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

Stéphane Bonhomme
CEMFI, Madrid[2,3]

Jean-Marc Robin
Paris School of Economics,
University Paris 1 - Pantheon - Sorbonne,[4]
and University College London[5]

Revised version: October 2008

[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.

[2]Centro de Estudios Monetarios y Financieros. Address: Casado del Alisal, 5, 28014 Madrid, Spain. E-mail: bonhomme@cemfi.es.

[3]Stéphane Bonhomme gratefully acknowledges the financial support from the Spanish Ministry of Science and Innovation through the Consolider-Ingenio 2010 Project "Consolidating Economics", and from the Spanish MEC, Grant SEJ2005-08880.

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

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

Abstract

In this paper, we construct a nonparametric estimator of the distributions of latent factors in linear independent multi-factor models under the assumption that factor loadings are known. Our approach allows to estimate the distributions of up to $L(L+1)/2$ factors given L measurements. The estimator uses empirical characteristic functions, like many available deconvolution estimators. We show that it is consistent, and derive asymptotic convergence rates. Monte-Carlo simulations show good finite-sample performance, less so if distributions are highly skewed or leptokurtic. We finally apply the generalized deconvolution procedure to decompose individual log earnings from the Panel Study of Income Dynamics (PSID) into permanent and transitory components.

JEL codes: C13, C14.

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

1 Introduction

In this paper, we consider linear multi-factor models of the form: $\mathbf{Y} = \mathbf{AX}$, where $\mathbf{Y} = (Y_1, ..., Y_L)^{\mathrm{T}}$ is a vector of L measurements, $\mathbf{X} = (X_1, ..., X_K)^{\mathrm{T}}$ is a vector of K unobserved and mutually independent latent factors, and $\mathbf{A}$ is a $L \times K$ matrix of parameters. The analysis is conducted assuming that the number of factors and the matrix of factor loadings $\mathbf{A}$ is known. The contribution of this paper is to provide a nonparametric estimator of the distribution function of $\mathbf{X}$ from an i.i.d. sample $\{\mathbf{Y}_n, n = 1, ..., N\}$ for up to $K = L(L+1)/2$ factors.

Applications of factor models are numerous in social sciences, and economics in particular. For example, the standard model of individual earnings dynamics is a linear multi-factor model with four additive components, or factors: a deterministic function of regressors, a fixed effect, a persistent autoregressive component, and a transitory moving-average component (e.g., Hall and Mishkin, 1982, Abowd and Card, 1989). Factor variances are usually estimated based on second-order moment restrictions. Estimating the whole distribution of factors is less common but still useful. Lillard and Willis (1978) compute transition probabilities into and out of poverty ("first passage times"). More recently, economists have shown considerable interest in estimating complete earnings models to feed life-cycle consumption models (e.g., Guvenen, 2007a, 2007b, Kaplan and Violante, 2008). Estimating factor distributions nonparametrically in those models is thus of substantial interest.[1]

Horowitz and Markatou (1996) were the first to propose fully nonparametric estimators for linear models with error components, with an application to earnings panel data. They show that, unlike in the standard semiparametric deconvolution problem,[2] *every* component of the convolution can be identified and nonparametrically estimated if repeated observations of the dependent variable are available. The basic setup that they study has two measurements (two observations per individual), one common factor and two independent errors with identical and symmetric distribution.[3]

Horowitz and Markatou consider several extensions of this simple framework, namely

[1] The usual approach is to adopt flexible parametric distributions for the individual effects and the innovations (e.g., Chamberlain and Hirano, 1999, Hirano, 2002, Geweke and Keane, 2000, 2007).

[2] For references on the classical deconvolution problem, see Carroll and Hall (1988), Zhang (1990), Fan (1991), and Carroll *et al.* (1995).

[3] Susko and Nadon (2002) consider the same setup and estimator as Horowitz and Markatou with an application to gene expression. More recently, Delaigle, Hall and Meister (2008) have proposed a modified kernel estimator that may achieve, under smoothness conditions, the same asymptotic performance as if the distribution of errors were known.

stationary AR or MA errors, or asymmetric errors.[4] However, it is not easy to see how to extend the basic approach in a systematic way to analyze more complex models. In this paper, we extend their approach and develop a general method for estimating factor distributions in linear factor models with many independent factors with different, unrestricted distributions.

Our estimator generalizes the one proposed in another important paper by Li and Vuong (1998). Li and Vuong consider the same basic setup as Horowitz and Markatou, with repeated measurements and independent errors. They allow measurement errors to display different, possibly asymmetric, distributions.[5] Their estimator of the densities of the common factor and the independent errors builds on an identification result due to Kotlarski (1967) (also stated by P. Rao, 1992, p. 21).

Székely and C. R. Rao (2000) generalize Kotlarski's identification result to the general case of linear multi-factor models $\mathbf{Y} = \mathbf{AX}$ with known $\mathbf{A}$ and unrestricted, independent unobserved factors $\mathbf{X}$. They show that the $L(L+1)/2$ second-order partial derivatives of the characteristic function of $\mathbf{Y}$ deliver a system of functional identifying restrictions allowing to identify a maximal number of $K = L(L+1)/2$ factors. Our estimator is based on this system of identifying restrictions, replacing the characteristic function of the vector of measurements by an empirical analog, and using a smoothing kernel with trimming. Our estimator requires no optimization, unlike parametric approaches. Moreover, we provide a simple method to choose the trimming parameter, inspired by the "plug-in" method proposed in Delaigle and Gijbels (2004).

We compute upper bounds to the rate of uniform convergence of the estimator, as is usual in the deconvolution literature. However, we depart from most of the previous literature, and in particular from Li and Vuong (1998), by allowing the supports of factor distributions to be unbounded. The rate of convergence of our estimator depends crucially on the smoothness of factor distributions, and may be very slow. When the characteristic function of the factor of interest has fatter tails than the characteristic functions of other factors, the rate may be logarithmic in N. These slow rates of convergence do not indicate a flaw in our approach but are instead a fundamental property of nonparametric deconvolution estimators. Indeed, logarithmic rates are the best convergence rates that

[4] These extensions are developed in Horowitz (1998, p. 125-136).

[5] Li and Vuong's estimator has been used by Li *et al.* (2000) in the context of a structural auction model, and in Li (2002) in a nonlinear errors-in-variables model. Hall and Yao (2003) proposed an estimator that is closely related to Li and Vuong (1998). Related methods have been used by Schennach (2004a, 2004b) in the context of nonlinear regression and nonparametric regression, respectively, when the regressors are measured with error, and by Hu and Ridder (2007) in order to deal with measurement error when marginal information is available.

can be attained in some circumstances (Carroll and Hall, 1988, Fan, 1991).

Despite the slow asymptotic rates of convergence, Monte Carlo simulations are encouraging. When the true factor distributions are normal or Laplace, we find moderate biases and tight confidence bands. Interestingly, our generalized deconvolution estimator has the same finite-sample bias and variance as a standard deconvolution estimator assuming that all factor densities are known except the one to be estimated. Moreover, it achieves comparable finite-sample performance to the estimators of Horowitz and Markatou (1996) and Li and Vuong (1998) in the basic setup with two measurements. We also find that the shape of factor distributions strongly influences the performance of the estimator. In particular, the estimation of factor distributions is more difficult when these distributions are skewed or leptokurtic.

We apply our methodology to individual earnings data from the PSID. We model the residuals of log earnings on individual covariates as the sum of an individual effect, a random walk and a white noise, and estimate the distributions of innovations from first differences, instead of earnings levels as in Horowitz and Markatou (1996).[6] Our results show that both shocks exhibit more kurtosis than the normal distribution. We use the model to analyze the respective roles of permanent and transitory shocks in earnings mobility, and to correlate the variance of earnings shocks to job mobility. In particular, 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. Section 2 presents the model and assumptions. In Section 3, we provide a simple proof of the identification result in Székely and Rao (2000), that we use in Section 4 to construct an estimator of factor densities. In Section 5, we prove the consistency of the estimator and discuss convergence speed. Sections 6 display Monte Carlo simulations. Section 7 presents an application to earnings dynamics. Section 8 concludes.

2 Model and assumptions

2.1 Model

The main features of the model are summarized in the following assumption ($\mathbf{A}^{\mathrm{T}}$ denotes the matrix transpose of matrix $\mathbf{A}$):

[6]Note that Horowitz and Markatou assume independent shocks with identical symmetric distributions. As they use CPS data, which is a two-year panel, possibilities for identifying complex error-component models are extremely limited.

Assumption A1 *We consider the following basic setup:*

1. $\mathbf{X}_n = (X_{n1}, ..., X_{nK})^{\mathrm{T}}, n = 1, ..., N$, *are* N *independent copies of a vector* $\mathbf{X} = (X_1, ..., X_K)^{\mathrm{T}}$ *of* K *real valued, mutually independent, and non degenerate random variables, with zero mean and finite variances. Vectors* $\mathbf{X}_1, ..., \mathbf{X}_N$ *are unobserved to the econometrician and are called factors.*

2. $\mathbf{Y}_n = (Y_{n1}, ..., Y_{nL})^{\mathrm{T}}, n = 1, ..., N$, *are* N *independent copies of a vector* $\mathbf{Y} = (Y_1, ..., Y_L)^{\mathrm{T}}$ *with zero mean. Vectors* $\mathbf{Y}_1, ..., \mathbf{Y}_N$ *are observed and are called measurements.*

3. $\mathbf{Y} = \mathbf{A}\mathbf{X}$ *, where* $\mathbf{A} = [a_{\ell k}]$ *is a known* $L \times K$ *matrix of scalar parameters and any two columns of* $\mathbf{A}$ *are linearly independent.*

Remark 1. It is usual to denote as U instead of X a factor variable that appears in only one equation (an error).

Remark 2. Measurements are demeaned. Obviously, factor distributions are only identified up to a location parameter. We therefore normalize the factor means to 0.

Remark 3. Assuming that factors have finite variances implies that the characteristic functions of factors X_k are twice differentiable (Lukacs, 1970, Theorem 2.3.1.). This property is instrumental in the construction of the characteristic functions of factors from that of the vector of measurements. However, it is not a necessary condition for identification, as shown by Székely and Rao (2000).

Remark 4. We assume that factor loadings are known to the researcher. Alternatively, one could assume that a root-N consistent estimator of $\mathbf{A}$ is available. The asymptotic results derived in this paper would remain unchanged, as we find convergence rates of density estimators that are slower than root-N.

With $K \leq L$, the distribution of $\mathbf{X}$ is trivially identified as that of $\mathbf{A}^{-}\mathbf{Y}$, where $\mathbf{A}^{-}$ is a pseudo inverse of $\mathbf{A}$. The aim of this paper is to propose a general method to estimate the distributions of factors when there are more factors than measurements ($K > L$). This situation arises naturally if there are L common factors and L errors, as in standard factor analysis.

The empirical application that we shall later consider deals with earnings dynamics. The standard model assumes that log earnings residuals can be decomposed into a fixed

effect, a persistent component and a transitory component (e.g., Hall and Mishkin, 1982, Abowd and Card, 1989):

$$
\begin{aligned}
w_{nt} &= f_n + y_{nt}^P + y_{nt}^T, \quad n = 1, ..., N, \quad t = 1, ..., T, && (1)\\
y_{nt}^P &= y_{nt-1}^P + \varepsilon_{nt}, \quad t \geq 2, && (2)\\
y_{nt}^T &= \eta_{nt}, && (3)\\
\eta_{n1} &= \eta_{nT} = 0, && (4)
\end{aligned}
$$

where w_{nt} is the residual of a regression of individual log earnings on a set of strictly exogenous regressors,[7] f_n is the fixed effect, y_{nt}^P is the persistent component, usually modelled as a random walk, and y_{nt}^T is the transitory shock/measurement error, modelled as a white noise. Innovations ε_{nt} and η_{nt} are mutually independent and independent over time.

The model considered in Horowitz and Markatou (1996) is a simplified version of model (1)-(4), without the permanent component ($\varepsilon_{nt} = 0$), and with i.i.d. transitory shocks η_{nt}, which in addition are assumed symmetrically distributed.

Unlike Horowitz and Markatou (1996), we shall not attempt to estimate the distribution of the fixed effect f_n. The existence of an autoregressive component makes earnings levels depend on an initial condition that may be correlated with f_n. As is usual in the literature on earnings dynamics (e.g., Abowd and Card, 1989, Meghir and Pistaferri, 2004), we instead difference out unobserved individual fixed effects. For example, setting $T = 4$ to simplify the presentation:

$$
\begin{pmatrix} w_{n2} - w_{n1} \\ w_{n3} - w_{n2} \\ w_{n4} - w_{n3} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \eta_{n2} \\ \eta_{n3} \end{pmatrix} + \begin{pmatrix} \varepsilon_{n2} \\ \varepsilon_{n3} \\ \varepsilon_{n4} \end{pmatrix}. \qquad (5)
$$

This model satisfies our setup with $L = T - 1 = 3$, $K = 2T - 3 = 5$, and

$$
\begin{aligned}
\mathbf{Y}_n &= (w_{n2} - w_{n1}, w_{n3} - w_{n2}, w_{n4} - w_{n3})^{\mathrm{T}}, \\
\mathbf{X}_n &= (\eta_{n2}, \eta_{n3}, \varepsilon_{n2}, \varepsilon_{n3}, \varepsilon_{n4})^{\mathrm{T}}, \\
\mathbf{A} &= \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ -1 & 1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 1 \end{pmatrix}.
\end{aligned}
$$

Most of the literature on earnings dynamics focuses on estimating the variances of permanent and transitory shocks in models similar to this one. In comparison, we here

[7] See Horowitz and Markatou (1996) for a formal treatment of cases where the dependent variable is the residual of a prior regression. Their analysis supposes strictly exogenous and bounded regressors, and root-N consistent estimates of the regression coefficients.

address the more difficult task of nonparametrically estimating the full distributions of these shocks.

2.2 Assumptions

Given the assumptions of linearity and independence, it will be convenient to work with characteristic functions. We make the following additional assumption.

Assumption A2 *The characteristic functions of factor variables $X_1, ..., X_K$ have no real zeros.*

This assumption is very common in the literature on nonparametric deconvolution (see Schennach, 2004a, and references therein). The characteristic functions may have complex zeros if factors have bounded support. Real zeros arise in the case of symmetric, bounded distributions, such as the uniform.

Next, let $\mathbf{A}_k$ denote the kth column of matrix $\mathbf{A}$, for $k \in \{1, ..., K\}$. Then,

$$\mathbf{Y} = \mathbf{AX} = \sum_{k=1}^{K} \mathbf{A}_k X_k,$$

and the variance-covariance matrix of $\mathbf{Y}$ is thus

$$\mathrm{Var}\left(\mathbf{Y}\right) = \mathbf{A}\,\mathrm{Var}\left(\mathbf{X}\right)\mathbf{A}^{\mathrm{T}} = \sum_{k=1}^{K} \mathrm{Var}(X_k)\mathbf{A}_k\mathbf{A}_k^{\mathrm{T}}, \tag{6}$$

as factors are independent, hence uncorrelated.

Let vech be the matrix operator that acts on symmetric matrices like the standard vec operator except that it only selects the components below or on the diagonal. For example, if $\mathbf{B} = [\mathbf{b}_{ij}]$ is 3×3 symmetric:

$$\begin{aligned} \mathrm{vec}\left(\mathbf{B}\right) &= \left(b_{11}, b_{12}, b_{13}, b_{12}, b_{22}, b_{23}, b_{13}, b_{23}, b_{33}\right), \\ \mathrm{vech}\left(\mathbf{B}\right) &= \left(b_{11}, b_{12}, b_{13}, b_{22}, b_{23}, b_{33}\right). \end{aligned}$$

As $\mathrm{Var}\left(\mathbf{Y}\right)$ is symmetric, one can reexpress the set of second-order restrictions in (6) as

$$\begin{aligned} \mathrm{vech}\left(\mathrm{Var}\left(\mathbf{Y}\right)\right) &= \sum_{k=1}^{K} \mathrm{Var}(X_k)\,\mathrm{vech}\left(\mathbf{A}_k\mathbf{A}_k^{\mathrm{T}}\right) \\ &= \mathbf{Q}\begin{pmatrix} \mathrm{Var}(X_1) \\ \vdots \\ \mathrm{Var}(X_K) \end{pmatrix}, \end{aligned} \tag{7}$$

where

$$\mathbf{Q} = \left[\text{vech}\left(\mathbf{A}_1\mathbf{A}_1^{\mathrm{T}}\right), ..., \text{vech}\left(\mathbf{A}_K\mathbf{A}_K^{\mathrm{T}}\right)\right].$$

Matrix $\mathbf{Q}$ has $L(L+1)/2$ rows and K columns. A generic row is $[a_{\ell 1}a_{m1}, ..., a_{\ell K}a_{mK}]$ for $\ell, m = 1, ..., L, \ell \leq m$.

Given $\mathbf{A}$, factor variances are obviously identifiable only if the following assumption holds true.

Assumption A3 *Matrix* $\mathbf{Q}$ *has full column rank* $K \leq L(L+1)/2$.

Note that $\mathbf{Q}$ could have rank less than K if two columns of $\mathbf{A}$ were proportional, a case that is ruled out by Assumption A1. Also, it is easily seen that Assumption A3 is equivalent to assuming that:

$$\text{rank}\left([\mathbf{A}_1 \otimes \mathbf{A}_1, ..., \mathbf{A}_K \otimes \mathbf{A}_K]\right) = K,$$

where $\otimes$ denotes the Kronecker product, and where matrix $[\mathbf{A}_1 \otimes \mathbf{A}_1, ..., \mathbf{A}_K \otimes \mathbf{A}_K]$ is sometimes referred to as the Kathri-Rao matrix product of $\mathbf{A}$ by itself (Kathri and C. R. Rao, 1968).

2.3 Examples

Example 1: The classical measurement error model.

$$\left\{\begin{array}{rcl} Y_1 &=& \alpha X + U_1, \\ Y_2 &=& X + U_2, \end{array}\right. \tag{8}$$

has

$$\begin{array}{rcl} \mathbf{Y} &=& (Y_1, Y_2)^{\mathrm{T}}, \quad \mathbf{X} = (X, U_1, U_2)^{\mathrm{T}}, \\ \mathbf{A} &=& \left(\begin{array}{ccc} \alpha & 1 & 0 \\ 1 & 0 & 1 \end{array}\right), \quad \mathbf{Q} = \left(\begin{array}{ccc} \alpha^2 & 1 & 0 \\ \alpha & 0 & 0 \\ 1 & 0 & 1 \end{array}\right). \end{array}$$

So $\mathbf{Q}$ has full rank 3 unless $\alpha = 0$, in which case the first and third columns of $\mathbf{A}$ are identical. The identification of α in model (8) was studied in Reiersol (1950).

