6	1.6	1.6
Family size	3.68	3.80	3.69	3.60
Family income ($1000)	33.7	33.5	33.6	33.9
North east	.15	.15	.13	.17
North central	.27	.29	.25	.27
South	.43	.44	.51	.36
SMSA	.59	.60	.55	.61
Number	659	152	240	267

Table 1: Means of variables

7 Application to earnings dynamics

In this section, we apply our methodology to estimate the distributions of shocks in a simple model of earnings dynamics.

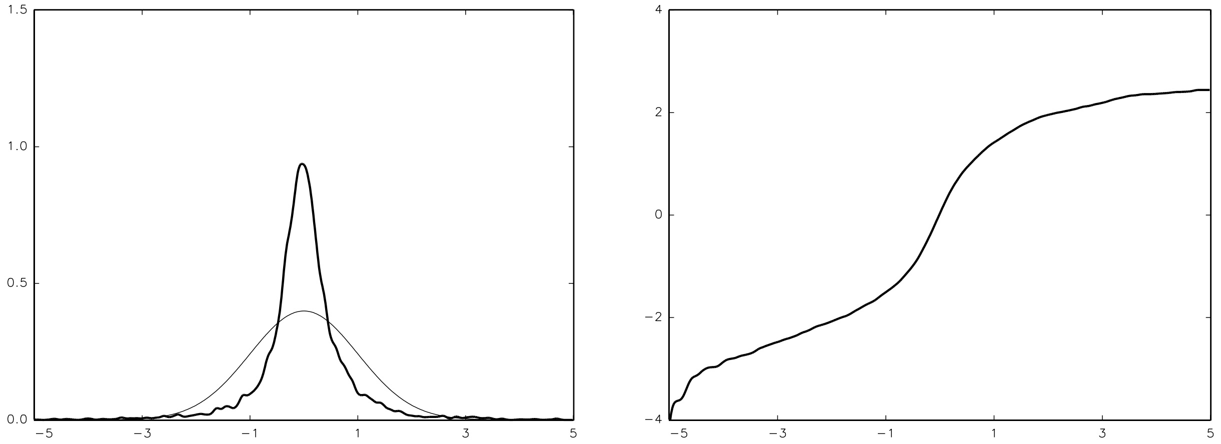
The data. We use PSID data, between 1978 to 1987. We select employed male workers who have non missing wage observations on the full period. We thus obtain a balanced panel of 659 individuals that we follow over 10 years. Descriptive statistics are presented in the first column of Table 1.

Let w_{it} denote the logarithm of annual wages, and let x_{it} be a vector of regressors, namely: education dummies, a quadratic polynomial in age, a race dummy, geographic indicators and year dummies. We shall focus on wage growth residuals, defined as the residuals of the OLS regression of $\Delta w_{it} = w_{it} - w_{it-1}$ on $\Delta x_{it} = x_{it} - x_{it-1}$. We denote these residuals as Δy_{it}. We shall also consider moving sums of wage growth residuals, defined as $\Delta_s y_{it} = y_{it} - y_{i,t-s} = \sum_{k=1}^{s} \Delta y_{i,t-k+1}$, for $s = 1, 2, ...$ Table 2 shows the marginal moments of these variables, as well as their first three autocorrelation coefficients. Focusing on the first row, we see that the variance of $\Delta_s y_{it}$ increases with s. This shows that wage differences between two points in time are more dispersed the longer the lag.

Another feature of Table 1 is the high kurtosis of wage growth residuals. Figure 7 confirms that the distribution of Δy_{it} is very different from the normal. On panel a), the thick line represents a kernel estimate of the density (standardized in order to have mean zero and variance one), and the thin one is the standard normal density. An alternative way of presenting the evidence of non-normality is to draw the normal probability plot

	Δy_{it}	$\Delta_2 y_{it}$	$\Delta_3 y_{it}$	$\Delta_T y_{it}$
Variance	.0915	.1206	.1374	.1804
Skewness	-.244	-.440	-.377	.059
Kurtosis	24.2	22.9	19.7	9.43
Autocorrelation 1	-.354	.171	.302	-
Autocorrelation 2	-.048	-.377	.057	-
Autocorrelation 3	-.019	-.065	-.404	-

Table 2: Moments of log wage differences



a) Density

b) Normal probability plot

Figure 7: Non normality of wage growth residuals

of Δy_{it}. If the data are normally distributed, then $\Phi^{-1}(F_N(\Delta y_{it}))$ is a straight line up to sampling error.[11] Panel b) in Figure 7 shows that this is not the case, as the c.d.f. of Δy_{it} has fatter tails than the normal. This evidence is similar to the one in Horowitz and Markatou (1996), who use data from the Current Population Survey.

The model. We consider the following model:

$$\begin{aligned} \Delta y_{it} &= \Delta p_{it} + \Delta r_{it}, \\ &= \varepsilon_{it} + r_{it} - r_{it-1}, \quad i = 1...N, \quad t = 2...T, \end{aligned} \tag{28}$$

where p_{it} follows a random walk: $p_{it} = p_{it-1} + \varepsilon_{it}$, where ε_{it} and r_{it} are white noise innovations with variances σ_ε^2 and σ_r^2. We shall refer to p_{it} as the permanent wage component and to r_{it} as the transitory component.

Permanent-transitory decompositions are very popular in the earnings dynamics literature, see among others Lillard and Willis (1978) and Abowd and Card (1989). There is a growing concern that the distributions of wage shocks might be non normal (see e.g.

[11] Recall that $F_N(\Delta y_{it})$ denotes the empirical c.d.f. of Δy_{it}.

Geweke and Keane, 2000). To assess this issue, Horowitz and Markatou (1996) estimate a model with an individual fixed effect and a transitory i.i.d. shock. There is no permanent shock in their model. Their estimation procedure is fully nonparametric. However, one particular implication of their model is that Δy_{it}, $\Delta_2 y_{it}$, ... are identically distributed. This is clearly at odds with the evidence presented in Table 2. However, the introduction of a permanent component easily permits to capture the increase in $\text{Var}(\Delta_s y_{it})$ when s increases.[12] The generalized deconvolution technique of this paper allows to conduct the same fully nonparametric analysis as in Horowitz and Markatou (1996) while allowing for a permanent component in wages.

We estimate σ_ε^2 and σ_r^2 in (28) by Equally Weighted Minimum Distance. Then, as the first and last permanent/transitory shocks are not separately identified, we treat $\varepsilon_{i2} - r_{i1}$ and $\varepsilon_{iT} + r_{iT}$ as additional factors. We end up with $K = 2T - 3$ factors. We estimate the c.g.f. of each ε_{it}, for $t = 3, ..., T-1$, and then use the average c.g.f. to obtain the density estimate by deconvolution. We proceed identically for each r_{it}, $t = 2, ..., T-1$.[13]

Estimation results. We estimate that the variance of permanent shocks is $\sigma_\varepsilon^2 = .02561$, and the transitory variance is $\sigma_r^2 = .03612$, with standard errors of .00497 and .00451, respectively.[14] According to these estimates, permanent shocks account for 26% of the total variance of wage growth residuals.

Figure 8 presents the density estimates. The permanent and temporary components are shown in panels a) and b), respectively. In each panel, the thick solid line represents the density of the shock, standardized to have unit variance, and the thin solid line represents the standard normal density, that we draw for comparison. The dashed lines delimit the bootstrapped 10%-90% confidence band.

