Recovering Latent Variables by Matching[*]

Manuel Arellano[†] Stéphane Bonhomme[‡]

Revised Draft: May 6, 2021

Abstract

We propose an optimal-transport-based matching method to nonparametrically estimate linear models with independent latent variables. The method consists in generating pseudo-observations from the latent variables, so that the Euclidean distance between the model's predictions and their matched counterparts in the data is minimized. We show that our nonparametric estimator is consistent, and we document that it performs well in simulated data. We apply this method to study the cyclicality of permanent and transitory income shocks in the Panel Study of Income Dynamics. We find that the dispersion of income shocks is approximately acyclical, whereas the skewness of permanent shocks is procyclical. By comparison, we find that the dispersion and skewness of shocks to hourly wages vary little with the business cycle.

KEYWORDS: Latent variables, nonparametric estimation, matching, factor models, optimal transport, income dynamics.

JEL CODES: C14, C33.

[*] We thank Colin Mallows, Tincho Almuzara, Alfred Galichon, Jiaying Gu, Kei Hirano, Pierre Jacob, Roger Koenker, Thibaut Lamadon, Guillaume Pouliot, Azeem Shaikh, Tim Vogelsang, Daniel Wilhelm, and audiences at various places for comments. Tincho Almuzara and Beatriz Zamorra provided excellent research assistance. Arellano acknowledges research funding from the Ministerio de Economía y Competitividad, Grant ECO2016-79848-P. Bonhomme acknowledges support from the NSF, Grant SES-1658920.

[†] CEMFI, Madrid.

[‡] University of Chicago.

1 Introduction

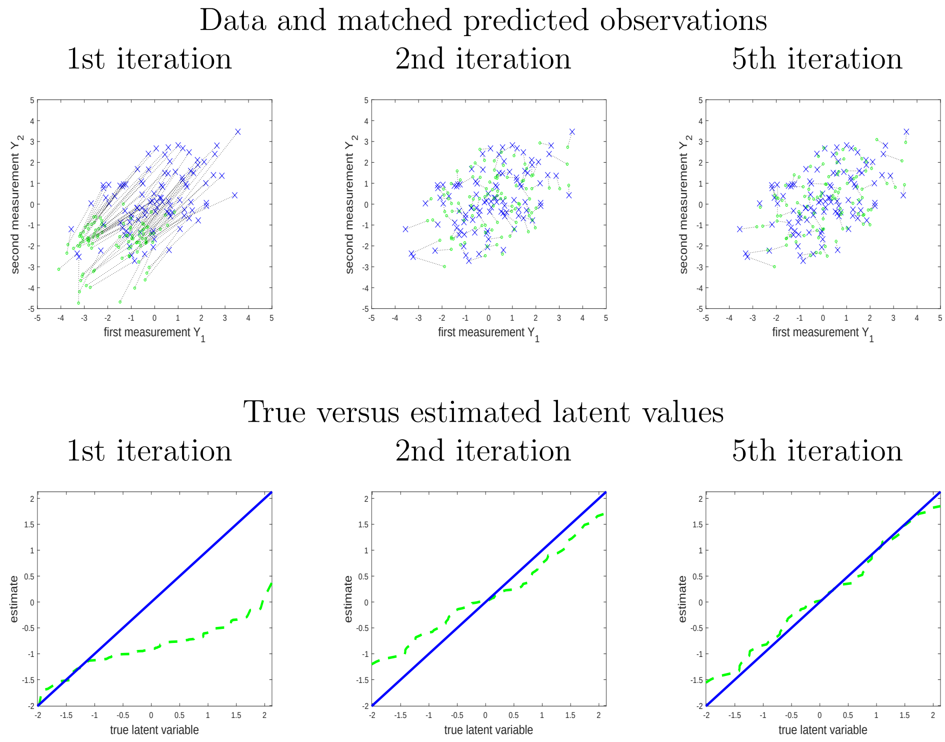
In this paper we propose a method to nonparametrically estimate a class of models with latent variables. We focus on linear factor models whose latent factors are mutually independent. These models have a wide array of applications, including measurement error models, models with repeated measurements, and error components models. We mention some applications in Section 2. In many empirical settings, such as in our economic application to the study of the cyclical behavior of income shocks, it is appealing not to restrict the functional form of the distributions of latent variables and adopt a nonparametric approach.

Nonparametric estimation based on empirical characteristic functions has been extensively studied in the literature; see Carroll and Hall (1988) and Stefanski and Carroll (1990), among many other contributions. However, while such Fourier-based methods apply to general mutivariate linear factor models with independent components (e.g., Horowitz and Markatou, 1996; Li and Vuong, 1998; Delaigle *et al.*, 2008, Bonhomme and Robin, 2010, Comte and Kappus, 2015), they tend to be sensitive to the choice of regularization parameters, and they do not guarantee that the estimated densities be non-negative and integrate to one. Recently, Efron (2016) motivated his "parametric g-modeling" approach by the difficulties of nonparametric estimation in this context.

In this paper we propose a novel nonparametric estimator, and provide evidence that it performs well even in relatively small samples. Our approach differs from the literature in two main aspects. First, we generate a sample of *pseudo-observations* that may be interpreted as the order statistics of the latent variables. Moments, densities, or other functionals can then be estimated based on them. In particular, densities will be non-negative and integrate to one by construction. Means or other features of the distribution of the latent variables conditional on the data, such as optimal predictors, can also be directly estimated.

The second main feature of our approach is that it is based on *matching*. Specifically, we generate pseudo-observations from the latent variables so that the Euclidean distance between the model's predictions and their matched counterparts in the data is minimized. The model predictions are computed as independent combinations of the pseudo latent observations. This "observation matching" estimation approach can be interpreted as a nonparametric counterpart to (simulated) method-of-moments estimators, which are commonly used in parametric econometric models. Our nonparametric approach, which amounts to minimizing a quadratic *Wasserstein* distance between empirical distribution functions, exploits

Figure 1: Illustration of the estimation algorithm



Data and matched predicted observations

1st iteration

2nd iteration

5th iteration

True versus estimated latent values

1st iteration

2nd iteration

5th iteration

Notes: The graphs correspond to one simulation from a model with two independent measurements $Y_1 = X_1 + X_2$, $Y_2 = X_1 + X_3$, *with* X_1, X_2, X_3 *mutually independent. The* X*'s are standardized Beta(2,2), and there are* $N = 100$ *observations. The top panel shows the observations* Y_1, Y_2 *(crosses) and the predicted observations* Y_1^{pred}, Y_2^{pred} *(circles), with a link between them when they are matched to each other. The bottom panel shows the estimates of* X_1 *values sorted in ascending order on the y-axis against the population values on the x-axis (dashed), and the 45 degree line (solid). See Sections 3 and 4 for details about the algorithm.*

linearity and independence to provide a tractable estimator.

As an illustration, in Figure 1 we show the results of several iterations of our algorithm, in a model with two independent measurements and 100 individuals. We start the algorithm from parameter values that are far from the true ones, in the left column. As shown in the top panel, the outcome observations in the data (in crosses) are first matched to model-based predictions (in circles). Pseudo-observations of the latent variables are then updated based on the matched outcome values. The objective function we aim to minimize is the sum of squares of the segments shown in the top panel. The bottom panel shows the estimates of the latent variables sorted in ascending order on the y-axis, against the true values on the x-axis. Within a few iterations, the model's predictions and the empirical observations tend to agree with each other, and the distribution of the pseudo latent observations gets close to the population distribution.[1]

Our approach builds on and generalizes an important idea due to Colin Mallows (2007), who proposed a "deconvolution by simulation" method based on iterating between sorts of the data and random permutations of pseudo-observations of a latent variable. Mallows (2007) focused on the classical deconvolution model with scalar outcome and known error distribution. Our main goal in this paper is to extend Mallows' insight and propose a framework to analyze estimators based on matching predicted values from the model to data observations.

In particular, as an extension of Mallows' (2007) original idea, we show how our method can handle multivariate outcomes, hence extending the scope of application to repeated measurements models and multi-factor models. While a number of estimation methods are available for nonparametric deconvolution with known error distribution (with contributions to multivariate deconvolution by Comte and Lacour, 2013, and Lepski and Willer, 2019, among others), the multivariate case with unknown error distribution — which is of interest in many applications — remains challenging. Our estimator exploits that the multi-factor models we consider are linear in the independent latent variables, even though they imply nonlinear restrictions on density functions.

A key step in our analysis is to relate the estimation problem to optimal transport; see for example Villani (2003). In our context, optimal transport provides a natural way to estimate models with multivariate outcomes via "generalized sorting" algorithms (i.e., matching algorithms) based on linear programming.

To establish the consistency of our estimator we use that, in large samples, it minimizes the Wasserstein distance between the population distribution of the data and the one implied by the model. This problem has a unique solution under suitable conditions on the characteristic functions of the factors (Székely and Rao, 2000). Consistency then follows from verifying the conditions for the consistency of sieve extremum estimators (e.g., Chen, 2007) in this setting. When analyzing the multivariate case, our arguments rely on properties of Wasserstein distances established in the optimal transport literature.

We illustrate the performance of our estimator on simulated data. Under various specifications of a nonparametric repeated measurements model, we find that it recovers accurately the true underlying quantile functions and densities, even for samples with only 100 indi-

[1] Codes to implement the estimator are available on the second author's webpage.

vidual observations. In addition, we find that our estimator performs comparably to or better than an oracle characteristic-function based estimator in our simulations. In contrast with Fourier methods, our estimator imposes that quantile functions be monotone, and that densities be non-negative. In the related problem of nonparametric instrumental variables estimation, Chetverikov and Wilhelm (2017) show that imposing monotonicity in estimation can help alleviating ill-posedness issues. We conjecture that this feature contributes to explain the finite-sample performance of our estimator.

We then apply our method to study the cyclicality of permanent and transitory income shocks in the United States. Answering this question is important, since a well-calibrated cyclical income process is a key input to many economic models of business cycle dynamics. Storesletten *et al.* (2004) estimate using the Panel Study of Income Dynamics (PSID) that the dispersion of persistent shocks is countercyclical. However, using a nonparametric descriptive analysis, Guvenen *et al.* (2014) find using administrative data that the dispersion of log-income growth is acyclical, whereas skewness is procyclical. Recently, Busch *et al.* (2018) find similar results using the PSID.

We revisit this debate by working with a permanent-transitory model of log-income dynamics, and estimating the annual densities of permanent and transitory shocks nonparametrically. Using the PSID, we estimate that income shocks are not normally distributed, confirming previous evidence using other nonparametric methods. Our main finding is that the dispersion of income shocks is approximately acyclical, whereas the skewness of permanent shocks is procyclical. By comparison, our nonparametric estimates suggest that the dispersion and skewness of shocks to hourly wages vary little with the business cycle.

Our matching-based, minimum Wasserstein distance estimator is related to recent work on the estimation of parametric generative models (see Bernton *et al.*, 2017; Genevay *et al.*, 2017; Bousquet *et al.*, 2017). In contrast with this emerging literature, the models we consider here are nonparametric. In an early theoretical contribution, Bassetti *et al.* (2006) study consistency in minimum Wasserstein distance estimation. Recently, Rigollet and Weed (2019) develop a minimum Wasserstein deconvolution approach for uncoupled isotonic regression, and Rigollet and Weed (2018) relate maximum-likelihood scalar deconvolution under Gaussian noise to entropic regularized optimal transport. Lastly, our general estimation strategy is also related to Galichon and Henry's (2011) analysis of partially identified models.

As we show in Section 8, our matching approach can be generalized to nonparametric estimation of other latent variables models. We describe how to extend our approach to estimate linear factor models where blocks of factors are independent of each other, but factors are not independent within blocks. To do so, we exploit the vector quantile representation of Carlier *et al.* (2016). In addition, in the appendix we outline several possible generalizations: to random coefficients models with exogenous covariates (Beran and Hall, 1992), nonparametric deconvolution under heteroskedasticity (Delaigle and Meister, 2008), and nonparametric finite mixture models (Hall and Zhou, 2003).

2 Independent factor models: some applications

We study linear independent factor models of the form $Y = AX$, where $Y = (Y_1, ..., Y_T)'$, $X = (X_1, ..., X_K)'$, A is a $T \times K$ matrix, and the components $X_1, ..., X_K$ are mutually independent. In this section we review several examples of models and applications that have such a structure.

We focus on the case $K > T$, so the system is singular and the realizations of the latent variables are not identifiable, although under suitable conditions their distributions will be.[2] We assume that the matrix A is known. However, it is sufficient that a consistent estimate of A be available. In applications with unknown A, a consistent estimate can be obtained using the variance restrictions $\text{Var}(Y) = A\,\text{Var}(X)A'$, where $\text{Var}(X)$ is diagonal, provided such restrictions are sufficient for identification. Depending on the setting, these variance restrictions may be complemented with higher-order moment restrictions (e.g., Bonhomme and Robin, 2009). When A is unknown, our recommendation is to estimate A in a first step, and then apply our method to estimate the distribution of X. Our consistency result is unaffected, provided the estimate of A is consistent.[3]

Nonparametric deconvolution. When $T = 1$, $Y = X_1 + X_2$, and X_2 has a known or consistently estimable distribution, one obtains the scalar nonparametric deconvolution model. Nonparametric deconvolution is often used to deal with the presence of measurement error. In such settings, Y is an error-ridden variable, X_1 is the true value of the variable, and X_2 is an independent, classical measurement error (e.g., Carroll *et al.*, 2006; Chen *et al.*, 2011;

[2] When $A'A$ is non-singular and A is known, $\widehat{X} = (A'A)^{-1}A'Y$ recovers X exactly. We are interested in situations, such as deconvolution and filtering, where exact recovery of the latent variables is not possible.

[3] An interesting possibility would be to jointly estimate A and the distribution of X. Although we do not study it formally, we comment on this possibility in Subsection 3.2.

Schennach, 2013a).[4] The nonparametric deconvolution problem has been extensively studied, so we necessarily need to provide only a selected set of references here; see Meister (2009) for a review. This literature has proposed adaptive and non-adaptive density estimators, and it has characterized minimax rates of convergence in various settings. For example, the case with known error distribution has been approached using kernel deconvolution estimators (e.g., Carroll and Hall, 1988; Fan, 1991), wavelet methods (e.g., Pensky and Vidakovic, 1999), regularization techniques (e.g., Carrasco and Florens, 2011), and nonparametric maximum likelihood methods (e.g., Gu and Koenker, 2017). Studies where the error distribution is unknown include Johannes (2009), Comte and Lacour (2011), and Kappus and Mabon (2014). While a large part of the literature focuses on estimating density functions, Dattner *et al.* (2011) propose an estimator of the distribution function.

Repeated measurements. A leading example of a linear independent factor model is:

$$Y_t = \underbrace{\alpha}_{\equiv X_1} + \underbrace{\varepsilon_t}_{\equiv X_{t+1}}, \quad t = 1, ..., T, \tag{1}$$

where $Y_1, ..., Y_T$ are observed outcomes and $\alpha, \varepsilon_1, ..., \varepsilon_T$ are latent and mutually independent. Working with $T = 2$, Kotlarski (1967) provided simple conditions under which the density functions of the latent factors are nonparametrically identified in model (1). This structure arises frequently in applications: α can be a latent skill of an individual measured with error, or a teacher- or bank-specific effect, for example. Compared to commonly used Gaussian specifications, a nonparametric estimator of the distribution of α in (1) will be robust to functional form assumptions under the assumption of mutual independence. Non-Gaussianity, such as skewness or fat tail behavior, is relevant in many empirical settings. In model (1), nonparametric estimators based on empirical characteristic functions can be constructed by mimicking and extending Kotlarski's proof (e.g., Li and Vuong, 1998; Li, 2002; Horowitz and Markatou, 1996; Comte and Kappus, 2015). Recently, Duval and Kappus (2017) propose an estimator for a related grouped data model, and study its properties.

Error components. A prominent error component model in economics is the permanent-transitory model for the dynamics of log-income: $Y_t = \eta_t + \varepsilon_t$, where $\eta_t = \eta_{t-1} + v_t$ is a

[4] Other applications in economics include the estimation of the heterogeneous effects of an exogenous binary treatment under the assumption that the potential outcome in the absence of treatment is independent of the gains from treatment (Heckman *et al.*, 1997), and the estimation of the distribution of time-invariant random coefficients of binary treatments in panel data models (Arellano and Bonhomme, 2012).

random walk with independent innovations, and all ε_t's and v_t's are independent over time and independent of each other and of the initial η_0 (e.g., Hall and Mishkin, 1982; Blundell *et al.*, 2008). Identification is established in Székely and Rao (2000). Bonhomme and Robin (2010) propose nonparametric characteristic-function based estimators of factor densities. In such settings, a nonparametric approach is able to capture the skewness and kurtosis of income shocks.[5]

Extensions to other classes of models. While we focus our formal analysis on linear models with independent factors, in Section 8 and Appendix E we outline how the main idea can be extended to estimate other models with latent variables. One such generalization, which we present in Section 8, is a model with blocks of factors that are independent, allowing factors to be dependent within blocks. Formally, suppose that a T-dimensional outcome vector Y can be written as $Y = \sum_{\ell=1}^{L} A_\ell X_\ell$, where $X_1, ..., X_L$ are mutually independent random *vectors*. For all $\ell \in \{1, ..., L\}$, X_ℓ has n_ℓ scalar components $X_{\ell k}$ — which are *not* assumed independent of one another — and A_ℓ is a known $T \times n_\ell$ matrix. Allowing for dependent components within independent blocks of latent factors is of interest in a variety of models, such as the following model for longitudinal data:

$$Y_{it} = \alpha_i + \beta_i t + \varepsilon_{it}, \quad i = 1, ..., N, \quad t = 1, ..., T,$$

where (α_i, β_i), ε_{i1}, ..., ε_{iT} are mutually independent, yet the unit-specific intercept α_i and slope β_i can depend on each other.[6] In addition, in Appendix E we describe how the same ideas can be applied to linear models with random coefficients, finite mixture models, and deconvolution models with heteroskedastic errors. Although studying each of these extensions in detail would require a separate analysis, we outline how the principle of latent variable estimation by matching can be applied in all these instances.

[5] See also Botosaru and Sasaki (2015). Our approach may be used to estimate linear autoregressive specifications of the form $\eta_t = \alpha + \rho\eta_{t-1} + v_t$, where we estimate (α, ρ) — i.e., the matrix A — in a first step. An important application of error components models is to relax independence in repeated measurements models such as (1). This can be done provided T is large enough. Modeling ε_t in (1) as a finite-order moving average or autoregressive process with independent innovations preserves the linear independent factor structure of the model (Arellano and Bonhomme, 2012; see also Hu *et al.*, 2019). In addition, in model (1) Schennach (2013b) points out that full independence between the factors is not necessary, and that sub-independence suffices to establish identification.

[6] Ben Moshe (2017) shows how to allow for arbitrary subsets of dependent factors, and proposes characteristic-function based estimators.

3 Latent variable estimation by matching

In this section, to introduce the main ideas we start by describing our estimator in the scalar nonparametric deconvolution model. We then show how the same approach can be used to estimate linear multi-factor models with independent factors.

3.1 Nonparametric deconvolution

Let $Y = X_1 + X_2$ be a scalar outcome, where X_1 and X_2 are independent, X_1 is unobserved to the analyst, and its distribution is unspecified. We assume that Y, X_1 and X_2 are continuously distributed, and postpone more specific assumptions until Section 5. Let F_Z denote the cumulative distribution function (cdf) of any random variable Z. We assume that two random samples, $Y_1, ..., Y_N$ and $X_{12}, ..., X_{N2}$, drawn from F_Y and F_{X_2}, respectively, are available.[7]

Our goal is to estimate a sample of *pseudo-observations* $\widehat{X}_{11}, ..., \widehat{X}_{N1}$, whose empirical cdf is asymptotically distributed as F_{X_1} as N tends to infinity. To do so, we minimize a distance between the sample of observed Y's and a sample of Y's predicted by the model. We rely on the quadratic Wasserstein distance (see Chapter 7 in Villani, 2003), which is the minimum Euclidean distance between observed Y's and predicted Y's with respect to all possible reorderings of the observations.

