Recovering Latent Variables by Matching*

Manuel Arellano† Stéphane Bonhomme‡

First draft: January 2018

Abstract

We propose a 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 the empirical distribution of those latent values is consistent for the population distribution. We illustrate the method on simulated data, and in two applications: nonparametric estimation of the densities of permanent and transitory earnings shocks on panel data from the PSID, and nonparametric estimation of school quality in the Spanish region of Madrid.

KEYWORDS: Unobserved heterogeneity, nonparametric estimation, matching, factor models, optimal transport.

*We thank Colin Mallows, Kei Hirano, Roger Koenker, Thibaut Lamadon, Guillaume Pouliot, Azeem Shaikh and Tim Vogelsang for comments. We thank Miguel Ruiz for sharing the school data. 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

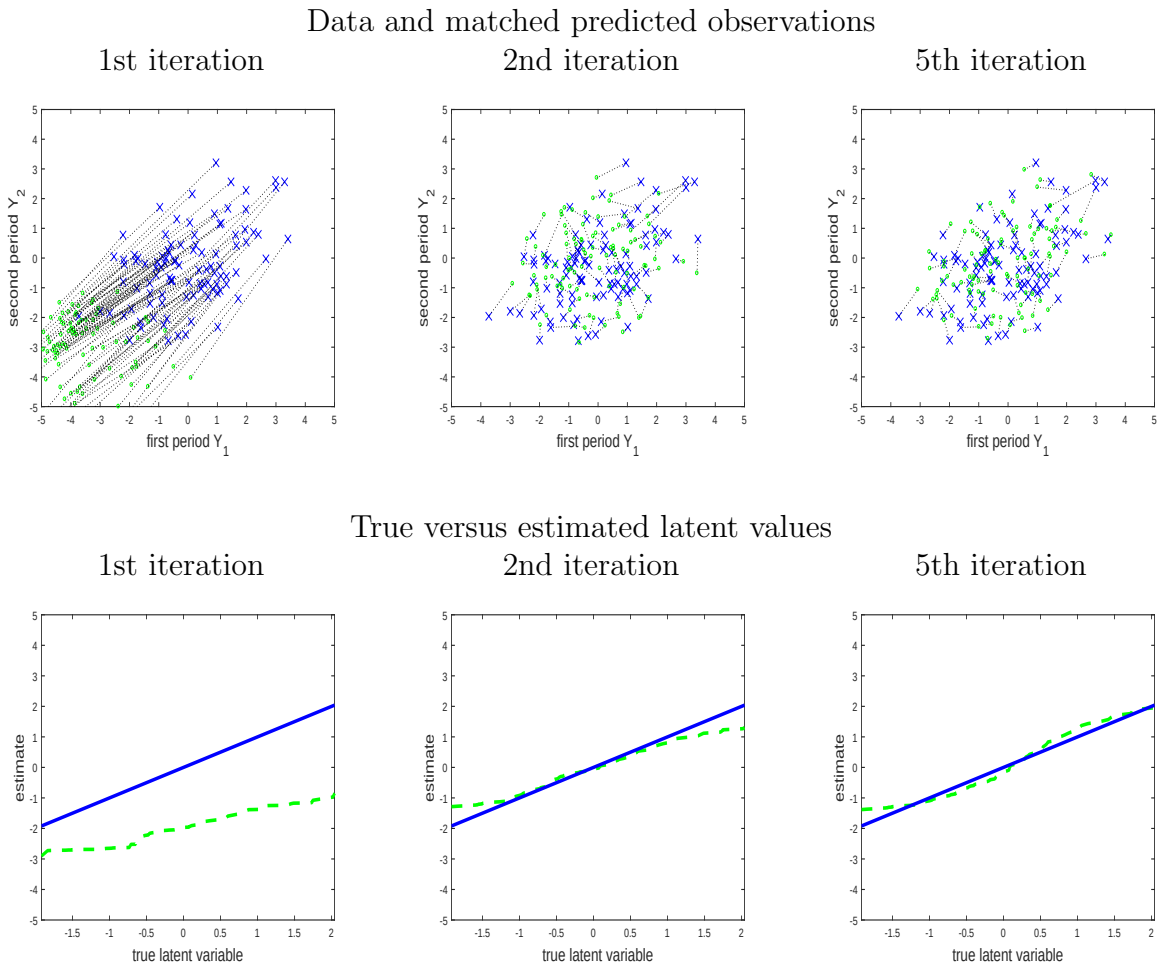
In this paper we propose a method to nonparametrically estimate a class of models with latent variables. We focus on linear factor models where the latent factors are mutually independent. These models have a wide array of economic applications, including measurement error models, fixed-effects models, and error components models. We review the existing literature on these models in Section 2. In many empirical settings, it is appealing not to restrict the functional form of the distributions of latent variables and follow a nonparametric approach.