Figure 8 shows that none of the two distributions is Gaussian. Both permanent and transitory shocks appear strongly leptokurtic. In particular, they have high modes and fatter tails than the normal. Moreover, the transitory part seems to have higher kurtosis than the permanent component. Consistently with the Monte Carlo simulations, the estimation is somewhat less precise in this case, especially in the left tail. Lastly, both

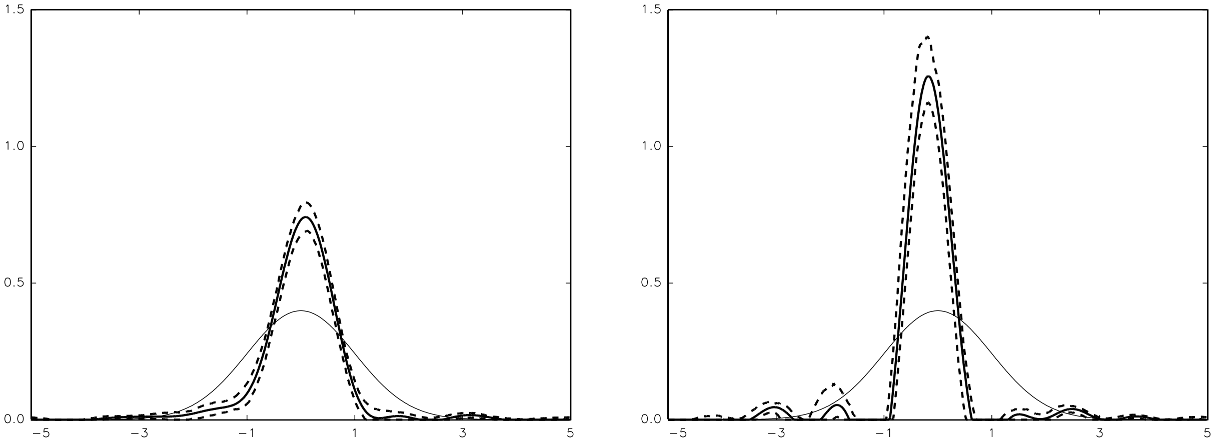
[12] Notice that model (28) implies that

$$\text{Var}\left(\Delta_s y_{it}\right) - \text{Var}\left(\Delta y_{it}\right) = (s-1)\,\sigma_\varepsilon^2.$$

The marginal distributions of Δy_{it} and $\Delta_2 y_{it}$ thus contain all the necessary information to identify σ_ε^2 and σ_η^2.

[13] Note that we do not use all the restrictions implied by stationarity. For instance, the distributions of the additional shocks at $t = 2$ and $t = T$ are convolutions of the permanent and stationary densities.

[14] Standard errors were computed by 1000 iterations of individual block bootstrap.



a) Permanent shock

b) Transitory shock

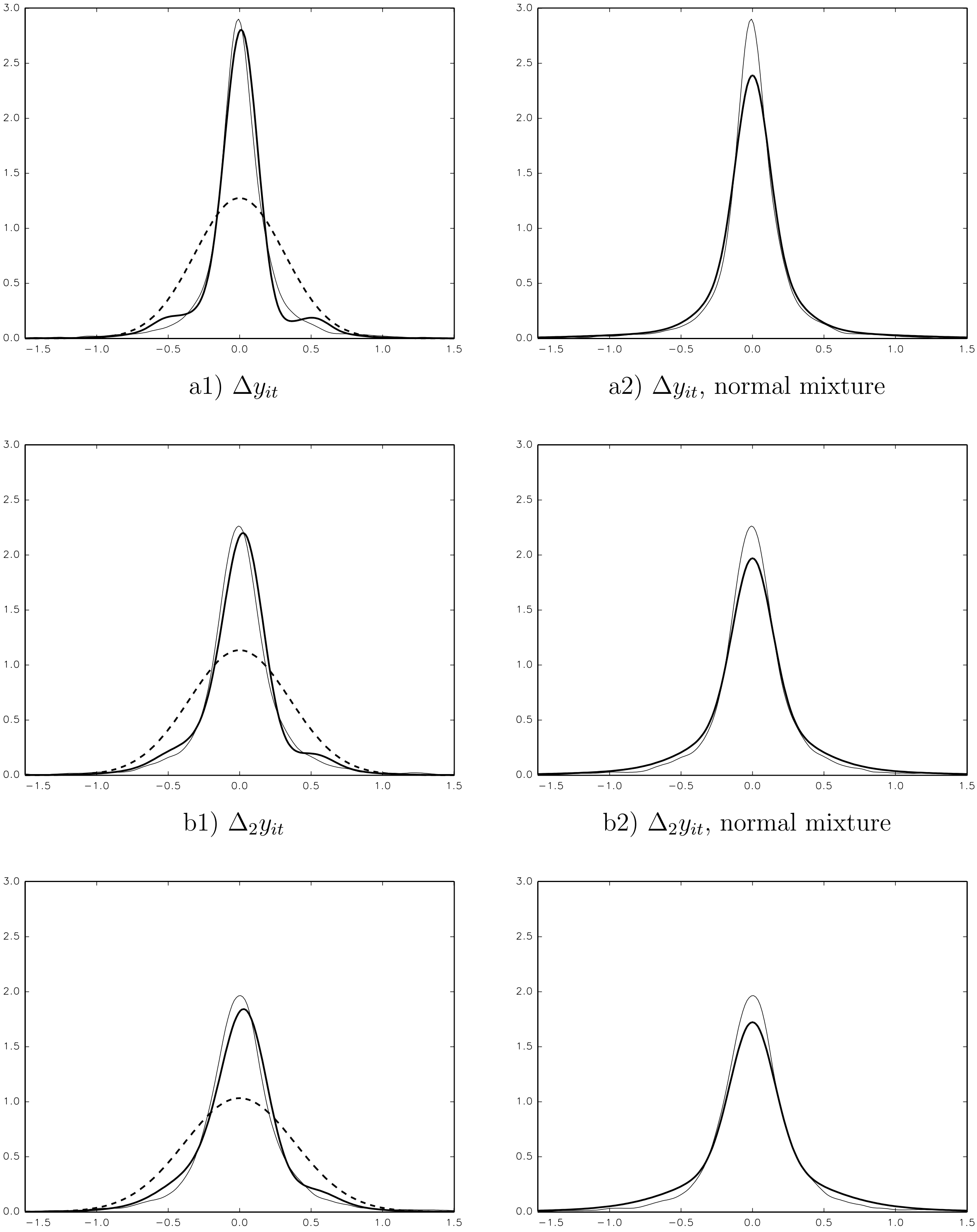
Figure 8: Nonparametric estimates of the densities of standardized permanent and transitory shocks (normal case: thin line; nonparametric estimate: thick line; bootstrapped 10%-90% confidence band: dashed lines).

densities are approximately symmetric.

Fit. Figure 9 compares our density estimates for $\Delta_s y_{it}$, $s = 1, 2, 3$, to actual p.d.f.'s. In panels a1) to c1), the thin line is a kernel estimator of the actual distribution's density. The thick line is the predicted density. The dashed line shows the density that is predicted under the assumption of normal innovations. The predicted densities of $\Delta_s y_{it}$, $s = 1, 2, 3$, where calculated analytically by convolution of the estimated densities of ε_{it} and r_{it}.