Let $\overline{C}_N > 0$ and $\underline{C}_N > 0$ be two constants. We define the parameter space $\mathcal{X}_N$ to be the set of vectors $X_1 = (X_{11}, ..., X_{N1}) \in \mathbb{R}^N$ such that $|X_{i1}| \leq \overline{C}_N$ and $\underline{C}_N \leq (N+1)(X_{i+1,1} - X_{i1}) \leq \overline{C}_N$ for all i. The constants $\underline{C}_N$ and $\overline{C}_N$ play a role in our consistency argument below, and we will study how their choice affects our estimator in simulations. Let Π_N denote the set of permutations π of $\{1, ..., N\}$. We propose to compute:

$$\widehat{X}_1 = \underset{X_1 \in \mathcal{X}_N}{\operatorname{argmin}} \left\{ \min_{\pi \in \Pi_N} \sum_{i=1}^{N} \left(Y_{\pi(i)} - X_{\sigma(i),1} - X_{i2} \right)^2 \right\}, \tag{2}$$

where σ is a random permutation in Π_N (i.e., a uniform draw on Π_N), independent of $Y_1, ..., Y_N, X_{12}, ..., X_{N2}$.

To interpret the objective function in the right-hand side of (2), note that, for any random permutation σ, $Z_i \equiv X_{\sigma(i),1} + X_{i2}$, $i = 1, ..., N$, are N draws from the model. Predicted values

[7] The sample sizes being the same for Y and X_2 is not essential and can easily be relaxed. In a setting where the cdf F_{X_2} is known, one can draw a sample from it, or alternatively work with an integral counterpart to our estimator.

could be generated in other ways. For example, one could instead compute $X_{i1}+\widetilde{X}_{i2}$, for $\widetilde{X}_{i2}$ i.i.d. draws from the empirical distribution of X_{i2}. Alternatively, one could generate $R>1$ predictions per observation i, although here we take $R=1$ to minimize computation cost.[8]

A simple way to reduce the dependence of the estimator on the random σ draw is to compute $\widehat{X}_{i1}^{(m)}$, for $i=1,...,N$ and $m=1,...,M$, where $\sigma^{(1)},...,\sigma^{(M)}$ are independent random permutations drawn from Π_N, and to report the averages: $\widehat{X}_{i1}=\frac{1}{M}\sum_{m=1}^{M}\widehat{X}_{i1}^{(m)}$, for $i=1,...,N$. For fixed M, such averages will be consistent as N tends to infinity under similar conditions as our baseline estimator.

The estimator $\widehat{X}_1$ in (2) minimizes the Wasserstein distance between the empirical distributions of the model predictions $Z_i=X_{\sigma(i),1}+X_{i2}$ and the outcome observations Y_i:

$$W_2(\widehat{F}_Y,\widehat{F}_Z)=\left\{\min_{\pi\in\Pi_N}\sum_{i=1}^{N}\left(Y_{\pi(i)}-Z_i\right)^2\right\}^{\frac{1}{2}}. \tag{3}$$

Since Y_i and Z_i are scalar, the Hardy-Littlewood-Pólya rearrangement inequality implies that the solution to (3) is to sort Y_i's and Z_i's in the same order. That is, if Y_i are sorted in ascending order, and letting $\widehat{\pi}$ denote the minimum argument in (3), $\widehat{\pi}(i)=\widehat{\text{Rank}}(Z_i)\equiv N\widehat{F}_Z(Z_i)$ is the rank of Z_i.

3.2 Nonparametric factor models

We now apply the same idea to a general linear independent multi-factor model $Y=AX$, where A is a $T\times K$ matrix with generic element a_{tk}, and $X=(X_1,...,X_K)'$ with $X_1,...,X_K$ mutually independent. For simplicity we assume that X and Y have zero mean.[9] We seek to compute pseudo-observations $\widehat{X}_{11},...,\widehat{X}_{N1}$, ..., $\widehat{X}_{1K},...,\widehat{X}_{NK}$ that minimize the Wasserstein distance between the sample of observed Y's, which here are $T\times 1$ vectors, and the sample of Y's predicted by the factor model.

As before, let $\overline{C}_N>0$ and $\underline{C}_N>0$ be two constants, and let $\mathcal{X}_N$ be the set of $(X_1,...,X_N)\in\mathbb{R}^{NK}$ such that $|X_{ik}|\leq\overline{C}_N$ and $\underline{C}_N\leq(N+1)(X_{i+1,k}-X_{ik})\leq\overline{C}_N$ for

[8] Specifically, one could compute $X_{\sigma(i,r),1}+X_{i2}$, with $\sigma(\cdot,1),...,\sigma(\cdot,R)$ being R independent permutations. In that case, π would be a generalized permutation, mapping $\{1,...,N\}^R$ to $\{1,...,N\}$.

[9] It is common in applications to assume that some of the X_k's have zero mean while leaving the remaining means unrestricted. For example, in the repeated measurements model, assuming that $\mathbb{E}(X_1)=0$ suffices for identification. Our algorithm can easily be adapted to such cases.

all i and k, and $\sum_{i=1}^{N} X_{ik} = 0$ for all k. We define:

$$\widehat{X} = \underset{X \in \mathcal{X}_N}{\text{argmin}} \left\{ \min_{\pi \in \Pi_N} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{\pi(i),t} - \sum_{k=1}^{K} a_{tk} X_{\sigma_k(i),k} \right)^2 \right\}, \tag{4}$$

where $\sigma_1, ..., \sigma_K$ are independent random permutations in Π_N, independent of $Y_{11}, ..., Y_{NT}$.[10]

As in the scalar case, $Z_{it} \equiv \sum_{k=1}^{K} a_{tk} X_{\sigma_k(i),k}$, $i = 1, ..., N$, $t = 1, ..., T$, are NT predicted values from the factor model. Hence, as before, the vector $\widehat{X}$ minimizes the Wasserstein distance between the empirical distributions of the data $(Y_{i1}, ..., Y_{iT})$ and model predictions $(Z_{i1}, ..., Z_{iT})$. When Y_i are multivariate, the minimization with respect to π inside the brackets in (4) does not have an explicit form in general. However, from optimal transport theory it is well-known that the solution can be obtained by solving a linear program. We will exploit this feature in our estimation algorithm.

Densities and expectations. In Section 5 we will provide conditions under which $\widehat{X}_{ik}$, $i = 1, ..., N$, consistently estimate the quantile function of X_k. More precisely, we will show that $\max_{i=1,...,N} |\widehat{X}_{ik} - F_{X_k}^{-1}(\frac{i}{N+1})|$ tends to zero in probability asymptotically. This provides uniformly consistent estimators of the quantile functions of the latent variables, which can in turn be used for density estimation under a slight modification of the parameter space $\mathcal{X}_N$. Indeed, let us restrict the parameter space to elements $X = (X_1, ..., X_N)$ in $\mathcal{X}_N$ which satisfy the following additional restrictions on second-order differences: $(N+1)^2 |X_{i+2,k} - 2X_{i+1,k} + X_{ik}| \leq \overline{C}_N$, for all i and k. Let us then define, for a bandwidth parameter $b > 0$ and a kernel function $\kappa \geq 0$ that integrates to one:

$$\widehat{f}_{X_k}(x) = \frac{1}{Nb} \sum_{i=1}^{N} \kappa \left(\frac{\widehat{X}_{ik} - x}{b} \right), \quad x \in \mathbb{R}. \tag{5}$$

We will show that $\widehat{f}_{X_k}$ is uniformly consistent for the density of X_k, under suitable conditions on the kernel κ and bandwidth b.

In addition, our estimator delivers simple consistent estimators of unconditional and conditional expectations, as we discuss in Appendix B. As an example of practical interest, in the repeated measurements model (1) the best predictor of X_1 under squared loss can be

[10] If $A = \{a_{tk}\}$ is unknown and $\widehat{A} = \{\widehat{a}_{tk}\}$ is a consistent estimate of it, we replace a_{tk} by $\widehat{a}_{tk}$ in (4). We proceed similarly in the algorithm we propose in the next section. Alternatively, one could jointly minimize the objective function on the right-hand side of (4) with respect to both X and $\{a_{tk}\}$. Here we do not study the formal properties of such a joint estimation method.

estimated as:

$$\widehat{\mathbb{E}}(X_1 \,|\, Y = Y_i) = \sum_{i=1}^{N} \widehat{\omega}_i \widehat{X}_{i1}, \tag{6}$$

where the weights $\widehat{\omega}_i$ are given by: $\widehat{\omega}_i = \frac{\prod_{t=1}^{T} \widehat{f}_{X_{t+1}}(Y_{it} - \widehat{X}_{i1})}{\sum_{j=1}^{N} \prod_{t=1}^{T} \widehat{f}_{X_{t+1}}(Y_{jt} - \widehat{X}_{j1})}$, for $i = 1, ..., N$.

4 Computation

The optimization problems in (2) and (4) are mixed integer quadratic programs. Since exact integer programming algorithms are currently limited in the dimensions they can allow for, here we describe a simple, practical method to minimize (2) and (4).

4.1 Algorithm

The algorithm we propose is based on the observation that, for given $X_{11}, ..., X_{NK}$ values, (4) is a discrete optimal transport problem, hence it can be solved by any linear programming routine. In turn, given π, (4) is a monotone least squares problem. Our estimation algorithm is as follows. Here we focus on the general form (4), since the estimator for the scalar deconvolution model (2) is a special case of it.

Algorithm.

- *Start with initial values* $\widehat{X}_1^{(1)}, ..., \widehat{X}_N^{(1)}$ *in* $\mathbb{R}^K$*. Iterate the following two steps on* $s = 1, 2, ...$ *until convergence.*
- *(Matching step) Given* $\widehat{X}_1^{(s)}, ..., \widehat{X}_N^{(s)}$*, compute:*[11]

$$\begin{aligned} \widehat{\pi}^{(s+1)} &= \underset{\pi \in \Pi_N}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{\pi(i),t} - \sum_{k=1}^{K} a_{tk} \widehat{X}^{(s)}_{\sigma_k(i),k} \right)^2 \\ &= \underset{\pi \in \Pi_N}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\sum_{k=1}^{K} a_{tk} \widehat{X}^{(s)}_{\sigma_k(i),k} \right) Y_{\pi(i),t}. \end{aligned} \tag{7}$$

- *(Update step) Compute:*

$$\widehat{X}^{(s+1)} = \underset{X \in \mathcal{X}_N}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{\widehat{\pi}^{(s+1)}(i),t} - \sum_{k=1}^{K} a_{tk} X_{\sigma_k(i),k} \right)^2. \tag{8}$$

[11] Notice that, since π is a permutation, $\sum_{i=1}^{N} \sum_{t=1}^{T} Y^2_{\pi(i),t} = \sum_{i=1}^{N} \sum_{t=1}^{T} Y^2_{it}$ does not depend on π.

Starting from a given set of parameter values, iterating between (7) and (8) is guaranteed to weakly decrease the value of the objective function in (4). In all our experiments we used a tolerance of at most 10^{-4} for the difference in objective functions, and we never observed any failure of convergence. Both steps in the algorithm are straightforward to implement. The matching step (7) can be computed by linear programming, due to the fact that the linear programming relaxation of a discrete optimal transport problem has integer-valued solutions.[12] Formally, $\widehat{\pi}^{(s+1)}$ in (7) is a solution to the following *linear program*:

$$\max_{P\in\mathcal{P}_N} \sum_{i=1}^N\sum_{t=1}^T\left(\sum_{k=1}^K a_{tk}\widehat{X}^{(s)}_{\sigma_k(i),k}\right)\left(\sum_{j=1}^N P_{ij}Y_{jt}\right),$$

where $\mathcal{P}_N$ denotes the set of $N\times N$ matrices with non-negative elements, whose rows and columns all sum to one. In the scalar nonparametric deconvolution case (2) with Y_i sorted in ascending order, this gives $\widehat{\pi}^{(s+1)}(i)=\widehat{\text{Rank}}\left(\widehat{X}^{(s)}_{\sigma(i),1}+X_{i2}\right)$ for all i.

Remark 1. *It is possible to write $\widehat{X}=(\widehat{X}_1,...,\widehat{X}_N)$ in (4) as the solution to a* quadratic program*:*

$$(\widehat{X},\widehat{P})=\underset{X\in\mathcal{X}_N,\,P\in\mathcal{P}_N}{\text{argmin}}\sum_{i=1}^N\sum_{t=1}^T\left\{\left(\sum_{k=1}^K a_{tk}X_{\sigma_k(i),k}\right)^2-2\left(\sum_{k=1}^K a_{tk}X_{\sigma_k(i),k}\right)\left(\sum_{j=1}^N P_{ij}Y_{jt}\right)\right\},$$

which is not convex in general. Our estimation algorithm is a method to solve this non-convex quadratic program. However, the algorithm is not guaranteed to reach a global minimum in (4). Our implementation is based on starting the algorithm from multiple random values. We will assess the impact of starting values on simulated data. Note an important difference with Fourier methods for deconvolution based on characteristic functions, which do not require nonlinear optimization and are typically very fast.

Remark 2. *In dual form, the matching step involves of the order of N parameters. In our implementation we use a standard linear programming solver (in Gurobi). In larger samples than the ones we consider in our simulations and application, an alternative possibility would be to rely instead on entropic-regularized optimal transport methods (Cuturi, 2013; Genevay et al., 2016), which can take advantage of smooth optimization techniques.*[13]

[12] See for example Chapter 3 in Galichon (2016) on discrete Monge-Kantorovitch problems.

[13] An entropic-regularized counterpart to (4) is, for $\epsilon_N>0$:

$$\widehat{X}=\underset{X\in\mathcal{X}_N}{\text{argmin}}\left\{\min_{P\in\mathcal{P}_N}\sum_{i=1}^N\sum_{j=1}^N P_{ij}\sum_{t=1}^T\left(Y_{jt}-\sum_{k=1}^K a_{tk}X_{\sigma_k(i),k}\right)^2+\epsilon_N\sum_{i=1}^N\sum_{j=1}^N P_{ij}\left(\ln(P_{ij})-1\right)\right\}.$$

4.2 Comparison to Mallows (2007)

Our algorithm may be seen as a generalization of Mallows' (2007) "deconvolution by simulation" method. To highlight the connection, consider the scalar nonparametric deconvolution model. The two steps in our algorithm take the following form (when Y_i are sorted):

$$\widehat{\pi}^{(s+1)}(i) = \widehat{\mathrm{Rank}}\left(\widehat{X}^{(s)}_{\sigma(i),1} + X_{i2}\right), \quad i = 1, ..., N,$$

$$\widehat{X}_1^{(s+1)} = \underset{X_1 \in \mathcal{X}_N}{\mathrm{argmin}} \sum_{i=1}^{N} \left(Y_{\widehat{\pi}^{(s+1)}(i)} - X_{\sigma(i),1} - X_{i2}\right)^2 .$$

The Mallows (2007) algorithm is closely related to this algorithm. The main difference is that, instead of minimizing an objective function for fixed values of the random permutation σ, random permutations are re-drawn in each step of the algorithm. In addition, the ordering of the X_{i1}'s is not restricted, and neither are the values and increments of the X_{i1}'s. Formally, the sub-steps of the Mallows algorithm are the following:

- Draw a random permutation $\sigma^{(s)} \in \Pi_N$.
- Compute $\widehat{\pi}^{(s+1)}(i) = \widehat{\mathrm{Rank}}\left(\widehat{X}^{(s)}_{\sigma^{(s)}(i),1} + X_{i2}\right)$, $i = 1, ..., N$.
- Compute $\widehat{X}^{(s+1)}_{\sigma^{(s)}(i),1} = Y_{\widehat{\pi}^{(s+1)}(i)} - X_{i2}$, $i = 1, ..., N$.[14]

To provide intuition about this algorithm, Mallows (2007) observes that, starting with draws from the true latent X_1, one expects the iteration to continue to draw from that distribution. However, starting from different values, the $\widehat{X}_1$ vectors implied by the algorithm will follow a complex N-dimensional Markov Chain. Moreover, the consistency properties of the Mallows estimator are currently unknown. Lastly, note that the methods introduced in this paper naturally deliver counterparts to the Mallows algorithm for general linear independent factor models.

5 Consistency analysis

In this section we provide conditions under which the estimators introduced in Section 3 are consistent. For $k \in \{1, ..., K\}$, let us denote the quantile function of X_k as:

$$F^{-1}_{X_k}(\tau) = \inf \left\{x \in \mathrm{Supp}(X_k) \,:\, F_{X_k}(x) \geq \tau\right\}, \text{ for all } \tau \in (0,1).$$

[14] Strictly speaking, Mallows (2007) redefines $\widehat{X}^{(s+1)}_{i1} \equiv \widehat{X}^{(s+1)}_{\sigma^{(s)}(i),1}$ for all $i = 1, ..., N$ at the end of step s, and then applies the random permutation $\sigma^{(s+1)}$ to the new $\widehat{X}^{(s+1)}$ values. This difference with the algorithm outlined here turns out to be immaterial, since the composition of $\sigma^{(s+1)}$ and $\sigma^{(s)}$ is also a random permutation of $\{1, ..., N\}$.

In addition, for any candidate quantile function H_k that maps the unit interval to the real line, let us define the following Sobolev sup-norms:

$$\|H_k\|_\infty = \sup_{\tau\in(0,1)} |H_k(\tau)|, \quad \text{and} \quad \|H_k\|_{1,\infty} = \max_{m\in\{0,1\}} \sup_{\tau\in(0,1)} |\nabla^m H_k(\tau)|,$$

where $\nabla^m H_k$ denotes the m-th derivative of H_k (when it exists). We will simply denote $\nabla = \nabla^1$ for the first derivative.

To a solution $\widehat{X}_k$ to (4),[15] we will associate an interpolating quantile function $\widehat{H}_k$ such that $\widehat{H}_k\left(\frac{i}{N+1}\right) = \widehat{X}_{ik}$ for all i. We will then show that $\|\widehat{H}_k - F_{X_k}^{-1}\|_\infty = o_p(1)$. This result will be obtained as an application of the consistency theorem for sieve extremum estimators in Chernozhukov *et al.* (2007).

We make the following assumptions.

Assumption 1.

(i) *(Continuity and support) Y and X have compact supports in $\mathbb{R}^T$ and $\mathbb{R}^K$, respectively, and admit absolutely continuous densities f_Y, f_X that are bounded away from zero and infinity. Moreover, f_Y is differentiable with bounded derivatives.*

(ii) *(Identification) The densities f_{X_k}, $k = 1, ..., K$, are identified given f_Y.*

(iii) *(Penalization) $\overline{C}_N$ is increasing and $\underline{C}_N$ is decreasing with $\lim_{N\to+\infty} \overline{C}_N = \overline{C}$ and $\lim_{N\to+\infty} \underline{C}_N = \underline{C}$, where $\overline{C}$ and $\underline{C} < \overline{C}$ are such that, for all k, $\|F_{X_k}^{-1}\|_{1,\infty} \leq \overline{C}$ and $\nabla F_{X_k}^{-1}(\tau) \geq \underline{C}$ for all $\tau \in (0,1)$.*

(iv) *(Sampling) $(Y_{i1}, ..., Y_{iT})$, $i = 1, ..., N$, are i.i.d.*

In part (i) we require the latent factors and observed measurements to be continuously distributed. This rules out discrete latent variables. Though convenient for the derivations, the compact supports assumption is strong. This could be relaxed by working with weighted norms, at the cost of achieving a weaker consistency result.[16] The simulation experiments we report below suggest that the estimator continues to perform well when supports are unbounded. For identification in part (ii), it suffices that the characteristic functions of X_k do not vanish on the real line, and the vectors $\operatorname{vec} A_k A_k'$ be linearly independent (Székely

[15] It is not necessary for $\widehat{X}_k$ to be an exact minimizer of (4). As we show in the proof, it suffices that the value of the objective function at $(\widehat{X}_1, ..., \widehat{X}_K)$ be in an ϵ_N-neighborhood of the global minimum, for ϵ_N tending to zero as N tends to infinity.