Example 2: A simple spatial model.

$$\left\{\begin{array}{rcl} Y_1 &=& X_1 + \rho X_2 + \rho X_3 + U_1, \\ Y_2 &=& \rho X_1 + X_2 + \rho X_3 + U_2, \\ Y_3 &=& \rho X_1 + \rho X_2 + X_3 + U_3, \end{array}\right. \tag{9}$$

has

$$\mathbf{Y} = (Y_1, Y_2, Y_3)^{\mathrm{T}}, \quad \mathbf{X} = (X_1, X_2, X_3, U_1, U_2, U_3)^{\mathrm{T}},$$

$$\mathbf{A} = \begin{pmatrix} 1 & \rho & \rho & 1 & 0 & 0 \\ \rho & 1 & \rho & 0 & 1 & 0 \\ \rho & \rho & 1 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{Q} = \begin{pmatrix} 1 & \rho^2 & \rho^2 & 1 & 0 & 0 \\ \rho & \rho & \rho^2 & 0 & 0 & 0 \\ \rho & \rho^2 & \rho & 0 & 0 & 0 \\ \rho^2 & 1 & \rho^2 & 0 & 1 & 0 \\ \rho^2 & \rho & \rho & 0 & 0 & 0 \\ \rho^2 & \rho^2 & 1 & 0 & 0 & 1 \end{pmatrix}.$$

One verifies that $\mathbf{Q}$ has rank 6 unless $\rho \in \{-2, 0, 1\}$. If $\rho = 0$ or $\rho = 1$ then some columns of $\mathbf{A}$ are proportional, so $\mathbf{Q}$ does not have full column rank. If $\rho = -2$, $\mathbf{Q}$ has rank strictly less than $K = 6$, although any two columns of $\mathbf{A}$ are linearly independent.

3 Identification of factor distributions

In this section, we derive the identifying restrictions that will be used for estimation in the next section. A sketch of the arguments below can be found in Székely and Rao (2000, remark 6, p. 200).

3.1 Notation

Let us denote the characteristic function (c.f.) of X_k as

$$\begin{aligned} \varphi_{X_k}(\tau) &= \mathbb{E}\left(e^{i\tau X_k}\right), \quad \tau \in \mathbb{R}, \\ &= \int e^{i\tau x} f_{X_k}(x)\, dx, \quad \tau \in \mathbb{R}, \end{aligned}$$

where f_{X_k} is the probability density function (p.d.f.) of X_k, and $i = \sqrt{-1}$.

As previously noted, X_k having finite variance, φ_{X_k} is well defined and everywhere twice differentiable. Moreover, as φ_{X_k} is nowhere vanishing, the cumulant generating function (c.g.f.) of X_k, i.e. the logarithm of its characteristic function, is also well defined and everywhere twice differentiable.[8] We denote the c.g.f. of X_k as

$$\begin{aligned} \kappa_{X_k}(\tau) &= \ln \varphi_{X_k}(\tau), \quad \tau \in \mathbb{R}, \\ &= \ln\left[\mathbb{E}\left(e^{i\tau X_k}\right)\right], \quad \tau \in \mathbb{R}. \end{aligned}$$

The density is uniquely determined by the characteristic function by the inverse Fourier transformation (for x in the support of X_k):

$$f_{X_k}(x) = \frac{1}{2\pi} \int e^{-i\tau x} \varphi_{X_k}(\tau)\, d\tau. \tag{10}$$

[8]Note that, for identification, it suffices to assume that the set of zeros of factor c.f.'s is Lebesgue-negligible, see e.g. Carrasco and Florens (2007).

In particular, it follows from (10) that if the c.f. of X_k is identified, then its p.d.f. is also identified.

We similarly denote the multivariate c.f. and c.g.f. of $\mathbf{Y}$, which are functions that map $\mathbb{R}^L$ into the complex plane, as

$$\begin{aligned}\varphi_{\mathbf{Y}}(\mathbf{t}) &\equiv \mathbb{E}\left(e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}}\right), \mathbf{t} \in \mathbb{R}^L, \\ \kappa_{\mathbf{Y}}(\mathbf{t}) &\equiv \ln\left[\mathbb{E}\left(e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}}\right)\right], \mathbf{t} \in \mathbb{R}^L.\end{aligned}$$

We will make extensive use of the derivatives of $\kappa_{\mathbf{Y}}$. Let $\partial_\ell \kappa_{\mathbf{Y}}(\mathbf{t})$ denote the ℓth partial derivative of $\kappa_{\mathbf{Y}}(\mathbf{t})$, and $\partial^2_{\ell m}\kappa_{\mathbf{Y}}(\mathbf{t})$ the second-order partial derivative of $\kappa_{\mathbf{Y}}(\mathbf{t})$ with respect to t_ℓ and t_m. We also denote as $\nabla\kappa_{\mathbf{Y}}(\mathbf{t}) = [\partial_\ell \kappa_{\mathbf{Y}}(\mathbf{t})]$ the gradient vector, and as $\nabla\nabla^{\mathrm{T}}\kappa_{\mathbf{Y}}(\mathbf{t}) = [\partial^2_{\ell m}\kappa_{\mathbf{Y}}(\mathbf{t})]$ the Hessian matrix.

3.2 Identifying restrictions

Because of a well-known property of characteristic functions, the c.g.f. of a linear combination of independent random variables is equal to the same linear combination of their c.g.f.'s. Specifically, as factors are assumed mutually independent, for all $\mathbf{t} = (t_1, ..., t_L) \in \mathbb{R}^L$,

$$\kappa_{\mathbf{Y}}(\mathbf{t}) = \sum_{k=1}^{K} \kappa_{X_k}\left(\mathbf{t}^{\mathrm{T}}\mathbf{A}_k\right). \tag{11}$$

First-differentiating equation (11) yields:

$$\nabla\kappa_{\mathbf{Y}}(\mathbf{t}) = \sum_{k=1}^{K} \kappa'_{X_k}\left(\mathbf{t}^{\mathrm{T}}\mathbf{A}_k\right)\mathbf{A}_k.$$

If $K > L$ there are more functions κ'_{X_k} than partial derivatives $\partial_\ell\kappa_{\mathbf{Y}}$. To obtain an invertible system, we differentiate once more:

$$\nabla\nabla^{\mathrm{T}}\kappa_{\mathbf{Y}}(\mathbf{t}) = \sum_{k=1}^{K} \kappa''_{X_k}\left(\mathbf{t}^{\mathrm{T}}\mathbf{A}_k\right)\mathbf{A}_k\mathbf{A}_k^{\mathrm{T}}.$$

As $\nabla\nabla^{\mathrm{T}}\kappa_{\mathbf{Y}}(\mathbf{t})$ is symmetric, we may as well rewrite this set of restrictions as

$$\begin{aligned}\mathrm{vech}\left(\nabla\nabla^{\mathrm{T}}\kappa_{\mathbf{Y}}(\mathbf{t})\right) &= \sum_{k=1}^{K} \kappa''_{X_k}\left(\mathbf{t}^{\mathrm{T}}\mathbf{A}_k\right)\mathrm{vech}\left(\mathbf{A}_k\mathbf{A}_k^{\mathrm{T}}\right) \\ &= \mathbf{Q}\begin{pmatrix}\kappa''_{X_1}\left(\mathbf{t}^{\mathrm{T}}\mathbf{A}_1\right) \\ \vdots \\ \kappa''_{X_K}\left(\mathbf{t}^{\mathrm{T}}\mathbf{A}_K\right)\end{pmatrix}. \end{aligned} \tag{12}$$

Note that, evaluated at $\mathbf{t} = 0$, equation (12) yields the covariance restrictions (7). The independence assumption on factor variables, which is more restrictive than uncorrelatedness, yields many more restrictions on factor c.g.f.'s, one for each value of $\mathbf{t} \in \mathbb{R}^L$.

Equation (12) shows that, if $\mathbf{Q}$ has full column rank (Assumption A3) and if factors are independent and not only uncorrelated, the second derivatives of the c.g.f.'s of factor variables are identified. Namely, inverting (12), we obtain

$$\begin{pmatrix} \kappa''_{X_1}\left(\mathbf{t}^{\mathrm{T}}\mathbf{A}_1\right) \\ \vdots \\ \kappa''_{X_K}\left(\mathbf{t}^{\mathrm{T}}\mathbf{A}_K\right) \end{pmatrix} = \mathbf{Q}^{-}\,\mathrm{vech}\left(\nabla\nabla^{\mathrm{T}}\kappa_{\mathbf{Y}}\left(\mathbf{t}\right)\right), \tag{13}$$

where $\mathbf{Q}^{-}$ is a pseudo inverse of $\mathbf{Q}$, e.g. $\mathbf{Q}^{-} = \left(\mathbf{Q}^{\mathrm{T}}\mathbf{Q}\right)^{-1}\mathbf{Q}^{\mathrm{T}}$.

Let $\tau \in \mathbb{R}$, and let $k \in \{1, ..., K\}$. System (13) offers many overidentifying restrictions for $\kappa''_{X_k}(\tau)$. Indeed, there are many ways to choose $\mathbf{t}$ such that $\mathbf{t}^{\mathrm{T}}\mathbf{A}_k = \tau$. A possible choice is to take $\mathbf{t} = \frac{\tau\mathbf{A}_k}{\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k}$. We will provide a motivation for this choice based on asymptotic arguments in Section 5. We shall refer to $\mathbf{A}_k$ as our "preferred direction of integration". However, this direction is by no means unique. Any choice of $\mathbf{t} = \frac{\tau\boldsymbol{\theta}}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k}$, for $\boldsymbol{\theta} \in \mathbb{R}^L \backslash \{0\}$, will also work.

Let $\mathbf{Q}_k^{-}$ denote the kth row of $\mathbf{Q}^{-}$. For any direction $\boldsymbol{\theta} \in \mathbb{R}^L \backslash \{0\}$ and $\tau \in \mathbb{R}$,

$$\kappa''_{X_k}(\tau) = \mathbf{Q}_k^{-}\,\mathrm{vech}\left(\nabla\nabla^{\mathrm{T}}\kappa_{\mathbf{Y}}\left(\frac{\tau\boldsymbol{\theta}}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k}\right)\right).$$

The c.g.f. of X_k then follows by integrating twice this equation. Two constants of integrations are readily available: $\kappa'_{X_k}(0) = i\mathbb{E}X_k = 0$, as factors have zero means, and $\kappa_{X_k}(0) = 0$, because a c.f. is equal to one at zero. Hence,

$$\kappa_{X_k}(\tau) = \int_0^{\tau}\int_0^{u} \mathbf{Q}_k^{-}\,\mathrm{vech}\left(\nabla\nabla^{\mathrm{T}}\kappa_{\mathbf{Y}}\left(\frac{v\boldsymbol{\theta}}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k}\right)\right) dvdu. \tag{14}$$

Equation (14) can be used for estimating factor characteristic functions and densities, as we explain in the next section.

3.3 Example 1: The measurement error model

In the case of model (8), we have:

$$\kappa_{\mathbf{Y}}(t_1, t_2) = \kappa_X(\alpha t_1 + t_2) + \kappa_{U_1}(t_1) + \kappa_{U_2}(t_2),$$

and

$$\mathrm{vech}\left(\nabla\nabla^{\mathrm{T}}\kappa_{\mathbf{Y}}\left(\mathbf{t}\right)\right) = \begin{pmatrix} \partial^2_{11}\kappa_{\mathbf{Y}}(t_1, t_2) \\ \partial^2_{12}\kappa_{\mathbf{Y}}(t_1, t_2) \\ \partial^2_{22}\kappa_{\mathbf{Y}}(t_1, t_2) \end{pmatrix} = \begin{pmatrix} \alpha^2 & 1 & 0 \\ \alpha & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \kappa''_X(\alpha t_1 + t_2) \\ \kappa''_{U_1}(t_1) \\ \kappa''_{U_2}(t_2) \end{pmatrix}.$$

This yields:

$$\begin{pmatrix} \kappa''_X(\alpha t_1 + t_2) \\ \kappa''_{U_1}(t_1) \\ \kappa''_{U_2}(t_2) \end{pmatrix} = \begin{pmatrix} 0 & \alpha^{-1} & 0 \\ 1 & -\alpha & 0 \\ 0 & -\alpha^{-1} & 1 \end{pmatrix} \begin{pmatrix} \partial^2_{11}\kappa_{\mathbf{Y}}(t_1, t_2) \\ \partial^2_{12}\kappa_{\mathbf{Y}}(t_1, t_2) \\ \partial^2_{22}\kappa_{\mathbf{Y}}(t_1, t_2) \end{pmatrix}. \tag{15}$$

The common factor. Let us set $\alpha = 1$ for comparability with previous results in the literature. Then, $\frac{\mathbf{A}_1}{\mathbf{A}_1^{\mathrm{T}}\mathbf{A}_1} = \left(\frac{1}{2}, \frac{1}{2}\right)^{\mathrm{T}}$ and $\kappa_X(\tau)$ can be thus represented as:

$$\kappa_X(\tau) = \int_0^\tau \int_0^u \partial^2_{12}\kappa_{\mathbf{Y}}\left(\frac{v}{2}, \frac{v}{2}\right) dvdu. \tag{16}$$

Alternatively, using $\boldsymbol{\theta} = (0, 1)$ as direction of integration yields:

$$\begin{aligned} \kappa_X(\tau) &= \int_0^\tau \int_0^u \partial^2_{12}\kappa_{\mathbf{Y}}(0, v)\, dvdu \\ &= \int_0^\tau \partial_1\kappa_{\mathbf{Y}}(0, u)\, du. \end{aligned} \tag{17}$$

This is the expression used in Li and Vuong (1998) and Schennach (2004a, 2004b), which requires only one single differentiation and integration. However, it is not true in general that the double integral in (14) can be simplified into a simple integral by using an appropriate direction of integration, as Example 2 below will show.

Remark that, for any α, our preferred direction of integration is $\mathbf{A}_1 = (\alpha, 1)^{\mathrm{T}}$. So, when $|\alpha|$ gets larger, Y_1 contributes more to κ_X relative to Y_2. This makes intuitive sense, as Y_1 becomes more informative about X.

Errors. For U_1, $\frac{\mathbf{A}_2}{\mathbf{A}_2^{\mathrm{T}}\mathbf{A}_2} = (1, 0)^{\mathrm{T}}$, and

$$\begin{aligned} \kappa_{U_1}(\tau) &= \int_0^\tau \int_0^u \left(\partial^2_{11} - \partial^2_{12}\right)\kappa_{\mathbf{Y}}(v, 0)\, dvdu \\ &= \kappa_{\mathbf{Y}}(\tau, 0) - \int_0^\tau \partial_2\kappa_{\mathbf{Y}}(u, 0)\, du \\ &= \kappa_{Y_1}(\tau) - \int_0^\tau \partial_2\kappa_{\mathbf{Y}}(u, 0)\, du. \end{aligned} \tag{18}$$

Li and Vuong use a slightly different formula:

$$\kappa_{U_1}(\tau) = \kappa_{Y_1}(\tau) - \int_0^\tau \partial_1\kappa_{\mathbf{Y}}(0, u)\, du. \tag{19}$$

3.4 Example 2: The spatial model

We then reconsider the case of model (9). We obtain, for the first factor:

$$\kappa''_{X_1}(t_1 + \rho t_2 + \rho t_3) = \frac{\partial^2_{12} + \partial^2_{13} - (\rho + 1)\partial^2_{23}}{(1 - \rho)(\rho + 2)\rho}\kappa_{\mathbf{Y}}(t_1, t_2, t_3),$$

and, for the first error:

$$\kappa''_{U_1}(t_1) = \frac{\partial^2_{11} - (\rho + \rho^2 + 1)(\partial^2_{12} + \partial^2_{13}) + (2\rho + 1)\partial^2_{23}}{(\rho + 2)\rho} \kappa_{\mathbf{Y}}(t_1, t_2, t_3).$$

Set $t_1 = \tau - \rho t_2 - \rho t_3$. Then,

$$\kappa_{X_1}(\tau) = \int_0^\tau \int_0^u \left[\frac{\partial^2_{12} + \partial^2_{13} - (\rho + 1)\partial^2_{23}}{(1 - \rho)(\rho + 2)\rho} \kappa_{\mathbf{Y}}(v - \rho t_2 - \rho t_3, t_2, t_3) \right] dvdu.$$

One easily verifies that, even when $t_2 = t_3 = 0$, the double integral does not simplify to a single integral.

Lastly, in this example our preferred solution is

$$\kappa_{X_1}(\tau) = \int_0^\tau \int_0^u \left[\frac{\partial^2_{12} + \partial^2_{13} - (\rho + 1)\partial^2_{23}}{(1 - \rho)(\rho + 2)\rho} \kappa_{\mathbf{Y}} \left(v \frac{(1, \rho, \rho)^{\mathrm{T}}}{1 + 2\rho^2} \right) \right] dvdu.$$

Here also, this solution has intuitive appeal when $|\rho|$ increases.

4 Estimation

We here introduce our estimator of factor densities. Asymptotic theory is in the next section.

4.1 Characteristic functions

Given an i.i.d. sample $\{\mathbf{Y}_1, ..., \mathbf{Y}_N\}$ of size N, with $\mathbf{Y}_n = (Y_{n1}, ..., Y_{nL})^{\mathrm{T}}$, we first estimate $\kappa_{\mathbf{Y}}$ and its derivatives by empirical analogs, replacing mathematical expectations by arithmetic means:

$$\widehat{\kappa}_{\mathbf{Y}}(\mathbf{t}) = \ln \left(\mathbb{E}_N \left[e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}} \right] \right),$$

$$\widehat{\partial_\ell \kappa_{\mathbf{Y}}}(\mathbf{t}) = i \frac{\mathbb{E}_N \left[Y_\ell e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}} \right]}{\mathbb{E}_N \left[e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}} \right]} = \partial_\ell \widehat{\kappa}_{\mathbf{Y}}(\mathbf{t}),$$

and

$$\widehat{\partial^2_{\ell m} \kappa_{\mathbf{Y}}}(\mathbf{t}) = -\frac{\mathbb{E}_N \left[Y_\ell Y_m e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}} \right]}{\mathbb{E}_N \left[e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}} \right]} + \frac{\mathbb{E}_N \left[Y_\ell e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}} \right]}{\mathbb{E}_N \left[e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}} \right]} \frac{\mathbb{E}_N \left[Y_m e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}} \right]}{\mathbb{E}_N \left[e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}} \right]} = \partial^2_{\ell m} \widehat{\kappa}_{\mathbf{Y}}(t),$$

where $\mathbb{E}_N$ denotes the empirical expectation operator: $\mathbb{E}_N g(\mathbf{Y}) = \frac{1}{N} \sum_{n=1}^N g(\mathbf{Y}_n)$. For example:

$$\exp\left(\widehat{\kappa}_{\mathbf{Y}}(\mathbf{t})\right) = \frac{1}{N} \sum_{n=1}^N e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}_n}, \quad \mathbf{t} \in \mathbb{R}^L,$$

is the empirical characteristic function of $\mathbf{Y}$.