Figure 9 shows that our specification reproduces two features apparent in Table 2: the high kurtosis of wage growth residuals, and the decreasing kurtosis when the time lag increases. Note that the high mode of the density is remarkably well captured by our nonparametric method, even in the case of $\Delta_3 y_{it}$. In contrast, the normal specification gives a rather poor fit. However, the density tails are more imprecisely estimated. A likely reason for this is the imprecise estimation of the tails of the highly leptokurtic transitory shock (see Figure 8, panel b)).

We then present in Table 3 the moments of wage growth residuals, as in the data and as predicted under normality and nonparametrically. We see that the bad estimation of the tails of factor densities has strong consequences on moment estimates. In particular, variances are severely underestimated. Moreover, the estimated kurtosis is greater than three (normal case) but very far from the kurtosis of the distribution to be fitted (around 20). Overall, our method captures the shapes of factor distributions well, but fails at fitting the tails, which produces a severe underestimation of higher moments.



a1) Δy_{it}

a2) Δy_{it}, normal mixture

b1) $\Delta_2 y_{it}$

b2) $\Delta_2 y_{it}$, normal mixture

c1) $\Delta_3 y_{it}$

c2) $\Delta_3 y_{it}$, normal mixture

Figure 9: Fit of the model, densities of wage growth residuals (data: thin line; nonparametric/normal mixture estimate: thick line; normal estimate: dashed line).

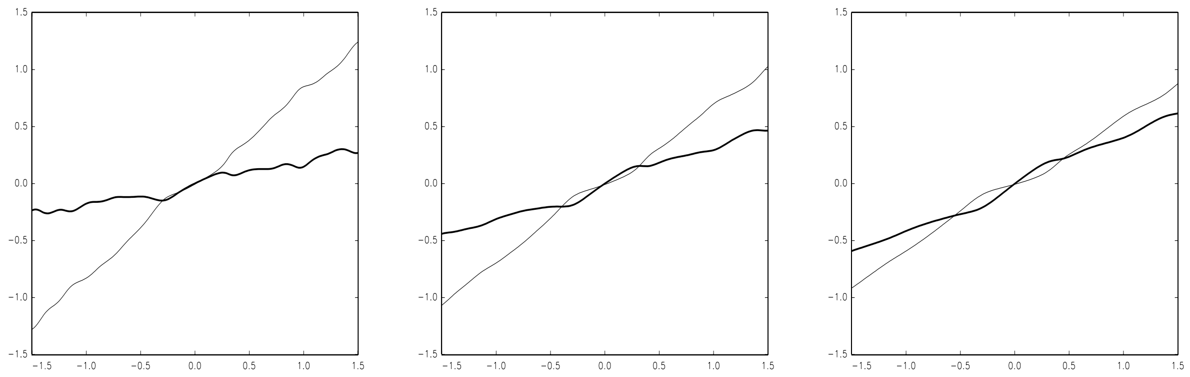
	Δy_{it}	$\Delta^2 y_{it}$	$\Delta^3 y_{it}$
		Data	
Variance	.0915	.1206	.1374
Skewness	-.244	-.440	-.377
Kurtosis	24.2	22.9	19.7
		Predicted, nonparametric	
Variance	.0579	.0734	.0892
Skewness	-.031	-.0034	-.0093
Kurtosis	6.46	5.46	4.87
		Predicted, normal	
Variance	.0978	.1235	.1492
Skewness	0	0	0
Kurtosis	3	3	3
		Predicted, normal mixture	
Variance	.1007	.1275	.1543
Skewness	0	0	0
Kurtosis	11.5	8.66	7.28

Table 3: Fit of the model, moments of wage growth residuals

To fit the moments better, we use our nonparametric estimates as a guide to find a convenient parametric form for factor densities. Figure 8 suggests that a mixture of two normals centered at zero may work well in practice. We thus estimate model (28) under this parametric specification for both ε_{it} and r_{it}. Parameters are estimated by Maximum Likelihood, using the EM algorithm of Dempster, Laird and Rubin (1977). Panels a2) to c2) in Figure 9 show the fit of the model. The high modes of the densities are slightly less precisely estimated. However, the tails seem better approximated. This visual impression is confirmed by the last three rows of Table 3, showing the first three predicted moments. The normal mixture specification yields much better estimates of variance and kurtosis. Notice that the normal mixture model was already used by Geweke and Keane (2000) to model earnings dynamics. Our results thus strongly support this modelling choice.

Wage mobility. We then use the model to weight the respective influence of permanent and transitory shocks in wage mobility. To this end, we compute the conditional expectations of the permanent and transitory components of $\Delta_s y_{it}$, $s = 1, 2, 3$: $\mathbb{E}(\sum_{r=0}^{s-1} \varepsilon_{it-r}|\Delta_s y_{it})$ and $\mathbb{E}(r_{it} - r_{it-s}|\Delta_s y_{it})$.

To do so, we first compute the conditional distribution of permanent and transitory shocks using Bayes rule. For instance the conditional density of the permanent shock



a) Δy_{it} b) $\Delta_2 y_{it}$ c) $\Delta_3 y_{it}$

Figure 10: Conditional expectation of shocks given wage growth residuals (permanent: thick line; transitory: thin line)

given wage observations is given by:

$$f(\varepsilon|\Delta y) = \frac{h(\varepsilon)f(\Delta y|\varepsilon)}{\int h(\widetilde{\varepsilon})f(\Delta y|\widetilde{\varepsilon})d\widetilde{\varepsilon}} = \frac{h(\varepsilon)\int g(r)g(\Delta y - \varepsilon - r)dr}{\int h(\widetilde{\varepsilon})\int g(r)g(\Delta y - \widetilde{\varepsilon} - r)drd\widetilde{\varepsilon}},$$

where h is the p.d.f. of ε and g is the p.d.f. of r. We proceed similarly for transitory shocks $r_{it} - r_{it-1}$.

Figure 10 plots these conditional expectations. We verify that the volatility of earnings is more likely to have a permanent origin if s is large. In panel a), we see for example that a log wage growth of ±1.5 has a transitory origin for more than ±1 and a permanent origin for less that ±0.5. In panel c), we see that a change $\Delta_3 y_{it}$ of ±1.5 is almost as likely to be transitory as permanent.

Job changes. Finally, we address the issue of the link between the degree of permanence of wage shocks and job-to-job mobility. It is notoriously difficult to identify job changes precisely in the PSID (see Brown and Light, 1992), so we tend to think of this exercise as tentative. We adopt the simplest criteria to identify job changes, setting the job change dummy equal to one if tenure is less than 12 months.[15] We then classify individuals into job stayers (no job change during the period), infrequent job changers (one or two job changes) and frequent job changers (more than three job changes). The last three columns of Table 1 give descriptive statistics for these three groups of individuals.

Then we compute the densities of permanent and transitory shocks given wage growth separately for each category of job changers by averaging within each group the con-

[15] Note that there were two "tenure" variables before 1987 in the PSID: time in position and time with employer. We take the former as our definition of tenure.