While the estimation of independent factor models often relies on parametric assumptions, based on Gaussian mixtures or other parametric families, there is also a large literature on nonparametric estimation based on empirical characteristic functions (e.g., Carroll and Hall, 1988, Stefanski and Carroll, 1990, Horowitz and Markatou, 1996, Li and Vuong, 1998, Bonhomme and Robin, 2010). However, those estimators tend to be highly sensitive to the choice of regularization parameters, and they do not guarantee that the estimated distribution functions be non-negative and integrate to one. The difficulties with existing nonparametric estimators are well-documented, see for example Efron (2016) and Chapter 21 in Efron and Hastie (2016).

In this paper we propose a nonparametric estimation approach that differs from the literature in two main aspects. First, we generate a sample of *pseudo-observations* from the latent variables. Such pseudo-observations may be interpreted as the order statistics of the latent variables. Moments, densities, or general functionals can then be estimated based on them. In particular, estimated densities will be proper 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 the 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. This amounts to minimizing a quadratic *Wasserstein* distance between empirical distribution functions. 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 moment matching estimators, which are commonly used in parametric econometric models.

Figure 1: Illustration of the estimation algorithm



Notes: The graphs correspond to one simulation from a fixed-effects model with two observation periods $Y_1 = X_1 + X_2$, $Y_2 = X_1 + X_3$, *with* X_1, X_2, X_3 *mutually independent (Kotlarski, 1967). In the data generating process the* X*'s are standardized Beta(2,2), and* $N = 100$. *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). Details on the algorithm are given in Section 3.*

As an illustration, in Figure 1 we show the results of several iterations of our algorithm, in a fixed-effects model with two observation periods and 100 individuals. We start the algorithm from parameter values that are far from the true ones. As shown on 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 on the top panel. The bottom panel shows the estimates of

the latent individual-specific effect, sorted in ascending order (on the y-axis), against the true values (on the x-axis). We see that within a few iterations the model's predictions and the empirical observations tend to agree with each other (on the top panel), and that the distribution of the pseudo latent observations gets close to the population distribution (on the bottom panel).

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. He focused on the classical deconvolution model with scalar outcome and known error distribution. Our main goal in this paper is to extend Mallows' insight by proposing a framework to analyze estimators based on matching predicted values from the model to data observations. This allows us to derive a well-defined population version of the estimation problem and establish consistency of our estimator. In addition, we show how the method can be generalized beyond scalar deconvolution, to models with multivariate outcomes such as fixed-effects models and other factor models.

A key step in our analysis is to relate the estimation problem to optimal transport theory. Optimal transport is the subject of active research in mathematics, see for example Villani (2003, 2008). Economic applications of optimal transport are many fold, as documented in Galichon (2016). 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. We also use properties of Wasserstein distances established in the optimal transport literature in our asymptotic analysis.

Our matching-based, minimum Wasserstein distance estimator is related to recent work in machine learning and statistics on the estimation of parametric generative models (see Bernton *et al.*, 2017, Genevay *et al.*, 2017, and Bousquet *et al.*, 2017). In contrast with this emerging literature, the models we consider here are nonparametric. An early theoretical contribution to minimum Wasserstein distance estimation by Bassetti *et al.* (2006) is more closely related to our consistency analysis. Our general estimation strategy is also related to Galichon and Henry's (2011) analysis of partially identified models. As we show at the end of the paper, our matching approach can be generalized to nonparametric estimation of other latent variables, such as nonparametric finite mixture models (Hall and Zhou, 2003).

We illustrate the performance of our estimator on simulated data. Under various spec-

ifications of the scalar nonparametric deconvolution model and the fixed-effects model, we find that our estimator recovers true underlying quantile functions and densities quite accurately, even for samples with only 100 individual observations. In addition, in our Monte Carlo designs we find our estimator outperforms characteristic-function based estimators, particularly due to improved estimation of the tails of the distributions.