Then, for any $\boldsymbol{\theta} \in \mathbb{R}^L$ (the direction of integration), we estimate factor cumulant generating functions as:

$$\widehat{\kappa}_{X_k}(\tau) = \int_0^\tau \int_0^u \mathbf{Q}_k^- \operatorname{vech}\left[\nabla\nabla^{\mathrm{T}}\widehat{\kappa}_{\mathbf{Y}}\left(\frac{v\boldsymbol{\theta}}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k}\right)\right] dvdu, \quad \tau \in \mathbb{R}. \tag{20}$$

Equivalently, the characteristic function of X_k is estimated as

$$\widehat{\varphi}_{X_k}(\tau) = \exp\left(\int_0^\tau \int_0^u \mathbf{Q}_k^- \operatorname{vech}\left[\nabla\nabla^{\mathrm{T}}\widehat{\kappa}_{\mathbf{Y}}\left(\frac{v\boldsymbol{\theta}}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k}\right)\right] dvdu\right), \quad \tau \in \mathbb{R}. \tag{21}$$

Our estimator of φ_{X_k} depends on the direction of integration $\boldsymbol{\theta}$. We suggest to choose $\boldsymbol{\theta} = \mathbf{A}_k$ (our preferred choice). More generally, it is possible to average various alternative estimators over a distribution W of $\boldsymbol{\theta}$'s:

$$\widehat{\varphi}_{X_k}(\tau) = \exp\left(\int_0^\tau \int_0^u \mathbf{Q}_k^- \operatorname{vech}\left[\int \nabla\nabla^{\mathrm{T}}\widehat{\kappa}_{\mathbf{Y}}\left(\frac{v\boldsymbol{\theta}}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k}\right) dW(\boldsymbol{\theta})\right] dvdu\right). \tag{22}$$

In the Monte-Carlo section, we will show the performance of an estimator based on $W = \sum_j \delta_{\boldsymbol{\theta}_j}$, where the $\boldsymbol{\theta}_j$'s are drawn from $\mathcal{N}(\mathbf{A}_k, \sigma^2\mathbf{I}_L)$ for some σ.

4.2 Density functions

The probability distribution function (p.d.f.) of X_k is obtained from its characteristic function by the inverse Fourier transformation (10). However, it is well-known that the integral in (10) does not converge when the characteristic function is replaced by its empirical analog, using $\widehat{\varphi}_{X_k}$ instead of φ_{X_k} (e.g., Horowitz, 1998, p. 104).

To ensure convergence we truncate the integral on a compact interval $[-T_N, T_N]$, where T_N tends to infinity with the sample size N at a rate that will be discussed in the next section. The p.d.f. of X_k is then estimated as

$$\widehat{f}_{X_k}(x) = \frac{1}{2\pi}\int \varphi_H\left(\frac{\tau}{T_N}\right) e^{-i\tau x}\widehat{\varphi}_{X_k}(\tau)d\tau, \tag{23}$$

where $\widehat{\varphi}_{X_k}(\tau)$ is given by (22). In equation (23), φ_H is a function supported on $[-1, 1]$ that is the Fourier transform of a kernel H of even order: $\varphi_H(u) = \int e^{iuv}H(v)dv$.[9]

The kernel H allows to smooth the estimation of the density, especially in the tails. We shall use the second-order kernel

$$H_2(v) = \frac{48\cos(x)}{\pi x^4}\left(1 - \frac{15}{x^2}\right) - \frac{144\sin(x)}{\pi x^5}\left(2 - \frac{5}{x^2}\right),$$

which corresponds to:

$$\varphi_{H_2}(u) = \left(1 - u^2\right)^3 \cdot \mathbf{1}\{u \in [-1, 1]\}.$$

[9] A kernel of order q is a function H, not necessarily nonnegative, such that $v^k H(v)$ is integrable for all $k \leq q$, $\int v^k H(v)dv = 0$ for all $k \leq q-1$, and $\int v^q H(v)dv \neq 0$. See, e.g., P. Rao (1983), p. 40.

The second-order kernel H_2 has often been used in the deconvolution literature (see, e.g., Delaigle and Gijbels, 2002, and references therein). Higher-order kernels may also be used in place of H_2, such as the infinite-order kernel $H_\infty(v) = \sin(v)/v$ used in Li and Vuong (1998), that yields $\varphi_{H_\infty}(u) = \mathbf{1}\{u \in [-1, 1]\}$. Higher-order kernels reduce the bias of the density estimate at the cost of higher variance.

Numerical issues. Computing the density estimator $\widehat{f}_{X_k}$ in practice requires integrating three times: twice to recover the cumulant generating function of X_k, and once to perform the inverse Fourier transformation. It is well-known that usual fast integration techniques such as Romberg's method may give very misleading results when computing the inverse Fourier transform, because the function to integrate is strongly oscillating (Delaigle and Gijbels, 2007). For this reason, we use as integration method the slow but reliable trapezoid rule, with a large number of nods. In our experiments, using 200 equidistant nods (over the interval $[-T_N, T_N]$, where the choice of T_N will be discussed below) gave very good approximations.

In addition, our estimator of the characteristic function of X_k given by (21) or (22) does not guarantee that $\left|\widehat{\varphi}_{X_k}\right|$ is less than one, as a proper c.f. should be. In practice, we suggest to set $\widehat{\varphi}_{X_k}(\tau) = 0$ whenever $\left|\widehat{\varphi}_{X_k}(\tau)\right| > 1.1$.

5 Asymptotic theory

In this section, we study the asymptotic properties of the estimator and show that $\widehat{f}_{X_k}$ given by (23) is a uniformly consistent estimator of f_{X_k}, for all $k = 1, ..., K$. All mathematical proofs are in the appendix.

The study of the asymptotic properties of our estimator is in two steps. First, we characterize the properties of the estimator of factor characteristic functions $\widehat{\varphi}_{X_k}(\tau)$. Then, we study their inverse Fourier transforms, that is the density estimators $\widehat{f}_{X_k}$.

We characterize upper bounds to the rates of uniform convergence of the estimators. Besides providing sufficient conditions for consistency, this asymptotic analysis is useful to understand the properties of factor distributions which improve convergence. Moreover, it will allow us to motivate our preferred direction of integration.

5.1 Characteristic functions

The c.f. estimator $\widehat{\varphi}_{X_k}$ involves means of functions of measurements of the form $e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}}$, $Y_\ell e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}}$, or $Y_\ell Y_m e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}}$. Because $\widehat{\varphi}_{X_k}(\tau)$ is then obtained by integration over $\mathbf{t}$, a uniform

consistency result is needed for $\mathbb{E}_N e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}}$, $\mathbb{E}_N Y_\ell e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}}$, and $\mathbb{E}_N Y_\ell Y_m e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}}$. The next lemma extends Horowitz and Markatou's Lemma 1, that deals with $\mathbb{E}_N e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}}$, to empirical means of $Y_\ell e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}}$ or $Y_\ell Y_m e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}}$.

Li and Vuong (1998)[10] assume that factor distributions have bounded support. This may be quite restrictive in some cases, as in the case of earnings data which display particularly wide ranges of values. Here we relax this assumption and allow for unbounded support.

For any vector $\mathbf{t} = (t_1, ..., t_L)^{\mathrm{T}} \in \mathbb{R}^L$ $(L \geq 1)$, we use the sup norm: $|\mathbf{t}| = \max_\ell |t_\ell|$.

Lemma 1 *Let X be a scalar random variable and let $\mathbf{Y}$ be a vector of L random variables. Define $Z = \left(X, \mathbf{Y}^{\mathrm{T}}\right)^{\mathrm{T}}$. 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 the moment generating functions of X^2 and $|X\mathbf{Y}|$ exist in some neighborhood of 0. Define $f_t(x, \mathbf{y}) = xe^{i\mathbf{t}^{\mathrm{T}}\mathbf{y}}$, for $\mathbf{t} \in \mathbb{R}^L$. Then,*

$$\sup_{|\mathbf{t}| \leq T_N} |\mathbb{E}_N f_{\mathbf{t}} - \mathbb{E} f_{\mathbf{t}}| = O(\varepsilon_N) \ a.s.$$

if T_N and ε_N are chosen such that

$$\begin{aligned} T_N &= BN^{\frac{\delta}{2}}, \ B, \delta > 0, \\ \varepsilon_N &= A\frac{\ln N}{\sqrt{N}}, \ A > 8\sqrt{2 + L(1+\delta)}. \end{aligned}$$

Lemma 1 shows that $(\sqrt{N}/\ln N)$ is an upper bound to the rate of convergence on $[-T_N, T_N]$, provided that T_N does not grow faster than some power of N. The proof requires that the moment generating functions (m.g.f.) of X^2 and $|X\mathbf{Y}|$ exist in some neighborhood of the origin. This implies in particular that every moment of X^2 and $|X\mathbf{Y}|$ is finite. This is a restrictive assumption, which could be relaxed by assuming the existence of the first $J \geq 2$ moments of $|X|$ and $|X\mathbf{Y}|$, at the cost of a lower rate of convergence in Lemma 1.[11]

Remark also that existence of the moment generating function is less restrictive than assuming that factors have bounded support. Under this assumption, Li and Vuong

[10] Horowitz and Markatou (1996) do not assume support boundedness. However, as pointed out by Hu and Ridder (2008), their reference (p. 164) to theorem 2.37 of Pollard (1984) is inexact, and support boundedness is implicitly needed.

[11] When only the existence of the first $J \geq 2$ moments of $|X|$ and $|X\mathbf{Y}|$ is assumed, a strict upper bound to the rate of convergence in Lemma 1 is given by: $N^{\frac{1}{2}-\frac{1}{2J}}$.

(1998) obtain a better bound $O\left[\left(\frac{\ln \ln N}{N}\right)^{\frac{1}{2}}\right]$, as support boundedness allows to use the law of iterated logarithm as in Csörgő (1981, Theorem 1).[12]

Lemma 1 applies to $\mathbb{E}_N\left[Y_\ell e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}}\right]$ and $\mathbb{E}_N\left[Y_\ell Y_m e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}}\right]$, for every $\ell, m = 1, ..., L$, if the following assumption holds.

Assumption A4 *All variables Y_ℓ^2, $Y_\ell Y_m$ and $Y_\ell Y_m Y_j$ for ℓ, m, j in $\{1, ..., L\}$, have their m.g.f.'s existing in some neighborhood around 0.*

Assuming that this assumption holds, the following uniform consistency result for the characteristic functions of factors follows.

Theorem 1 *Suppose that there exists an integrable, decreasing function $g : \mathbb{R}^+ \to [0, 1]$, such that $|\varphi_{\mathbf{Y}}(\mathbf{t})| \geq g(|\mathbf{t}|)$ for $|\mathbf{t}|$ large enough. Then:*

$$\sup_{|\tau| \leq T_N} \left|\widehat{\varphi}_{X_k}(\tau) - \varphi_{X_k}(\tau)\right| = \left[\int g\left(T_N \left|\frac{\boldsymbol{\theta}}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k}\right|\right)^{-3} dW(\boldsymbol{\theta})\right] T_N^2 \varepsilon_N O(1) \quad a.s. \tag{24}$$

with ε_N and T_N as in Lemma 1, and with the additional restriction that the right-hand side in (24) is $o(1)$ for consistency.

Theorem 1 provides an argument for our preferred choice for the directions of integration $\boldsymbol{\theta} = \mathbf{A}_k$. As norms are equivalent in finite dimensional spaces and as g is decreasing, one can replace the sup norm of $\frac{\boldsymbol{\theta}}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k}$ by its Euclidian norm in $g\left(T_N \left|\frac{\boldsymbol{\theta}}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k}\right|\right)^{-3}$. This quantity is minimum for $\boldsymbol{\theta} = \mathbf{A}_k$, which minimizes the Euclidian norm of $\frac{\boldsymbol{\theta}}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k}$.[13] So, the fastest rate is attained when W assigns all mass to $\boldsymbol{\theta} = \mathbf{A}_k$. In Section 6, we will also provide finite sample evidence that supports this choice of direction of integration.[14]

In the rest of this section, for expositional simplicity, we consider the special case where W assigns all mass to one single $\boldsymbol{\theta}$, and redefine g such that (24) becomes

$$\sup_{|\tau| \leq T_N} \left|\widehat{\varphi}_{X_k}(\tau) - \varphi_{X_k}(\tau)\right| = \frac{T_N^2 \varepsilon_N}{g(T_N)^3} O(1) \quad \text{a.s.} \tag{25}$$

with the additional restriction that $\frac{T_N^2 \varepsilon_N}{g(T_N)^3}$ is $o(1)$ for consistency.

[12] Li and Vuong's argument, p. 146, that boundedness is not a strong assumption as it can be achieved by transforming $\mathbf{Y}$ suitably and assuming that the transformed $\mathbf{Y}$ follows the linear factor model, is a bit contrived. Economic theory usually strongly conditions the form of the model and statistical theory has little say in that construction process.

[13] Indeed, Cauchy-Schwarz inequality implies that $(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k)^2 \leq (\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta})(\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k)$, so $\frac{\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}}{(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k)^2} \geq \frac{\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k}{(\mathbf{A}_k^{\mathrm{T}}\mathbf{A}_k)^2}$.

[14] However, a drawback of our preferred direction of integration is that it is not invariant to a linear transformation of the model (as $\mathbf{BY} = \mathbf{BAX}$) unless the transformation is a rotation ($\mathbf{B}$ orthogonal). Hence, our preferred direction of integration depends on the model representation, which is arbitrary.

Theorem 1 shows that the rate of uniform convergence for $\widehat{\varphi}_{X_k}$ depends on the tail of the characteristic function of the vector of measurements, as characterized by function g, which measures the smoothness of the distribution of $\mathbf{Y}$.[15] Furthermore, the estimation error increases with T_N because of the integration step. More estimation errors add up if one has to integrate an estimate of κ''_{X_k} over a wider range. Hence, the rate of convergence decreases when T_N grows, both because of a wider range of integration and of division by small values of the characteristic function.

Lastly, it is instructive to compare the convergence rate of Theorem 1 to the convergence rates obtained in simpler models by other authors. The general form of our estimator proceeds from the second-order differentiation of $\kappa_{\mathbf{Y}}(\mathbf{t})$ and a subsequent double-integration.[16] If no differentiation were required (as in Horowitz and Markatou, 1996, or Delaigle *et al.*, 2008) or if first-order differentiation sufficed (as in Li and Vuong, 1998), then both $g(T_N)$ and T_N would be raised to a smaller power. This is why these authors obtain *faster* convergence rates of the estimators of factor characteristic functions. Note, however, that previous work on deconvolution focuses on much simpler models than the general multi-factor models that we consider. In the present case, differentiation and integration are likely to be necessary steps in the construction of the estimator.

Special cases. In a seminal contribution to the standard deconvolution theory with only one unknown factor distribution, Fan (1991) distinguished two classes of distributions: ordinary smooth, for which the c.f. converges to zero at a polynomial rate (e.g., Laplace or Gamma), and supersmooth distributions, for which the c.f. converges to zero at an exponential rate (e.g., normal). Here we illustrate the previous results in the case of ordinary smooth and supersmooth factor distributions.

Let us first consider the case where all factor distributions are ordinary smooth:

$$g(t) = t^{-\beta},\ t > 0,\ \beta > 1. \tag{26}$$

Taking $T_N = N^{\frac{\delta}{2}}$ and $\varepsilon_N = A \ln N/\sqrt{N}$, as in Lemma 1, we obtain

$$\frac{T_N^2 \varepsilon_N}{g(T_N)^3} = A \frac{\ln N}{N^{\frac{1}{2}-\left(1+\frac{3}{2}\beta\right)\delta}}, \tag{27}$$

[15] Technically, the estimation error on φ_{X_k} decreases with $g(T_N)$ because the estimator is based on the differentiation of $\kappa_{\mathbf{Y}}(\mathbf{t}) = \ln \varphi_{\mathbf{Y}}(\mathbf{t})$. A term $1/\left|\varphi_{\mathbf{Y}}(\mathbf{t})\right|$ thus appears, which is bounded by $1/g(|\mathbf{t}|)$, hence by $1/g(T_N)$ if $|\mathbf{t}| \leq T_N$ and $|\mathbf{t}|$ large enough.

[16] Hence, T_N is squared in (25) because of the double integration, and $g(T_N)$ is cubed because of the second-order differentiation (+2) of the logarithm (+1) of the characteristic function of $\mathbf{Y}$.

and for any $\delta < \frac{1}{2+3\beta}$, the estimator of factor c.f. converges uniformly on $[-T_N, T_N]$, an upper bound to its rate of convergence being given by (27). Hence, a smoother distribution of $\mathbf{Y}$ (a larger value of β) requires more trimming (a lower δ).

Let us then consider the case where factor distributions are supersmooth, so:

$$g(t) = t^{-\beta_0} e^{-\beta t^{1/\beta_1}}, \ t > 0, \ \beta, \beta_1 > 0. \tag{28}$$

Then, because of the restriction that $\frac{T_N^2 \varepsilon_N}{g(T_N)^3} = o(1)$ in Theorem 1, T_N cannot increase at polynomial rate any longer. One has to restrict T_N to a logarithmic function of N in order to ensure uniform convergence of $\widehat{\varphi}_{X_k}$ on $[-T_N, T_N]$. For example, taking $T_N = (\delta \ln N)^{\beta_1}$, the rate becomes:

$$\frac{T_N^2 \varepsilon_N}{g_Y(T_N)^3} = \left[A\delta^{(2+3\beta_0)\beta_1}\right] \frac{(\ln N)^{1+(2+3\beta_0)\beta_1}}{N^{\frac{1}{2}-3\beta\delta}}, \tag{29}$$

and the estimator is consistent if $\delta < \frac{1}{6\beta}$. Again, a smoother distribution (β larger) requires more trimming (lower δ).

5.2 Density functions

The following theorem gives conditions under which $\widehat{f}_{X_k}$ converges uniformly to f_{X_k} when the sample size tends to infinity.