Job changes	0	1/2	3+
		Δy_{it}	
total	.04048	.05334	.10929
permanent	.01244	.01400	.02195
transitory	.02804	.03934	.08734
		$\Delta^2 y_{it}$	
total	.04793	.07045	.13705
permanent	.02239	.02858	.04702
transitory	.02554	.04187	.09003
		$\Delta^3 y_{it}$	
total	.06140	.08155	.15740
permanent	.03310	.04170	.07253
transitory	.02830	.03985	.08487

Table 4: Variances of the shocks by categories of job changers

ditional densities that we have already calculated. Table 4 presents the variances of permanent and transitory shocks for each mobility group. Focusing on the first three rows we see that wage volatility, as measured by the variance, is higher for frequent job changers. Moreover, these individuals are more likely to experience both permanent and transitory wage changes. The transitory variance is about 40% higher for infrequent job movers than for job stayers (.039 versus .028), and about three times higher for frequent job movers (.087). At the same time, the permanent variance is about 10% higher for infrequent job movers than for job stayers (.014 versus .012), and about 70% higher for frequent job movers (.022). As permanent shocks accumulate over time while transitory shocks do not, the difference in wage growth volatility increases with the length of time over which wage growth is computed. For example, the variance of wage growth over ten years is .15 ($= .028 + 10 * .012$) for an individual who stayed with the same employer over the whole period, while it is about .31 ($= .087 + 10 * .022$) for an individual who has changed job three times or more.

These results give some basis to the interpretation of permanent shocks as resulting for a large part from job changes. Nevertheless, identifying permanent wage shocks with job changes is likely to be wrong for two reasons. First, part of the shocks faced by job stayers are permanent. Indeed, the share of permanent variance in total variance is higher for job stayers (more than one third two years apart) than for frequent job changers (less than one fourth). This finding suggests that there might be other permanent wage movements, caused for instance by within-job promotions. Second, job changers also face more transitory shocks. Describing precisely these effects requires modelling job change decisions together with wage profiles.

8 Conclusion

This paper provides a generalization of the nonparametric estimator of Li and Vuong (1998) to the case of a general linear independent factor structure, allowing for any number of measurements, L, and at most $\frac{L(L+1)}{2}$ factors (including errors). The main lessons of the standard deconvolution literature carry over to the more general context that we consider in this paper. In particular, asymptotic convergence rates are low, and it is more difficult to estimate the distribution of one factor if the characteristic functions of the other factors have thinner tails.

Our Monte Carlo results yield interesting insights. The finite-sample performance of our estimator critically depends on the shape of the distributions to be estimated (smoothness and tails properties). We also find that it is easier to estimate distributions with little skewness or kurtosis excess. In Bonhomme and Robin (2006), we show that skewness and peakedness are required for the matrix of factor loadings to be identified from higher-order moments. There is thus a tension between obtaining a precise estimate of factor loadings and a precise estimate of the distribution of factors.

In any case, identifying the distributions of more factors than measurements should be viewed as a considerably more difficult problem than the prototypical measurement error case. Given the difficulty of the problem at hand, we view the results of our simulations and the application as a confirmation that the nonparametric deconvolution approach can be successfully applied to a wide range of distributions.

The empirical application shows that the permanent and transitory components of individual earnings dynamics are clearly non normal. Predicting transitory and permanent shocks for the individuals in the sample, we see that frequent job changers face more permanent and transitory earnings shocks than job stayers. This result has important consequences for welfare analysis. Savings and insurance should be very different if the risk of large deviations is much higher than is usually assumed with normal shocks. Of course, the model of wage dynamics that we have considered is very limited. We might want to add non i.i.d. transitory shocks and yet allow for measurement error (as in Abowd and Card, 1989). We experimented with a MA(1) transitory shock without much success. It seems very difficult to nonparametrically identify the MA(1) component from the PSID data. Thus, maybe the sample we have studied is not appropriate, or a single non normal MA(0) transitory shock/measurement error is enough to describe the PSID data.

Another interesting issue is the assumption of independence between factors that we maintain throughout this analysis. Meghir and Pistaferri (2004) show evidence of autoregressive conditional heteroskedasticity in permanent and transitory components. It is not straightforward at all to extend the study of the nonparametric identification and estimation of factor densities in conditionally heteroskedastic factor models like:

$$y_{it} = A\varepsilon_{it}, \quad \varepsilon_{it}^{k} = \sigma(\varepsilon_{it-1})\eta_{it}^{k}, \quad k = 1, ..., K,$$

where $\eta_{it} = (\eta_{it}^{1}, ..., \eta_{it}^{K})^{T}$ is a $K \times 1$ vector of i.i.d. random variables. But this is a very exciting problem for future research.

APPENDIX

A Proof of Lemma 4

1. First, remark that

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

and, for any $T > 0$,

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

It will thus suffice to show that the proposition is true for the family of functions $\operatorname{Re}(f_t)(x, y) = x \cos(t^T y)$, $t \in \mathbb{R}$, 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 $f_t(x, y) = x \cos(t^T y)$, using the same notation for f_t and its real part. The proof uses the techniques exposed in Chapter II of Pollard (1984). See also Mendelson (2003).

2. Firstly, the integrability of X allows us to choose a constant K, for any $\varepsilon > 0$, such that $\mathbb{E}\left[|X| \mathbf{1}\{|X| > K\}\right] \leq \varepsilon$. Then, writing $\mathbb{E}_N f_t$ for the sample mean $\frac{1}{N}\sum_{n=1}^N f_t(X_n, Y_n)$,

$$\begin{aligned}
\sup_{|t| \leq T} |\mathbb{E}_N f_t - \mathbb{E} f_t| \quad &\leq \quad \sup_{|t| \leq T} \left|\mathbb{E}_N \left[f_t \mathbf{1}\{|X| \leq K\}\right] - \mathbb{E}\left[f_t \mathbf{1}\{|X| \leq K\}\right]\right| \\
&\quad + \sup_{|t| \leq T} \mathbb{E}_N \left[|f_t| \mathbf{1}\{|X| > K\}\right] + \sup_{|t| \leq T} \mathbb{E}\left[|f_t| \mathbf{1}\{|X| > K\}\right] \\
&\leq \quad \sup_{|t| \leq T} \left|\mathbb{E}_N \left[f_t \mathbf{1}\{|X| \leq K\}\right] - \mathbb{E}\left[f_t \mathbf{1}\{|X| \leq K\}\right]\right| \\
&\quad + \mathbb{E}_N \left[|X| \mathbf{1}\{|X| > K\}\right] + \mathbb{E}\left[|X| \mathbf{1}\{|X| > K\}\right] .
\end{aligned}$$

The last two terms converge almost surely to $2\mathbb{E}\left[|X| \mathbf{1}\{|X| > K\}\right]$, which is less than 2ε.

3. From now on, one may as well consider that the support of X is absolutely bounded by K (i.e. $|X| \leq K$ almost surely). Let $\mathbf{Z}_N = (Z_1, ..., Z_N)$ be an i.i.d. sample of random variables with distribution F. The two symmetrization steps of the proof of the Glivenko-Cantelli Theorem provide a first bound. The first symmetrization step replaces $F_N - F$ by $F_N - F'_N$, where F'_N (resp. $\mathbb{E}'_N$) is the empirical distribution (resp. the empirical mean operator) of another *i.i.d.* sample $\mathbf{Z}'_N = (Z'_1, ..., Z'_N)$ of random variables with distribution F, independent of $\mathbf{Z}_N$. Specifically, the symmetrization lemma in section 3 of chapter II of Pollard (1984) shows that

$$\Pr\left\{\sup_{|t| \leq T} |\mathbb{E}_N f_t - \mathbb{E} f_t| \geq \varepsilon\right\} \leq 2 \Pr\left\{\sup_{|t| \leq T} \left|\mathbb{E}_N f_t - \mathbb{E}'_N f_t\right| \geq \frac{1}{2}\varepsilon\right\}, \tag{A1}$$

if $\Pr\left\{|\mathbb{E}_N f_t - \mathbb{E} f_t| \leq \frac{1}{2}\varepsilon\right\} \geq \frac{1}{2}$ for all $|t| \leq T$. Chebyshev inequality shows that the latter inequality holds whenever $N \geq \frac{8 \operatorname{Var} f_t(X_n, Y_n)}{\varepsilon^2}$. As

$$\operatorname{Var} f_t(X_n, Y_n) = \operatorname{Var}\left[X_n \cos\left(t^T Y_N\right)\right] \leq \mathbb{E} X_n^2 \leq M_1,$$

inequality (A1) is true for $N \geq \frac{8M_1}{\varepsilon^2}$.