[16] Part (i) ensures that $F_{X_k}^{-1}$ belongs to an $\|\cdot\|_{1,\infty}$-ball, which is compact under $\|\cdot\|_\infty$ (Gallant and Nychka, 1987). Compactness can be preserved when norms are replaced by weighted norms (e.g., using polynomial or exponential weights); see for example Theorem 7 in Freyberger and Masten (2019), and the analysis in Newey and Powell (2003).

and Rao, 2000); moreover, the assumption that characteristic functions are non-zero can be relaxed (Evdokimov and White, 2012). The constants $\underline{C}_N$ and $\overline{C}_N$ appearing in part (iii) ensure that the $\widehat{X}_{ik}$ values are bounded and of bounded variation. In Section 6 we assess the impact of $\underline{C}_N$ and $\overline{C}_N$ in simulations, and we provide a simple data-driven approach. Lastly, the independent random permutations $\sigma_1, ..., \sigma_K$ in (4) depend on N, although we have omitted this dependence for conciseness.

Consistency is established in the following theorem. Proofs are in Appendix A.

Theorem 1. *Consider the independent factor model $Y = AX$. Let Assumption 1 hold. Then, as N tends to infinity:*

$$\max_{i\in\{1,...,N\}} \left| \widehat{X}_{ik} - F_{X_k}^{-1}\left(\frac{i}{N+1}\right) \right| = o_p(1), \qquad \textit{for all } k = 1, ..., K.$$

While Theorem 1 does not formally cover the scalar deconvolution model, the same proof arguments can be used to show the following result, under assumptions that mimic those of Theorem 1.

Corollary 1. *Consider the scalar deconvolution model $Y = X_1 + X_2$, where one observes two samples $Y_1, ..., Y_N$ and $X_{21}, ..., X_{2N}$ from Y and X_2, respectively. Let Assumption A1 in Appendix A hold. Then, as N tends to infinity:*

$$\max_{i\in\{1,...,N\}} \left| \widehat{X}_{i1} - F_{X_1}^{-1}\left(\frac{i}{N+1}\right) \right| = o_p(1).$$

An important step in the proof of Theorem 1 is to define the population counterpart to the estimation problem (4). Let μ_Y denote the population measure of Y. Moreover, for any candidate quantile functions $H = (H_1, ..., H_K)$, let μ_{AH} denote the population measure of the random vector $Z \equiv \sum_{k=1}^K A_k H_k(V_k)$, where $V_1, ..., V_K$ are independent standard uniform random variables on the unit interval. Finally, let $\mathcal{M}(\mu_Y, \mu_{AH})$ denote the set all possible joint distributions of the random vectors Y and $\sum_{k=1}^K A_k H_k(V_k)$, with marginals μ_Y and μ_{AH}. The population objective function is then:

$$Q(H) \equiv \inf_{\pi\in\mathcal{M}(\mu_Y,\mu_{AH})} \mathbb{E}_\pi \left[\sum_{t=1}^T \left(Y_t - \sum_{k=1}^K a_{tk} H_k(V_k) \right)^2 \right], \tag{9}$$

which is the quadratic Wasserstein distance between the population distribution of the data and the one implied by the model. Under part (ii) in Assumption 1, $Q(H)$ is uniquely minimized at the true quantile functions $H_k = F_{X_k}^{-1}$. In the scalar deconvolution model, the population objective takes the explicit form:

$$Q(H_1) \equiv \mathbb{E}\left[\left(F_Y^{-1}\left(\int_0^1 F_{X_2}\left(H_1(V_1) + F_{X_2}^{-1}(V_2) - H_1(\tau)\right) d\tau\right) - H_1(V_1) - F_{X_2}^{-1}(V_2)\right)^2\right],$$

where the expectation is taken with respect to independent standard uniform random variables V_1 and V_2, and the integral is simply the population rank of $H_1(V_1) + F_{X_2}^{-1}(V_2)$.

Densities and expectations. Under slightly stronger assumptions, Theorem 1 can be modified to obtain consistent estimators of both $F_{X_k}^{-1}$ and its derivative, which can then be used for density estimation. To see this, let us denote as $\mathcal{X}_N^{(2)}$ the set of X in $\mathcal{X}_N$ which satisfy the restrictions on second-order differences: $(N+1)^2 \left|X_{i+2,k} - 2X_{i+1,k} + X_{ik}\right| \leq \overline{C}_N$, for all i and k, and replace the minimization in (4) by a minimization with respect to $X \in \mathcal{X}_N^{(2)}$.

We then have the following result.[17]

Corollary 2. *Consider the independent factor model $Y = AX$. Suppose that Assumption 1 holds, and that, for all k, $\max_{m\in\{0,1,2\}} \sup_{\tau\in(0,1)} \left|\nabla^m (F_{X_k}^{-1})(\tau)\right| \leq \overline{C}$. Let b in (5) be such that $b \to 0$ and $Nb^2 \to +\infty$ as N tends to infinity. Let κ be a Lipschitz kernel that integrates to one and has finite first moment. Then we have:*

$$\sup_{x\in\mathbb{R}} \left|\widehat{f}_{X_k}(x) - f_{X_k}(x)\right| = o_p(1), \qquad \textit{for all } k = 1, ..., K. \tag{10}$$

Lastly, given Corollary 2 one can check that conditional expectations estimators, such as (6) and those in Appendix B, are consistent in sup-norm for their population counterparts.

Remark 3. *It follows from existing convergence rates in nonparametric deconvolution models (e.g., Fan, 1991; Hall and Lahiri, 2008) that neither $\widehat{X}_{ik}$ (as an estimator of the quantile function of X_k) nor its functionals will converge at the root-N rate in general. Bertail et al. (1999) propose an inference method under the condition that the estimator is N^β-consistent with a continuous asymptotic distribution, for some $\beta > 0$. Their rate-adaptive method is attractive in our setting, although polynomial convergence rates may rule out cases of severe ill-posedness. Completing the characterization of the asymptotic behavior of our estimator is an important task for future work.*

[17] The alternative density estimator $\widetilde{f}_{X_k}(x) \equiv 1/\nabla \widehat{H}_k(\widehat{H}_k^{-1}(x))$ can be shown to be uniformly consistent for f_{X_k} as N tends to infinity under the same conditions.

6 Performance on simulated data

In this section we illustrate the finite-sample performance of our estimator on data simulated from a nonparametric model with two independent measurements.

6.1 Setup

Let $Y_1 = X_1 + X_2$, $Y_2 = X_1 + X_3$, where X_1, X_2, X_3 are independent of each other and have identical distributions. We consider four specifications for the distribution of X_k for all k: Beta$(2,2)$, Beta$(5,2)$, normal, and log-normal, all standardized so that X_k has mean zero and variance one. To restrict the maximum values of $\widehat{X}_{ik}$, its increments, and its second-order differences, we consider two choices for the penalization constants: $(\underline{C}_N, \overline{C}_N) = (.1, 10)$ ("strong constraint"), and $(\underline{C}_N, \overline{C}_N) = (0, 10000)$ ("weak constraint"). To minimize the objective function in (4) we start with 10 randomly generated starting values, drawn from widely dispersed mixtures of five Gaussian distributions, and keep the solution corresponding to the minimum value of the objective. Lastly, we draw $M = 10$ independent random permutations in Π_N, and average the resulting M sets of estimates $\widehat{X}_{i1}^{(m)}$, for $i = 1, ..., N$.

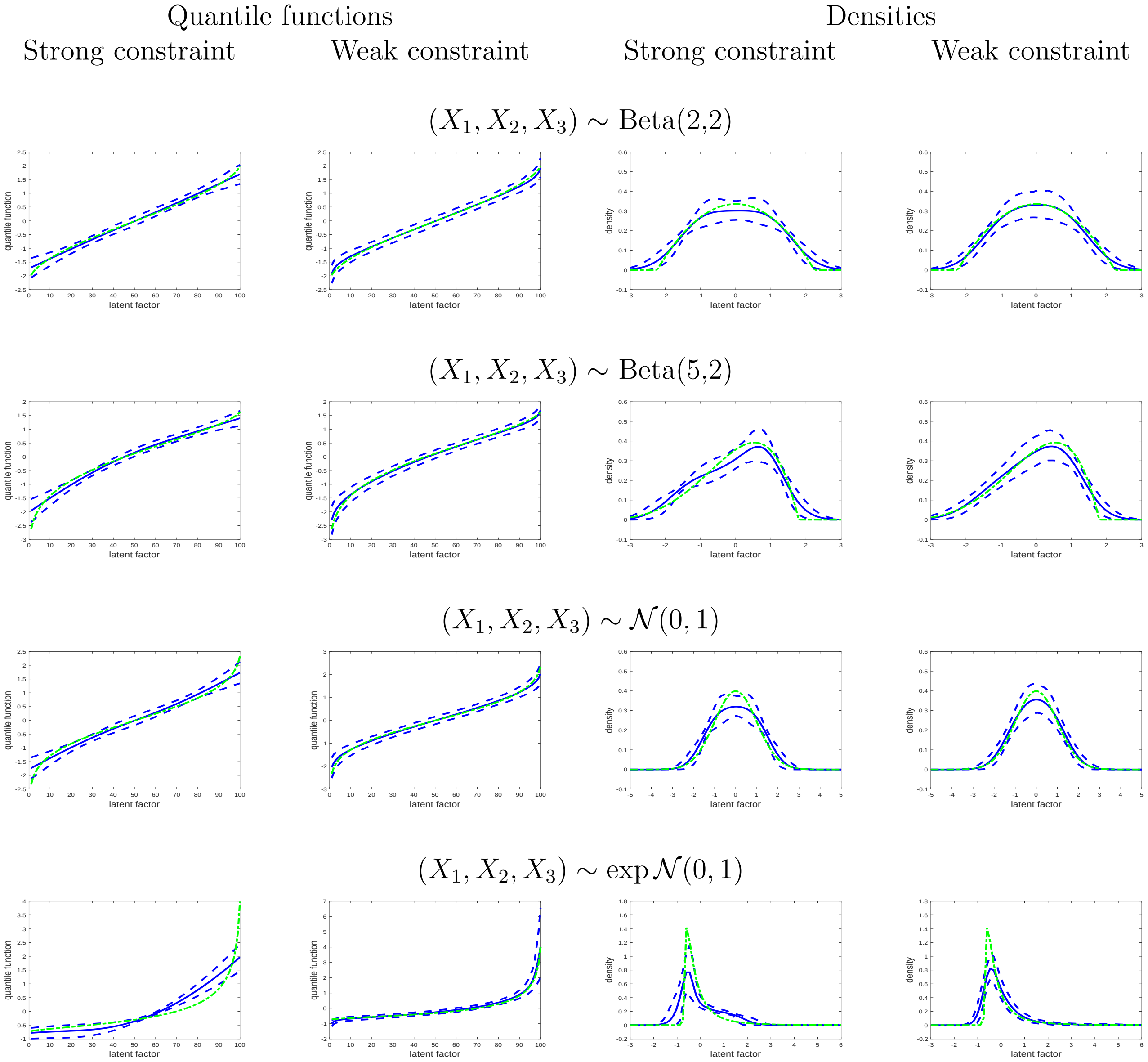
In Appendix C we study the sensitivity of the estimates to the penalization constants, the starting values, and the number M of σ draws, in a nonparametric deconvolution model. We find that the estimator is quite robust to these choices. In particular, we document that taking conservative choices for $\underline{C}_N$ and $\overline{C}_N$ (such as in the "weak constraint" case) results in a well-behaved estimator, suggesting that our matching procedure induces an implicit regularization, even in the absence of additional constraints on parameters. At the same time, we find that such a conservative choice may not be optimal in terms of mean squared errors of quantile estimates. The optimal choice of penalization constants is an interesting question for future work.[18]

6.2 Results

In the first two columns in Figure 2 we show the estimates of the quantile functions $\widehat{X}_{i1} = \widehat{F}_{X_1}^{-1}\left(\frac{i}{N+1}\right)$, for the four specifications and both penalization parameters. The results for the

[18] A simple recommendation for practice is based on a truncated normal distribution. Let $\widehat{\sigma}_k$ denote a consistent estimate of the standard deviation of X_k, e.g. obtained by covariance-based minimum distance, and let $c > 0$ be a tuning parameter. Possible penalization constants are: $2.3c\widehat{\sigma}_k$ (upper bound on quantile values), $2.5c^{-1}\widehat{\sigma}_k$ and $37c\widehat{\sigma}_k$ (lower and upper bounds for first derivatives), and $3300c\widehat{\sigma}_k$ (upper bound on second derivatives). When $c = 1$, these constants are binding when X_k follows a normal truncated at the 99th percentiles. As a default choice one may take $c = 2$.

Figure 2: Monte Carlo results in the model with repeated measurements, $N = 100$, $T = 2$



Quantile functions — Strong constraint, Weak constraint

Densities — Strong constraint, Weak constraint

$(X_1, X_2, X_3) \sim$ Beta(2,2)

$(X_1, X_2, X_3) \sim$ Beta(5,2)

$(X_1, X_2, X_3) \sim \mathcal{N}(0, 1)$

$(X_1, X_2, X_3) \sim \exp \mathcal{N}(0, 1)$

Notes: Simulated data from the repeated measurements model, results for the first factor X_1*. The mean across simulations is shown in solid, 10 and 90 percent pointwise quantiles are shown in dashed, and the true quantile function or density is shown in dashed-dotted.* 1000 *simulations.* 10 *random starting values.* $M = 10$ *averages over* σ *draws.*

other two factors are similar and omitted for brevity. The solid and dashed lines correspond to the mean and 10 and 90 percentiles across 1000 simulations, respectively, while the dashed-dotted line is the true quantile function. The sample size is $N = 100$. Even for such a small sample size, our nonparametric estimator performs well, especially under a weaker constraint

Table 1: Monte Carlo simulation, mean integrated squared and absolute errors of density estimators in the repeated measurements model

	MISE	MIAE	MISE	MIAE	MISE	MIAE
	$(X_1, X_2, X_3) \sim$ Beta(2,2)					
	Strong constraint		Weak constraint		Fourier	
	0.0089	0.1724	0.0088	0.1669	0.0145	0.2493
	$(X_1, X_2, X_3) \sim$ Beta(5,2)					
	Strong constraint		Weak constraint		Fourier	
	0.0132	0.2129	0.0125	0.2001	0.0199	0.2695
	$(X_1, X_2, X_3) \sim \mathcal{N}(0,1)$					
	Strong constraint		Weak constraint		Fourier	
	0.0155	0.2364	0.0109	0.1882	0.0142	0.2479
	$(X_1, X_2, X_3) \sim \exp[\mathcal{N}(0,1)]$					
	Strong constraint		Weak constraint		Fourier	
	0.2527	0.6632	0.1625	0.4605	0.2028	0.5715

Notes: Mean integrated squared and absolute errors across 1000 *simulations from the repeated measurements model.* $N = 100$, $T = 2$. *"Fourier" is the characteristic-function based estimator of Comte and Kappus (2015), for an oracle choice of the tuning parameter. Results for the first factor* X_1.

on the parameters (second column). In the last two columns of Figure 2 we show density estimates for the same specifications. We take a Gaussian kernel and set the bandwidth based on Silverman's rule. Although there are some biases in the strong constraint case, our nonparametric estimator reproduces the shape of the unknown densities well.

In Table 1 we report the mean integrated squared and absolute errors (MISE and MIAE, respectively) of our density estimators, for the four distributional specifications and $N = 100$. We see that the estimator performs better under weak constraint. In the last two columns of Table 1 we report the MISE and MAE of the Fourier-based estimator proposed by Comte and Kappus (2015), for an oracle choice of the tuning parameter.[19] We see that, while our estimator under strong constraint performs slightly worse than the oracle Fourier estimator for the normal and log-normal specifications, it performs better for the two beta specifications, and the matching estimator under weak constraint performs best in all specifications. From results in Chetverikov and Wilhelm (2017), we conjecture that finite-sample performance may benefit from the fact that our estimator directly enforces monotonicity of quantile functions and non-negativity of densities. However, proving this conjecture would

[19] When implementing the Fourier estimator we enforce the non-negativity and integral constraints *ex-post*. To select the tuning parameter, we minimize the Monte Carlo MISE of the estimator on a grid of values.

require deriving additional theoretical results beyond our consistency analysis.

Lastly, in Appendix C we present numerical calculations of the rate of convergence of our estimator of latent quantiles, in data simulated from a nonparametric scalar deconvolution model. The results suggest the rate ranges between $N^{-\frac{3}{10}}$ and $N^{-\frac{7}{10}}$ in the data generating processes that we study. We also compare the performance of our method to Mallows' (2007) "deconvolution by simulation" estimator.

7 Empirical application: income risk over the business cycle

In this section we use our method to study the cyclical behavior of income risk in the US.

7.1 Setup

In an influential contribution, Storesletten *et al.* (2004) report using the PSID that the dispersion of idiosyncratic income shocks increases substantially in recessions. Guvenen *et al.* (2014) re-examine this finding, using US administrative data and focusing on log-income growth. They find that the dispersion of log-income growth is acyclical, and that its skewness is procyclical. Recently, Busch *et al.* (2018) find similar results using the PSID and data from Sweden and Germany. Nakajima and Smyrnyagin (2019) use an approach similar to the one in Storesletten *et al.* (2004), using a larger PSID sample and different measures of income, and find that log-income shocks exhibit countercyclical dispersion and procyclical skewness. This literature is motivated by the key quantitative role of the cyclical behavior of the income process when calibrating models of business cycle dynamics.

Here we revisit this question, by estimating a nonparametric permanent-transitory model using the PSID, in the period 1969–1991. We model log-income, net of the effect of some covariates, as the sum of a random walk $\eta_{it} = \eta_{i,t-1} + v_{it}$ and an independent innovation ε_{it}. In first-differences we have, denoting log-income growth as $\Delta Y_{it} = Y_{it} - Y_{i,t-1}$:

$$\Delta Y_{it} = v_{it} + \varepsilon_{it} - \varepsilon_{i,t-1}, \quad t = 1, ..., T. \tag{11}$$

Model (11) is a linear factor model with $2T-1$ independent factors.[20] We leave the distributions of v_{it} and ε_{it} unrestricted. Our aim is to document the behavior of these distributions over the business cycle. Compared to the existing literature, a substantive difference is that we estimate the densities of the shocks nonparametrically, as opposed to relying on a parametric model.[21] This is important, since estimates of non-Gaussian models (e.g., Horowitz and Markatou, 1996; Geweke and Keane, 2000; Bonhomme and Robin, 2010; Arellano *et al.*, 2017) and descriptive evidence (e.g., Guvenen *et al.*, 2014; Guvenen *et al.*, 2016) both suggest that income shocks are strongly non-Gaussian in the US.

Studying aggregate dynamics using survey panel data like the PSID is complicated by attrition and confounding age effects. To minimize the impact of these factors, we follow the approach pioneered by Storesletten *et al.* (2004) and construct a sequence of balanced, four-year subpanels. In every subpanel, we require that households have non-missing data on income and demographics and comply with standard selection criteria: the household has positive annual labor income during the four years, the head is between 23 and 60 years old, and is not part of the SEO low-income sample or the immigrant sample. We estimate model (11) on 20 subpanels, whose base years range between 1969 and 1988. Log-household income growth is net of indicators for age (of head), education, gender, race, marital status, state of residence, number of children, and family size. We provide descriptive statistics on the 20 four-year subpanels in Appendix D.