Theorem 2 *Suppose that there exists an integrable, decreasing function* $\widetilde{g} : \mathbb{R}^+ \to [0,1]$ *such that* $|\varphi_{\mathbf{X}}(\boldsymbol{\tau})| = \left|\prod_{k=1}^K \varphi_{X_k}(\tau_k)\right| \geq \widetilde{g}(|\boldsymbol{\tau}|)$ *for* $\boldsymbol{\tau} = (\tau_1, ..., \tau_K)^{\mathrm{T}}$ *and for* $|\boldsymbol{\tau}| = \max(|\tau_k|)$ *large enough. Suppose also that there exist* K *integrable functions* $h_k : \mathbb{R}^+ \to [0,1]$ *such that* $h_k(|\tau|) \geq \left|\varphi_{X_k}(\tau)\right|$ *for all* τ. *Lastly, let* H *be a kernel of even order* $q \geq 2$ *with Fourier transform satisfying* $\varphi_H(t) = 0$ *for* $|t| > 1$. *Then:*

$$\begin{aligned} \sup_x \left|\widehat{f}_{X_k}(x) - f_{X_k}(x)\right| \ &= \ \frac{T_N^3 \varepsilon_N}{g(T_N)^3} O(1) \\ &\quad + \frac{C}{T_N^q} \int_{-T_N}^{T_N} v^q h_k(|v|) dv + 2 \int_{T_N}^{+\infty} h_k(v) dv \quad a.s. \end{aligned} \tag{30}$$

where $g(t) = \widetilde{g}(L|\mathbf{A}|t)$, *with* $|\mathbf{A}| = \max_{i,j}(|a_{ij}|)$, C *is a positive constant, and where* ε_N *and* T_N *are as in Lemma 1, with the additional restriction that* $\frac{T_N^3 \varepsilon_N}{g(T_N)^3} = o(1)$ *for consistency.*

The estimation error on f_{X_k}, in sup norm, has two components. The first component directly results from the estimation error on the characteristic function φ_{X_k}.[17] The second

[17] In particular, T_N is cubed in (30) instead of squared because of the integral in the inverse Fourier transform.

component results from the fact that the inverse Fourier transform yielding $\widehat{f}_{X_k}$ depends on the smoothing kernel φ_H. For example, with $\varphi_H(u) = \varphi_{H_\infty}(u) = \mathbf{1}\{u \in [-1,1]\}$, then

$$\begin{aligned} f_{X_k}(x) &= \frac{1}{2\pi}\int e^{-ivx}\varphi_{X_k}(v)dv, \\ \widehat{f}_{X_k}(x) &= \frac{1}{2\pi}\int_{-T_N}^{T_N} e^{-ivx}\widehat{\varphi}_{X_k}(v)dv, \end{aligned}$$

and thus

$$\widehat{f}_{X_k}(x) - f_{X_k}(x) = \frac{1}{2\pi}\int_{-T_N}^{T_N} e^{-ivx}\left[\widehat{\varphi}_{X_k}(v) - \varphi_{X_k}(v)\right]dv - \frac{1}{\pi}\int_{T_N}^{\infty} e^{-ivx}\,\mathrm{Re}\left[\varphi_{X_k}(v)\right]dv.$$

One thus has to ensure that $\int_{T_N}^{\infty} e^{-ivx}\,\mathrm{Re}\left[\varphi_{X_k}(v)\right]dv$ tends to zero when N tends to infinity. Interestingly, this term is decreasing in T_N, and it is smaller the thinner the tails of φ_{X_k}.

The presence of the second component on the right-hand side of (30) has two consequences. First, although we emphasized in the previous section that *more* trimming (a lower value of T_N) is necessary to ensure a faster convergence of the estimator of φ_{X_k} on $[-T_N, T_N]$, *less* trimming (a larger T_N) is required to reduce the second component of (30). This tension between two conflicting objectives arises in standard nonparametric deconvolution, and explains why asymptotic convergence rates are often slow. Second, although the presence of smoother factor distributions makes deconvolution more difficult, a smoother distribution of X_k also makes the second component in (30) smaller, improving the convergence rate of the density estimator.[18]

Special cases. We illustrate the results in the case of ordinary smooth and supersmooth factor distributions. To minimize the complexity of the discussion, we focus on the kernel $\varphi_H(u) = \varphi_{H_\infty}(u) = \mathbf{1}\{u \in [-1,1]\}$. Remark that φ_{H_∞} is an infinite-order kernel ($q = \infty$), so that the term $\frac{1}{T_N^q}\int_{-T_N}^{T_N} v^q h_k(|v|)dv$ is zero in (30).

Let us start with ordinary smooth factors. As X_k is ordinary smooth one can choose:

$$h_k(t) = t^{-\alpha},\ \alpha > 1.$$

Let $T_N = N^{\frac{\delta}{2}}$, $\delta > 0$. Then,

$$\int_{T_N}^{+\infty} h_k(v)dv = \frac{N^{-\frac{\delta}{2}(\alpha-1)}}{\alpha - 1}.$$

[18] This also happens in the standard deconvolution problem: $Y = X + U$, U known, where the performance of the estimator of the density of X improves when the smoothness of U decreases, and when the smoothness of X increases.

So the convergence rate of $\widehat{f}_{X_k}$ satisfies, using (27):

$$\sup_x \left| \widehat{f}_{X_k}(x) - f_{X_k}(x) \right| = O\left(\frac{\ln N}{N^{\frac{1}{2}-\frac{3}{2}(1+\beta)\delta}} \right) + O\left(\frac{1}{N^{\frac{\delta}{2}(\alpha-1)}} \right). \tag{31}$$

Remark that the first term on the right-hand side of (31) increases when δ increases, as a larger T_N reduces the convergence rate of the characteristic function. However, the second term on the right-hand side of (31) decreases when δ increases, as less trimming decreases the quantity $|\int_{T_N}^{\infty} e^{-ivx} \operatorname{Re}\left[\varphi_{X_k}(v)\right] dv|$. Hence, intuitively, there should exist an optimal degree of trimming.

Remark also that the convergence rate of $\widehat{f}_{X_k}$ is polynomial in N. For example, by choosing δ arbitrarily close to $1/(2+3\beta+\alpha)$, we see that an upper bound to the convergence rate of the density estimator is given by:

$$\sup_x \left| \widehat{f}_{X_k}(x) - f_{X_k}(x) \right| = O\left(\frac{\ln N}{N^{\frac{\alpha-1}{4+6\beta+2\alpha}}} \right).$$

So the rate is faster when α increases (smoother X_k), as long as β (the smoothness of $\mathbf{Y}$) stays constant. Moreover, it is easy to see that, with $\beta \geq \alpha$, the convergence rate of $\widehat{f}_{X_k}$ is never faster than $(N^{\frac{1}{8}}/\ln N)$ (obtained when $\alpha = \beta \to +\infty$). This rate is lower than the best rate of standard nonparametric deconvolution (which, according to Fan, 1991, p. 1265, is $N^{\frac{1}{2}}$), and it is also lower than the rate derived by Li and Vuong (1998), who find a strict upper bound of $N^{\frac{1}{6}}$ in the ordinary smooth case.[19] Slower rates are the price to pay for allowing for many unobservable components.

Interestingly, when factor X_k is ordinary smooth but $\mathbf{Y}$ is supersmooth we obtain much lower rates of convergence. This case may arise if one of the factor variables (different from X_k) has a supersmooth distribution.[20]

Let $T_N = (\delta \ln N)^{\beta_1}$. Then:

$$\int_{T_N}^{+\infty} h_k(v)dv = \frac{(\delta \ln N)^{-\beta_1(\alpha-1)}}{\alpha - 1},$$

which yields, using (29):

$$\sup_x \left| \widehat{f}_{X_k}(x) - f_{X_k}(x) \right| = O\left(\frac{(\ln N)^{1+3(1+\beta_0)\beta_1}}{N^{\frac{1}{2}-3\beta\delta}} \right) + O\left(\frac{1}{(\delta \ln N)^{\beta_1(\alpha-1)}} \right). \tag{32}$$

This rate is *logarithmic* in N, because of the presence of the second term on the right-hand side of (32).

[19] Li and Vuong's estimator requires one first-order differentiation and one integration. This case yields $\frac{T_N^2 \varepsilon_N}{g(T_N)^2} O(1)$ instead of $\frac{T_N^3 \varepsilon_N}{g(T_N)^3} O(1)$ in (30). This explains the difference between $N^{1/8}$ and $N^{1/6}$.

[20] For example, in the simple deconvolution model $Y = X + U$ with X ordinary smooth and U supersmooth, $\varphi_Y(t) = \varphi_X(t)\varphi_U(t)$ has thin tails because of the presence of φ_U, hence Y is supersmooth.

Logarithmic rates of convergence are also obtained in the classical deconvolution problem with one unknown factor, when the distribution of the factor of interest is ordinary smooth while the distribution of the error is supersmooth (Carroll and Hall, 1988, Fan, 1991). Finding optimal convergence rates for the multi-factor model we consider in this paper is a difficult task, which we do not address.

Before ending this discussion, note that additional convergence rates can be derived in the cases where X_k follows a supersmooth distribution, although the expressions are more involved.

5.3 Practical choice of the trimming parameter T_N

We use a method recently developed in deconvolution kernel density estimation to choose the trimming parameter T_N. In the context of the deconvolution problem with known error distribution, Delaigle and Gijbels (2002, 2004) propose to base the choice of the bandwidth on an approximation of the Mean Integrated Squared Error of the kernel density estimator. Comparing different approaches, they find that a "plug-in" method works well in many simulation designs.

We adapt Delaigle and Gijbels' method to the case of a multi-factor model $\mathbf{Y} = \mathbf{AX}$ as follows. For $k \in \{1, ..., K\}$, let $\boldsymbol{\theta}$ be a direction of integration. Then

$$\frac{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{Y}}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k} = \frac{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{AX}}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k} = X_k + \sum_{m \neq k} \frac{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_m}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k} X_m. \tag{33}$$

We treat the distribution of $\sum_{m \neq k} \frac{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_m}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k} X_m$ in (33) as if it were known. In this case, the problem of estimating the density of factor X_k boils down to a deconvolution problem with known error distribution, and the method of Delaigle and Gijbels (2004) can be applied. A detailed presentation of the "plug-in" method and of our bandwidth selection procedure is given in Appendix D.

6 Monte-Carlo simulations

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

6.1 Measurement error model

We start with the estimation of the density of X in the measurement error model (8) with $\alpha = 1$. Namely:

$$\begin{cases} Y_1 = X + U_1 \\ Y_2 = X + U_2, \end{cases} \tag{34}$$

where U_1, U_2 and X are mutually independent, and have mean zero and variance one.

We first consider the case of normal errors U_1 and U_2, and various choices of distribution for X. In Figure 1 we report the outcomes of 100 simulations of samples of size $N = 1000$. In the first column we estimate the density of X using our method, assuming that all three distributions are unknown, and with our preferred choice of direction of integration $\left(\frac{1}{2}, \frac{1}{2}\right)$. In the second column we estimate the density of X from:

$$\frac{Y_1 + Y_2}{2} = X + \frac{U_1 + U_2}{2},$$

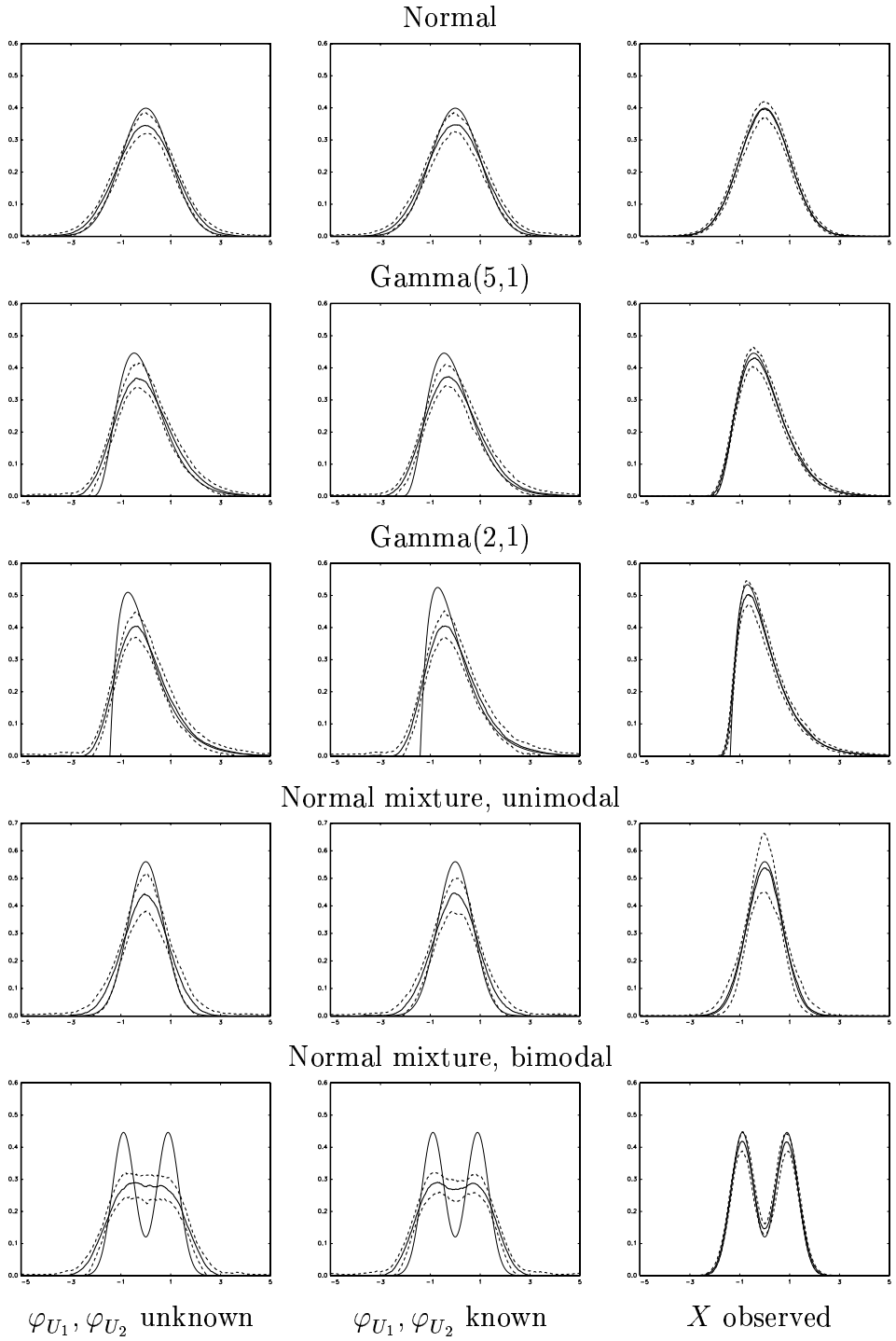
assuming that $\frac{U_1+U_2}{2}$ has known c.f. $\varphi_{\frac{U_1+U_2}{2}}(u) = \exp\left(-\frac{1}{4}u^2\right)$. We use kernel deconvolution for estimation, with the second-order kernel H_2 for smoothing, and choose the trimming parameter T_N using the "plug-in" method of Delaigle and Gijbels (2004). Lastly, in the third column we show the Gaussian kernel density estimator of X for comparison, using Silverman's rule of thumb for choosing the bandwidth. Obviously, this last estimator cannot be computed with real data as X is unobserved. On each graph, the thin solid line represents the population density of X, and the thick solid line is the pointwise median of simulations. The dashed lines delimit the 10%-90% pointwise confidence bands.

Both nonparametric deconvolution methods estimate normal factor distributions well. However, the density at the mode is significantly biased–the true value being systematically outside the confidence band. They both display very similar biases, and the same confidence bands, only moderately wider than when X is observed without error. This suggests that repeated measurements can be very effective at providing information on the distributions of unknown latent variables. Also, the informal choice of bandwidth that we use appears to give very good results, as good as for the deconvolution problem with known error distribution for which it was initially designed.

For non Gaussian factor distributions, we observe that the deconvolution estimators have some difficulty to capture skewness and kurtosis. The Gamma$(5, 1)$ and Gamma$(2, 1)$ distributions have skewness .9 and 1.4, and kurtosis 4.2 and 6, respectively. We see that the bias is larger in the second case. To further study the impact of factor kurtosis on estimation we consider for X a two-components normal mixture that has excess kurtosis equal to 100, that is: $X \sim \frac{400}{403}\mathcal{N}(0, \frac{1}{2}) + \frac{3}{403}\mathcal{N}(0, \frac{406}{6})$. The bias is also larger than in the case where X is normal, although the estimator does a good job at capturing the peak of the density. Lastly, we generate a bimodal distribution as a two-component mixture of normals with different means: $X \sim \frac{1}{2}\mathcal{N}(-2, 1) + \frac{1}{2}\mathcal{N}(2, 1)$. The estimator fails to capture the bimodality.[21]

[21]It is worth noting that in these various designs, we experimented increasing the sample size to

Figure 1: Monte Carlo estimates of f_X in the measurement error model with normal errors



Note: Density of X in model (34). Thin line=true; thick=median of 100 simulations; dashed=10%-90% confidence bands. "Normal mixture, unimodal" is $\frac{400}{403}\mathcal{N}(0, \frac{1}{2}) + \frac{3}{403}\mathcal{N}(0, \frac{406}{6})$, "Normal mixture, bimodal" is $\frac{1}{2}\mathcal{N}(-2, 1) + \frac{1}{2}\mathcal{N}(2, 1)$. $N = 1000$.

Table 1: MISE of various density estimators in the measurement error model

	$\widehat{f}_X$			$\widehat{f}_{U_1}$
	Standard normal errors			
	(1)	(2)	(3)	(4)
Normal	.0011	.0052	.0050	.0069
Laplace	.0030	.025	.026	.0045
Gamma(2,1)	.0036	.026	.024	.0044
Gamma(5,1)	.0012	.011	.010	.0062
Log-normal	.099	.29	.24	.0032
Normal mixture, unimodal	.0076	.020	.022	.0040
Normal mixture, bimodal	.0024	.041	.049	.0097
	Standard Laplace errors			
	(1)	(2)	(3)	(4)
Normal	.0012	.0047	.0042	.034
Laplace	.0027	.018	.020	.020
Gamma(2,1)	.0030	.018	.018	.019
Gamma(5,1)	.0012	.0069	.0070	.027
Log-normal	.061	.22	.16	.011
Normal mixture, unimodal	.0074	.016	.016	.021
Normal mixture, bimodal	.0027	.029	.038	.035

Note: See the note to Figure 1. (1) refers to the case where X is observed, (2) to the deconvolution estimator with known error distributions, and (3) to our generalized deconvolution estimator of f_X. (4) refers to our estimator of f_{U_1}. $N = 1000$, 100 simulations.

In Table 1 we report the Mean Integrated Squared Error of various estimators $\widehat{f}_X$ of f_X, given by:

$$MISE = \mathbb{E}\left[\int \left(\widehat{f}_X(x) - f_X(x)\right)^2 dx\right].$$

In addition to the case of standard normal errors, we also report the results for Laplace-distributed errors. X follows one of the five distributions of Figure 1, and may also be Laplace or log-normally distributed. The estimators we consider are: a kernel density estimate of the density when the factor is observed (column 1), the deconvolution estimator with known error distributions (column 2), and our generalized deconvolution estimator (column 3). Our estimator of f_{U_1} is reported in column (4).

Table 1 confirms that the performance of our estimator is comparable to that of the deconvolution estimator with known error distributions. When errors are distributed as Laplace random variables, theory suggests that the deconvolution problem should be less difficult, and the MISE is indeed slightly lower than in the case of normal errors. Still,

$N = 10000$, and still obtained a sizeable bias (although reduced compared to the case $N = 1000$).

the differences between the cases of ordinary smooth and supersmooth errors do not seem very large.