We then apply our method to two empirical illustrations. In the first one, we estimate quantile functions and densities of permanent and transitory earnings shocks in the PSID. We find strong evidence of non-Gaussianity in both types of shocks characterized by excess kurtosis, confirming results obtained by Horowitz and Markatou (1996), Geweke and Keane (2000), and Bonhomme and Robin (2010). In the second illustration we estimate the distribution of school fixed-effects, net of transitory fluctuations, in the Madrid region in Spain. We find that there is substantial year-to-year variation in school outcomes, and that controlling for it makes a larger difference at the bottom of the distribution of students' performance than at the top.

The outline of the paper is as follows. In Section 2 we describe linear independent factor models, and we briefly review applications and existing estimation approaches. In Section 3 we introduce our matching estimator. In Sections 4 and 5 we focus on computation and consistency, respectively. In Sections 6 and 7 we present the simulation exercises and empirical illustrations. In Section 8 we outline an extension to finite mixture models. Lastly, we conclude in Section 9. Proofs and additional material are collected in the appendix.

2 Independent factor models

We focus on linear independent factor models of the form $Y = AX$, where $Y = (Y_1, ..., Y_T)'$, $X = (X_1, ..., X_K)'$, A is a known $T \times K$ matrix, and the components $X_1, ..., X_K$ are mutually independent.[1] In this section we review several examples of models and applications which have such a structure. We focus on the case $K > T$, so the system is singular and the latent variables themselves are not identifiable, although under suitable conditions their distributions will be.

Nonparametric deconvolution. The classical nonparametric deconvolution model obtains when $T = 1$ and $Y = X_1 + X_2$, under the assumption that X_2 has a known distribution.

[1]The analysis is unchanged in case A is not known but a consistent estimator of it is available.

This model has been extensively studied in statistics and econometrics. Nonparametric deconvolution is often used to deal with the presence of measurement error. In such settings Y is an error-ridden variable, X_1 the true value of the variable, and X_2 an independent, classical measurement error.[2]

Another economic application of nonparametric deconvolution is to the estimation of the heterogeneous effects of an exogenous binary treatment $D \in \{0, 1\}$. Under the assumption that the potential outcome $Y(0)$ in the absence of treatment is independent of the gains from treatment $Y(1) - Y(0)$, Heckman, Smith and Clements (1997) noted that:

$$\underbrace{Y(1)}_{\text{level (D=1)}} = \underbrace{Y(0)}_{\text{level (D=0)}} + \underbrace{Y(1) - Y(0)}_{gain}$$

is a classical deconvolution model, since the distributions of $Y(1)$ and $Y(0)$ are both consistently estimable under exogeneity. Building on this observation they provide conditions under which the joint distribution of potential outcomes $(Y(0), Y(1))$ can be consistently estimated.[3]

The random coefficients panel data models studied in Arellano and Bonhomme (2012) provide a third application. Consider a model with a time-varying binary treatment D_t and time-invariant treatment effect, whose outcome is: $Y_t = \alpha + \beta D_t + \varepsilon_t$, where ε_t are i.i.d., independent of $(\alpha, \beta, D_1, ..., D_T)$, but the dependence of (α, β) on $D_1, ..., D_T$ is unrestricted ("fixed-effects endogeneity"). Consider as an example a sequence of three outcomes corresponding to $D_1 = 0, D_2 = 0, D_3 = 1$. Arellano and Bonhomme observed that the following two equations:

$$Y_2 - Y_1 = \underbrace{\varepsilon_2 - \varepsilon_1}_{=\widetilde{X}_2}, \qquad Y_3 - Y_2 = \underbrace{\beta}_{=X_1} + \underbrace{\varepsilon_3 - \varepsilon_2}_{=X_2},$$

can be interpreted as a classical deconvolution model, since X_2 and $\widetilde{X}_2$ have the same distribution and X_1, X_2 are independent. They showed how to non-parametrically estimate the distribution of treatment effects β.

The literature on nonparametric deconvolution provides conditions under which the distribution of X_1 is nonparametrically identified in $Y = X_1 + X_2$. Approaches to estimation are numerous.[4]

[2]Carroll, Ruppert, Stefanski and Crainiceanu (2006), Chen, Hong and Nekipelov (2011), and Schennach (2013a) provide comprehensive reviews of the measurement error literature.

[3]Wu and Perloff (2006) propose an estimation method in this setup based on moment restrictions and entropy maximization.