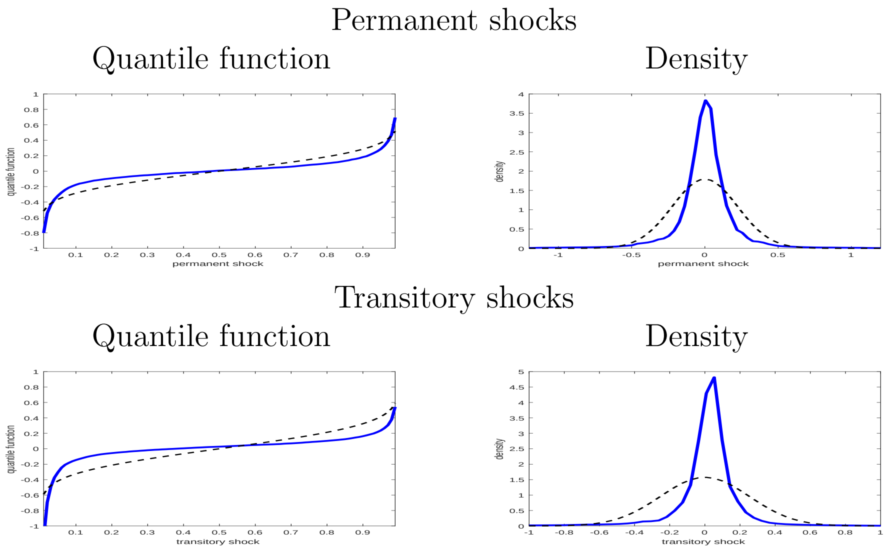
7.2 Results

To estimate the model, as suggested by our Monte Carlo simulations, we set conservative values for the penalization constants (that is, we use the "weak constraint" values of the simulation section), we use a single starting value in the algorithm, and we average the results of $M = 10$ draws. Our first finding is that income shocks are strongly non-Gaussian.

[20] Indeed, we have: $$\underbrace{\begin{pmatrix} \Delta Y_1 \\ \Delta Y_2 \\ \Delta Y_3 \\ \dots \\ \Delta Y_T \end{pmatrix}}_{\equiv Y} = \underbrace{\begin{pmatrix} 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & -1 & 1 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & -1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 & 0 & \dots & -1 \end{pmatrix}}_{\equiv A} \underbrace{\begin{pmatrix} v_1 - \varepsilon_0 \\ v_2 \\ \dots \\ v_T + \varepsilon_T \\ \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_{T-1} \end{pmatrix}}_{\equiv X}.$$

[21] For example, Storesletten *et al.* (2004) estimate an AR(1) process for the persistent component, whose baseline value for the autoregressive coefficient is 0.96. While they estimate the model in levels, our motivation for estimating (11) in first-differences is that differences are robust to heterogeneity between cohorts.

Figure 3: Quantile functions and densities of income shocks, averaged over years


Permanent shocks

Quantile function

Density

Transitory shocks

Quantile function

Density

Notes: Sequence of balanced four-year subpanels from the PSID, 1969-1991. Nonparametric estimates of the quantile functions and densities of permanent and transitory income shocks to log-household annual labor income residuals, averaged over years. Normal fits are shown in dashed.

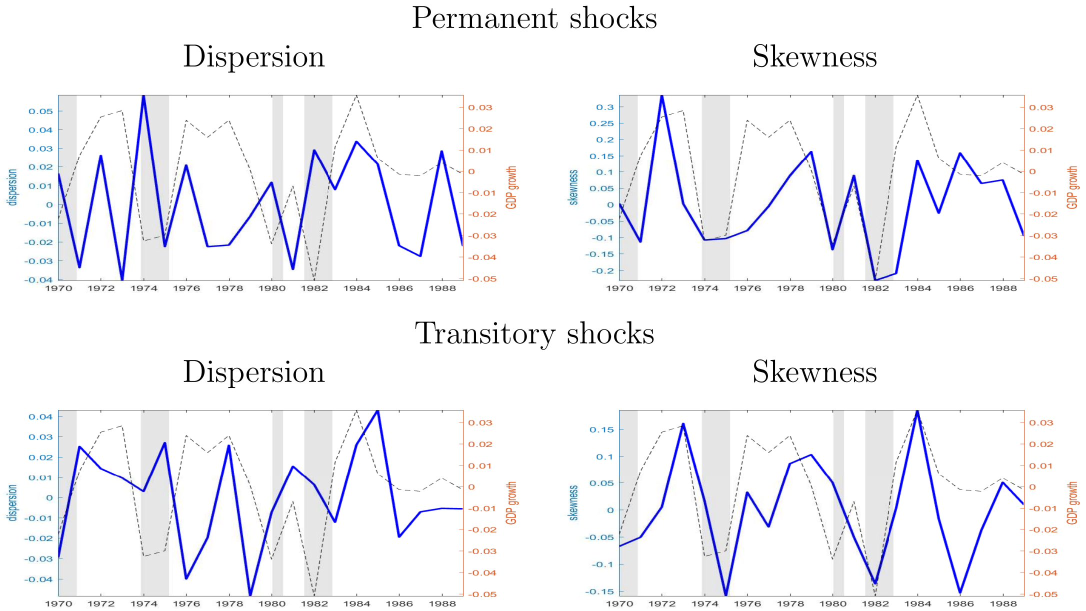
In Figure 3 we report the estimated quantile functions and densities of permanent shocks v_{it} and transitory shocks ε_{it}, averaged over years (in solid), together with normal fits (in dashed). The excess kurtosis of both shocks is in line with previous evidence reported in the literature (e.g., Geweke and Keane, 2000; Bonhomme and Robin, 2010).

We are interested in how features of these distributions vary with the business cycle. In the left column of Figure 4 we plot the percentile difference of log-income $P_{90} - P_{10}$ (a common measure of dispersion, in solid) together with log-GDP growth (in dashed), both of them net of a linear time trend. While permanent and transitory shocks tend to move countercyclically in the first part of the period, the relationship tends to become procyclical in the 1980's. As we report in Table 2, the coefficient of log-GDP growth in a regression of the dispersion of permanent income shocks on log-GDP growth and a time trend is -0.25, with a Newey-West standard error of 0.30.[22] Hence, we do not find significant evidence that the dispersion of permanent shocks varies systematically with the business cycle. In addition, we neither find that the dispersion of transitory shocks varies with the cycle.

Next, in the right column of Figure 4 we plot the Bowley-Kelley quantile measure of skew-

[22] We compute the Newey-West formula with one lag. Using two or three lags instead has little impact. In this calculation we do not account for the fact that the quantiles are estimated, our rationale being that the cross-sectional sizes are large relative to the length of the time series.

Figure 4: Dispersion and skewness of income shocks over the business cycle



Permanent shocks

Dispersion

Skewness

Transitory shocks

Dispersion

Skewness

Notes: See notes to Figure 3. Dispersion ($P_{90} - P_{10}$) and skewness (Bowley-Kelley) are indicated in solid, log-real GDP growth is in dashed.

ness $[(P_{90} - P_{50}) - (P_{50} - P_{10})]/(P_{90} - P_{10})$. The graphs of permanent and transitory income shocks suggest that skewness is procyclical. This is confirmed in Table 2, which shows that the coefficient of log-GDP growth in the skewness regression is 3.07 for permanent shocks, and 2.36 for transitory shocks, significant at the 5% level in both cases. Our nonparametric estimates of a permanent-transitory model of income dynamics thus suggest that dispersion is approximately acyclical, and skewness is procyclical, in line with the conclusions of the descriptive evidence in Guvenen *et al.* (2014) and Busch *et al.* (2018).

We performed several exercises to probe the robustness of these findings: using the "strong constraint" penalization of Section 6, measuring business cycle conditions using the unemployment rate instead of log-GDP growth, and varying the choice of starting values in the algorithm. While we found the year-to-year variation in Figure 4 to depend on the chosen specification, in all our checks we found a lack of systematic cyclical variability of the dispersion of income shocks, and a significant procyclicality of the skewness of permanent shocks. Among the results reported in Table 2, we found the procyclicality of the skewness of transitory shocks to be most sensitive to specification changes.

Finally, we use the information in the PSID about hours worked to compute similar

Table 2: Cyclicality of the distributions of income shocks

	Annual income							
	Permanent				Transitory			
	Dispersion	Skewness	Upper	Lower	Dispersion	Skewness	Upper	Lower
Coeff.	-0.2528	3.0752	0.4647	-0.7175	0.0752	2.3612	0.4133	-0.3381
St. Er.	0.3011	0.7576	0.2023	0.2167	0.2380	0.6239	0.1536	0.1568

	Hourly wages							
	Permanent				Transitory			
	Dispersion	Skewness	Upper	Lower	Dispersion	Skewness	Upper	Lower
Coeff.	0.1629	0.5235	0.1680	-0.0051	0.2374	0.7793	0.2594	-0.0220
St. Er.	0.3750	0.5558	0.2627	0.1295	0.2453	0.7093	0.2150	0.1351

Notes: See notes to Figure 3. The coefficients are obtained from a regression of $P_{90} - P_{10}$ dispersion (respectively, Bowley-Kelley skewness, upper tail $P_{90} - P_{50}$, or lower tail $P_{50} - P_{10}$) on log-real GDP growth and a linear time trend. Newey-West standard errors (one lag).

measures of cyclicality based on hourly wages of household heads. Evidence from Italy and France (Hoffmann and Malacrino, 2019; Pora and Wilner, 2019) suggests that days and hours worked may contribute significantly to the observed cyclical patterns of skewness. For the US, Nakajima and Smyrnyagin (2019) obtain similar conclusions. In contrast, Busch *et al.* (2018) find a moderate role of hours worked in Germany. In the bottom panel of Table 2 we see that the skewnesses of permanent and transitory shocks to hourly wages do not vary significantly with the cycle, and that the point estimates are greatly reduced compared to the case of total income. This suggests that hours worked largely contribute to the distributional income dynamics that we document.

8 Extensions

Consider a block-independent factor model $Y = \sum_{\ell=1}^{L} A_\ell X_\ell$, where $X_1, ..., X_L$ are mutually independent random vectors. To extend our approach to this case, we exploit the vector quantile representation given in Carlier *et al.* (2016). Under regularity conditions, one has $X_\ell = G_\ell(V_\ell)$, where V_ℓ is a vector of n_ℓ independent standard uniform random variables, and G_ℓ, which is a Brenier map obtained by optimal transport, is the gradient of a convex function. The functions G_ℓ satisfy a *cyclical monotonicity* condition analogous to the usual monotonicity of univariate quantile functions (see Chapter 2 in Villani, 2003):

$$\text{For all } m \geq 2 \text{ and } v_1, ..., v_m, v_{m+1} = v_1 \in [0,1]^{n_\ell} : \sum_{j=1}^{m} G_\ell(v_j)' \, (v_{j+1} - v_j) \leq 0.$$

Using the vector quantile representation, we define the population objective function for block-independent factor models in a similar way to the independent case (9); that is:

$$Q(G_1, ..., G_L) \equiv \inf_{\pi \in \mathcal{M}\left(\mu_Y, \mu_{\sum_{\ell=1}^L A_\ell G_\ell}\right)} \mathbb{E}_\pi \left[\sum_{t=1}^T \left(Y_t - \sum_{\ell=1}^L A'_{\ell t} G_\ell \left(V_\ell \right) \right)^2 \right], \tag{12}$$

where $A'_{\ell t}$ is the t-th row of A_ℓ, and V_ℓ are n_ℓ-dimensional vectors of independent standard uniform random variables.

An empirical counterpart to (12) then leads to the estimator:

$$(\widehat{X}_1, ..., \widehat{X}_L) = \underset{(X_1,...,X_L) \in \mathcal{X}_N^{\text{blocks}}}{\operatorname{argmin}} \left\{ \min_{\pi \in \Pi_N} \sum_{i=1}^N \sum_{t=1}^T \left(Y_{\pi(i),t} - \sum_{\ell=1}^L \sum_{k=1}^{n_\ell} a_{\ell t k} X_{\ell,i,k} \right)^2 \right\}, \tag{13}$$

where the vectors $(X_1, ..., X_L) \in \mathcal{X}_N^{\text{blocks}}$ are restricted to be cyclically monotone in the following sense (for $\sigma_{\ell k}$ independent random permutations of $\{1, ..., N\}$):

$$\text{For all } m \leq N \text{ and } i_1, ..., i_m, i_{m+1} = i_1 \in \{1, ..., N\} : \sum_{j=1}^m \sum_{k=1}^{n_\ell} X_{\ell, i_j, k} \left(\sigma_{\ell k}(i_{j+1}) - \sigma_{\ell k}(i_j) \right) \leq 0, \tag{14}$$

which generalizes the univariate monotonicity condition we imposed on X_{ik} in (4).[23]

For all $\ell \in \{1, ..., L\}$, $\widehat{X}_\ell$ can then be interpreted as an empirical counterpart to the vector quantile representation of X_ℓ based on the function G_ℓ, subject to a suitable rearrangement. More formally, a consistency statement for $\widehat{X}_\ell$ will take the form:

$$\max_{i \in \{1,...,N\}} \left\| \widehat{X}_{\ell i} - G_\ell \left(\frac{\sigma_{\ell 1}(i)}{N+1}, ..., \frac{\sigma_{\ell K}(i)}{N+1} \right) \right\| = o_p(1), \quad \text{for all } \ell = 1, ..., L.$$

As in the independent case, one can compute a local minimum in (13) using an algorithm that iterates between matching and update steps.[24]

Finally, we note that our approach can be extended to estimate other models, beyond linear independent and block-independent factor models. In Appendix E we outline several possible extensions, to random coefficients models with exogenous covariates, nonparametric deconvolution under heteroskedasticity, and nonparametric finite mixture models. We leave the formal analysis of these extensions to future work.

[23] Indeed, in the univariate case $X_{i+1,k} \geq X_{ik}$ for all i is equivalent to $\sum_{j=1}^m X_{\sigma_k(i_j),k} \left(\sigma_k(i_{j+1}) - \sigma_k(i_j) \right) \leq 0$ for all $m \leq N$ and length-m cycle $i_1, ..., i_m, i_{m+1} = i_1$.

[24] Note that (14) is linear in $X_{\ell,i,k}$'s. However, it may be impractical to enforce all restrictions in (14) in the update step. In applications, a possibility is to select S_N restrictions at random, where S_N depends on the sample size.

9 Conclusion

In this paper we have proposed an approach to nonparametrically estimate linear models with independent latent variables. The method is based on matching predicted values from the model to the empirical observations. We have provided a simple algorithm for computation, and established consistency. We have also documented good performance of our estimator in small samples, and we have used it to shed new light on the cyclicality of permanent and transitory shocks to income and wages in the US. An important question for future work will be to characterize rates of convergence and confidence sets.

References

[1] Arellano, M., and S. Bonhomme (2012): "Identifying Distributional Characteristics in Random Coefficients Panel Data Models", *Review of Economic Studies*, 79, 987–1020.

[2] Bassetti, F., A. Bodini, and E. Regazzini (2006): "On Minimum Kantorovich Distance Estimators," *Statistics and probability letters*, 76(12), 1298–1302.

[3] Ben-Moshe, D. (2017): "Identification of Joint Distributions in Dependent Factor Models," to appear in *Econometric Theory.*

[4] Beran, R., and P. Hall (1992): "Estimating Coefficient Distributions in Random Coefficient Regressions," *Annals of Statistics*, 20(4), 1970–1984.

[5] Bernton, E., P. E. Jacob, M. Gerber, and C. P. Robert (2017): "Inference in Generative Models Using the Wasserstein Distance," arXiv preprint arXiv:1701.05146.

[6] Bertail, P., D. N. Politis, and J. P. Romano (1999): "On Subsampling Estimators with Unknown Rate of Convergence," *J. of the Am. Stat. Ass.*, 94(446), 569–579.

[7] Blundell, R., L. Pistaferri, and I. Preston (2008): "Consumption Inequality and Partial Insurance," *American Economic Review*, 98(5): 1887–1921.

[8] Bonhomme, S., and J. M. Robin (2009): "Consistent Noisy Independent Component Analysis," *Journal of Econometrics*, 149(1), 12–25.

[9] Bonhomme, S., and J. M. Robin (2010): "Generalized Nonparametric Deconvolution with an Application to Earnings Dynamics," *Review of Economic Studies*, 77(2), 491–533.

[10] Bousquet, O., S. Gelly, I. Tolstikhin, C. J. Simon-Gabriel, and B. Schoelkopf (2017): "From Optimal Transport to Generative Modeling: The VEGAN Cookbook," arXiv preprint arXiv:1705.07642.

[11] Botosaru, I., and Y. Sasaki (2015): "Nonparametric Heteroskedasticity in Persistent Panel Processes: An Application to Earnings Dynamics," unpublished manuscript.

[12] Busch, C., D. Domeij, F. Guvenen, and R. Madera (2018): "Asymmetric Business-Cycle Risk and Social Insurance" (No. w24569). National Bureau of Economic Research.

[13] Carlier, G., V. Chernozhukov, and A. Galichon (2016): "Vector Quantile Regression: An Optimal Transport Approach," *The Annals of Statistics*, 44(3), 1165–1192.

[14] Carrasco, M., and J.P. Florens (2011): "Spectral Method for Deconvolving a Density," *Econometric Theory*, 27(3), 546–581.

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

[16] Carroll, R. J., D. Ruppert, L. A. Stefanski, C. M. Crainiceanu (2006): *Measurement Error in Nonlinear Models: A Modern Perspective.* CRC press.

[17] Chen, X. (2007): "Sieve Methods in Econometrics," *Handbook of Econometrics*, vol. 6, 5549–5632.

[18] Chen, X., H. Hong, H., and D. Nekipelov, D. (2011): "Nonlinear Models of Measurement Errors," *Journal of Economic Literature*, 49(4), 901–937.

[19] Chernozhukov, V., G. W. Imbens, and W. K. Newey (2007): "Instrumental Variable Estimation of Nonseparable Models," *Journal of Econometrics*, 139(1), 4–14.

[20] Chetty, R., and N. Hendren (2018): "The Impacts of Neighborhoods on Intergenerational Mobility: County-Level Estimates," *Quarterly Journal of Economics*, 133(2), 1163-1228.

[21] Chetverikov, D., and D. Wilhelm (2017): "Nonparametric Instrumental Variable Estimation under Monotonicity," *Econometrica*, 85(4), 1303–1320.

[22] Comte, F., and J. Kappus (2015): "Density Deconvolution from Repeated Measurements without Symmetry Assumption on the Errors," *Journal of Multivariate Analysis*, 140, 31–46.

[23] Comte, F., and C. Lacour (2011): "Data-Driven Density Estimation in the Presence of Additive Noise with Unknown Distribution," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4), 601–627.

[24] Comte, F., and C. Lacour (2013): "Anisotropic Adaptive Kernel Deconvolution," In *Annales de l'IHP Probabilités et Statistiques*, 49(2), 569–609.

[25] Csörgö, M. (1983): *Quantile Processes with Statistical Applications*, SIAM.

[26] Dattner, I., A. Goldenshluger, and A. Juditsky (2011): "On Deconvolution of Distribution Functions," *The Annals of Statistics*, 2477–2501.

[27] Delaigle, A., P. Hall, and A. Meister (2008): "On Deconvolution with Repeated Measurements," *Annals of Statistics*, 36, 665-685.

[28] Delaigle, A., and A. Meister (2008): "Density Estimation with Heteroscedastic Error," *Bernoulli*, 14(2), 562–579.

[29] Duval, C., and J. Kappus (2017): "Nonparametric Adaptive Estimation for Grouped Data," *Journal of Statistical Planning and Inference*, 182, 12–28.

[30] Efron, B. (2016): "Empirical Bayes Deconvolution Estimates," *Biometrika*, 103(1), 1–20.

[31] Evdokimov, K., and H. White (2012): "Some Extensions of a Lemma of Kotlarski," *Econometric Theory*, 28(04), 925–932.