6.2 Comparison with other estimators

Here we still consider the measurement error model (34), using the same setup as above, and compare our estimator to the ones of Horowitz and Markatou (1996, HM hereafter) and Li and Vuong (1998, LV hereafter). For all estimators we use the smoothing kernel H_2 and select the trimming parameter T_N using the "plug-in" method of Delaigle and Gijbels (2004).

The first two columns of Table 2 display the MISE of the HM and LV estimators. Column (3) gives the MISE of our estimator using our preferred direction of integration $(\frac{1}{2}, \frac{1}{2})$ (LV use $(0, 1)$). Lastly, column (4) shows the MISE of an estimator that averages over the directions of integration, see (22), using ten draws of a bivariate normal distribution with mean $(\frac{1}{2}, \frac{1}{2})$ and standard deviation .10. We show the MISE of the estimators of the density of X (the factor) and of that of U_1 (the first error).

Our estimator performs very well compared to HM and LV. The MISE of $\widehat{f}_X$ is consistently lower for our estimator, except when X is lognormally distributed, in which case all estimators do badly. For the error U_1, the MISE of our estimator is comparable to that of HM, and generally lower than that of LV. Also, the estimator that averages over various directions of integration performs similarly or slightly worse than the one using our preferred direction of integration.[22]

In Section 5 we argued that choosing the direction of integration that minimizes the norm of $\frac{\boldsymbol{\theta}}{\boldsymbol{\theta}^T \mathbf{A}_k}$ yields faster convergence rates. We now provide some direct Monte Carlo evidence in support of this claim.

Figure 2 presents Monte Carlo simulations for estimates of the c.g.f.'s of $\mathbf{Y}$ and X in the measurement error model (8). The setup is as before (100 simulations). In panels a) and b), we plot empirical estimates of $\mathrm{Re}\,\kappa_{\mathbf{Y}}(0, \tau)$ and $\mathrm{Re}\,\kappa_{\mathbf{Y}}(\frac{\tau}{2}, \frac{\tau}{2})$, for $\tau \in \mathbb{R}^+$. We set the scale on the x-axis equal to τ^2. The c.f. of the standard normal distribution being $e^{-t^2/2}$, the true value of the c.g.f. is then a straight line with slope -1 in panel a), and $-3/4$ in panel b).

The c.g.f. of $\mathbf{Y}$ is well estimated over a wide range but the precision is worse at higher frequencies (see also Diggle and Hall, 1993). This explains why $\mathrm{Re}\,\kappa_{\mathbf{Y}}(\frac{\tau}{2}, \frac{\tau}{2})$ is better estimated, for given τ, than $\mathrm{Re}\,\kappa_{\mathbf{Y}}(0, \tau)$. In panel c), we report estimates of the

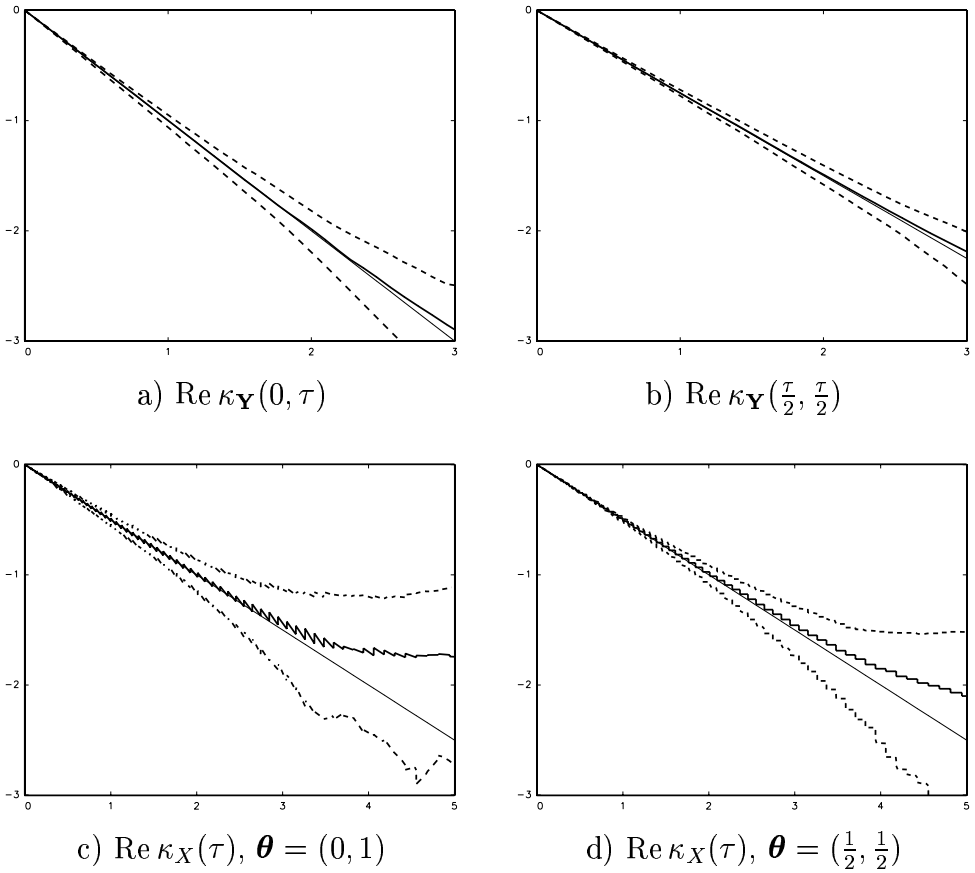
[22] We also tried to average directions of integration around the origin, and obtained much larger MISE.

Table 2: Comparison with other estimators in the measurement error model

	MISE of $\widehat{f}_X$			
	(1)	(2)	(3)	(4)
Normal	.0090	.0079	.0049	.0056
Laplace	.035	.037	.026	.026
Gamma(2,1)	.031	.035	.024	.028
Gamma(5,1)	.016	.013	.011	.0087
Log-normal	.24	.23	.35	.25
Normal mixture, unimodal	.030	.024	.019	.020
Normal mixture, bimodal	.062	.063	.048	.051
	MISE of $\widehat{f}_{U_1}$			
	(1)	(2)	(3)	(4)
Normal	.0062	.0096	.0066	.0072
Laplace	.0042	.0056	.0046	.0051
Gamma(2,1)	.0043	.0051	.0046	.0057
Gamma(5,1)	.0056	.0069	.0054	.0064
Log-normal	.0039	.0027	.0026	.0041
Normal mixture, unimodal	.0048	.0058	.0047	.0055
Normal mixture, bimodal	.0084	.017	.0099	.0084

Note: See the note to Figure 1. (1) refers to the estimator in Horowitz and Markatou (1996), (2) to the estimator of Li and Vuong (1998), (3) to our generalized deconvolution estimator, and (4) to a modification of our estimator which averages 10 different directions of integration, see the text. Errors are standard normal. $N = 1000$, 100 *simulations.*

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



a) $\operatorname{Re} \kappa_{\mathbf{Y}}(0, \tau)$ b) $\operatorname{Re} \kappa_{\mathbf{Y}}(\frac{\tau}{2}, \frac{\tau}{2})$

c) $\operatorname{Re} \kappa_X(\tau)$, $\boldsymbol{\theta} = (0, 1)$ d) $\operatorname{Re} \kappa_X(\tau)$, $\boldsymbol{\theta} = (\frac{1}{2}, \frac{1}{2})$

Note: Estimates of $\kappa_{\mathbf{Y}}$ and κ_X is the measurement error model (34). τ^2 is plotted on the x-axis, c.g.f.'s on the y-axis. Thin line=true; thick=median of 100 simulations; dashed=10%-90% confidence bands. $N = 1000$.

c.g.f. of X obtained by Li and Vuong's method; that is: integrating along the direction $(0,1)$. Panel d) shows the results of our preferred method, integrating along the direction $(\frac{1}{2},\frac{1}{2})$. The c.g.f. of X is better estimated (more precisely and with less bias) by the second method.

6.3 Spatial model

We then consider the spatial model with $L=3$ and $K=6$:

$$\left\{\begin{array}{rcccccccc} Y_1 & = & 2X_1 & + & X_2 & + & X_3 & + & U_1 \\ Y_2 & = & X_1 & + & 2X_2 & + & X_3 & + & U_2 \\ Y_3 & = & X_1 & + & X_2 & + & 2X_3 & + & U_3, \end{array}\right. \tag{35}$$

where X_k, $k=1,...,3$, and U_k, $k=1,...,3$, are mutually independent. This corresponds to model (9) with $\rho=1/2$.

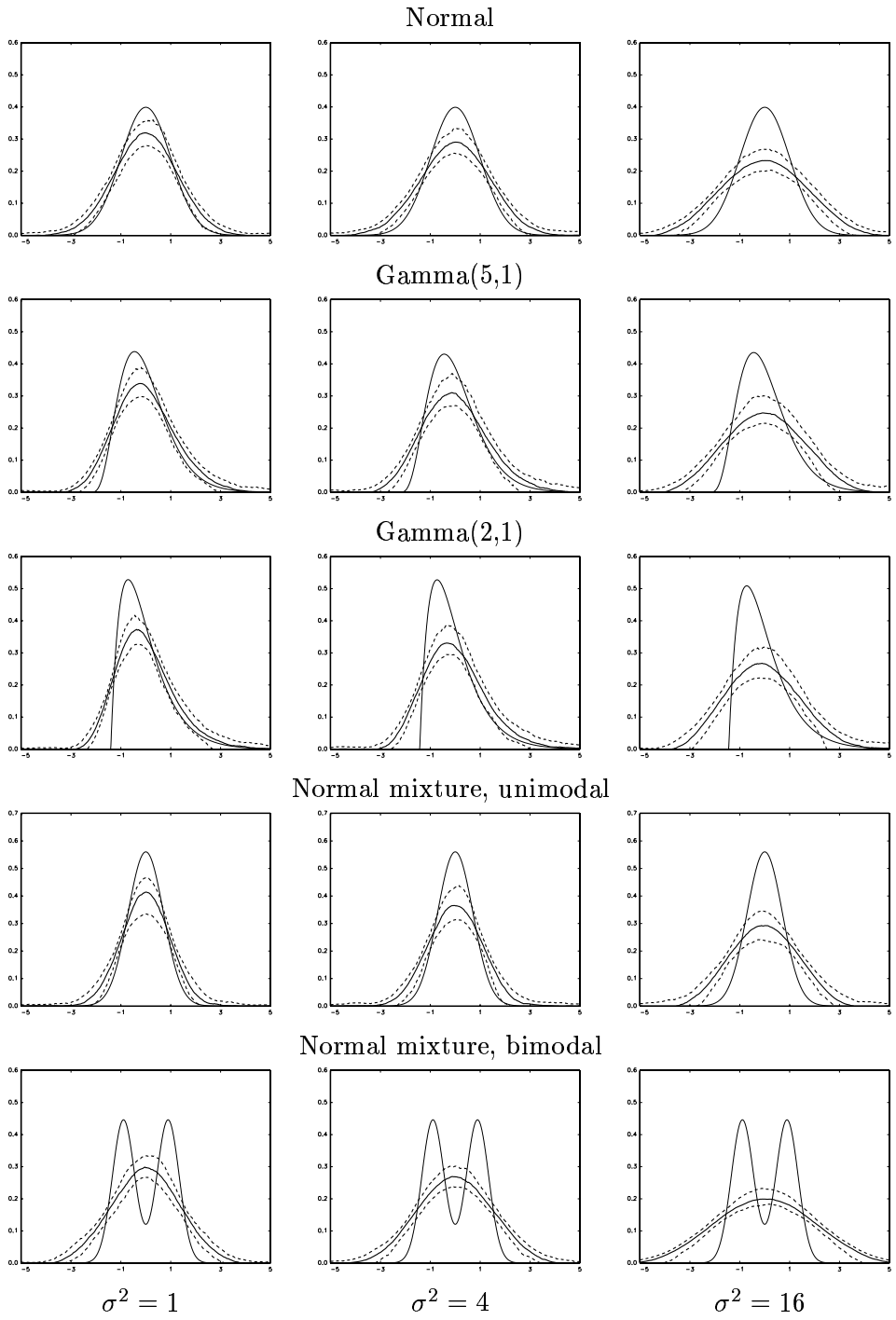
All factor densities belong to the same parametric family. We only let their variances differ: the variances of X_1, X_2 and X_3 are equal to 1, while U_1, U_2 and U_3 have either variance 1 (first column in the figure), 4 (second column), or 16 (third column). The sample size is $N=1000$, and the number of simulations and the conventions used in graphical display are the same as for the measurement error model.

Figure 3 presents the results. We see that when errors U_1, U_2 and U_3 have moderate variance (1 or 4) the density of X_1 is well estimated. The results are comparable to the ones obtained in Figure 1, with a slightly larger bias. When error variances increase to 16, the density of X_1 becomes badly estimated.

For other distributions, namely Gamma or mixture of normals, we generally obtain worse results than for the measurement error model (34). This is confirmed by Table 3, which shows the MISE of the density estimates of X_1 and U_1. Remark that, when σ^2 increases, the performance of $\widehat{f}_{X_1}$ worsens while that of $\widehat{f}_{U_1}$ improves.

Estimating 6 factor densities using 3 measurements is of course more difficult than estimating 3 factor densities using 2 measurements. Yet, in the case of moderate error variances the shapes of the densities are reasonably well reproduced. This suggests that nonparametric deconvolution techniques can be successfully applied to difficult problems, where the number of factors one is trying to extract is large relative to the number of available measurements.

Figure 3: Monte Carlo estimates of f_{X_1} in model (9)



Note: Density of X_1 in model (35). X_k, $k = 1, 2, 3$, are drawn from the same distribution with mean zero and variance 1. U_k, $k = 1, 2, 3$, are drawn from the same distribution as X_1, X_2, X_3 with mean zero and variance σ^2. "Normal mixture, unimodal" is $\frac{400}{403}\mathcal{N}(0, \frac{1}{2}) + \frac{3}{403}\mathcal{N}(0, \frac{406}{6})$, "Normal mixture, bimodal" is $\frac{1}{2}\mathcal{N}(-2, 1) + \frac{1}{2}\mathcal{N}(2, 1)$. $N = 1000$, 100 simulations.

Table 3: MISE of the generalized deconvolution estimator in the spatial model

	MISE of $\widehat{f}_{X_1}$			MISE of $\widehat{f}_{U_1}$		
	$\sigma^2 = 1$	$\sigma^2 = 4$	$\sigma^2 = 16$	$\sigma^2 = 1$	$\sigma^2 = 4$	$\sigma^2 = 16$
Normal	.013	.022	.047	.040	.0096	.0049
Laplace	.039	.059	.098	.081	.038	.028
Gamma(2,1)	.036	.053	.092	.072	.039	.027
Gamma(5,1)	.018	.028	.058	.050	.019	.0094
Log-normal	.35	.30	.40	.32	.15	.12
Normal mixture, unimodal	.030	.044	.10	.078	.028	.022
Normal mixture, bimodal	.074	.077	.098	.098	.068	.055

Note: All factors follow the same distribution up to scale: X_1 to X_3 have unitary variance, while U_1 to U_3 have variance σ^2. $N = 1000$, 100 simulations.

7 Application to earnings dynamics

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

7.1 The data

We use PSID data, between 1978 and 1987. Let y_{nt} denote the logarithm of annual earnings, and let x_{nt} be a vector of regressors, namely: education dummies, a quadratic polynomial in age, a race dummy, geographic indicators and year dummies. We compute the residuals of the OLS regression of $\Delta y_{nt} = y_{nt} - y_{nt-1}$ on $\Delta x_{nt} = x_{nt} - x_{nt-1}$, and denote them as Δw_{nt}. In the sequel we shall refer to Δw_{nt} as wage growth residuals, while keeping in mind that they reflect changes in wage rates and hours worked. We select employed male workers who have non missing observations of Δw_{nt} for the whole period, and for whom wage growth does not exceed 150% in absolute value. We obtain a balanced panel of 624 individuals, with 9 observations of wage growth per individual. Descriptive statistics are presented in the first column of Table 4.

Wage growth residuals Δw_{nt} are the measurements that we use in this application. We shall also consider moving sums of wage growth residuals, defined as

$$\Delta_s w_{nt} = w_{nt} - w_{n,t-s} = \sum_{k=1}^{s} \Delta w_{n,t-k+1}, \text{ for } s = 1, 2, ...$$

Table 5 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 $\Delta_s w_{nt}$ increases with s. This indicates that wage differences between two points in time are

Table 4: Means of variables

Job changes	All	None	One/two	Three/more
Annual earnings (/1000)	36.4	35.3	36.7	37.0
Age	37.4	39.6	37.3	36.1
High school dropout	.21	.22	.23	.16
High school graduate	.54	.59	.51	.54
Hours	2194	2191	2199	2191
Married	.85	.84	.84	.85
White	.70	.63	.69	.75
North east	.15	.15	.13	.17
North central	.26	.30	.24	.27
South	.43	.44	.51	.36
SMSA	.59	.60	.55	.61
Number	624	150	234	240

Note: Balanced subsample of 624 individuals extracted from the PSID, 1978-1987. "None"=no job change; "One/two"= one or two job changes; "Three/more"= more than three job changes.

Table 5: Moments of wage growth residuals

Wage growth	$t/t+1$	$t/t+2$	$t/t+3$	$t/t+T$
Variance	.055	.073	.086	.137
Skewness	-.077	.062	-.073	.457
Kurtosis	10.3	11.2	8.0	4.8
Autocorrelation 1	-.33	.21	.35	-
Autocorrelation 2	-.06	-.34	.08	-
Autocorrelation 3	-.02	-.06	-.34	-

Note: Balanced subsample of 624 individuals extracted from the PSID, 1978-1987. Wage growth residuals are the OLS residuals of first-differenced log earnings on regressors. Wage growth between t and $t+s$ is obtained as the sum of s consecutive wage growth residuals.

more dispersed the longer the lag. Another feature of Table 5 is the high kurtosis of wage growth residuals. This evidence of non-normality is consistent with previous findings on U.S. data.

7.2 Model and estimation

We consider the model outlined in Section 2:

$$\begin{aligned} \Delta w_{nt} &= \Delta y^P_{nt} + \Delta y^T_{nt}, \\ &= \varepsilon_{nt} + \eta_{nt} - \eta_{n,t-1}, \quad i = 1, ..., N, \quad t = 2, ..., T, \end{aligned} \tag{36}$$

where y_{nt}^P follows a random walk: $y_{nt}^P = y_{nt-1}^P + \varepsilon_{nt}$, where ε_{nt} and η_{nt} are white noise innovations with variances σ_ε^2 and σ_η^2. As η_{n1} and ε_{n2} are not separately identified, we normalize η_{n1} to zero. Likewise, we set η_{nT} to zero. We shall refer to y_{nt}^P as the permanent component and to η_{nt} as the transitory component.