4. The second symmetrization step uses an i.i.d. sample of Rademacher random variables $\boldsymbol{\sigma}_N = (\sigma_1, ..., \sigma_N)$, where $\sigma_n = 1$ or -1 with the same probability $\frac{1}{2}$, $n = 1, ..., N$, independent

of $\mathbf{Z}_N$ and $\mathbf{Z}'_N$. The sequence of random variables $\sigma_n\left[f_t(Z_n) - f_t(Z'_n)\right]$ then has the same joint distribution as the original sequence $f_t(Z_n) - f_t(Z'_n)$. It follows that

$$\begin{aligned}
\Pr\left\{\sup_{|t|\le T}\left|\mathbb{E}_N f_t - \mathbb{E}'_N f_t\right| \ge \frac{1}{2}\varepsilon\right\} &= \Pr\left\{\sup_{|t|\le T}\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n f_t(Z_n) - \frac{1}{N}\sum_{n=1}^{N}\sigma_n f_t(Z'_n)\right| \ge \frac{1}{2}\varepsilon\right\}, \\
&\le \Pr\left\{\sup_{|t|\le T}\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n f_t(Z_n)\right| \ge \frac{1}{4}\varepsilon\right\} \\
&\quad + \Pr\left\{\sup_{|t|\le T}\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n f_t(Z'_n)\right| \ge \frac{1}{4}\varepsilon\right\}, \\
&= 2\Pr\left\{\sup_{|t|\le T}\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n f_t(Z_n)\right| \ge \frac{1}{4}\varepsilon\right\}, \\
&= 2\mathbb{E}\Pr\left\{\sup_{|t|\le T}\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n f_t(Z_n)\right| \ge \frac{1}{4}\varepsilon \,\middle|\, \mathbf{Z}_N\right\}. \qquad \text{(A2)}
\end{aligned}$$

5. A maximal inequality then follows from the following finite-covering argument. For any couple (t_1, t_2),

$$\begin{aligned}
\left|x\cos(t_1^T y) - x\cos(t_2^T y)\right| &\le \left|x(t_1^T y - t_2^T y)\right| \\
&\le \sum_{\ell}\left|xy_\ell\left(t_{1\ell} - t_{2\ell}\right)\right| \\
&\le \sum_{\ell}\left|xy_\ell\right| \cdot \left|t_1 - t_2\right| \\
&\le L\left|x\right|\left|y\right| \cdot \left|t_1 - t_2\right|.
\end{aligned}$$

Fix $\mathbf{Z}_N$, and define $M_{2,N} = \frac{1}{N}\sum_n |Y_n|$. Partition $[-T,T]^L$ into $r_N = \left(2T\frac{8LKM_{2,N}}{\varepsilon}\right)^L$ adjacent hypercubes of side length $\frac{\varepsilon}{8KM_{2,N}}$. Lastly, let $\{t_k; k = 1, ..., r_N\}$ be the set of all cube corners. Then, for any $t \in [-T,T]^L$ there exists k such that

$$\begin{aligned}
\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n\left(f_{t_k} - f_t\right)(Z_n)\right| &\le \frac{1}{N}\sum_{n=1}^{N}\left|\left(f_{t_k} - f_t\right)(Z_n)\right| \\
&\le LKM_{2,N}\left|t_k - t\right| \\
&\le LKM_{2,N}\frac{\varepsilon}{8LKM_{2,N}} = \frac{1}{8}\varepsilon,
\end{aligned}$$

as $|X_n| \le K$. Hence, for all $t \in [-T,T]^L$ such that $\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n f_t(Z_n)\right| \ge \frac{1}{4}\varepsilon$,

$$\begin{aligned}
\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n f_{t_k}(Z_n)\right| &\ge \left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n f_t(Z_n)\right| - \left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n\left(f_{t_k} - f_t\right)(Z_n)\right|, \\
&\ge \frac{1}{8}\varepsilon.
\end{aligned}$$

One can thus refine the bound further as:

$$\begin{aligned}\Pr\left\{\sup_{|t|\leq T}\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n f_t(Z_n)\right|\geq\frac{1}{4}\varepsilon\middle|\mathbf{Z}_N\right\} &\leq \Pr\left\{\max_k\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n f_{t_k}(Z_n)\right|\geq\frac{1}{8}\varepsilon\middle|\mathbf{Z}_N\right\}\\ &\leq \sum_{k=1}^{r_N}\Pr\left\{\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n f_{t_k}(Z_n)\right|\geq\frac{1}{8}\varepsilon\middle|\mathbf{Z}_N\right\},\\ &\leq r_N\max_k\Pr\left\{\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n f_{t_k}(Z_n)\right|\geq\frac{1}{8}\varepsilon\middle|\mathbf{Z}_N\right\}. \qquad \text{(A3)}\end{aligned}$$

6. Lastly, applying Hoeffding's inequality to the sequence $[\sigma_n f_{t_k}(Z_n)]$, bounded by K, yields:

$$\Pr\left\{\sup_k\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n f_{t_k}(Z_n)\right|\geq\frac{1}{8}\varepsilon\middle|\mathbf{Z}_N\right\}\leq 2\exp\left[-\frac{N\varepsilon^2}{128K^2}\right]. \qquad \text{(A4)}$$

7. At this stage, we have thus shown that, for $N\geq\frac{8M_1}{\varepsilon^2}$,

$$\begin{aligned}\Pr\left\{\sup_{|t|\leq T}|\mathbb{E}_N f_t-\mathbb{E}f_t|\geq\varepsilon\right\} &\leq 2^{4L+3}\mathbb{E}\left(M_{2,N}^L\right)\left(\frac{LKT}{\varepsilon}\right)^L\exp\left[-\frac{N\varepsilon^2}{128K^2}\right],\\ &= 2^{4L+3}L^L\mathbb{E}\left(M_{2,N}^L\right)\exp\left[L\ln\left(\frac{KT}{\varepsilon}\right)-\frac{N\varepsilon^2}{128K^2}\right]. \qquad \text{(A5)}\end{aligned}$$

Assuming that $\mathbb{E}\left|Y_N\right|^i<+\infty$ for all $i\leq\{1...L\}$, then one can find $M_2<+\infty$ such that $\mathbb{E}\left|Y_N\right|^i<M_2^i$, $\forall i\leq L$. Since (Y_N) is an i.i.d. sequence we thus have

$$\begin{aligned}\mathbb{E}\left(M_{2,N}^L\right) &= \mathbb{E}\left[\left(\frac{1}{N}\sum_{n=1}^{N}|Y_n|\right)^L\right]\\ &\leq (M_2)^L.\end{aligned}$$