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

[33] Freyberger, J., and M. Masten (2019): "A Practical Guide to Compact Infinite Dimensional Parameter Spaces," *Econometric Reviews*, 38(9), 979–1006.

[34] Galichon, A. (2016): *Optimal Transport Methods in Economics.* Princeton University Press.

[35] Galichon, A., and M. Henry (2011): "Set Identification in Models with Multiple Equilibria," *Review of Economic Studies*, 78(4), 1264–1298.

[36] Gallant, A. R., and D. W. Nychka (1987): "Semi-nonparametric Maximum Likelihood Estimation," *Econometrica*, 55(2), 363–90.

[37] Genevay, A., M. Cuturi, G. Peyré, and F. Bach (2016): "Stochastic Optimization for Large-Scale Optimal Transport," In *Advances in neural information processing systems*, 3440–3448.

[38] Genevay, A., G. Peyré, and M. Cuturi (2017): "Sinkhorn-AutoDiff: Tractable Wasserstein Learning of Generative Models," arXiv preprint arXiv:1706.00292.

[39] Geweke, J., and M. Keane (2000): "An Empirical Analysis of Earnings Dynamics Among Men in the PSID: 1968-1989," *Journal of Econometrics*, 96(2), 293–356.

[40] Gu, J., and R. Koenker (2017): "Empirical Bayesball Remixed: Empirical Bayes Methods for Longitudinal Data," *Journal of Applied Econometrics*, 32(3), 575–599.

[41] Guvenen, F., S. Ozkan, and J. Song (2014): "The Nature of Countercyclical Income Risk," *Journal of Political Economy*, 122(3), 621–660.

[42] Guvenen, F., F. Karahan, S. Ozkan, and J. Song (2016): "What Do Data on Millions of US Workers Reveal about Life-Cycle Earnings Dynamics?" to appear in *Econometrica.*

[43] Hall, P., and X. H. Zhou (2003): "Nonparametric Estimation of Component Distributions in a Multivariate Mixture," *Annals of Statistics*, 201–224.

[44] Hall, P., and S. N. Lahiri (2008): "Estimation of Distributions, Moments and Quantiles in Deconvolution Problems," *Annals of Statistics*, 36(5) 2110–2134.

[45] Hall, R. E., and F. S. Mishkin (1982): "The Sensitivity of Consumption to Transitory Income: Estimates from Panel Data on Households," *Econometrica*, 50(2), 461–481.

[46] Heckman, J. J., J. Smith, and N. Clements (1997): "Making the Most out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *Review of Economic Studies*, 64(4), 487–535.

[47] Hoffmann, E. B., and D. Malacrino (2019): "Employment Time and the Cyclicality of Earnings Growth," *Journal of Public Economics*, 169, 160–171.

[48] Horowitz, J. L., and M. Markatou (1996): "Semiparametric Estimation of Regression Models for Panel Data", *Review of Economic Studies*, 63, 145–168.

[49] Hu, Y., R. Moffitt, and Y. Sasaki (2019): "Semiparametric Estimation of the Canonical Permanent-Transitory Model of Earnings Dynamics," to appear in *Quantitative Economics.*

[50] Johannes, J. (2009): "Deconvolution with Unknown Error Distribution," *The Annals of Statistics*, 37(5A), 2301–2323.

[51] Kappus, J., and G. Mabon (2014): "Adaptive Density Estimation in Deconvolution Problems with Unknown Error Distribution," *Electronic Journal of Statistics*, 8(2), 2879–2904.

[52] Koenker, R., and J. Gu (2019): "Comment: Minimalist g-Modeling," *Stat. Sci.*, 34.2, 209–213.

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

[54] Lepski, O. V., and T. Willer (2019): "Oracle Inequalities and Adaptive Estimation in the Convolution Structure Density Model," *Annals of Statistics*, 47(1), 233–287.

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

[56] Li, T., and Q. Vuong (1998): "Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators," *Journal of Multivariate Analysis*, 65, 139–165.

[57] Mallows, C. (2007): "Deconvolution by Simulation," in: Liu, R., Strawderman, W., and C.H. Zhang (Eds.), *Complex Datasets and Inverse Problems: Tomography, Networks and Beyond*, Beachwood, Ohio, USA: Institute of Mathematical Statistics.

[58] Meister, A. (2009): *Deconvolution Problems in Nonparametric Statistics.* Springer Science & Business Media (Vol. 193).

[59] Nakajima, M., and V. Smirnyagin (2019): "Cyclical Labor Income Risk," SSRN 3432213.

[60] Pensky, M., and B. Vidakovic (1999): "Adaptive Wavelet Estimator for Nonparametric Density Deconvolution," *Annals of Statistics*, 27(6), 2033–2053.

[61] Peyré, G., and M. Cuturi (2019): "Computational Optimal Transport." *Foundations and Trends in Machine Learning*, 11(5–6), 355–607.

[62] Pora, P., and L. Wilner (2019): "Decomposition of Labor Earnings Growth: Recovering Gaussianity?" Unpublished manuscript.

[63] Rigollet, P., and J. Weed (2018): "Entropic Optimal Transport is Maximum-Likelihood Deconvolution," *Comptes Rendus Mathematique*, 356(11–12), 1228–1235.

[64] Rigollet, P., and J. Weed (2019): "Uncoupled Isotonic Regression via Minimum Wasserstein Deconvolution," arXiv preprint arXiv:1806.10648.

[65] Schennach, S. M. (2013a): "Measurement Error in Nonlinear Models: A Review," in *Advances in Economics and Econometrics: Econometric theory*, ed. by D. Acemoglu, M. Arellano, and E. Dekel, Cambridge University Press, vol. 3, 296–337.

[66] Schennach, S. (2013b): *Convolution Without Independence*, Cemmap WP No. CWP46/13.

[67] Stefanski, L. A., and R. J. Carroll (1990): "Deconvolving Kernel Density Estimators," *Statistics*, 21, 169–184.

[68] Storesletten, K., C. I. Telmer, and A. Yaron (2004): "Cyclical Dynamics in Idiosyncratic Labor Market Risk," *Journal of Political Economy*, 112(3), 695–717.

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

[70] Villani, C. (2003): *Topics in Optimal Transportation.* No. 58. American Mathematical Soc.

APPENDICES

A Proofs

A.1 Proofs of Theorem 1 and Corollary 1

Before proving Theorem 1 for multi-factor models, we first prove Corollary 1 for the scalar deconvolution case where explicit expressions for Wasserstein distances are available.

A.1.1 Scalar deconvolution: Corollary 1

We first state the following assumption, where for conciseness we denote $H \equiv H_1$.

Assumption A1.

(i) (Continuity and support) Y, X_1 and X_2 have compact supports in $\mathbb{R}$, and admit absolutely continuous densities f_Y, f_{X_1}, f_{X_2} that are bounded away from zero and infinity. Moreover, f_Y is differentiable with bounded derivative.

(ii) (Identification) The density f_{X_1} is identified given f_Y and f_{X_2}.

(iii) (Penalization) $\overline{C}_N$ is increasing and $\underline{C}_N$ is decreasing with $\lim_{N\to+\infty} \overline{C}_N = \overline{C}$ and $\lim_{N\to+\infty} \underline{C}_N = \underline{C}$, where $\overline{C}$ and $\underline{C} < \overline{C}$ are such that $\|F_{X_1}^{-1}\|_{1,\infty} \leq \overline{C}$ and $\nabla F_{X_1}^{-1}(\tau) \geq \underline{C}$ for all $\tau \in (0,1)$.

(iv) (Sampling) $Y_1, ..., Y_N$ and $X_{12}, ..., X_{N2}$ are i.i.d.

A sufficient condition for Assumption A1 (ii) is that the characteristic function of X_2 does not vanish on the real line; moreover, this condition can be relaxed by allowing for the presence of isolated zeros in the characteristic function (Carrasco and Florens, 2011).

We now prove Corollary 1. Define the empirical objective function, for any candidate quantile function H, as:

$$\begin{aligned}\widehat{Q}(H) &= \min_{\pi\in\Pi_N} \frac{1}{N}\sum_{i=1}^{N}\left(Y_{\pi(i)} - H\left(\frac{\sigma(i)}{N+1}\right) - X_{i2}\right)^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{F}_Y^{-1}\left(\frac{1}{N}\widehat{\text{Rank}}\left(H\left(\frac{\sigma(i)}{N+1}\right)+X_{i2}\right)\right) - H\left(\frac{\sigma(i)}{N+1}\right) - X_{i2}\right)^2,\end{aligned}$$

where $\widehat{F}_Y^{-1}(\tau) = \inf\{y \in \text{Supp}(Y) \,:\, \widehat{F}_Y(y) \geq \tau\}$, $\widehat{\text{Rank}}(Z_i) = N\widehat{F}_Z(Z_i)$, and $\sigma \in \Pi_N$. The second equality follows from Hardy-Littlewood-Pólya. With some abuse of notation, for all $X \in \mathbb{R}^N$ we will denote $\widehat{Q}(X) = \widehat{Q}(H)$ for any function H such that $H\left(\frac{i}{N+1}\right) = X_i$ for all i.

Define the population counterpart to $\widehat{Q}$, for any $H \in \mathcal{H}$, as:

$$Q(H) = \mathbb{E}\left[\left(F_Y^{-1}\left(\int_0^1 F_{X_2}\left(H(V) + X_2 - H(\tau)\right) d\tau\right) - H(V) - X_2\right)^2\right],$$

where the expectation is taken with respect to pairs (V, X_2) of independent random variables, for V standard uniform and $X_2 \sim F_{X_2}$.

Parameter space. Let $\mathcal{H}$ be the closure of the set $\{H \in \mathcal{C}^1 : \nabla H \geq \underline{C}, \|H\|_{1,\infty} \leq \overline{C}\}$ under the norm $\|\cdot\|_\infty$. $\mathcal{H}$ is compact under $\|\cdot\|_\infty$ (Gallant and Nychka, 1987).

Sieve construction. For any N, let us define the *sieve space*:

$$\mathcal{H}_N = \left\{H \in \mathcal{H} : \left|H\left(\frac{i}{N+1}\right)\right| \leq \overline{C}_N, \underline{C}_N \leq (N+1)\left(H\left(\frac{i+1}{N+1}\right) - H\left(\frac{i}{N+1}\right)\right) \leq \overline{C}_N\right\}.$$

Let $\widehat{X} \in \mathcal{X}_N$ be such that, for $\epsilon_N = o_p(1)$:

$$\widehat{Q}(\widehat{X}) \leq \min_{X \in \mathcal{X}_N} \widehat{Q}(X) + \epsilon_N.$$

We first note that there exists an $\widehat{H} \in \mathcal{H}_N$ such that $\widehat{H}\left(\frac{i}{N+1}\right) = \widehat{X}_i$ for all i.[25] Hence:

$$\widehat{Q}(\widehat{H}) = \widehat{Q}(\widehat{X}) \leq \min_{X \in \mathcal{X}_N} \widehat{Q}(X) + \epsilon_N \leq \inf_{H \in \mathcal{H}_N} \widehat{Q}(H) + \epsilon_N. \tag{A1}$$

Let $H_0 = F_{X_1}^{-1}$. To show Corollary 1 it is thus sufficient to show that, when $\widehat{H}$ satisfies (A1), we have $\|\widehat{H} - H_0\|_\infty = o_p(1)$. This will follow from verifying the conditions of Lemma A1 in Chernozhukov *et al.* (2007, p. 11), which we restate here for completeness.

Lemma A1. *(Chernozhukov et al., 2007) Suppose (i) $Q(H)$ is uniquely minimized at H_0 on $\mathcal{H}$; (ii) $Q(H)$ is continuous, $\mathcal{H}$ is compact, and $\sup_{H \in \mathcal{H}} |\widehat{Q}(H) - Q(H)| = o_p(1)$; (iii) $\mathcal{H}_N \subset \mathcal{H}$ and there exists $H_N \in \mathcal{H}_N$ such that $\|H_N - H_0\|_\infty = o_p(1)$. Then, if $\widehat{Q}(\widehat{H}) \leq \inf_{H \in \mathcal{H}_N} \widehat{Q}(H) + o_p(1)$, it follows that $\|\widehat{H} - H_0\|_\infty = o_p(1)$.*

Condition (i): $\boldsymbol{Q(H)}$ is uniquely minimized at $\boldsymbol{H_0}$ on $\boldsymbol{\mathcal{H}}$. We have $Q(H) \geq Q(H_0) = 0$ for all $H \in \mathcal{H}$. Suppose that $Q(H) = 0$. Then, (V, X_2)-almost surely we have:

$$F_Y^{-1}\left(\int_0^1 F_{X_2}\left(H(V) + X_2 - H(\tau)\right) d\tau\right) = H(V) + X_2. \tag{A2}$$

[25] Take a smooth interpolating function of the $\widehat{X}_i$'s, arbitrarily close in sup-norm to the piecewise-linear interpolant of the $\widehat{X}_i$'s extended to have slope $(\underline{C} + \overline{C})/2$ on the intervals $[0, 1/(N+1)]$ and $[N/(N+1), 1]$. This is always possible since $\overline{C}_N < \overline{C}$ and $\underline{C}_N > \underline{C}$.

Since the left-hand side in (A2) is distributed as F_Y, it follows that: $F_{H(V)+X_2}\left(H(V)+X_2\right) = F_Y\left(H(V)+X_2\right)$ almost surely, so $F_{H(V)+X_2} = F_Y$. Since Y and X_2 have densities f_Y and f_{X_2}, this implies that $f_Y(y) = \int_0^1 f_{X_2}(y - H(\tau))d\tau$. Now, since $H \in \mathcal{H}$, the function $f_{\widetilde{X}}(x) \equiv 1/\nabla H(H^{-1}(x))$ is continuous and bounded. We then have by a change of variables, $f_Y(y) = \int f_{X_2}(y-x)f_{\widetilde{X}}(x)dx$. Since f_{X_1} is identified given f_Y and f_{X_2}, it thus follows that $f_{X_1} = f_{\widetilde{X}}$, hence that $H = H_0$.

Condition (ii): $Q(H)$ is continuous on $\mathcal{H}$ under $\|\cdot\|_\infty$, $\mathcal{H}$ is compact under $\|\cdot\|_\infty$, and $\text{plim}_{N\to+\infty}\sup_{H\in\mathcal{H}}|\widehat{Q}(H) - Q(H)| = 0$. Compactness holds as indicated above. To show that $Q(H)$ is continuous on $\mathcal{H}$ under $\|\cdot\|_\infty$, let H_1, H_2 in $\mathcal{H}$. By Assumption A1 (i), F_Y^{-1} and F_{X_2} are Lipschitz. It follows that, for some constant $\widetilde{C}$ independent of H_1, H_2, $|Q(H_2) - Q(H_1)| \leq \widetilde{C}\|H_2 - H_1\|_\infty$. This implies Lipschitz continuity of Q.

Let us now show that $\sup_{H\in\mathcal{H}}|\widehat{Q}(H) - Q(H)| = o_p(1)$. Let:

$$G_H(v,x) \equiv \left(F_Y^{-1}\left(\int_0^1 F_{X_2}\left(H(v) + x - H(\tau)\right)d\tau\right) - H(v) - x\right)^2.$$

First, notice that, for all $H \in \mathcal{H}$ and as N tends to infinity:

$$\frac{1}{N}\sum_{i=1}^N G_H\left(\frac{\sigma(i)}{N+1}, X_{i2}\right) = \int_0^1 \mathbb{E}\left(G_H\left(\tau, X_2\right)\right)d\tau + o_p(1) = Q(H) + o_p(1). \tag{A3}$$

Hence, since $\mathcal{H}$ is compact for $\|\cdot\|_\infty$, and since $H \mapsto G_H$ is Lipschitz on $\mathcal{H}$, we have:

$$\sup_{H\in\mathcal{H}}\left|\frac{1}{N}\sum_{i=1}^N G_H\left(\frac{\sigma(i)}{N+1}, X_{i2}\right) - Q(H)\right| = o_p(1). \tag{A4}$$

Next, we are going to show that:

$$\sup_{H\in\mathcal{H}}\left|\frac{1}{N}\sum_{i=1}^N \frac{1}{N}\widehat{\text{Rank}}\left(H\left(\frac{\sigma(i)}{N+1}\right) + X_{i2}\right) - \int_0^1 F_{X_2}\left(H\left(\frac{\sigma(i)}{N+1}\right) + X_{i2} - H(\tau)\right)d\tau\right| = o_p(1). \tag{A5}$$

From (A4), (A5), and the fact that F_Y^{-1} is Lipschitz, we will then have:

$$\sup_{H\in\mathcal{H}}\left|\frac{1}{N}\sum_{i=1}^N\left(F_Y^{-1}\left(\frac{1}{N}\widehat{\text{Rank}}\left(H\left(\frac{\sigma(i)}{N+1}\right) + X_{i2}\right)\right) - H\left(\frac{\sigma(i)}{N+1}\right) - X_{i2}\right)^2 - Q(H)\right| = o_p(1). \tag{A6}$$

To show (A5) we are going to show that:

$$\sup_{H\in\mathcal{H},\, a\in\mathbb{R}}\left|\frac{1}{N}\sum_{i=1}^N \mathbf{1}\left\{H\left(\frac{\sigma(i)}{N+1}\right) + X_{i2} \leq a\right\} - \int_0^1 F_{X_2}\left(a - H(\tau)\right)d\tau\right| = o_p(1). \tag{A7}$$

Pointwise convergence in (A7) is readily verified (similarly to (A3)). To show uniform convergence, we will show that $\mathcal{G} = \{g_{H,a} : H \in \mathcal{H}, a \in \mathbb{R}\}$, where $g_{H,a}(v,u) \equiv \mathbf{1}\{H(v) + u \leq a\}$, has finite bracketing entropy for the L^1 norm. Let $C_2 > 0$ such that $f_{X_2}(u) \leq C_2$ for all u. Fix an $\epsilon > 0$. Since $\mathcal{H}$ has finite ϵ-bracketing entropy for $\| \cdot \|_\infty$, there exists a set of functions H_j, $j = 1, ..., J$, such that for all $H \in \mathcal{H}$ there is a j such that $H_j(\tau) \leq H(\tau) \leq H_{j+1}(\tau)$ for all τ, and $\|H_{j+1} - H_j\|_\infty \leq \frac{\epsilon}{2C_2}$ for all j. Moreover, since X_2 has compact support, there exists a set of scalars $-\infty = a_0 < a_1 < ... < a_K < a_{K+1} = +\infty$ such that $a_{k+1} - a_k \leq \frac{\epsilon}{2C_2}$ for $k = 1, ..., K-1$, $F_{X_2}(a_1 + \overline{C}) = 0$, and $F_{X_2}(a_K - \overline{C}) = 1$. Hence for all H and a there exist j and k such that $\mathbf{1}\{H_{j+1}(v) + u \leq a_k\} \leq g_{H,a}(v, u) \leq \mathbf{1}\{H_j(v) + u \leq a_{k+1}\}$ for all (u, v), and $\iint [\mathbf{1}\{H_j(v) + u \leq a_{k+1}\} - \mathbf{1}\{H_{j+1}(v) + u \leq a_k\}] f_{X_2}(u) du dv \leq C_2(\frac{\epsilon}{2C_2} + \frac{\epsilon}{2C_2}) = \epsilon$. Hence $\mathcal{G}$ has finite bracketing entropy, and (A7) follows.

Lastly, since f_Y is bounded away from zero and infinity and differentiable, the empirical quantile function of Y is such that (see Corollary 1.4.1 in Csörgö, 1983): $\|\widehat{F}_Y^{-1} - F_Y^{-1}\|_\infty = o_p(1)$. Hence, from (A6):

$$\sup_{H \in \mathcal{H}} \left| \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{F}_Y^{-1} \left(\frac{1}{N} \widehat{\text{Rank}} \left(H \left(\frac{\sigma(i)}{N+1} \right) + X_{i2} \right) \right) - H \left(\frac{\sigma(i)}{N+1} \right) - X_{i2} \right)^2 - Q(H) \right| = o_p(1),$$

which shows that $\sup_{H \in \mathcal{H}} |\widehat{Q}(H) - Q(H)| = o_p(1)$.