Permanent-transitory decompositions are very popular in the earnings dynamics literature, see among others Hall and Mishkin (1982) and Abowd and Card (1989). There is a growing concern that the distributions of wage shocks might be non normal (e.g., Geweke and Keane, 2000). To assess this issue, Horowitz and Markatou (1996) estimate a model of earnings levels 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 non-parametric. However, one particular implication of their model is that Δw_{nt}, $\Delta_2 w_{nt}$, ... are identically distributed. This is clearly at odds with the evidence presented in Table 5. The introduction of a permanent component easily permits to capture the increase in $\text{Var}(\Delta_s w_{nt})$ when s increases.[23]

Turning to estimation, the earnings dynamics model (36) is a linear factor model with $L = T - 1$ equations and $K = 2T - 3$ factors (see Section 2). The estimator given by (23) thus yields consistent estimates of the densities of all shocks. We estimate these densities using the second-order kernel H_2 and Delaigle and Gijbels' (2004) method to select the trimming parameter T_N. We then average all estimated densities.[24] This has the advantage that, if the stationarity restrictions do not hold, one still estimates consistently the average of shock densities over the period. This is appealing in our context, as there is ample evidence that the U.S. economy was subject to large changes in the wage distribution at the beginning of the 1980's. Likewise, different variances can be estimated for all shocks, and averaged *ex post*, yielding estimates $\widehat{\sigma}_\varepsilon^2$ and $\widehat{\sigma}_\eta^2$. We use Equally Weighted Minimum Distance to estimate those variances.

7.3 Results

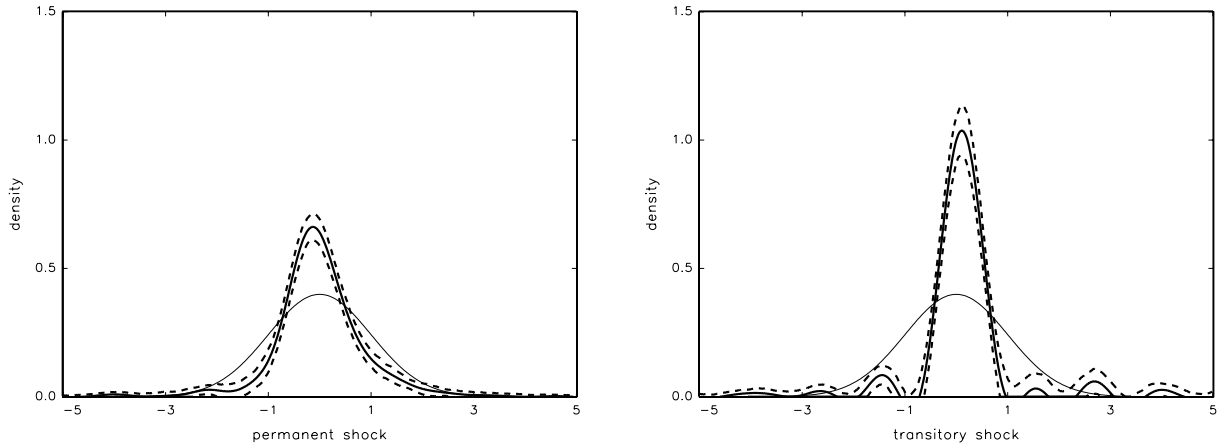
The estimated average variance of permanent shocks is $\widehat{\sigma}_\varepsilon^2 = .0208$, and the estimated average variance of transitory shocks is $\widehat{\sigma}_\eta^2 = .0185$, with standard errors of .0029 and .0017, respectively.[25] According to these estimates, permanent shocks account for 36% of the total variance of wage growth residuals.

[23] Notice that model (36) implies that: $\text{Var}(\Delta_s w_{nt}) - \text{Var}(\Delta w_{nt}) = (s-1)\sigma_\varepsilon^2$. The marginal distributions of Δw_{nt} and $\Delta_2 w_{nt}$ thus contain all the necessary information to identify σ_ε^2 and σ_η^2.

[24] We verified that averaging c.g.f's instead yielded very similar results.

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

Figure 4: Nonparametric estimates of the densities of standardized permanent and transitory shocks.



a) Permanent shock

b) Transitory shock

Note: Density estimates of ε_{nt} and η_{nt}, both standardized to have unit variance. Density estimate (thick); 10%-90% confidence bands of 100 bootstrap simulations (dashed); standard normal density (thin).

Figure 4 presents the estimated densities, obtained by averaging density estimates over the period. The permanent and transitory 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.[26]

Figure 4 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.[27] Lastly, both densities are approximately symmetric.

As noted above, we estimate the densities of permanent and transitory shocks by averaging the period-specific density estimates. Figure 5 shows the density estimates of the non-standardized permanent and transitory shocks in every period. In particular, we see that the density of permanent shocks tends to be more peaked in the second half of

[26]Remark that, as we do not derive the asymptotic distribution of the nonparametric estimator, the validity of the bootstrap in our context is difficult to verify.

[27]We checked that varying the trimming parameter T_N around the value that we obtained using Delaigle and Gijbels' (2004) method had little effect on the estimate $\widehat{f}_{\varepsilon}$, but a stronger effect on $\widehat{f}_{\eta}$, tail oscillations increasing with T_N.

the period, suggesting an increase in kurtosis, although the density shapes are not well estimated enough to be conclusive.

7.4 Fit

Figure 6 compares the predicted densities of $\Delta_s w_{nt}$, $s = 1, 2, 3$, using the model and the estimated densities of permanent and transitory shocks, to kernel density estimates. In panels a1) to c1), the thin line is a kernel estimator of the density $\Delta_s w_{nt}$ $(s = 1, 2, 3)$. The thick line is the predicted density. The dashed line shows the density that is predicted under the assumption that shocks are normally distributed. The predicted densities of $\Delta_s w_{nt}$, $s = 1, 2, 3$, were calculated analytically by convolution of the estimated densities of ε_{nt} and η_{nt}.

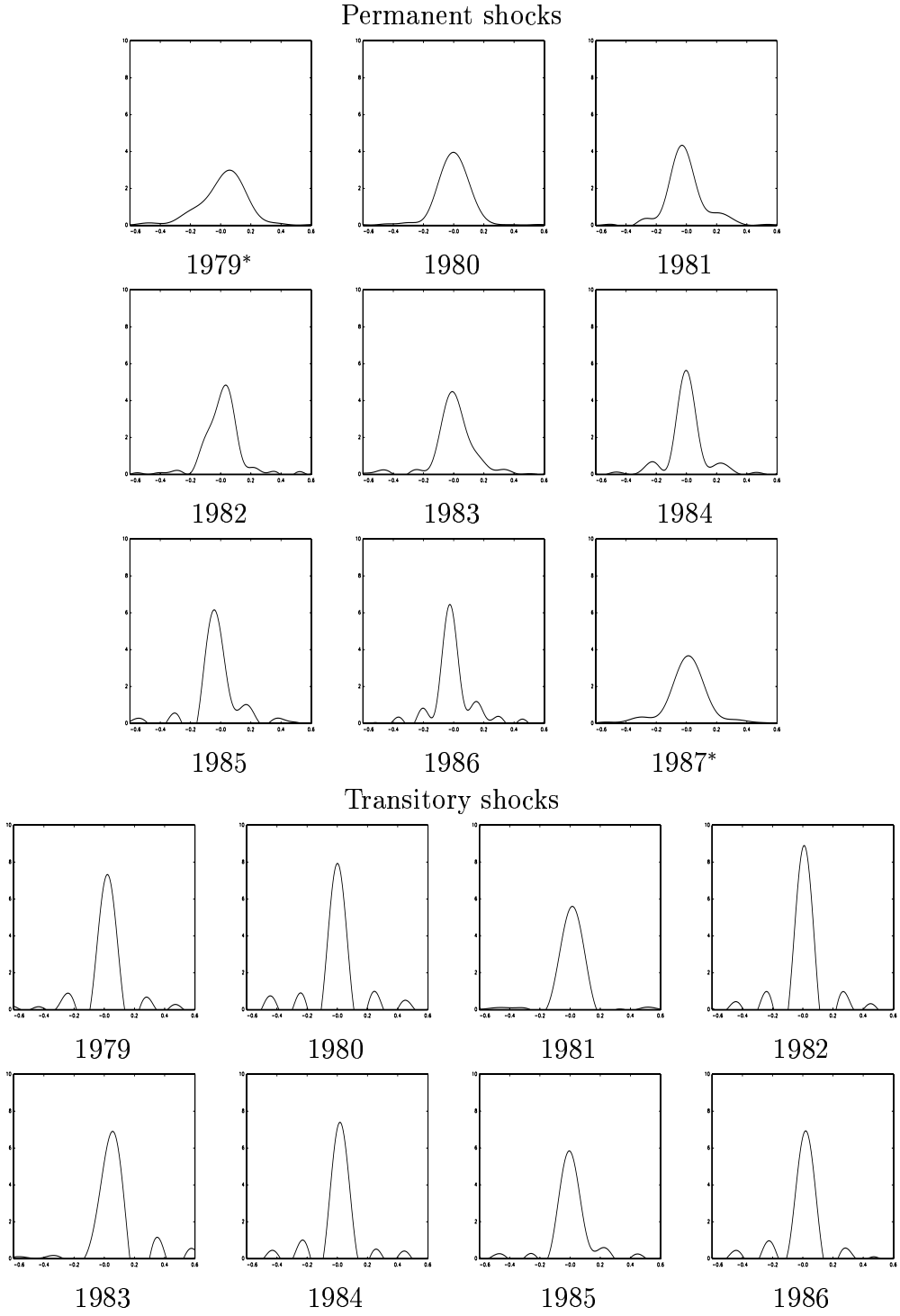
Figure 6 shows that our specification reproduces two features apparent in Table 5: the high kurtosis of wage growth residuals, and the decreasing kurtosis when the time lag increases. Note that the high mode of the density is remarkably well captured by our nonparametric method, even in the case of $\Delta_3 w_{nt}$. In contrast, the normal specification gives a rather poor fit.

We then present in Table 6 the moments of wage growth residuals, as in the data and as predicted under normality and nonparametrically. We see that variances are severely underestimated, reflecting a rather bad estimation of the density in the tails. Moreover, the estimated kurtosis is 5.6, that is significantly non-normal but very different from the kurtosis of the distribution to be fitted (10.3). Overall, our method captures the shapes of the densities of wage growth variables very well, but fails at fitting the tails, which leads to underestimating higher moments.

To fit the moments better, we use the nonparametric estimates $\widehat{f}_\varepsilon$ and $\widehat{f}_\eta$ as a guide to find a convenient parametric form for factor densities. Figure 4 suggests that a mixture of two normals centered at zero may work well in practice. We thus estimate model (36) under this parametric specification for both ε_{nt} and η_{nt}. Parameters are estimated by Maximum Likelihood, using the EM algorithm of Dempster, Laird and Rubin (1977). Panels a2) to c2) in Figure 6 show the fit of the model. The shape of the densities is very well reproduced. Moreover, the last three rows of Table 6 show that the normal mixture specification yields much better estimates of the variance and kurtosis of wage growth residuals.

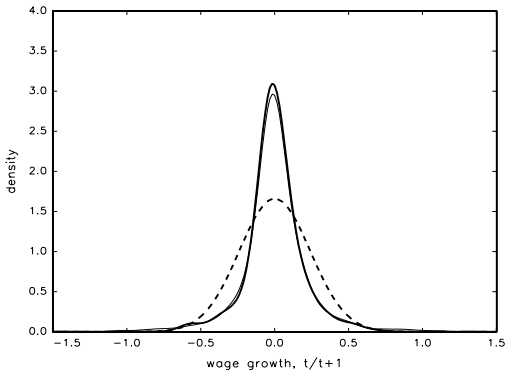
Notice that the normal mixture model was already used by Geweke and Keane (2000) to model earnings dynamics. Our results strongly support this modelling choice.

Figure 5: Density estimates of the non-standardized permanent and transitory shocks in every period

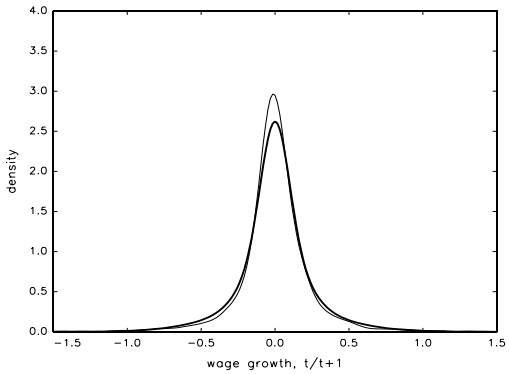


*Note: Density estimates of ε_{nt} and η_{nt} in every period. In the first and last periods, the "permanent" shock includes the permanent and transitory components (indicated by *).*

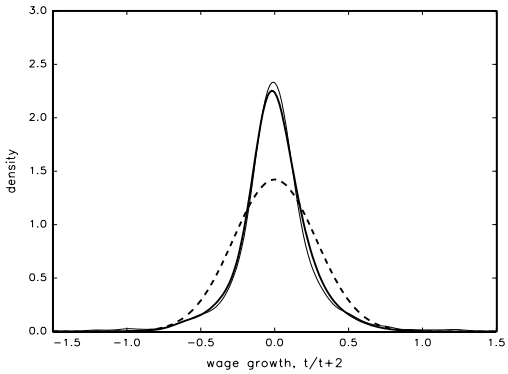
Figure 6: Fit of the model, densities of wage growth residuals.



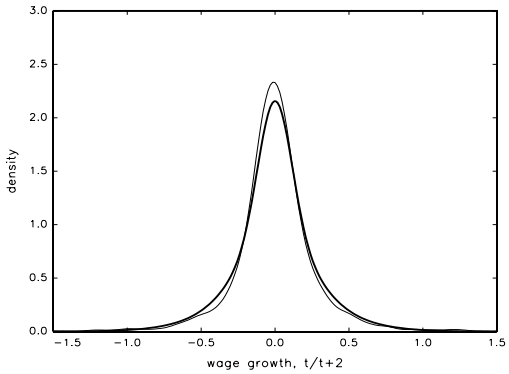
a1) wage growth $t/t+1$



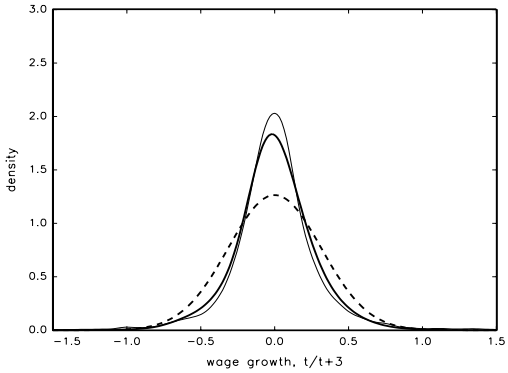
a2) wage growth $t/t+1$, normal mixture



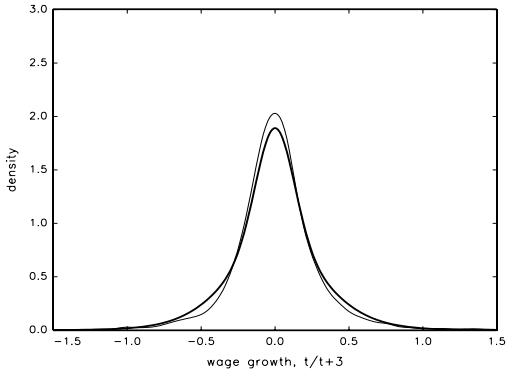
b1) wage growth $t/t+2$



b2) wage growth $t/t+2$, normal mixture



c1) wage growth $t/t+3$



c2) wage growth $t/t+3$, normal mixture

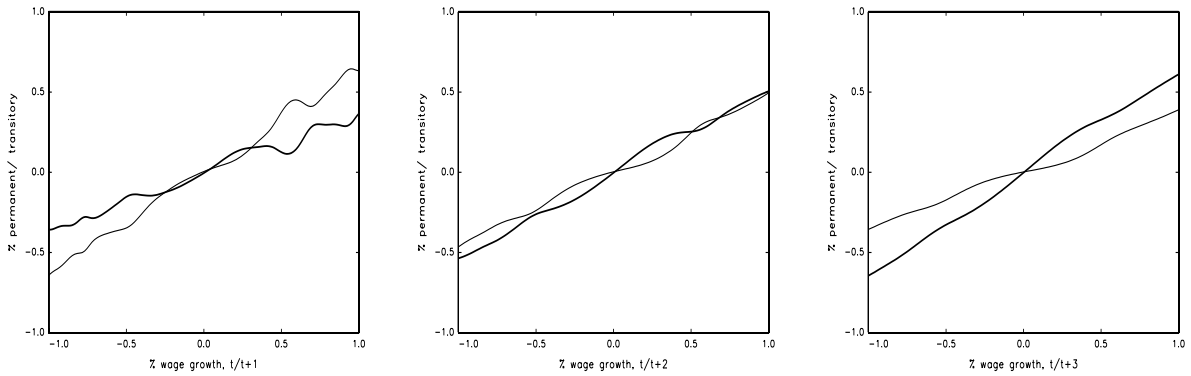
Note: Graphs a1), b1) and c1) show the fit of wage growth residuals calculated over one, two and three years, respectively, using the generalized deconvolution estimator. Graphs a2), b2) and c2): densities are estimated by Maximum Likelihood, where shocks follow two-component mixtures of zero mean normals. Predicted density (thick); kernel density estimate (thin); normal (dashed).

Table 6: Fit of the model, moments of wage growth residuals

Wage growth	$t/t+1$	$t/t+2$	$t/t+3$
	Data		
Variance	.055	.073	.086
Skewness	-.08	.06	-.07
Kurtosis	10.3	11.2	8.0
	Predicted, nonparametric		
Variance	.037	.053	.069
Skewness	-.02	-.02	-.02
Kurtosis	5.6	4.6	4.2
	Predicted, normal		
Variance	.057	.076	.096
Skewness	0	0	0
Kurtosis	3	3	3
	Predicted, normal mixture		
Variance	.058	.072	.086
Skewness	0	0	0
Kurtosis	6.3	5.3	4.8

Note: See the note to Figure 6. Moments are predicted using the predicted densities shown in Figure 6, by computing the integrals numerically.

Figure 7: Conditional expectations of shocks given wage growth residuals



a) Wage growth, $t/t+1$ b) Wage growth, $t/t+2$ c) Wage growth, $t/t+3$

Note: See the note to Figure 6. a): conditional expectation of ε_{nt} *(thick) and* $\eta_{nt} - \eta_{n,t-1}$ *(thin) given* Δw_{nt}*; b):* $\varepsilon_{nt} + \varepsilon_{n,t-1}$ *(thick) and* $\eta_{nt} - \eta_{n,t-2}$ *(thin) given* $\Delta_2 w_{nt}$*; c):* $\varepsilon_{nt} + \varepsilon_{n,t-1} + \varepsilon_{n,t-2}$ *(thick) and* $\eta_{nt} - \eta_{n,t-3}$ *(thin) given* $\Delta_3 w_{nt}$*.*

7.5 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 $\Delta_s w_{nt}$, $s = 1, 2, 3$: $\mathbb{E}\left(\sum_{r=0}^{s-1} \varepsilon_{nt-r} | \Delta_s w_{nt}\right)$ and $\mathbb{E}\left(\eta_{nt} - \eta_{nt-s} | \Delta_s w_{nt}\right)$.