Let (ε_N) be a sequence of positive numbers converging to zero and let (K_N) and (T_N) be two diverging sequences of positive numbers. If ε_N tends to zero slowly enough for $\varepsilon_N^2\geq\frac{8M_1}{N}$ and so that $\sum_N\exp\left[L\ln\left(\frac{K_NT_N}{\varepsilon_N}\right)-\frac{N\varepsilon_N^2}{128K_N^2}\right]<\infty$, then

$$\sum_N\Pr\left\{\sup_{|t|\leq T_N}|\mathbb{E}_N f_t-\mathbb{E}f_t|\geq\varepsilon_N\right\}<\infty.$$

The Borel-Cantelli Lemma then implies that only a finite number of events are such that

$$\sup_{|t|\leq T_N}|\mathbb{E}_N f_t-\mathbb{E}f_t|\geq\varepsilon_N.$$

Hence,

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

8. The last step of the proof characterizes T_N, K_N and ε_N further. The series

$$\sum_N\exp\left[\ln\left(\frac{K_NT_N}{\varepsilon_N}\right)-\frac{N\varepsilon_N^2}{128K_N^2}\right]$$

converges if $N\left(\frac{\varepsilon_N}{K_N}\right)^2$ increases faster than $\ln N$ and $\ln\left(\frac{K_N T_N}{\varepsilon_N}\right)$ not faster, i.e. $\frac{K_N}{\varepsilon_N}=o\left[\left(\frac{N}{\ln N}\right)^{\frac{1}{2}}\right]$ and $\ln\left(\frac{K_N T_N}{\varepsilon_N}\right)=O(\ln N)$, which in turn holds if $\ln T_N=O(\ln N)$: T_N tends to infinity not faster than polynomial rate.

Let $K_{|X|}(\varepsilon)$ be implicitly defined by the equality:

$$\mathbb{E}\left[|X|\,\mathbf{1}\{|X|>K\}\right]=\int_K^\infty u f_{|X|}(u)\,du=\varepsilon.$$

It is required that $K_N\geq K_{|X|}(\varepsilon_N)$. To choose ε_N as small as possible given $\frac{K_N}{\varepsilon_N}$, then it is best to set $K_N=K_{|X|}(\varepsilon_N)$.

This achieves to prove Lemma 4.

B Proof of Theorem 5

(i) Fix any $t\in\mathbb{R}^L$, let $\varphi(t)\equiv\varphi_Y(t)=\mathbb{E}\left[e^{it^TY}\right]$, $\psi_\ell(t)=\mathbb{E}\left[Y_\ell e^{it^TY}\right]$ and $\xi_{\ell m}(t)=\mathbb{E}\left[Y_\ell Y_m e^{it^TY}\right]$, for any $\ell,m=1,...,L$. Then, Lemma 4 defines $\varepsilon_N\downarrow 0$ and $T_N\to\infty$ such that (all convergence statements are implicitly holding almost surely)

$$\begin{aligned}\sup_{|t|\leq T_N}|\widehat{\varphi}(t)-\varphi(t)|&=O(\varepsilon_N),\\ \sup_{|t|\leq T_N}\left|\widehat{\psi}_\ell(t)-\psi_\ell(t)\right|&=O(\varepsilon_N),\\ \sup_{|t|\leq T_N}\left|\widehat{\xi}_{\ell m}(t)-\xi_{\ell m}(t)\right|&=O(\varepsilon_N),\end{aligned}$$

hold simultaneously. One can take the largest ε_N and the smallest T_N.

In addition, we shall require below that $\frac{T_N^2\varepsilon_N}{g(T_N)^3}=o(1)$. As $h(t)=\frac{t^2}{g(t)^3}$ is an increasing function, one can redefine T_N –if necessary– as $T_N=h^{-1}\left(\varepsilon_N^{\gamma-1}\right)$, with $0<\gamma<1$.

(ii) Removing the subscript Y from φ_Y and g_Y to simplify the notations, as $|\varphi(t)|\geq g(|t|)$ when $|t|\to\infty$, and as φ is nonvanishing everywhere, then for T_N large enough

$$\inf_{|t|\leq T_N}|\varphi(t)|\geq g(T_N),$$

and

$$\sup_{|t|\leq T_N}\left|\frac{\widehat{\varphi}(t)-\varphi(t)}{\varphi(t)}\right|=\frac{O(\varepsilon_N)}{g(T_N)}=o(1).$$

The last equality follows from the fact that $\frac{T_N^2\varepsilon_N}{g(T_N)^3}\geq\frac{\varepsilon_N}{g(T_N)}$ for N large enough.

(iii) We have

$$\frac{\partial\kappa_Y(t)}{\partial t_\ell}=i\frac{\psi_\ell(t)}{\varphi(t)}=i\frac{\mathbb{E}\left[Y_\ell e^{it^TY}\right]}{\mathbb{E}\left[e^{it^TY}\right]},$$

and

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

One can bound $\widehat{\psi}_\ell(t)$ as follows:

$$\begin{aligned}\sup_{|t|\leq T_N}\left|\widehat{\psi}_\ell(t)\right| &\leq \sup_{|t|\leq T_N}\left|\widehat{\psi}_\ell(t)-\psi_\ell(t)\right|+\sup_{t\in[-T_N,T_N]}|\psi_\ell(t)| \\ &\leq \sup_{|t|\leq T_N}\left|\widehat{\psi}_\ell(t)-\psi_\ell(t)\right|+\mathbb{E}\,|Y_\ell| = O(1),\end{aligned}$$

as $\mathbb{E}\,|Y_\ell|<\infty$ if $\mathbb{E}Y_\ell^2\leq M_1<\infty$.

It follows that

$$\sup_{|t|\leq T_N}\left|\frac{\widehat{\psi}_\ell(t)}{\widehat{\varphi}(t)}-\frac{\psi_\ell(t)}{\varphi(t)}\right| = \frac{O(\varepsilon_N)}{g(T_N)^2} = o\left(1\right).$$

The same argument applies to show that

$$\sup_{|t|\leq T_N}\left|\frac{\widehat{\xi}_{\ell m}(t)}{\widehat{\varphi}(t)}-\frac{\xi_{\ell m}(t)}{\varphi(t)}\right| = \frac{O(\varepsilon_N)}{g(T_N)^2} = o\left(1\right)$$

for all ℓ, m.

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

$$\begin{aligned}\zeta_{\ell m}(t) &\equiv \frac{\partial^2\kappa_Y}{\partial t_\ell\partial t_m}(t) \\ &= -\frac{\mathbb{E}\left[Y_\ell Y_m e^{it^TY}\right]}{\mathbb{E}\left[e^{it^TY}\right]}+\frac{\mathbb{E}\left[Y_\ell e^{it^TY}\right]}{\mathbb{E}\left[e^{it^TY}\right]}\frac{\mathbb{E}\left[Y_m e^{it^TY}\right]}{\mathbb{E}\left[e^{it^TY}\right]} \\ &= -\frac{\xi_{\ell m}(t)}{\varphi(t)}+\frac{\psi_\ell(t)}{\varphi(t)}\frac{\psi_m(t)}{\varphi(t)}.\end{aligned}$$