Condition (iii): $\boldsymbol{\mathcal{H}_N \subset \mathcal{H}}$ for all $\boldsymbol{N}$, and there exists a sequence $\boldsymbol{H_N \in \mathcal{H}_N}$ such that $\boldsymbol{\|H_N - H_0\|_\infty = o_p(1)}$. Since $\overline{C} > \underline{C}$ there is an $\epsilon > 0$ such that $\overline{C} > \underline{C} + \epsilon$. Let $G_0 : (0, 1) \to \mathbb{R}$ be a linear function with slope $\underline{C} + \epsilon$, such that $G_0(1/2) = 0$. For an increasing sequence λ_N tending to one as N tends to infinity, let $H_N = \lambda_N H_0 + (1 - \lambda_N) G_0$. Taking $1 - \lambda_N \geq \max \left\{ \frac{\underline{C}_N - \underline{C}}{\epsilon}, \frac{\overline{C} - \overline{C}_N}{\overline{C} - (\underline{C} + \epsilon)} \right\}$, we have $|H_N| \leq \overline{C}_N$ and $\underline{C}_N \leq \nabla H_N \leq \overline{C}_N$, hence $H_N \in \mathcal{H}_N$. Moreover:

$$\|H_N - H_0\|_\infty \leq (1 - \lambda_N)\|H_0\|_\infty + (1 - \lambda_N)\|G_0\|_\infty = o(1).$$

A.1.2 Factor models: Theorem 1

We now prove Theorem 1. For any $H = (H_1, ..., H_K)$, denote the empirical objective function as:

$$\widehat{Q}(H) = \min_{\pi \in \Pi_N} \frac{1}{N} \sum_{i=1}^{N} \left\| Y_{\pi(i)} - \sum_{k=1}^{K} A_k H_k \left(\frac{\sigma_k(i)}{N+1} \right) \right\|^2,$$

where $Y_i = (Y_{i1}, ..., Y_{iT})'$ is a $T \times 1$ vector for all i, $A = (A_1, ..., A_K)$ with A_k a $T \times 1$ vector for all k, and $\sigma_1, ..., \sigma_K$ are independent permutations in Π_N. Denote as $\widehat{\mu}_Y$ the empirical measure of Y_i, $i = 1, ..., N$, with population counterpart μ_Y, and as $\widetilde{\mu}_{AH}$ the empirical measure of $\sum_{k=1}^{K} A_k H_k \left(\frac{\sigma_k(i)}{N+1} \right)$, $i = 1, ..., N$, with population counterpart μ_{AH}. Then $\widehat{Q}(H)^{\frac{1}{2}} = W_2 \left(\widehat{\mu}_Y, \widetilde{\mu}_{AH} \right)$ is the quadratic

Wasserstein distance between $\widehat{\mu}_Y$ and $\widetilde{\mu}_{AH}$. Likewise, let us define the population counterpart to $\widehat{Q}$, for any $H = (H_1, ..., H_K)$, as:

$$Q(H) = \inf_{\pi \in \mathcal{M}(\mu_Y, \mu_{AH})} \mathbb{E}_\pi \left[\left\| Y - \sum_{k=1}^K A_k H_k \left(V_k\right) \right\|^2 \right],$$

where the infimum is taken over all possible joint distributions of the random vectors Y and $\sum_{k=1}^K A_k H_k \left(V_k\right)$, with marginals μ_Y and μ_{AH}. Then $Q(H)^{\frac{1}{2}} = W_2\left(\mu_Y, \mu_{AH}\right)$ is the Wasserstein distance between the two population marginals.

The proof follows the steps of the proof of Corollary 1. The differences are as follows.

Parameter space. Let $\mathcal{H}$ be the closure of the set $\{H \in \mathcal{C}^1 : \nabla H \geq \underline{C}, \|H\|_{1,\infty} \leq \overline{C}\}$ under $\|\cdot\|_\infty$. Then, let us define:

$$\mathcal{H}_K \equiv \left\{ (H_1, ..., H_K) : H_k \in \mathcal{H} \text{ and } \sum_{i=1}^N H_k\left(\frac{i}{N+1}\right) = 0 \text{ for all } k \right\}.$$

$\mathcal{H}_K$ is compact under $\|\cdot\|_\infty$. The sieve construction is then similar to the scalar case.

$Q(H)$ is uniquely minimized at H_0 on $\mathcal{H}_K$. Let H be such that $Q(H) = 0$. Then $W_2\left(\mu_Y, \mu_{AH}\right) = 0$. By Theorem 7.3 in Villani (2003) this implies that $\mu_Y = \mu_{AH}$. Hence $Y = \sum_{k=1}^K A_k H_{0k}\left(V_k\right)$ and $\sum_{k=1}^K A_k H_k\left(V_k\right)$ have the same distributions. By Assumption 1 *(ii)*, it follows that $H_k = H_{0k}$ for all k.

$Q(H)$ is continuous on $\mathcal{H}_K$. Let H_1 and H_2 in $\mathcal{H}_K$. Since Y has bounded support, and H_{1k} and H_{2k} are bounded for all k, we have, for some constant $\widetilde{C} > 0$:

$$\left|Q(H_2) - Q(H_1)\right| \leq \widetilde{C} \left| Q(H_2)^{\frac{1}{2}} - Q(H_1)^{\frac{1}{2}} \right| = \widetilde{C} \left| W_2\left(\mu_Y, \mu_{AH_2}\right) - W_2\left(\mu_Y, \mu_{AH_1}\right) \right|,$$

Hence, since W_2 satisfies the triangle inequality (see Theorem 7.3 in Villani, 2003):

$$\left|Q(H_2) - Q(H_1)\right| \leq \widetilde{C} W_2\left(\mu_{AH_1}, \mu_{AH_2}\right).$$

Next, since supports are bounded, $W_2\left(\mu_{AH_1}, \mu_{AH_2}\right)$ is bounded, up to a multiplicative constant, by the Kantorovich-Rubinstein distance:

$$W_1(\mu_{AH_1}, \mu_{AH_2}) = \inf_{\pi \in \mathcal{M}(\mu_{AH_1}, \mu_{AH_2})} \mathbb{E}_\pi \left(\left\| \sum_{k=1}^K A_k H_{1k}\left(V_{1k}\right) - \sum_{k=1}^K A_k H_{2k}\left(V_{2k}\right) \right\| \right).$$

Now, using the dual representation of the Kantorovich-Rubinstein distance (see Theorem 1.14 in Villani, 2003), W_1 can be equivalently written as:

$$W_1(\mu_{AH_1},\mu_{AH_2}) = \sup_{\varphi \textit{ 1-Lipschitz}} \mathbb{E}\left(\varphi\left(\sum_{k=1}^{K} A_k H_{1k}\left(V_{1k}\right)\right)\right) - \mathbb{E}\left(\varphi\left(\sum_{k=1}^{K} A_k H_{2k}\left(V_{2k}\right)\right)\right),$$

where φ are 1-Lipschitz functions; that is, such that $|\varphi(y_2)-\varphi(y_1)| \leq \|y_2-y_1\|$ for all y_1, y_2.

Hence:

$$\begin{aligned} W_1(\mu_{AH_1},\mu_{AH_2}) &= \sup_{\varphi \textit{ 1-Lipschitz}} \int ... \int \left[\varphi\left(\sum_{k=1}^{K} A_k H_{1k}\left(\tau_k\right)\right) - \varphi\left(\sum_{k=1}^{K} A_k H_{2k}\left(\tau_k\right)\right)\right] d\tau_1 ... d\tau_K \\ &\leq \int ... \int \left\|\sum_{k=1}^{K} A_k H_{1k}\left(\tau_k\right) - \sum_{k=1}^{K} A_k H_{2k}\left(\tau_k\right)\right\| d\tau_1 ... d\tau_K \\ &\leq \sum_{k=1}^{K} \|A_k\| \, \|H_{1k}-H_{2k}\|_\infty, \end{aligned}$$

which implies that $H \mapsto Q(H)$ is continuous on $\mathcal{H}_K$.

$\mathbf{plim}_{N\to+\infty} \mathbf{sup}_{H\in\mathcal{H}_K} |\widehat{Q}(H) - Q(H)| = 0.$ Using similar arguments to the ones we used to show the continuity of $Q(H)$, we have, for $\widehat{C} = O_p(1)$:

$$\begin{aligned} \sup_{H\in\mathcal{H}_K} |\widehat{Q}(H) - Q(H)| &\leq \widehat{C} \sup_{H\in\mathcal{H}_K} \left|W_2\left(\widehat{\mu}_Y, \widetilde{\mu}_{AH}\right) - W_2\left(\mu_Y, \mu_{AH}\right)\right| \\ &\leq \widehat{C} \sup_{H\in\mathcal{H}_K} \left(W_2\left(\mu_Y, \widehat{\mu}_Y\right) + W_2\left(\mu_{AH}, \widetilde{\mu}_{AH}\right)\right), \end{aligned}$$

where in the last line we have used the triangle inequality (Theorem 7.3 in Villani, 2003).

Now, there is a positive constant $\widetilde{C}$ such that:

$$W_2\left(\mu_Y, \widehat{\mu}_Y\right) \leq \widetilde{C} W_1\left(\mu_Y, \widehat{\mu}_Y\right) = \widetilde{C} \sup_{\varphi \textit{ 1-Lipschitz}} \left[\mathbb{E}\left(\varphi\left(Y\right)\right) - \frac{1}{N}\sum_{i=1}^{N} \varphi\left(Y_i\right)\right] = o_p(1),$$

where the last equality holds since the set of 1-Lipschitz functions on a compact set (that we can assume to be bounded without loss of generality) is compact under $\|\cdot\|_\infty$.

Next, we have:

$$\begin{aligned} &\sup_{H\in\mathcal{H}_K} W_2\left(\mu_{AH}, \widetilde{\mu}_{AH}\right) \leq \widetilde{C} \sup_{H\in\mathcal{H}_K} W_1\left(\mu_{AH}, \widetilde{\mu}_{AH}\right) \\ &= \widetilde{C} \sup_{H\in\mathcal{H}_K} \sup_{\varphi \textit{ 1-Lipschitz}} \left[\mathbb{E}\left(\varphi\left(\sum_{k=1}^{K} A_k H_k\left(V_k\right)\right)\right) - \frac{1}{N}\sum_{i=1}^{N} \varphi\left(\sum_{k=1}^{K} A_k H_k\left(\frac{\sigma_k(i)}{N+1}\right)\right)\right] = o_p(1), \end{aligned}$$

where the last equality follows since $(\varphi, H) \mapsto K_{\varphi,H}$, where $K_{\varphi,H}(v_1,...,v_K) \equiv \varphi(\sum_{k=1}^{K} A_k H_k(v_k))$, defined for 1-Lipschitz functions φ and $H \in \mathcal{H}_K$, is Lipschitz.

This concludes the proof of Theorem 1.

A.2 Proof of Corollary 2

Let $\mathcal{H}_K^{(2)}$ denote the set of functions $(H_1, ..., H_K) \in \mathcal{H}_K$ which additionally satisfy $\|\nabla^2 H_k\|_\infty \leq \overline{C}$ for all k. Note that $\mathcal{H}_K^{(2)}$ is compact under $\|\cdot\|_{1,\infty}$. We construct the sieve space as:

$$\mathcal{H}_N^{(2)} = \left\{ H \in \mathcal{H}_K^{(2)} \, : \, \left\{ H_k\left(\frac{i}{N+1}\right) \, : \, i = 1, ..., N, \, k = 1, ..., K \right\} \in \mathcal{X}_N^{(2)} \right\}.$$

Let $k \in \{1, ..., K\}$. Let $\widehat{H}_k \in \mathcal{H}_N^{(2)}$ be such that $\widehat{H}_k\left(\frac{i}{N+1}\right) = \widehat{X}_{ik}$ for all i. The proof of Theorem 1, replacing the norm $\|\cdot\|_\infty$ by the norm $\|\cdot\|_{1,\infty}$, implies that $\|\widehat{H}_k - H_{0k}\|_{1,\infty} = o_p(1)$, which in turn implies that $\|\widehat{H}_k - H_{0k}\|_\infty = o_p(1)$ and $\|\nabla\widehat{H}_k - \nabla H_{0k}\|_\infty = o_p(1)$.

We then have:

$$\begin{aligned}
&\left| \frac{1}{Nb}\sum_{i=1}^N \kappa\left(\frac{\widehat{H}_k\left(\frac{i}{N+1}\right) - x}{b}\right) - \frac{1}{b}\int_0^1 \kappa\left(\frac{\widehat{H}_k(u) - x}{b}\right) du \right| \\
&= \left| \frac{1}{b}\sum_{i=1}^N \int_{\frac{i-1}{N}}^{\frac{i}{N}} \left[\kappa\left(\frac{\widehat{H}_k\left(\frac{i}{N+1}\right) - x}{b}\right) - \kappa\left(\frac{\widehat{H}_k(u) - x}{b}\right) \right] du \right| \\
&\leq \frac{C}{b^2}\sum_{i=1}^N \int_{\frac{i-1}{N}}^{\frac{i}{N}} \left| \widehat{H}_k\left(\frac{i}{N+1}\right) - \widehat{H}_k(u) \right| du \\
&\leq \frac{\widetilde{C}}{b^2}\sum_{i=1}^N \int_{\frac{i-1}{N}}^{\frac{i}{N}} \left| \frac{i}{N+1} - u \right| du = O_p(N^{-1}b^{-2}) = o_p(1),
\end{aligned}$$

where $C = O_p(1)$, $\widetilde{C} = O_p(1)$, and we have used that κ is Lipschitz, $\nabla\widehat{H}_k$ is uniformly bounded, and $Nb^2 \to +\infty$.

Now, using the change of variables $\omega = \frac{\widehat{H}_k(u) - x}{b}$, we obtain:

$$\frac{1}{b}\int_0^1 \kappa\left(\frac{\widehat{H}_k(u) - x}{b}\right) du = \int \kappa(\omega)\frac{1}{\nabla\widehat{H}_k\left(\widehat{H}_k^{-1}(x + b\omega)\right)} d\omega = \frac{1}{\nabla\widehat{H}_k\left(\widehat{H}_k^{-1}(x)\right)} + o_p(1),$$

where we have used that $x \mapsto 1/\nabla\widehat{H}_k(\widehat{H}_k^{-1}(x))$ is differentiable with uniformly bounded derivative, κ has finite first moments, $b \to 0$, and κ integrates to one.

Lastly, note that $f_{X_k}(x) = 1/\nabla H_{0k}(H_{0k}^{-1}(x))$, where by the above we have $\|\widehat{H}_k - H_{0k}\|_\infty = o_p(1)$, $\|\widehat{H}_k^{-1} - H_{0k}^{-1}\|_\infty = o_p(1)$, and $\|\nabla\widehat{H}_k - \nabla H_{0k}\|_\infty = o_p(1)$.

This shows Corollary 2.

B Expectations

For any Lipschitz function h, $\mathbb{E}(h(X_k))$ can be consistently estimated as: $\frac{1}{N}\sum_{i=1}^N h(\widehat{X}_{ik})$. Likewise, for all t, $\mathbb{E}(h(X_k, Y_t))$ is consistently estimated as: $\frac{1}{N}\sum_{i=1}^N h(\widehat{X}_{\sigma_k(i),k}, \sum_{\ell=1}^K a_{t\ell}\widehat{X}_{\sigma_\ell(i),\ell})$, for independent random permutations $\sigma_1, ..., \sigma_K$ in Π_N. Moreover, given the $\widehat{X}_{ik}$'s and the $\widehat{f}_{X_k}$'s, a consistent estimator of $\mathbb{E}\left(X_k \,|\, Y = y\right)$ is readily constructed. To see this, suppose the matrix formed by all the columns of A except the k-th one has rank T (which ensures that the conditional density of Y given X_k is not degenerate). Partition A into a $T \times (K-T)$ submatrix B_k and a non-singular $T \times T$ submatrix C_k, where the k-th column of A is one of the columns of B_k. Denote as X^{B_k} (resp., $\widehat{X}^{B_k}_{\sigma(i)}$) and X^{C_k} (resp., $\widehat{X}^{C_k}_{\sigma(i)}$) the subvectors of X (resp., $(\widehat{X}_{\sigma_1(i)}, ..., \widehat{X}_{\sigma_K(i)})'$) corresponding to B_k and C_k. An estimator of $\mathbb{E}\left(X_k \,|\, Y = y\right)$ is then:

$$\widehat{\mathbb{E}}\left(X_k \,|\, Y = y\right) = \frac{\sum_{i=1}^N \widehat{f}_{X^{B_k}}\left(\widehat{X}^{B_k}_{\sigma(i)}\right) \widehat{f}_{X^{C_k}}\left(C_k^{-1}\left[y - B_k \widehat{X}^{B_k}_{\sigma(i)}\right]\right) \widehat{X}_{\sigma_k(i),k}}{\sum_{i=1}^N \widehat{f}_{X^{B_k}}\left(\widehat{X}^{B_k}_{\sigma(i)}\right) \widehat{f}_{X^{C_k}}\left(C_k^{-1}\left[y - B_k \widehat{X}^{B_k}_{\sigma(i)}\right]\right)}. \tag{B8}$$

As an example, in the repeated measurements model (1), a consistent estimator of $\mathbb{E}\left(X_1 \,|\, Y = y\right)$ is, for $y = (y_1, ..., y_T)$:

$$\widehat{\mathbb{E}}\left(X_1 \,|\, Y = y\right) = \frac{\sum_{i=1}^N \prod_{t=1}^T \widehat{f}_{X_{t+1}}\left(y_t - \widehat{X}_{\sigma_1(i),1}\right) \widehat{X}_{\sigma_1(i),1}}{\sum_{i=1}^N \prod_{t=1}^T \widehat{f}_{X_{t+1}}\left(y_t - \widehat{X}_{i1}\right)} = \frac{\sum_{i=1}^N \prod_{t=1}^T \widehat{f}_{X_{t+1}}\left(y_t - \widehat{X}_{i1}\right) \widehat{X}_{i1}}{\sum_{i=1}^N \prod_{t=1}^T \widehat{f}_{X_{t+1}}\left(y_t - \widehat{X}_{i1}\right)}. \tag{B9}$$

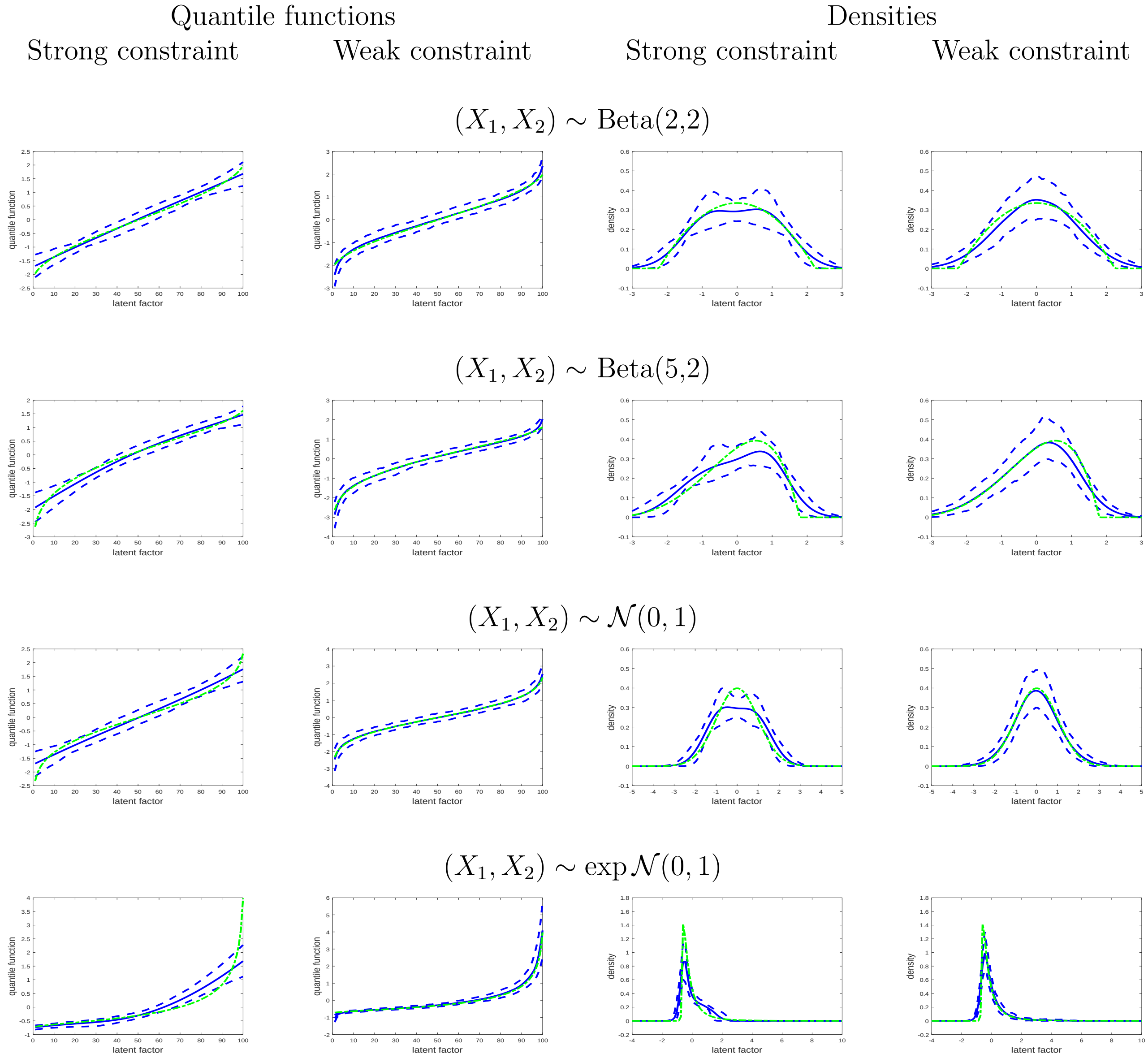
More generally, the densities $\widehat{f}_{X^{B_k}}$ and $\widehat{f}_{X^{C_k}}$ in (B8) are products of marginal densities of individual latent factors.