To do so, we first compute the conditional distribution of permanent and transitory shocks using Bayes rule. For example, the conditional density of the permanent shock given wage observations is given by:

$$f(\varepsilon|\Delta w) = \frac{f_\varepsilon(\varepsilon) f(\Delta w|\varepsilon)}{\int f_\varepsilon(\widetilde{\varepsilon}) f(\Delta w|\widetilde{\varepsilon}) d\widetilde{\varepsilon}} = \frac{f_\varepsilon(\varepsilon) \int f_\eta(\eta) f_\eta(\Delta w - \varepsilon + \eta) d\eta}{\int f_\varepsilon(\widetilde{\varepsilon}) \int f_\eta(\eta) f_\eta(\Delta w - \widetilde{\varepsilon} + \eta) d\eta d\widetilde{\varepsilon}},$$

where f_ε is the p.d.f. of ε and f_η is the p.d.f. of η. We proceed similarly for transitory shocks $\eta_{nt} - \eta_{nt-1}$.

Figure 7 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 $\pm 100\%$ has a transitory origin for more than $\pm 60\%$ and a permanent origin for less that $\pm 30\%$. In panel c), we see that a change $\Delta_3 w_{nt}$ of $\pm 100\%$ is almost twice more likely to be permanent than transitory.

Table 7: Variances of the shocks by categories of job changers

Job changes	None	One/two	Three/more
	wage growth, $t/t+1$		
total	.034	.039	.068
permanent	.014	.016	.022
transitory	.020	.023	.046
	wage growth, $t/t+2$		
total	.041	.053	.089
permanent	.025	.032	.053
transitory	.016	.021	.036
	wage growth, $t/t+3$		
total	.054	.063	.108
permanent	.037	.044	.076
transitory	.017	.019	.032

Note: See the note to Figure 6. "None"=no job change in the observation period; "One/two"= one or two job changes; "Three/more"= more than three job changes. Variances of wage growth residuals ("Total") and the variances of the permanent and transitory parts, conditional on having experienced a given number of job changes.

7.6 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.[28] 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 4 in Appendix give descriptive statistics for these three groups of individuals.

Then we compute the densities of permanent and transitory shocks given wage growth residuals, separately for each category of job changers by averaging within each group the conditional densities that we have already calculated. Table 7 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

[28]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.

transitory wage changes. The transitory variance is about 15% higher for infrequent job movers than for job stayers (.023 versus .020), and about 2.3 times higher for frequent job movers (.046). At the same time, the permanent variance is about 15% higher for infrequent job movers than for job stayers (.016 versus .014), and about 60% 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 .16 ($= .020 + 10 * .014$) for an individual who stayed with the same employer over the whole period, while it is about .27 ($= .046 + 10 * .022$) for an individual who has changed job three times or more.

These results give some basis to the interpretation of permanent shocks to log earnings as resulting for a large part from job changes.[29] Nevertheless, identifying permanent shocks with job changes is likely to be wrong for two reasons. First, part of the shocks faced by job stayers is permanent. Indeed, the share of permanent variance in total variance is higher for job stayers (40%) than for frequent job changers (30%). This finding suggests that there might be other permanent wage movements, caused for example 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 linear independent factor models, allowing for any number of measurements, L, and at most $K = \frac{L(L+1)}{2}$ latent factors. On the theoretical side, 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 slow, 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 seems rather good, remarkably similar to the performance of the kernel deconvolution estimator that assumes that the distributions of all factors but one are known, and at least as good as alternative estimators proposed by Horowitz and Marka-

[29] Note that we do not identify the part of the wage growth variance that comes from differences in hours worked from the one coming from differences in wage rates. Nor are we able to tell whether job or individual-specific components are mostly responsible for the results.

tou (1996) and Li and Vuong (1998) in the measurement error model. Moreover, the performance critically depends on the shape of the distributions to be estimated, as we find that it is easier to estimate distributions with little skewness or excess kurtosis.[30]

In any case, identifying the distributions of more factors than measurements should be viewed as considerably more difficult than the classical nonparametric deconvolution problem. Given the difficulty of the problem at hand, we view the results of our simulations and the application as a confirmation that the generalized nonparametric deconvolution approach that we propose 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. These results have important consequences for welfare analysis. For example, savings and insurance could be very different if the risk of large deviations is much higher than is usually assumed with normal shocks. Of course, the model of earnings dynamics that we have considered is very limited. One 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 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) shows 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:

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

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

[30] In Bonhomme and Robin (2008), 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 models where second-order information is not sufficient to ensure the identification of the factor loadings.

References

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

[2] BONHOMME, S., and J.-M. ROBIN (2008): "Consistent Noisy Independent Component Analysis," *mimeo.*

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

[4] CARRASCO, M., and J.P. FLORENS (2007): "Spectral Method for Deconvolving a Density," *mimeo.*

[5] CARROLL, R.J., D. RUPPERT, and L.A. STEFANSKI (1995): *Measurement Error in Nonlinear Models.* New York: Chapman & Hall.

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

[7] CHAMBERLAIN, G., and K. HIRANO (1999): "Predictive Distributions Based on Longitudinal Earnings Data ," *Annales d'Économie et de Statistiques*, 55-56, 211-242.

[8] CSÖRGŐ, S. (1981): "Limit Behaviour of the Empirical Characteristic Function," *The Annals of Probability*, Vol 9, No. 1, 130-144.

[9] DELAIGLE, A., and I. GIJBELS (2002): "Estimation of Integrated Squared Density Derivatives From a Contaminated Sample," *Journal of the Royal Statistical Society*, B, 64, 869-886.

[10] DELAIGLE, A., and I. GIJBELS (2004): "Comparison of Data-Driven Bandwidth Selection Procedures in Deconvolution Kernel Density Estimation," *Computational Statistics and Data Analysis*, 45, 249-267.

[11] DELAIGLE, A., and I. GIJBELS (2007): "Frequent Problems in Calculating Integrals and Optimizing Objective Functions: A Case Study in Density Deconvolution," *Statistics and Computing*, 17, 349-355.

[12] DELAIGLE, A., P. HALL, and A. MEISTER (2008): "On Deconvolution with Repeated Measurements," *Annals of Statistics*, 36(2), 665-685.

[13] 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.

[14] 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.

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

[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] GEWEKE, J., and M. KEANE (2007): "Smoothly Mixing Regressions," *Journal of Econometrics*, 138, 252-290.

[18] GUVENEN, F. (2007a): "Learning Your Earning: Are Labor Income Shocks Really Very Persistent?" *American Economic Review*, 97, 687-712.

[19] GUVENEN, F. (2007b): "An Empirical Investigation of Labor Income Processes," *mimeo.*

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

[21] HALL, R., and F. MISHKIN (1982): "The Sensitivity of Consumption to Transitory Income: Estimates from Panel Data on Households," *Econometrica*, 50, 461-481.

[22] HIRANO, K. (2002): "Semiparametric Bayesian Inference in Autoregressive Panel Data Models," *Econometrica*, 70, 780-799.

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

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

[25] HU, Y., and G. RIDDER (2007): "Estimation of Nonlinear Models with Mismeasured Regressors Using Marginal Information", *mimeo*, available at /http://www.econ.jhu.edu/People/Hu/EIV-marg-2007.pdf.

[26] HU, Y., and G. RIDDER (2008): "On Deconvolution as a First Stage Nonparametric Estimator", *mimeo.*

[27] KAPLAN, G., and G. VIOLANTE (2008): "How Much Insurance in Bewley Models?" *mimeo.*

[28] KHATRI, C.G., and C.R. RAO (1968): "Solutions to Some Functional Equations and their Applications to Characterization of Probability Distributions," *Sankhyä*, 30, 167-180.

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

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

[31] LI, T., I. PERRIGNE, and Q. VUONG (2000): "Conditionally Independent Private Information in OSC Wildcat Auctions," *Journal of Econometrics*, 98, 129-161.

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

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

[34] LUKACS, E. (1970): *Characteristic functions*, second ed. Griffin, London.

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

[36] POLLARD, D. (1984): *Convergence of Stochastic Processes.* Springer: New-York.

[37] POLLARD, D. (2002): *A User's Guide to Measure Theoretical Probability.* Cambridge University Press.

[38] RAO, P. (1983): *Nonparametric Functional Estimation.* New-York: Academic Press.

[39] RAO, P. (1992): *Identifiability in Stochastic Models.* New-York: Academic Press.

[40] REIERSOL, O. (1950): "Identifiability of a Linear Relation Between Variables Which Are Subject to Error," *Econometrica*, 18(4), 375-389.

[41] SCHENNACH, S. (2004a): "Estimation of Nonlinear Models with Measurement Error," *Econometrica*, 72, 33-75.

[42] SCHENNACH, S. (2004b): "Nonparametric Estimation in the Presence of Measurement Error," *Econometric Theory*, 20, 1046-94.

[43] SUSKO, E. and NADON, R. (2002): "Estimation of a residual distribution with small numbers of repeated measurements," *Canadian Journal of Statistics*, 30, 383–400.

[44] SZÉKELY, G.J., and C.R. RAO (2000): "Identifiability of Distributions of Independent Random Variables by Linear Combinations and Moments," *Sankhyä*, 62, 193-202.

[45] ZHANG, C. (1990): "Fourier methods for estimating mixing densities and distributions," *Annals of Statistics*, 18, 806-831.

APPENDIX

A Proof of Lemma 1

(i) 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 $\operatorname{Re}(f_t)(x, \mathbf{y}) = x\cos(\mathbf{t}^{\mathrm{T}}\mathbf{y})$, $\mathbf{t} \in \mathbb{R}^L$, 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 $f_t(x, \mathbf{y}) = x\cos(\mathbf{t}^{\mathrm{T}}\mathbf{y})$, using the same notation for f_t and its real part.

(ii) For any T, let $\mathcal{G} = \{f_{\mathbf{t}}(x, \mathbf{y}), |\mathbf{t}| \leq T\}$. The first step of the proof is to find the L_1-covering number of $\mathcal{G}$.[31] For any couple $(\mathbf{t}_1, \mathbf{t}_2)$,

$$\begin{aligned}
\left|x\cos(\mathbf{t}_1^T\mathbf{y}) - x\cos(\mathbf{t}_2^T\mathbf{y})\right| &\leq \left|x(\mathbf{t}_1^T\mathbf{y} - \mathbf{t}_2^T\mathbf{y})\right| \\
&\leq \sum_\ell |xy_\ell\,(t_{1\ell} - t_{2\ell})| \\
&\leq \sum_\ell |xy_\ell| \cdot |\mathbf{t}_1 - \mathbf{t}_2| \\
&\leq L\,|x\mathbf{y}| \cdot |\mathbf{t}_1 - \mathbf{t}_2|.
\end{aligned}$$

Discretize $(-T, T)^L$ into $\left(\frac{2TL\mathbb{E}_N|X\mathbf{Y}|}{\varepsilon} - 1\right)^L$ points $\mathbf{t}_j$ by cutting $[-T, T]$ into equidistant segments of length $\frac{\varepsilon}{L\mathbb{E}_N|X\mathbf{Y}|}$. Let $g_j(x, \mathbf{y}) = x\cos(\mathbf{t}_j^{\mathrm{T}}\mathbf{y})$. Then, for all $\mathbf{t} \in [-T, T]^L$, there exists j such that

$$\begin{aligned}
\mathbb{E}_N\left|X\cos(\mathbf{t}^T\mathbf{Y}) - X\cos(\mathbf{t}_j^T\mathbf{Y})\right| &\leq L\mathbb{E}_N\,|X\mathbf{Y}| \cdot |\mathbf{t} - \mathbf{t}_j| \\
&\leq \varepsilon.
\end{aligned}$$

It follows that the L_1-covering number of $\mathcal{G}$ satisfies

$$\mathcal{N}_1(\varepsilon, P_N, \mathcal{G}) \leq C\left(\frac{T\mathbb{E}_N|X\mathbf{Y}|}{\varepsilon}\right)^L,$$

where P_N is the probability measure obtained by independent sampling from F, and $C = (2L)^L$.

Note that although the covering number is indeed inversely proportional to a power of ε, Theorem 2.37 of Pollard (1984, p. 34) cannot be applied for three reasons. First, the upper bound to the covering number depends on the sample P_N via $\mathbb{E}_N|X\mathbf{Y}|$. Second, the functions in $\mathcal{G}$ are not bounded because the support of X is unbounded. Third, eventually, one will index

[31] Let Q be a probability measure on S and $\mathcal{F}$ be a class of functions in $\mathcal{L}_1(Q)$. For $\varepsilon > 0$, the covering number $\mathcal{N}_1(\varepsilon, Q, \mathcal{F})$ is the smallest integer m such that there exists functions $g_1, ..., g_m$ in $\mathcal{L}_1(Q)$ such that $\min_j \mathbb{E}_Q \|f - g_j\| \leq \varepsilon$ for all $f \in \mathcal{F}$ (Pollard, 1984, p. 25).

T on N to make it go to infinity with N. A specific proof therefore has to be tailored to adjust Pollard's proof of Theorem 2.37 to our setup.

(iii) Equations (30) and (31) in Pollard (1984, p. 31) imply that, for a given sample $\mathbf{Z}_N$,

$$\Pr\left\{\sup_{|\mathbf{t}|\leq T}|\mathbb{E}_N f_{\mathbf{t}} - \mathbb{E}f_{\mathbf{t}}| > \varepsilon \middle| \mathbf{Z}_N\right\} \leq 8C\left(\frac{T\mathbb{E}_N|X\mathbf{Y}|}{\varepsilon}\right)^L \exp\left[-\frac{N\varepsilon^2}{128}\Big/\mathbb{E}_N X^2\right], \tag{A1}$$

as $\mathbb{E}_N g_j^2 \leq \mathbb{E}_N X^2$ for all j, and provided that $\operatorname{Var}\mathbb{E}_N f_{\mathbf{t}} \leq \frac{\varepsilon^2}{8}$. Now,

$$\begin{aligned} N\operatorname{Var}\mathbb{E}_N f_{\mathbf{t}} &= \operatorname{Var}\left[X\cos\left(\mathbf{t}^T\mathbf{Y}\right)\right] \\ &= \mathbb{E}\left[X^2\cos\left(\mathbf{t}^T\mathbf{Y}\right)^2\right] - \left[\mathbb{E}X\cos\left(\mathbf{t}^T\mathbf{Y}\right)\right]^2 \\ &\leq \mathbb{E}X^2 \equiv M_1 < \infty. \end{aligned}$$

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

Then, for all $k > 0$:

$$\begin{aligned} \Pr\left\{\sup_{|\mathbf{t}|\leq T}|\mathbb{E}_N f_{\mathbf{t}} - \mathbb{E}f_{\mathbf{t}}| > \varepsilon\right\} &= \Pr\left\{\sup_{|\mathbf{t}|\leq T}|\mathbb{E}_N f_{\mathbf{t}} - \mathbb{E}f_{\mathbf{t}}| > \varepsilon \middle| \mathbb{E}_N X^2 < k, \mathbb{E}_N|X\mathbf{Y}| < k\right\} \\ &\quad \times \Pr\left\{\mathbb{E}_N X^2 < k, \mathbb{E}_N|X\mathbf{Y}| < k\right\} \\ &\quad + \Pr\left\{\sup_{|\mathbf{t}|\leq T}|\mathbb{E}_N f_{\mathbf{t}} - \mathbb{E}f_{\mathbf{t}}| > \varepsilon \middle| \mathbb{E}_N X^2 \geq k \text{ or } \mathbb{E}_N|X\mathbf{Y}| \geq k\right\} \\ &\quad \times \Pr\left\{\mathbb{E}_N X^2 \geq k \text{ or } \mathbb{E}_N|X\mathbf{Y}| \geq k\right\} \\ &\leq \Pr\left\{\sup_{|\mathbf{t}|\leq T}|\mathbb{E}_N f_{\mathbf{t}} - \mathbb{E}f_{\mathbf{t}}| > \varepsilon \middle| \mathbb{E}_N X^2 < k, \mathbb{E}_N|X\mathbf{Y}| < k\right\} \\ &\quad + \Pr\left\{\mathbb{E}_N X^2 \geq k \text{ or } \mathbb{E}_N|X\mathbf{Y}| \geq k\right\}, \end{aligned}$$

where the last inequality results from bounding two of the four probabilities above by one.

To obtain a final inequality, use a general Chernoff bound:

$$\begin{aligned} \Pr\left\{\mathbb{E}_N X^2 \geq k \text{ or } \mathbb{E}_N|X\mathbf{Y}| \geq k\right\} &= \Pr\left\{\exp\left(\mathbb{E}_N X^2\right) \geq e^k \text{ or } \exp\left(\mathbb{E}_N|X\mathbf{Y}|\right) \geq e^k\right\} \\ &\leq \frac{\mathbb{E}\left[\exp\left(\mathbb{E}_N X^2\right)\right] + \mathbb{E}\left[\exp\left(\mathbb{E}_N|X\mathbf{Y}|\right)\right]}{e^k}. \end{aligned} \tag{A2}$$

Now,

$$\begin{aligned} \mathbb{E}\left[\exp\left(\mathbb{E}_N X^2\right)\right] &= \mathbb{E}\left[\exp\left(\frac{1}{N}\sum_{n=1}^N X_n^2\right)\right] \\ &= \prod_{n=1}^N \mathbb{E}\left[\exp\left(\frac{X_n^2}{N}\right)\right] \\ &= \left(\mathbb{E}e^{X^2/N}\right)^N \\ &= \left[M_{X^2}\left(\frac{1}{N}\right)\right]^N, \end{aligned}$$

denoting as M_{X^2} the moment generating function of X^2. By assumption, $M_{X^2}(t)$ exists for all t in a neighborhood around 0. Hence all moments of X^2 are finite, and

$$M_{X^2}(1/N) = 1 + \mathbb{E}X^2/N + O(1/N^2),$$

so that

$$\lim_{N\to\infty} M_{X^2}\left(1/N\right)^N = \exp\left(\mathbb{E}X^2\right).$$

One can thus bound $\mathbb{E}\left[\exp\left(\mathbb{E}_N X^2\right)\right]$ by $2\exp\left(\mathbb{E}X^2\right)$ for N large enough. The same argument applies to $\mathbb{E}\left[\exp\left(\mathbb{E}_N|X\mathbf{Y}|\right)\right]$.