Let $\widehat{\zeta}_{\ell m}(t) = -\frac{\widehat{\xi}_{\ell m}(t)}{\widehat{\varphi}(t)}+\frac{\widehat{\psi}_\ell(t)}{\widehat{\varphi}(t)}\frac{\widehat{\psi}_m(t)}{\widehat{\varphi}(t)}$. Then,

$$\begin{aligned}\widehat{\zeta}_{\ell m}(t)-\zeta_{\ell m}(t) &= -\left[\frac{\widehat{\xi}_{\ell m}(t)}{\widehat{\varphi}(t)}-\frac{\xi_{\ell m}(t)}{\varphi(t)}\right] \\ &\quad+\left[\frac{\widehat{\psi}_\ell(t)}{\widehat{\varphi}(t)}-\frac{\psi_\ell(t)}{\varphi(t)}\right]\frac{\psi_m(t)}{\varphi(t)}+\left[\frac{\widehat{\psi}_m(t)}{\widehat{\varphi}(t)}-\frac{\psi_m(t)}{\varphi(t)}\right]\frac{\psi_\ell(t)}{\varphi(t)} \\ &\quad+\left[\frac{\widehat{\psi}_\ell(t)}{\widehat{\varphi}(t)}-\frac{\psi_\ell(t)}{\varphi(t)}\right]\left[\frac{\widehat{\psi}_m(t)}{\widehat{\varphi}(t)}-\frac{\psi_m(t)}{\varphi(t)}\right].\end{aligned}$$

Since

$$\sup_{|t|\leq T_N}\left|\frac{\psi_\ell(t)}{\varphi(t)}\right|\leq\frac{\mathbb{E}\,|Y_\ell|}{g(T_N)}$$

for all ℓ, it follows that

$$\sup_{|t|\leq T_N}\left|\widehat{\zeta}_{\ell m}(t)-\zeta_{\ell m}(t)\right| = \frac{O(\varepsilon_N)}{g(T_N)^2}+\frac{O(\varepsilon_N)}{g(T_N)^3}+\left(\frac{O(\varepsilon_N)}{g(T_N)^2}\right)^2 = \frac{O(\varepsilon_N)}{g(T_N)^3}$$

because

$$\frac{\varepsilon_N}{g(T_N)^3}>\frac{\varepsilon_N^2}{g(T_N)^4}\Leftrightarrow 1>\frac{\varepsilon_N}{g(T_N)}$$

for N large enough.

(v) For any vector $t = (t_1, ..., t_L)^T \in \mathbb{R}^L$ and $\tau \in \mathbb{R}$, then

$$\begin{aligned} B_\ell(t) &= \sup_{\tau\in[-T_N,T_N]} \left| \int_0^\tau \frac{\widehat{\psi}_\ell(ut)}{\widehat{\varphi}(ut)} du - \int_0^{t_\ell} \frac{\psi_\ell(ut)}{\varphi(ut)} du \right| \\ &\leq \sup_{\tau\in[-T_N,T_N]} \left(\tau \sup_{|t|\leq T_N} \left| \frac{\widehat{\psi}_\ell(t)}{\widehat{\varphi}(t)} - \frac{\psi_\ell(t)}{\varphi(t)} \right| \right) \\ &\leq T_N \sup_{|t|\leq T_N} \left| \frac{\widehat{\psi}_\ell(t)}{\widehat{\varphi}(t)} - \frac{\psi_\ell(t)}{\varphi(t)} \right| \\ &= \frac{T_N}{g(T_N)^2} O(\varepsilon_N). \end{aligned}$$

Similarly,

$$\begin{aligned} C_{\ell m}(t) &= \sup_{\tau\in[-T_N,T_N]} \left| \int_0^\tau \int_0^u \widehat{\zeta}_{\ell m}(vt) dv du - \int_0^\tau \int_0^u \zeta_{\ell m}(vt) dv du \right| \\ &\leq \sup_{\tau\in[-T_N,T_N]} \left(\frac{\tau^2}{2} \sup_{|t|\leq T_N} \left| \widehat{\zeta}_{\ell m}(t) - \zeta_{\ell m}(t) \right| \right) \\ &\leq T_N^2 \sup_{|t|\leq T_N} \left| \widehat{\zeta}_{\ell m}(t) - \zeta_{\ell m}(t) \right| \\ &= \frac{T_N^2}{g(T_N)^3} O(\varepsilon_N). \end{aligned}$$

Moreover, for any distribution W on $\mathcal{T}_k$,

$$\int B_\ell(t)\, dW(t) \leq \sup_{|t|\leq T_N} B_\ell(t) \cdot \int dW(t) = \frac{T_N}{g(T_N)^2} O(\varepsilon_N)$$

and

$$\int C_{\ell m}(t)\, dW(t) \leq \sup_{|t|\leq T_N} C_{\ell m}(t) \cdot \int dW(t) = \frac{T_N^2}{g(T_N)^3} O(\varepsilon_N).$$

(vi) It easily follows from the previous step that:

$$\sup_{\tau\in[-T_N,T_N]} \left| \widehat{\kappa}_{X_k}(\tau) - \kappa_{X_k}(\tau) \right| = \frac{T_N^2}{g(T_N)^3} O(\varepsilon_N) = o(1).$$

In particular, $\sup_{\tau\in[-T_N,T_N]} \left| \widehat{\kappa}_{X_k}(\tau) - \kappa_{X_k}(\tau) \right| < 1$ for N large enough. Therefore, for N large enough

$$\begin{aligned} \sup_{\tau\in[-T_N,T_N]} \left| \widehat{\varphi}_{X_k}(\tau) - \varphi_{X_k}(\tau) \right| &= \sup_{\tau\in[-T_N,T_N]} \left| \exp\left(\widehat{\kappa}_{X_k}(\tau)\right) - \exp\left(\widehat{\kappa}_{X_k}(\tau)\right) \right|, \\ &\leq \sup_{\tau\in[-T_N,T_N]} \left| \widehat{\kappa}_{X_k}(\tau) - \kappa_{X_k}(\tau) \right|, \end{aligned}$$

from which it follows that

$$\sup_{\tau\in[-T_N,T_N]} \left| \widehat{\varphi}_{X_k}(\tau) - \varphi_{X_k}(\tau) \right| = \frac{T_N^2}{g(T_N)^3} O(\varepsilon_N).$$

This ends the proof of Theorem 5.

C Proof of Theorem 6

For all x in the support of X_k:

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

Note that

$$\begin{aligned}|\varphi_Y(t)| &= \left|\mathbb{E}\left[e^{it^TY}\right]\right| = \left|\mathbb{E}\left[e^{it^TAX}\right]\right| \\ &= |\varphi_X(A^Tt)| \\ &\geq g_X(\left|A^Tt\right|) \\ &\geq g_X(L\,|A|\,|t|),\end{aligned}$$

where $|A| = \max_{i,j}(|a_{ij}|)$. Moreover, function g_Y inherits g_X's properties: it maps $\mathbb{R}^+$ onto $[0,1]$, it is decreasing and it is integrable, so that in particular $g_Y(|t|) \to 0$ when $|t| \to \infty$. We can thus apply Theorem 5 and obtain:

$$\sup_x \left|\widehat{f}_{X_k}(x) - f_{X_k}(x)\right| = \frac{T_N^3}{g(T_N)^3}O(\varepsilon_N) + O\left(\int_{T_N}^{+\infty} h_{X_k}(v)dv\right).$$

where $g(|t|) = g_X(L\,|A|\,|t|)$ and $\int_{T_N}^{+\infty} h_X(v)dv = o(T_N)$, as h_{X_k} is integrable. This ends the proof of Theorem 6.