C Additional simulation results

In this section of the appendix we show simulation results for a scalar nonparametric deconvolution model. Consider the model $Y = X_1 + X_2$, where X_1 and X_2 are scalar, independent, and follow identical distributions. We assume that an i.i.d. sample of size N from X_2 is available. As for the repeated measurements model in the main text, we consider four specifications: Beta$(2, 2)$, Beta$(5, 2)$, normal, and log-normal, and we consider two choices for the penalization constants: $(\underline{C}_N, \overline{C}_N) = (.1, 10)$ ("strong constraint"), and $(\underline{C}_N, \overline{C}_N) = (0, 10000)$ ("weak constraint"). We use 10 randomly generated starting values, and average $M = 10$ sets of estimates.

In the first two columns in Figure C1 we show the estimates of the quantile functions $\widehat{X}_{i1} = \widehat{F}^{-1}_{X_1}\left(\frac{i}{N+1}\right)$, for the four specifications and both penalization parameters. The solid and dashed lines correspond to the mean, 10 and 90 percentiles across 100 simulations, respectively, while the dashed-dotted line corresponds to the true quantile function. The sample size is $N = 100$. In the last two columns of Figure C1 we show density estimates for the same specifications. The results reproduce the shape of the unknown quantile functions and densities rather well.

Figure C1: Monte Carlo results, deconvolution model, $N = 100$



Quantile functions

Densities

Strong constraint

Weak constraint

Strong constraint

Weak constraint

$(X_1, X_2) \sim$ Beta(2,2)

$(X_1, X_2) \sim$ Beta(5,2)

$(X_1, X_2) \sim \mathcal{N}(0,1)$

$(X_1, X_2) \sim \exp \mathcal{N}(0,1)$

Notes: Simulated data from the deconvolution model $Y = X_1 + X_2$. *The mean across simulations is in solid, 10 and 90 percent pointwise quantiles are in dashed, and the true quantile function or density of* X_1 *is in dashed-dotted.* 100 *simulations. 10 averages over* σ *draws.*

In Figure C2 we report additional results for the Beta$(2, 2)$ specification, for $N = 100$ (columns 1 and 3) and $N = 500$ (columns 2 and 4). In the first two rows we report the results based on a single σ draw per estimate (i.e., $M = 1$), whereas in the next two rows we show the results for the estimator averaged over $M = 10$ different σ draws. While we see that averaging seems to slightly increase the precision of estimated quantile functions and densities, the results based on one σ draw are comparable to the ones based on 10 draws. In the last row of Figure C2 we show results when using a single starting parameter value in our algorithm, instead of 10 values in our baseline estimates. We see that the results are little affected, suggesting that the impact of starting values on the performance of the estimator is moderate.

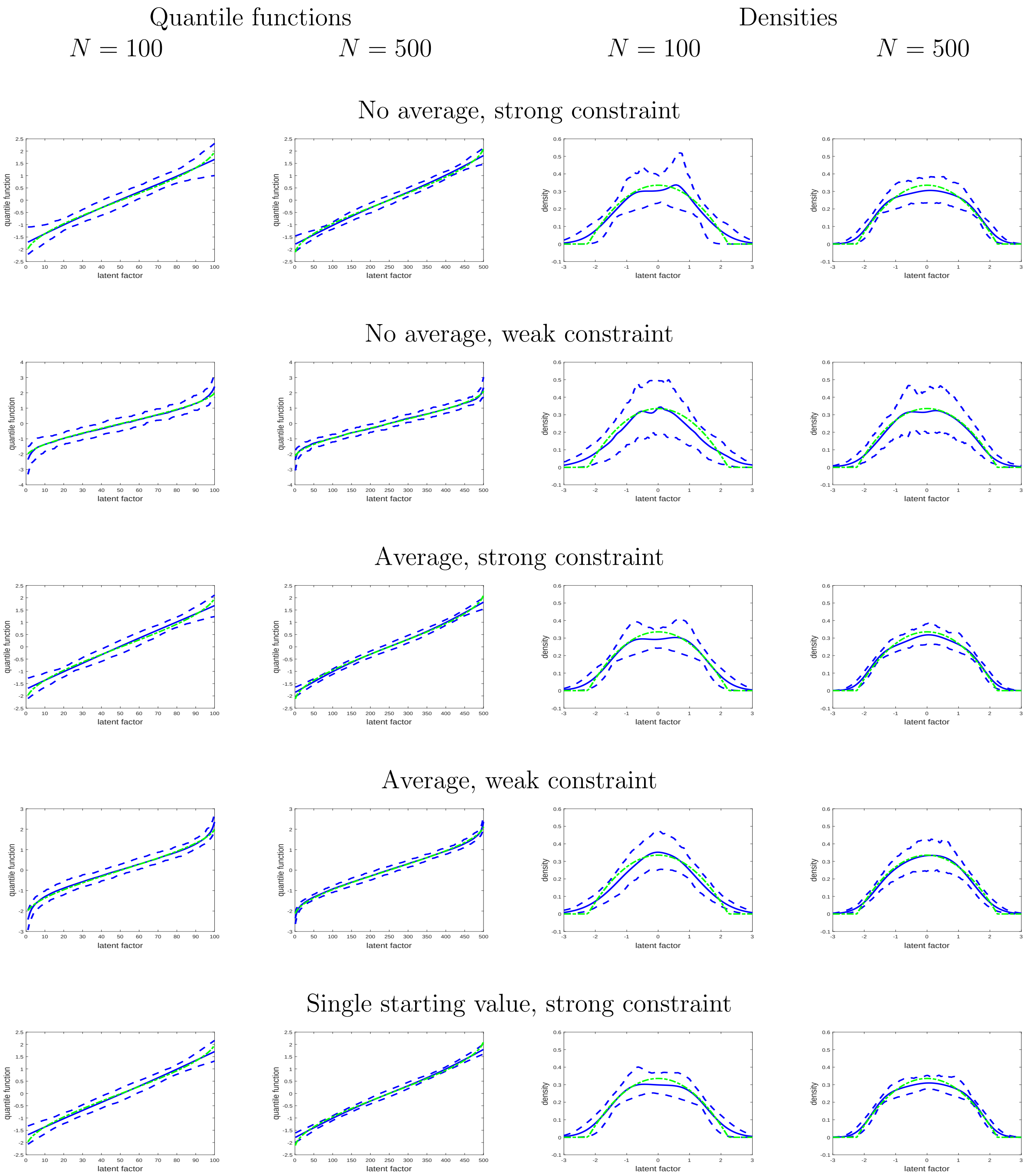
In Table C1 we attempt to quantify the rate of convergence of our quantile function estimator in a simulation experiment. We report the mean squared error at various quantiles (25%, median, and 75%) for the four distributional specifications. We focus on the weak constraint case, and rely on a single σ draw and single starting parameter value in each replication. We report the results of 500 simulations. In the last column of Table C1 we report a numerical rate of convergence based on these results, which we compute by regressing the log-mean squared error on the log-sample size. The results suggest the rate ranges between $N^{-\frac{3}{10}}$ and $N^{-\frac{7}{10}}$.[26]

Next, we assess the impact of the penalization parameters $\overline{C}_N$ and $\underline{C}_N$ on the mean squared error of quantile estimates, at the median and 25% and 75% percentiles. In Figure C3 we show the results for the four specifications, when varying the logarithm of $\overline{C}_N$ between 0 and 150 and setting $\underline{C}_N = \overline{C}_N^{-1}$, for two sample sizes: $N = 100$ (top panel) and $N = 500$ (bottom panel). Two features emerge. First, setting $\overline{C}_N$ to a very large number, which essentially fully relaxes the constraints, still results in a well-behaved estimator. This is in contrast with popular regularization methods for ill-posed inverse problems such as Tikhonov regularization or spectral cut-off, for which decreasing the amount of penalization typically causes large increases in variance. The sensitivity of characteristic-function based estimators to the choice of regularization parameters is also well documented. We interpret this feature of our estimator as reflecting the fact that the matching-based procedure induces an implicit regularization, even in the absence of additional constraints on parameters. Second, the results show that fully removing the penalization may not be optimal in terms of mean squared error. This raises the question of the optimal choice of the penalization parameters, which exceeds the scope of this paper.

We next consider a data generating process (DGP) which has been previously used to assess the finite-sample behavior of several estimators in the nonparametric deconvolution model. This DGP was used in Koenker and Gu (2019), and it is a slight variation of a DGP introduced by Efron

[26] From Theorem 3.7 in Hall and Lahiri (2008), when characteristic functions of X_1 and X_2 are converging at polynomial rates of order b and a, respectively, the optimal rate of convergence for quantile estimation is $N^{-\frac{2b}{2a+2b-1}}$. As an example, in the case of the Beta(2,2) and Beta(5,2) distributions, the corresponding rate is $N^{-\frac{4}{7}}$.

Figure C2: Monte Carlo results, deconvolution model, Beta(2,2), $N = 100, 500$



Quantile functions

Densities

$N = 100$ $N = 500$ $N = 100$ $N = 500$

No average, strong constraint

No average, weak constraint

Average, strong constraint

Average, weak constraint

Single starting value, strong constraint

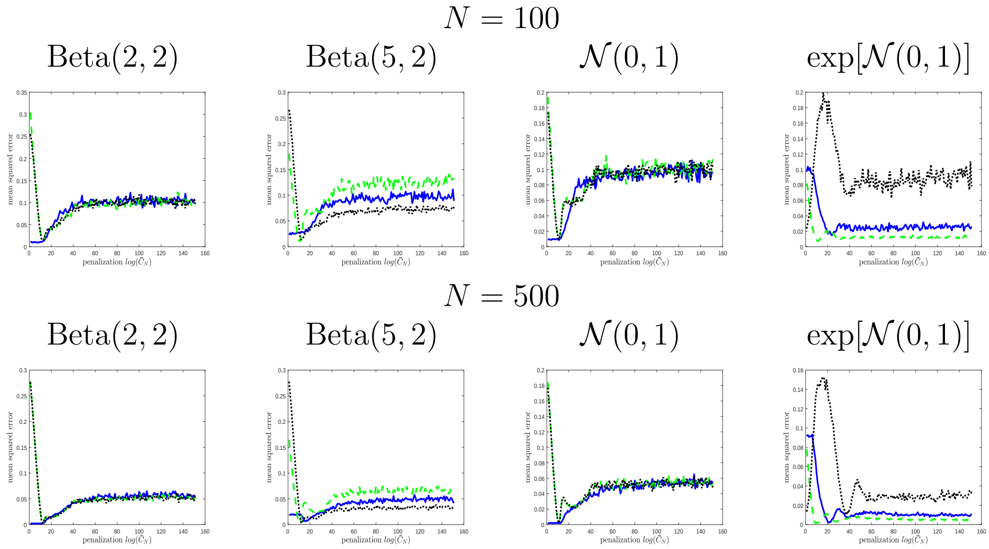
Notes: Simulated data from the deconvolution model $Y = X_1 + X_2$*. The mean across simulations is in solid, 10 and 90 percent pointwise quantiles are in dashed, and the true quantile function or density of* X_1 *is in dashed-dotted.* 100 *simulations.*

Table C1: Monte Carlo simulation, mean squared error of estimated quantiles of X_1 in the deconvolution model: 25%, 50%, and 75%

$N =$	100	200	300	400	500	600	700	800	900	1000	Implied rate
					Beta(2,2)						
25% perc.	0.1019	0.0695	0.0649	0.0514	0.0436	0.0472	0.0443	0.0431	0.0387	0.0457	-0.3866
Median	0.1086	0.0863	0.0625	0.0624	0.0540	0.0609	0.0481	0.0494	0.0499	0.0468	-0.3622
75% perc.	0.0946	0.0671	0.0642	0.0565	0.0466	0.0454	0.0449	0.0415	0.0392	0.0444	-0.3660
					Beta(5,2)						
25% perc.	0.1188	0.0916	0.0711	0.0635	0.0628	0.0625	0.0582	0.0587	0.0547	0.0572	-0.3237
Median	0.0927	0.0757	0.0516	0.0532	0.0466	0.0456	0.0386	0.0372	0.0357	0.0387	-0.4219
75% perc.	0.0732	0.0503	0.0417	0.0334	0.0347	0.0319	0.0260	0.0239	0.0249	0.0246	-0.4888
					$\mathcal{N}(0,1)$						
25% perc.	0.1146	0.0674	0.0650	0.0559	0.0549	0.0518	0.0514	0.0396	0.0463	0.0464	-0.3789
Median	0.0892	0.0789	0.0596	0.0584	0.0427	0.0520	0.0477	0.0444	0.0450	0.0416	-0.3411
75% perc.	0.1036	0.0745	0.0663	0.0605	0.0599	0.0461	0.0516	0.0459	0.0430	0.0451	-0.3704
					$\exp[\mathcal{N}(0,1)]$						
25% perc.	0.0114	0.0086	0.0052	0.0052	0.0050	0.0048	0.0049	0.0050	0.0042	0.0047	-0.3925
Median	0.0265	0.0195	0.0133	0.0108	0.0085	0.0118	0.0070	0.0070	0.0071	0.0084	-0.5885
75% perc.	0.0761	0.0586	0.0384	0.0333	0.0289	0.0243	0.0227	0.0186	0.0196	0.0192	-0.6555

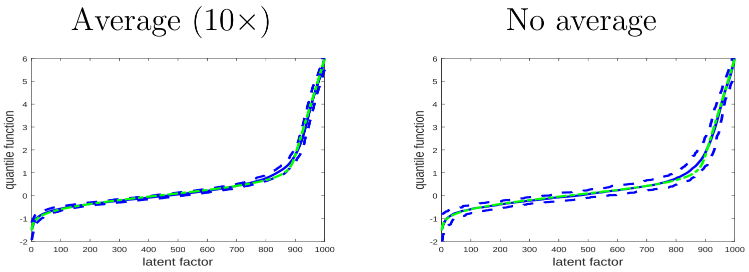
Notes: Mean squared error across 500 *simulations from the deconvolution model* $Y = X_1 + X_2$. *No average, single starting value, weak constraint. The implied rate in the last column is the regression coefficient of the log-mean squared error on the log-sample size.*

Figure C3: Monte Carlo simulation, mean squared error of estimated quantiles of X_1 as a function of the penalization parameter



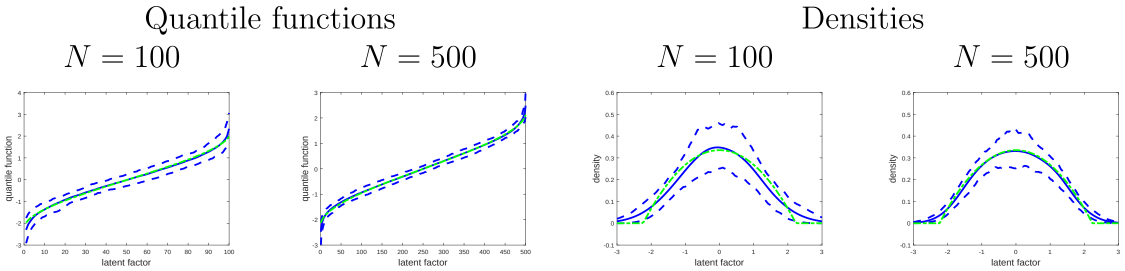
Notes: Simulated data from the deconvolution model $Y = X_1 + X_2$*. Log of penalization* $\overline{C}_N$ *(x-axis) against mean squared error (y-axis).* $\underline{C}_N$ *is set to* $\overline{C}_N^{-1}$*. Solid corresponds to the median, dashed to the 25% quantile, dotted to the 75% quantile. No average, single starting value, weak constraint.* $N = 100$ *(top panel) and* $N = 500$ *(bottom panel),* 500 *simulations.*

Figure C4: Monte Carlo results, deconvolution model, Efron-Koenker-Gu specification, $N = 1000$



Notes: Simulated data from the specification of the deconvolution model $Y = X_1 + X_2$ *used in Koenker and Gu (2019), which is a slight variation on a DGP used in Efron (2016). The mean across simulations is in solid, 10 and 90 percent pointwise quantiles are in dashed, and the true quantile function of* X_1 *is in dashed-dotted. Weak constraint.* 100 *simulations.*

Figure C5: Monte Carlo results, deconvolution model, Beta(2,2), Mallows' (2007) algorithm


Quantile functions

$N = 100$ $N = 500$

Densities

$N = 100$ $N = 500$

Notes: Simulated data from the deconvolution model. The mean across simulations is in solid, 10 and 90 percent pointwise quantiles are in dashed, and the true density is in dashed-dotted. Mallows' (2007) algorithm. 100 *simulations.*

(2016). Let $Y = X_1 + X_2$, where X_2 is distributed as a standard normal, and X_1 is distributed as a mixture of two distributions: a normal $\left(0, \frac{1}{2}\right)$ with probability $\frac{6}{7}$, and a uniform on the $[0, 6]$ interval with probability $\frac{1}{7}$. Koenker and Gu report that the Stefanski and Carroll (1990) characteristic-function based estimator performs quite poorly on this DGP, distribution functions estimated on a sample of 1000 observations showing wide oscillations. In Figure C4 we apply our estimator to this DGP, and report the results of 100 simulations. In the left graph we show quantile function estimates averaged 10 times, whereas in the right the results correspond to a single σ draw per estimation. We see that nonparametric estimates are very close to the true quantile function. This performance stands in sharp contrast with that of characteristic-function based estimates, and is similar to the performance of the parametric estimator analyzed in Efron (2016).