Therefore,

$$\Pr\left\{\sup_{|\mathbf{t}|\leq T}|\mathbb{E}_N f_{\mathbf{t}} - \mathbb{E}f_{\mathbf{t}}| > \varepsilon\right\} \leq 8C\left(\frac{Tk}{\varepsilon}\right)^L \exp\left[-\frac{N\varepsilon^2}{128k}\right] + 2\frac{\exp\left(\mathbb{E}X^2\right)+\exp\left(\mathbb{E}|X\mathbf{Y}|\right)}{e^k}, \quad \text{(A3)}$$

for any N large enough that satisfies $N \geq \frac{8M_1}{\varepsilon^2}$.

(iv) Lastly, index ε, T and k by N, and suppose that ε_N tends to 0 and that T_N and k_N tend to infinity in such a way that

$$\sum_N \frac{1}{e^{k_N}} < \infty, \quad \text{(A4)}$$

and

$$\sum_N \exp\left\{L\ln\left(\frac{T_N k_N}{\varepsilon_N}\right) - \frac{N\varepsilon_N^2}{128k_N}\right\} < \infty. \quad \text{(A5)}$$

A standard application of the Borel-Cantelli lemma then implies that

$$\sup_{|\mathbf{t}|\leq T_N}|\mathbb{E}_N f_{\mathbf{t}} - \mathbb{E}f_{\mathbf{t}}| < \varepsilon_N, \qquad \text{a.s.}$$

Details about the Borel-Cantelli argument for almost sure convergence can be found in Pollard (2002, p. 34-35).

Condition (A5) is satisfied if $\exp\left[L\ln\left(\frac{T_N k_N}{\varepsilon_N}\right) - \frac{N\varepsilon_N^2}{128k_N}\right]$ decreases faster than $1/N$. Let

$$\begin{aligned}\varepsilon_N &= A\frac{\ln N}{\sqrt{N}},\ A>0,\\ k_N &= (1+\alpha)\ln N,\end{aligned}$$

with $\alpha > 0$ to satisfy condition (A4). Hence

$$\frac{N\varepsilon_N^2}{128k_N} = \frac{A^2}{128(1+\alpha)}\ln N.$$

Let also

$$T_N = BN^{\frac{\delta}{2}},\ B,\delta>0.$$

Then,

$$L\ln\left(\frac{T_N k_N}{\varepsilon_N}\right) - \frac{N\varepsilon_N^2}{128k_N} = L\ln\left(\frac{B(1+\alpha)}{A}\right) + \left(\frac{L}{2}(1+\delta) - \frac{A^2}{128(1+\alpha)}\right)\ln N$$

decreases faster than $-\ln N$ if

$$A^2 > 64\left[L\left(1+\delta\right)+2\right](1+\alpha).$$

Whatever $\delta > 0$, one can thus choose any $A > 8\sqrt{2+L(1+\delta)}$.

This achieves to prove Lemma 1.

Remark. Compared to the proof of the law of the iterated logarithm (also referred to as the "log log law"), it is the additional term $L \ln\left(\frac{T_N k_N}{\varepsilon_N}\right)$ that makes all the difference. This term arises from the necessity to cover the set of functions $\mathcal{G}$. If X and $\mathbf{Y}$ are bounded, then one can proceed differently, and adapt the proof of Theorem 1 in Csörgő (1981) that uses the law of the iterated logarithm.

B Proof of Theorem 1

In this proof and the next, all convergence statements are implicitly understood to hold almost surely.

Here, we aim at bounding $\sup_{|\tau|\leq T_N} \left|\widehat{\varphi}_{X_k}(\tau) - \varphi_{X_k}(\tau)\right|$, where

$$\widehat{\varphi}_{X_k}(\tau) = \exp\left(\int_0^\tau \int_0^u \mathbf{Q}_k^- \operatorname{vech}\left[\int \nabla\nabla^{\mathrm{T}} \widehat{\kappa}_{\mathbf{Y}}\left(\frac{v\boldsymbol{\theta}}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k}\right) dW(\boldsymbol{\theta})\right] dvdu\right)$$

for some distribution W. This will easily follow from bounding, for any $\ell, m = 1, ..., L$,

$$C_{\ell m}(\boldsymbol{\theta}) = \sup_{|\tau|\leq T_N} \left|\int_0^\tau \int_0^u \frac{\partial^2 \widehat{\kappa}_{\mathbf{Y}}}{\partial t_\ell \partial t_m}\left(\frac{v\boldsymbol{\theta}}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k}\right) dvdu - \int_0^\tau \int_0^u \frac{\partial^2 \kappa_{\mathbf{Y}}}{\partial t_\ell \partial t_m}\left(\frac{v\boldsymbol{\theta}}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k}\right) dvdu\right|.$$

(i) Fix $\mathbf{t} \in \mathbb{R}^L$. Denote $\varphi(\mathbf{t}) \equiv \varphi_{\mathbf{Y}}(\mathbf{t}) = \mathbb{E}\left[e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}}\right]$, $\psi_\ell(\mathbf{t}) = \mathbb{E}\left[Y_\ell e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}}\right]$ and $\xi_{\ell m}(\mathbf{t}) = \mathbb{E}\left[Y_\ell Y_m e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}}\right]$, for any $\ell, m = 1, ..., L$. Then, Lemma 1 implies that, for all function f in $\{\varphi, \{\psi_\ell\}_\ell, \{\xi_{\ell m}\}_{\ell,m}\}$:

$$\sup_{|\mathbf{t}|\leq T_N} \left|\widehat{f}(\mathbf{t}) - f(\mathbf{t})\right| = O(\varepsilon_N),$$

with

$$\begin{aligned} T_N &= BN^{\frac{\delta}{2}},\ B, \delta > 0, \\ \varepsilon_N &= A\frac{\ln N}{\sqrt{N}},\ A > 8\sqrt{2 + L(1+\delta)}. \end{aligned}$$

(ii) There exists c such that $|\varphi(\mathbf{t})| \geq g(|\mathbf{t}|)$ when $|\mathbf{t}| > c$. As g is decreasing, then for all $c < |\mathbf{t}| \leq T_N$,

$$|\varphi(\mathbf{t})| \geq g(|\mathbf{t}|) \geq g(T_N).$$

Hence,

$$\inf_{|\mathbf{t}|\leq T_N} |\varphi(\mathbf{t})| \geq \min\left\{g(T_N), \inf_{|\mathbf{t}|\leq c} |\varphi(\mathbf{t})|\right\}.$$

Notice that, because of Assumption A2 and the continuity of φ:[32]

$$\inf_{|\mathbf{t}|\leq c} |\varphi(\mathbf{t})| > 0.$$

So, as $\lim_{|\mathbf{t}|\to\infty} g(|\mathbf{t}|) = 0$, it follows that $\min\left\{g(T_N), \inf_{|\mathbf{t}|\leq c} |\varphi(\mathbf{t})|\right\} = g(T_N)$ for T_N large enough.

[32] Remark that $|\varphi_{\mathbf{Y}}(\mathbf{t})| = |\varphi_{\mathbf{X}}(\mathbf{A}^{\mathrm{T}}\mathbf{t})| > 0$ for all $\mathbf{t}$ by Assumption A2.

Consequently, for N large enough,

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

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, and that, by assumption, $\frac{T_N^2\varepsilon_N}{g(T_N)^3} = o(1)$.

(iii) We have

$$\frac{\partial\kappa_{\mathbf{Y}}(\mathbf{t})}{\partial t_\ell} = i\frac{\psi_\ell(\mathbf{t})}{\varphi(\mathbf{t})} = i\frac{\mathbb{E}\left[Y_\ell e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}}\right]}{\mathbb{E}\left[e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}}\right]},$$

and

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

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

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

as $\mathbb{E}|Y_\ell| < \infty$.

It follows that

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

The same argument applies to show that

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

for all ℓ, m, if $\mathbb{E}|Y_\ell Y_m| < \infty$.

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

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

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

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

Since

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

for all ℓ, it follows that

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

for N large enough such that $g(T_N) < 1$.

(v) Fix a direction of integration $\boldsymbol{\theta} \in \mathbb{R}^L \backslash \{0\}$, and $\tau \in \mathbb{R}$. Then:

$$\begin{aligned}C_{\ell m}(\boldsymbol{\theta}) &= \sup_{\tau\in[-T_N,T_N]}\left|\int_0^\tau\int_0^u \widehat{\zeta}_{\ell m}\left(\frac{v\boldsymbol{\theta}}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k}\right)dvdu - \int_0^\tau\int_0^u \zeta_{\ell m}\left(\frac{v\boldsymbol{\theta}}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k}\right)dvdu\right| \\ &\leq \sup_{\tau\in[-T_N,T_N]}\left(\frac{\tau^2}{2}\sup_{|v|\leq T_N}\left|\widehat{\zeta}_{\ell m}\left(\frac{v\boldsymbol{\theta}}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k}\right) - \zeta_{\ell m}\left(\frac{v\boldsymbol{\theta}}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k}\right)\right|\right) \\ &\leq T_N^2 \sup_{|\mathbf{t}|\leq T_N\left|\frac{\boldsymbol{\theta}}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k}\right|}\left|\widehat{\zeta}_{\ell m}(\mathbf{t}) - \zeta_{\ell m}(\mathbf{t})\right| \\ &= \frac{T_N^2}{g\left(T_N\left|\frac{\boldsymbol{\theta}}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k}\right|\right)^3}O(\varepsilon_N).\end{aligned}$$

Hence, for any distribution W:[33]

$$\int C_{\ell m}(\boldsymbol{\theta})\,dW(\boldsymbol{\theta}) \leq \left[\int g\left(T_N\left|\frac{\boldsymbol{\theta}}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k}\right|\right)^{-3}dW(\boldsymbol{\theta})\right]T_N^2 O(\varepsilon_N).$$

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

$$\sup_{\tau\in[-T_N,T_N]}\left|\widehat{\kappa}_{X_k}(\tau) - \kappa_{X_k}(\tau)\right| = \left[\int g\left(T_N\left|\frac{\boldsymbol{\theta}}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k}\right|\right)^{-3}dW(\boldsymbol{\theta})\right]T_N^2 O(\varepsilon_N) = o(1).$$

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

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

[33] Technically, we need some support conditions on W that ensure that the statement $O(\varepsilon_N)$ above is uniform in $\boldsymbol{\theta}$.

from which it follows that

$$\sup_{\tau\in[-T_N,T_N]}\left|\widehat{\varphi}_{X_k}(\tau)-\varphi_{X_k}(\tau)\right| = \left[\int g\left(T_N\left|\frac{\boldsymbol{\theta}}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k}\right|\right)^{-3} dW(\boldsymbol{\theta})\right] T_N^2 O(\varepsilon_N).$$

This ends the proof of Theorem 1.

C Proof of Theorem 2

For all x in the support of X_k:

$$\begin{aligned}\widehat{f}_{X_k}(x)-f_{X_k}(x) &= \frac{1}{2\pi}\int\varphi_H\left(\frac{v}{T_N}\right)e^{-ivx}\left(\widehat{\varphi}_{X_k}(v)-\varphi_{X_k}(v)\right)dv \\ &\quad+\frac{1}{2\pi}\int\left(\varphi_H\left(\frac{v}{T_N}\right)-1\right)e^{-ivx}\varphi_{X_k}(v)dv,\end{aligned}$$

where φ_H is the c.f. of a smoothing kernel that is equal to 0 outside $[-1,1]$. So, for N large enough:

$$\begin{aligned}\left|\widehat{f}_{X_k}(x)-f_{X_k}(x)\right| &\leq \frac{1}{2\pi}\Big(\int_{-T_N}^{T_N}\left|\varphi_H\left(\frac{v}{T_N}\right)\right|\left|\widehat{\varphi}_{X_k}(v)-\varphi_{X_k}(v)\right|dv \\ &\qquad+\int\left|\varphi_H\left(\frac{v}{T_N}\right)-1\right|h_k(|v|)dv\Big) \\ &\leq \frac{T_N}{\pi}\sup_{|\tau|\leq T_N}\left|\widehat{\varphi}_{X_k}(\tau)-\varphi_{X_k}(\tau)\right|+\frac{1}{2\pi}\int\left|\varphi_H\left(\frac{v}{T_N}\right)-1\right|h_k(|v|)dv,\end{aligned} \tag{C6}$$

where we have used that $|\varphi_H|\leq 1$ (as φ_H is a c.f.), and that $|\varphi_{X_k}(v)|\leq h_k(|v|)$ for all $|v|$.

Note that

$$|\varphi_{\mathbf{Y}}(\mathbf{t})| = \left|\mathbb{E}\left[e^{i\mathbf{t}^{\mathrm{T}}\mathbf{Y}}\right]\right| = \left|\mathbb{E}\left[e^{i\mathbf{t}^{\mathrm{T}}\mathbf{A}\mathbf{X}}\right]\right| = \left|\prod_{k=1}^{K}\varphi_{X_k}\left(\mathbf{t}^{\mathrm{T}}\mathbf{A}_k\right)\right| \geq \widetilde{g}(\left|\mathbf{t}^{\mathrm{T}}\mathbf{A}\right|) \geq \widetilde{g}(L\,|\mathbf{A}|\,|\mathbf{t}|),$$

where $|\mathbf{A}| = \max_{i,j}(|a_{ij}|)$. In the last inequality we have used that $\widetilde{g}$ is decreasing, and that

$$\left|\mathbf{t}^{\mathrm{T}}\mathbf{A}\right| = \max_i\left(\left|\sum_{j=1}^{L}a_{ij}t_j\right|\right) \leq L\,|\mathbf{A}|\,|\mathbf{t}|\,.$$

Define $g(t)=\widetilde{g}\,(L\,|\mathbf{A}|\,t)$. Function g inherits $\widetilde{g}$'s properties: it maps $\mathbb{R}^+$ onto $[0,1]$, it is decreasing and it is integrable, so that in particular $g\,(|\mathbf{t}|)\to 0$ when $|\mathbf{t}|\to\infty$. We can thus apply Theorem 1 and obtain:

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

with ε_N and T_N as in Lemma 1.

If H is a higher-order kernel of order $q \geq 2$, then there exists a function m such that $\varphi_H(v) = 1 + m(v)v^q$ for all $v \in [-1,1]$, and $\varphi_H(v) = 0$ for $v \notin [-1,1]$, where m is continuous on $[-1,1]$. So the last term on the right-hand side of (C6) is:

$$\begin{aligned}\int \left|\varphi_H\left(\frac{v}{T_N}\right) - 1\right| h_k(|v|)dv &= \int_{-T_N}^{T_N} \left|m\left(\frac{v}{T_N}\right)\right| \left(\frac{v}{T_N}\right)^q h_k(|v|)dv + 2\int_{T_N}^{+\infty} h_k(|v|)dv \\ &= \sup_{v\in[-1,1]} |m(v)| \cdot \left(\frac{1}{T_N^q}\int_{-T_N}^{T_N} v^q h_k(|v|)dv\right) + 2\int_{T_N}^{+\infty} h_k(|v|)dv,\end{aligned}$$

where $\sup_{v\in[-1,1]} |m(v)| = O(1)$ since m is continuous on $[-1,1]$.

This ends the proof of Theorem 2.

D "Plug-in" bandwidth selection

We here present the "plug-in" method of Delaigle and Gijbels (2004) to choose the bandwidth in deconvolution kernel density estimation. We focus on second-order kernels in the presentation.

Known error distribution. To present the method, let us consider the deconvolution problem with known error distribution $Y = X + U$, where f_U, or equivalently φ_U, is known. Based on a random sample $Y_1, ..., Y_N$, the deconvolution kernel density estimator of f_X is given by:

$$\widehat{f}_X(x) = \frac{1}{2\pi}\int \varphi_H\left(\frac{v}{T_N}\right) e^{-ivx}\frac{\widehat{\varphi}_Y(v)}{\varphi_U(v)}dv,$$

where $\widehat{\varphi}_Y(v) = \mathbb{E}_N e^{ivY}$ is the empirical characteristic function of Y.

Let the Mean Integrated Squared Error (MISE) of $\widehat{f}_X$ be:

$$\text{MISE}(T_N) = \mathbb{E}\left(\int \left(\widehat{f}_X(x) - f_X(x)\right)^2 dx\right).$$

The choice of T_N relies on the following approximation of the MISE:

$$\text{MISE}(T_N) \approx \frac{1}{2\pi N}\int \left|\varphi_H\left(\frac{v}{T_N}\right)\right|^2 |\varphi_U(v)|^{-2} dv + \frac{\mu_{H,2}^2 R\left(f_X''\right)}{4T_N^4}. \tag{D7}$$

where

$$\begin{aligned}\mu_{H,2} &= \int v^2 H(v)dv, \\ R\left(f_X''\right) &= \int \left[f_X''(x)\right]^2 dx.\end{aligned}$$

For example, $\mu_{H_2,2} = 6$.

The plug-in method estimates $R\left(f_X''\right)$ by the following algorithm.

1. Estimate $R\left(f_X''''\right)$ as if X was normally distributed:

$$\widehat{R}\left(f_X''''\right) = \frac{8!}{2^9 4!\sqrt{\pi}\left[\widehat{\text{Var}}(X)\right]^{\frac{9}{2}}}.$$

2. Minimize the following quantity with respect to T:

$$-\frac{\mu_{H,2}\widehat{R}\left(f_X''''\right)}{T^2}+\frac{1}{2\pi N}\int v^6\left|\varphi_H\left(\frac{v}{T}\right)\right|^2|\varphi_U(v)|^{-2}\,dv.$$

This quantity can be interpreted as the squared asymptotic bias of $\widehat{R}\left(f_X'''\right)$. This step yields $\widehat{T}$.

3. Compute:

$$\widehat{R}\left(f_X'''\right)=\frac{1}{2\pi}\int v^6\left|\varphi_H\left(\frac{v}{\widehat{T}}\right)\right|^2\left|\frac{\widehat{\varphi}_Y(v)}{\varphi_U(v)}\right|^2 dv.$$

4. Iterate one more time steps 2 and 3. This yields $\widehat{R}\left(f_X''\right)$.

Finally, once $R\left(f_X''\right)$ has been estimated, $\widehat{T}_N$ is obtained as the minimizer of the approximated MISE given by the right-hand side of (D7).

Unknown error distribution. In practice, we replace $\varphi_U(v)$ in the above expressions by an estimate of the c.f. of $\sum_{m\neq k}\frac{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_m}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k}X_m$, as explained in 5.3. Because of (33), a consistent estimate of that c.f. is given by

$$\widehat{\varphi}_U(v)=\frac{\widehat{\varphi}_{\mathbf{Y}}\left(\frac{v\boldsymbol{\theta}}{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}_k}\right)}{\widehat{\varphi}_{X_k}(v)}, \tag{D8}$$

where $\widehat{\varphi}_{\mathbf{Y}}$ is the empirical c.f. of $\mathbf{Y}$, and $\widehat{\varphi}_{X_k}$ is the estimate of the c.f. of X_k given by (22).