D Proof of Corollary 7

The distribution of factor X_k is smooth and there exists $1 < \beta_k < \alpha_k$ such that

$$|\tau|^{-\alpha_k} \leq \left|\varphi_{X_k}(\tau)\right| \leq |\tau|^{-\beta_k}, \quad |\tau| \to \infty.$$

This allows to set $h_{X_k}(|\tau|) = |\tau|^{-\beta_k}$. Moreover, as factors X_k are mutually independent,

$$|\varphi_X(\tau)| = \prod_{k=1}^K \left|\varphi_{X_k}(\tau_k)\right| \geq \prod_{k=1}^K |\tau_k|^{-\alpha_k} \geq |\tau|^{-\sum_{k=1}^K \alpha_k}.$$

One can thus take $g_X(|\tau|) = |\tau|^{-\alpha}$, with $\alpha = \sum_{k=1}^K \alpha_k$.

Minimizing

$$\frac{T_N^3}{g_X(T_N)^3}\Delta_N + \int_{T_N}^{+\infty} h_{X_k}(v)dv = T_N^{3(1+\alpha)}\Delta_N + \frac{1}{\beta_k - 1}T_N^{1-\beta_k}$$

with respect to T_N yields:

$$T_N = (3(1+\alpha)\Delta_N)^{-1/(2+3\alpha+\beta_k)}.$$

If factor p.d.f.'s have Pareto-tails with $K_{|X_k|}(\varepsilon) \leq (1/\varepsilon)^{\frac{1}{a-1}}$, Lemma 4 shows that one can choose

$$\Delta_N = \left(\frac{\ln N}{N}\right)^{(1-1/a)(1/2-\gamma)},$$

where $0 < \gamma < 1/2$ should be chosen close to zero.

It thus follows that one can choose:

$$T_N = \left(\frac{N}{\ln N}\right)^{\frac{(1-1/a)(1/2-\gamma)}{2+3\alpha+\beta_k}}.$$

Note that T_N tends to infinity with N at a polynomial rate, so $\ln T_N = O(\ln N)$.

Then, up to a multiplicative constant

$$\frac{T_N^3}{g_X(T_N)^3}\Delta_N = T_N^{3(1+\alpha)}\Delta_N = \left(\frac{\ln N}{N}\right)^{\frac{\beta_k-1}{2+3\alpha+\beta_k}(1-1/a)(1/2-\gamma)}.$$

As this last term is $o(1)$, so is $\frac{T_N^2}{g_X(L|A|T_N)^3}\Delta_N$. Moreover:

$$\int_{T_N}^{+\infty}|t|^{-\beta_k}dv = \frac{1}{\beta_k-1}T_N^{1-\beta_k} = \left(\frac{\ln N}{N}\right)^{\frac{\beta_k-1}{2+3\alpha+\beta_k}(1-1/a)(1/2-\gamma)}.$$

This implies that:

$$\sup_x\left|\widehat{f}_{X_k}(x) - f_{X_k}(x)\right| = O\left(\left(\frac{\ln N}{N}\right)^{\frac{\beta_k-1}{2+3\alpha+\beta_k}(1-1/a)(1/2-\gamma)}\right) \quad \text{a.s.}$$

Note that, as $\beta_k > 1$ this quantity is $o(1)$. This ends the proof.

References

[1] ABOWD, J., and D. CARD (1989): "On the Covariance Structure of Earnings and Hours Changes," *Econometrica,* 57, 411-445.

[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] BONHOMME, S. and J.-M. ROBIN (2006), "Using High-Order Moments to Estimate Linear Independent Factor Models," *mimeo.*

[4] BROWN, J., and A. LIGHT (1992): "Interpreting Panel Data on Job Tenure," *Journal of Labor Economics,* 10, 219-257.

[5] 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.

[6] CARRASCO, M., and J.P. FLORENS (2005): "Spectral Method for Deconvolving a Density," *mimeo.*

[7] CARRASCO, M., J.P. FLORENS, and E. RENAULT (2006): "Linear Inverse Problems in Structural Econometrics: Estimation Based on Spectral Decomposition and Regularization," *Handbook of Econometrics*, J.J. Heckman and E.E Leamer (eds.), vol.6, North Holland.

[8] CARROLL, R. J., and P. HALL (1988): "Optimal rates of Convergence for Deconvoluting a Density," *Journal of the American Statistical Association*, 83, 1184-1186.

[9] COMTE F., Y. ROZENHOLC and M.-L. TAUPIN (2006): "Penalized contrast estimator for adaptive density deconvolution," *Canadian Journal of Statistics*, Vol. 34, 3, forthcoming.

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

[11] DEMPSTER, A. P., N. M. LAIRD and D. B. RUBIN (1977): "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, B 39(1), 1-38.

[12] 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.

[13] FAN, J. Q. (1991): "On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems," *Annals of statistics*, 19, 1257-1272.

[14] FAN, J. and J.-Y. KOO (2002): "Wavelet Deconvolution," *IEEE transactions on Information Theory*, Vol. 48, 3, 734-747.

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

[16] GEWEKE, J., and M. KEANE (2000): "An Empirical Analysis of Earnings Dynamics Among Men in the PSID: 1968-1989," *Journal of Econometrics,* 96, 293-356.

[17] HALL, P., and Q. YAO (2003): "Inference in Components of Variance Models with Low Replications," *Annals of Statistics*, 31, 414-441.

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

[19] HOROWITZ, J. L. (1998): *Semiparametric Methods in Econometrics.* New-York: Springer-Verlag.

[20] HOROWITZ, J. L., and M. MARKATOU (1996): "Semiparametric Estimation of Regression Models for Panel Data," *Review of Economic Studies*, 63, 145-168.

[21] HU, Y. and G. RIDDER (2005), "Estimation of Nonlinear Models with Mismeasured Regressors Using Marginal Information", *mimeo.*

[22] KOTLARSKI, I. (1967): "On Characterizing the Gamma and Normal Distribution," *Pacific Journal of Mathematics*, 20, 69-76.

[23] LI, T. (2002): "Robust and Consistent Estimation of Nonlinear Errors-in-Variables Models, " *Journal of Econometrics*, 110, 1-26.

[24] LI, T., and Q. VUONG (1998): "Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators," *Journal of Multivariate Analysis*, 65, 139-165.

[25] LILLARD, L., and R. WILLIS (1978): “Dynamic Aspects of Earnings Mobility,” *Econometrica*, 46, 985-1012.

[26] LINTON, O., and Y. J. WHANG (2002): “Nonparametric Estimation with Aggregated Data,” *Econometric Theory*, 18, 420-468.

[27] MEGHIR, C., and L. PISTAFERRI (2004): “Income Variance Dynamics and Heterogeneity,” *Econometrica,* 72, 1-32.

[28] MENDELSON, S. (2003): “A few notes on Statistical Learning Theory,” in *Advanced Lectures in Machine Learning*, (S. Mendelson, A.J. Smola Eds), LNCS 2600, 1-40, Springer.

[29] 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 accoustics, speech and signal processing*, 3617-3620.

[30] PENSKY, M., and B. VIDAKOVIC (1999): “Adaptive wavelet estimator for nonparametric density deconvolution,” *The Annals of Statistics*, 27(6), 2033-2053.

[31] POLLARD, D. (1984): “Convergence in Distributions of Stochastic Processes.” Springer, New-York.

[32] RAO, P. (1992): “Identifiability in Stochastic Models.” New-York, Academic Press.

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

[34] SCHENNACH, S. (2004): “Estimation of Nonlinear Models with Measurement Error,” *Econometrica*, 72, 33-75.

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

[36] STEFANSKI, L., and R. J. CARROLL (1990): “Deconvoluting Kernel Density Estimators,” *Statistics*, 2, 169-184.