Lastly, in Figure C5 we report simulation results for Mallows' (2007) stochastic estimator, in the case of the Beta$(2, 2)$ specification. As we pointed out in Section 4, this algorithm is closely related to ours, with the difference that new random permutations are re-drawn in every step. We draw 100 such permutations, and keep the results corresponding to the last 50. The results are similar to the ones obtained using our estimator under weak constraint, as can be seen by comparing Figures C2 and C5.

D Complements for the empirical application

We provide descriptive statistics on the 20 four-year subpanels from the PSID in Table D2. In the last column, we report the number of recession months during the corresponding period according to the NBER classification. Dispersion of log-income growth residuals, as measured by the quantile

Table D2: Descriptive statistics

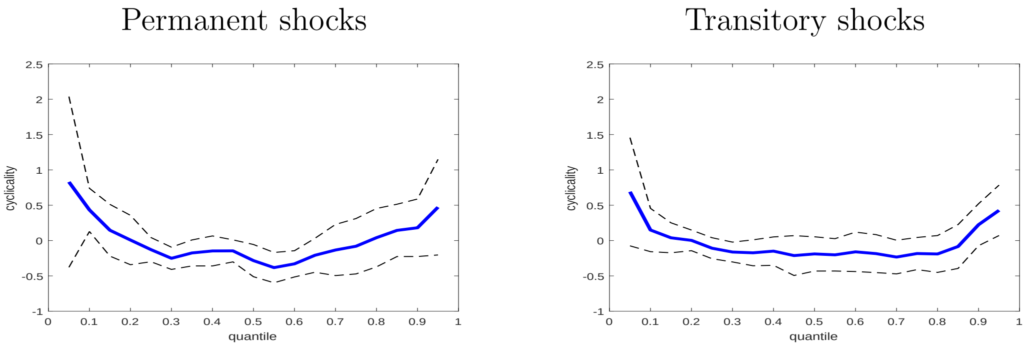
Period	Observations	Dispersion	Skewness	Kurtosis	Covariance	Recession months
1969–1972	1451	0.6495	-0.0303	6.7512	-0.0400	12
1970–1973	1511	0.7055	0.0222	7.0672	-0.0651	13
1971–1974	1555	0.7167	0.0304	7.1692	-0.0662	13
1972–1975	1605	0.7347	0.0239	6.5708	-0.0553	17
1973–1976	1649	0.7469	-0.0270	6.8749	-0.0506	17
1974–1977	1680	0.7493	-0.0500	6.8490	-0.0564	16
1975–1978	1776	0.7448	-0.0140	7.7408	-0.0808	4
1976–1979	1820	0.7108	0.0586	7.5271	-0.0627	0
1977–1980	1883	0.7308	0.0460	7.1757	-0.0467	7
1978–1981	1942	0.7383	0.0031	6.9209	-0.0674	12
1979–1982	2000	0.7584	-0.0557	7.0484	-0.0615	24
1980–1983	2038	0.7898	-0.0807	7.6125	-0.0619	24
1981–1984	2040	0.8049	-0.0782	8.0327	-0.0831	17
1982–1985	2062	0.8481	-0.0337	7.6789	-0.0824	12
1983–1986	2077	0.8314	0.0094	7.7976	-0.0877	0
1984–1987	2137	0.8196	0.0234	7.3038	-0.0703	0
1985–1988	2191	0.7829	0.0177	7.6650	-0.0659	0
1986–1989	2189	0.7655	0.0326	7.8191	-0.0472	0
1987–1990	2212	0.7342	0.0307	7.9466	-0.0670	5
1988–1991	2227	0.7494	-0.0229	7.8631	-0.0686	9

Notes: PSID, 1969–1991. Log-household annual income growth net of indicators for age (of head), education, gender, race, marital status, state of residence, number of children, and family size. Dispersion is the quantile difference $P_{90}-P_{10}$, Bowley-Kelley skewness is $[(P_{90}-P_{50})-(P_{50}-P_{10})]/(P_{90}-P_{10})$, Crow-Siddiqui kurtosis is $(P_{97.5}-P_{2.5})/(P_{75}-P_{25})$, and covariance is the first-order autocovariance of log-income growth. Recession months are computed according to the classification of the National Bureau of Economic Research (NBER).

difference $P_{90} - P_{10}$, tends to increase around the two recession periods in the mid-1970's and early 1980's. Skewness, as measured by the Bowley-Kelley ratio $[(P_{90} - P_{50}) - (P_{50} - P_{10})]/(P_{90} - P_{10})$, tends to become negative in recessions. Kurtosis, as measured by the Crow-Siddiqui ratio $(P_{97.5} - P_{2.5})/(P_{75} - P_{25})$, suggests substantial excess kurtosis relative to the Gaussian throughout the period. Lastly, the first-order autocovariance is consistently negative, and it tends to be larger in absolute value during recessions.

As a graphical way to illustrate the distributional dynamics of income over the business cycle, in Figure D6 we plot the coefficients of log-GDP growth in regressions of the quantiles of permanent or transitory income shocks on log-GDP growth and a time trend. The estimates suggest a U-shape pattern along the distribution, both for permanent and transitory shocks. Expansions are associated with increases at the top and bottom of the distribution, while recessions are associated

Figure D6: Quantiles over the business cycle

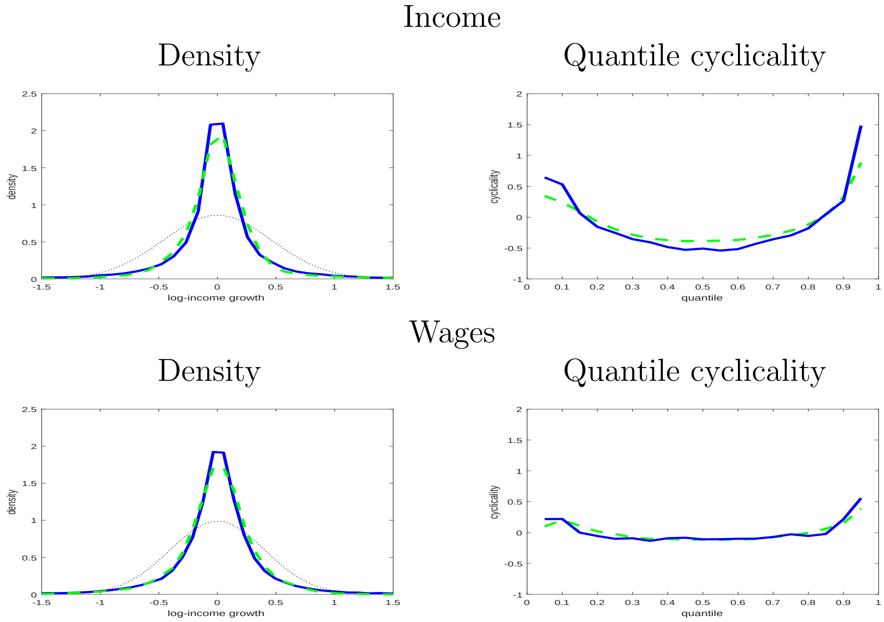


Permanent shocks

Transitory shocks

Notes: See notes to Figure 3. On the y-axis we report estimates of the coefficient of log-real GDP growth in the regression of quantiles of permanent or transitory shocks in a regression that includes a time trend. The quantiles are shown on the x-axis. Newey-West 95% confidence intervals are shown in dashed.

Figure D7: Fit to densities and quantile cyclicality of log-income/wage growth



Income

Density

Quantile cyclicality

Wages

Density

Quantile cyclicality

Notes: See notes to Figure 3. In the upper left panel we show the density of log-income growth in the data (in solid), and as predicted by our model (in dashed), with a normal fit (in dotted). In the upper right panel we show a measure of quantile cyclicality similar to the one in Figure D6 for log-income growth, in the data (in solid), and as predicted by the model (in dashed). In the bottom panels we show results for hourly wages.

with the opposite pattern and a relative increase of the middle quantiles.

In the upper panel of Figure D7 we show how the model fits the distributions of log-income growth, suggesting that our model is able to reproduce the density and quantile cyclicality of log-income growth that we observe in the data. Lastly, in the lower panel of Figure D7 we show the model fit to log-hourly wage growth. The estimates show that quantiles of log-hourly wage growth vary little with the business cycle in our sample, and that our model is able to reproduce this pattern.

E Extensions

In this section of the appendix we outline several generalizations of our estimation approach.

E.1 Random coefficients

Consider the linear cross-sectional random coefficients model:

$$Y = X_1 + \sum_{k=2}^{K} W_k X_k, \tag{E10}$$

where $(W_2, ..., W_K)$ and $(X_1, ..., X_K)$ are independent, the scalar outcome Y and the covariates $W_2, ..., W_K$ are observed, and $(X_1, ..., X_K)$ is a latent vector with an unrestricted joint distribution (e.g., Beran and Hall, 1992). To construct a matching estimator, one can augment (E10) with: $W_k = V_k$, $k = 2, ..., K$, where the V_k's are *auxiliary latent variables* independent of the X_k's. In this augmented model, the joint distributions of $(X_1, ..., X_K)$ and $(V_2, ..., V_K)$ can be estimated by minimizing the Euclidean distance between the model's predictions of Y, W observations, and their matched values in the data; that is,

$$(\widehat{X}, \widehat{V}) = \underset{(X,V)\in\mathcal{X}_N\times\mathcal{V}_N}{\operatorname{argmin}} \left\{ \min_{\pi\in\Pi_N} \sum_{i=1}^{N} \left(Y_{\pi(i)} - X_{\sigma_1(i),1} - \sum_{k=2}^{K} V_{\widetilde{\sigma}_k(i),k} X_{\sigma_k(i),k} \right)^2 + \lambda \sum_{k=2}^{K} (W_{\pi(i),k} - V_{\widetilde{\sigma}_k(i),k})^2 \right\}, \tag{E11}$$

where σ_k and $\widetilde{\sigma}_k$ are random permutations of $\{1, ..., N\}$, $\mathcal{V}_N$ is the parameter space for $V_2, ..., V_K$, and $\lambda > 0$ is a constant.

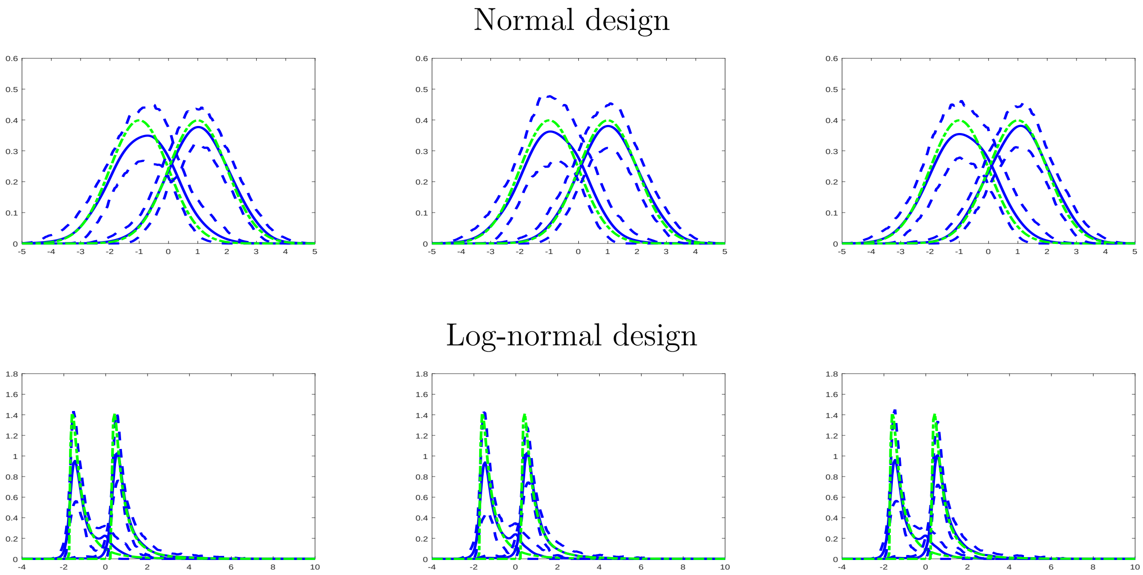
E.2 Finite mixtures

Consider next a finite mixture model with G groups, for a T-dimensional outcome Y:

$$Y_t = \sum_{g=1}^{G} Z_g X_{gt}, \quad t = 1, ..., T, \tag{E12}$$

where $Z_1, ..., Z_G$ and $X_{11}, ..., X_{GT}$ are unobserved, $Z_g \in \{0, 1\}$ with $\sum_{g=1}^{G} Z_g = 1$, and $(Z_1, ..., Z_G)$ and all X_{11}, ..., X_{GT} are mutually independent (e.g., Hall and Zhou, 2003). To construct a matching estimator, let $\mu = (\mu_1, ..., \mu_{G-1})$ and V standard uniform such that $Z_g = Z_g(V, \mu)$, where $Z_1(V, \mu) = 1$ if $V \leq \mu_1$, $Z_g(V, \mu) = 1$ if $\mu_{g-1} < V \leq \mu_g$ for $g = 2, ..., G-1$, and $Z_G(V, \mu) = 1$ if $\mu_{G-1} < V$.

Figure E8: Monte Carlo results, finite mixture model with two components



Normal design

Log-normal design

Notes: Simulated data from a finite mixture model with $G = 2$ components. The mean across simulations is in solid, 10 and 90 percent pointwise quantiles are in dashed, and the true density is in dashed-dotted. The two components have means -1 and 1 and unitary variances. Gaussian (top panel) and log-Gaussian (bottom panel) components. $N = 100$, $T = 3$, 100 simulations. $R = 10$ simulations per observation.

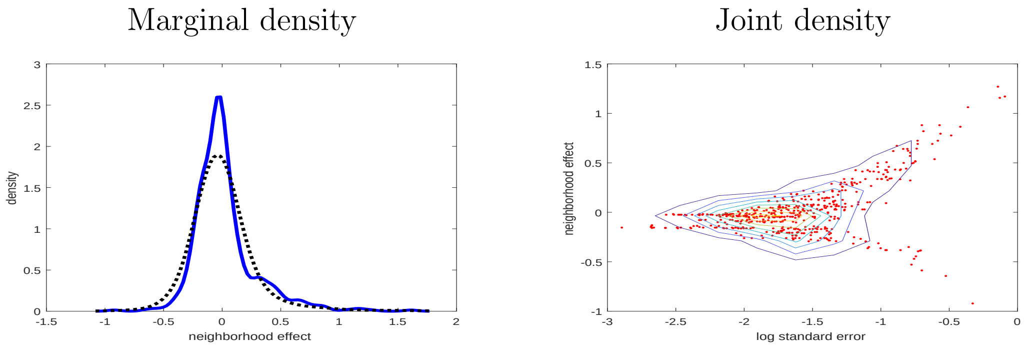
Let $\mathcal{M}_{G-1}$ be the set of vectors $\mu \in \mathbb{R}^{G-1}$ such that $0 \leq \mu_1 \leq \mu_2 \leq ... \leq \mu_{G-1} \leq 1$. We define the following estimator:

$$(\widehat{X}, \widehat{\mu}) = \underset{X \in \mathcal{X}_N,\, \mu \in \mathcal{M}_{G-1}}{\operatorname{argmin}} \left\{ \min_{\pi \in \Pi_N} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{\pi(i),t} - \sum_{g=1}^{G} Z_g(V_i, \mu) X_{\sigma_{gt}(i),gt} \right)^2 \right\}, \tag{E13}$$

where $V_1, ..., V_N$ are standard uniform draws, and σ_{gt} are random permutations in Π_N for all $g = 1, ..., G$, $t = 1, ..., T$, all independent of each other. For given μ, one can use an algorithm analogous to the one described in Section 4 to compute $\widehat{X}$. The outer minimization with respect to μ can be performed using simulated annealing or other methods to minimize non-differentiable objective functions. In the illustration below grid search is a viable option.

In Figure E8 we report the results of 100 simulations, for two DGPs, both of which are finite mixture models with $G = 2$ components with independent measurements. We consider a normal DGP and a log-normal DGP. To fix the labeling across simulations, we order the components by increasing means. We use a version of (E13) with multiple draws $\sigma_{gt}(i, r)$ for all i, with $R = 10$ simulations by observation (see footnote 8). We use 3 starting values in every inner loop, and perform an outer loop for 10 equidistant values of the first group's probability. The results in Figure E8 are encouraging, and suggest that matching estimators can perform well in nonparametric finite

Figure E9: Density of neighborhood effects



Notes: In the left graph we show the density of commuting zone effects X_{i1} in model (E14) in solid, and the density of neighborhood estimates Y_i in dashed. In the right graph we show contour plots of the joint density of (X_{i1}, S_i), where S_i is the standard deviation of Y_i. Calculations are based on statistics available on the Equality of Opportunity website.

mixture models too.

E.3 Heteroskedastic deconvolution

Finally, consider the model

$$Y = X_1 + SX_2, \tag{E14}$$

where (X_1, S) is independent of X_2, and $X_2 \sim F$, where F is known and has zero mean. The analyst observes a sample $Y_1, \widetilde{S}_1, ..., Y_N, \widetilde{S}_N$ from $(Y, \widetilde{S})$, where $\widetilde{S}_i$ is a consistent estimator of S_i for all i. To motivate this setup, consider the estimation of income neighborhood effects in Chetty and Hendren (2018), where i is a commuting zone or county, and Y_i is a neighborhood-specific estimate of the "causal effect" of place i. Within-i, a central limit theorem-type argument suggests that Y_i is approximately normally distributed, with mean X_{i1} and standard deviation S_i. Chetty and Hendren report, alongside Y_i estimates, standard deviation estimates $\widetilde{S}_i$. In this example F is the standard normal distribution. To estimate the distribution of X_1 by matching, we minimize the following objective:

$$(\widehat{X}_1, \widehat{S}) = \underset{(X_1,S)\in\mathcal{X}_N\times\mathcal{S}_N}{\operatorname{argmin}} \left\{ \min_{\pi\in\Pi_N} \sum_{i=1}^{N} \left(Y_{\pi(i)} - X_{i1} - S_i X_{\sigma(i),2}\right)^2 + \lambda \left(\widetilde{S}_{\pi(i)} - S_i\right)^2 \right\}, \tag{E15}$$

where σ is a random permutation of $\{1, ..., N\}$, $\mathcal{S}_N$ is the parameter space for S, and $\lambda > 0$ is a constant. The algorithm again consists in alternating optimal transport steps and least squares steps.

As an illustration, we estimate the density of neighborhood effects across US commuting zones using data made available by Chetty and Hendren (2018). For every commuting zone, Chetty and Hendren report an estimate Y_i of the causal income effect of i, alongside an estimate $\widetilde{S}_i$ of its standard error. We compute a heteroskedastic Gaussian deconvolution estimator of the density of the latent neighborhood effects. As a by-product, we obtain an estimate of the joint density of neighborhood effects and their standard errors. To implement the calculation we set $\lambda = 10$, trim the top 1% percentile of $\widetilde{S}_i$, and weigh all results by population weights. To accommodate the presence of weights in a simple way, we draw subsamples of 500 observations from the weighted empirical distribution of $(Y_i, \widetilde{S}_i)$. We then average the results across $M = 10$ subsamples.

We show the results in Figure E9. We see that neighborhood effects are not normally distributed. They show right skewness, and excess kurtosis. Estimates of Bowley-Kelley skewness and Crow-Siddiqui kurtosis of X_{i1} are 0.33 and 4.75, respectively. The joint density of neighborhood effects and standard errors suggests that less populated commuting zones with less precise estimates tend to have higher income premia. The rank correlation between neighborhood effects and standard errors is 0.39. Lastly, the joint density also shows a high degree of non-Gaussianity.