[4]Examples are kernel deconvolution estimators (Carroll and Hall, 1988, Delaigle and Gijbels, 2002, Fan,

Nonparametric distribution of fixed effects. A leading example of a linear independent factor model is the fixed-effects model:

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

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 fixed-effects structure arises frequently in economic applications. As an example, α can be a latent skill of an individual, measured with error (as in Cunha, Heckman and Schennach, 2010). In other applications researchers may be interested in estimating the distribution of worker, teacher, firm, or bank-specific fixed-effects, for example. Compared to standard Gaussian specifications, a nonparametric estimator of the distribution of α in (1) will be robust to functional form assumptions. Non-Gaussianity, such as skewness or fat tail behavior for example, may be relevant in many empirical settings. The fixed-effects model (1) and its generalizations are often estimated using flexible parametric specifications such as finite Gaussian mixtures (e.g., Carneiro, Hansen and Heckman, 2003). Nonparametric estimators based on empirical characteristic functions have been constructed by mimicking and extending the original proof due to Kotlarski.[5]

Error components: generalized nonparametric deconvolution. A prominent error component model is the permanent-transitory model for the dynamics of log-earnings: $Y_t = \eta_t + \varepsilon_t$, where $\eta_t = \eta_{t-1} + v_t$ is a random walk with independent innovations, and all ε_t's and v_t's are independent over time and independent of each other (e.g., Hall and Mishkin, 1982, Blundell, Pistaferri and Preston, 2008). This model is a special case of a linear independent factor model $Y = AX$, where $Y = (Y_1, ..., Y_T)'$ are observed outcomes, $X = (X_1, ..., X_K)'$ are mutually independent latent factors, and A is a known $T \times K$ matrix. Identification of such generalized deconvolution models was studied in Székely and Rao (2000). Bonhomme and Robin (2010) proposed nonparametric characteristic-function based estimators of factor densities.[6] In such settings a nonparametric approach allows one to capture the skewness or

1991), wavelet methods (Pensky and Vidakovic, 1999, Fan and Koo, 2002), regularization techniques (Carrasco and Florens, 2011), and nonparametric maximum likelihood methods (e.g., Gu and Koenker, 2017).

[5]See Li and Vuong (1998) and Li (2002). See also Horowitz and Markatou (1996).

[6]Botosaru and Sasaki (2015) showed how to allow in addition for nonparametric heteroskedasticity. Quantile-based estimation in linear and nonlinear factor models was introduced by Arellano and Bonhomme

kurtosis of earnings shocks.

An important application of error components models is to relax independence in fixed-effects models such as (1). This can be done provided T is large enough.[7] Such specifications can be estimated using the methods we introduce here.

3 Latent variable estimation by matching

In this section we start by describing our estimator in the classical nonparametric deconvolution model, and then turn to linear multi-factor models with independent factors.

3.1 Nonparametric deconvolution

Let $Y = X_1 + X_2$, where X_1 and X_2 are independent. X_1 is unobserved to the econometrician and its distribution is left 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 (c.d.f.) 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.[8] 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 use a minimum-distance estimator based on a particular distance between the sample of observed Y's and the sample of Y's predicted by the model. The distance we consider is the *Wasserstein distance* (see, e.g., 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.

Assume without loss of generality that $Y_i \leq Y_{i+1}$ and $X_{i2} \leq X_{i+1,2}$ for all i. Let Π_N denote the set of permutations $\pi : \{1, ..., N\} \to \{1, ..., N\}$, such that $\sum_{i=1}^N \mathbf{1}\{\pi(i) = j\} = 1$ for all j, and $\sum_{j=1}^N \mathbf{1}\{\pi(i) = j\} = 1$ for all i. Moreover, let $\overline{C}_N > 0, \underline{C}_N > 0$ be two constants, and let $\mathcal{X}_N$ be the set of parameter vectors $X_1 = (X_{11}, ..., X_{N1}) \in \mathbb{R}^N$ such that $|X_{i1}| \leq \overline{C}_N$

(2016) and applied by Arellano, Blundell and Bonhomme (2017) to document the dynamics of earnings in the PSID.

[7] Modeling X_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). Ben Moshe (2017) showed how to allow for arbitrary subsets of dependent factors, and proposed characteristic-function based estimators. In addition, in model (1) Schennach (2013b) pointed out that full independence between the factors is in fact not necessary, and that sub-independence suffices to establish identification.

[8] The sample size being the same for Y and X_2 is not essential and could easily be relaxed. In fact, in a setting where the c.d.f. F_{X_2} is known one could alternatively work with an integral counterpart to our estimator.

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. We will study the sensitivity of our estimator to those constants in the simulation section.

We propose to compute:

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

where σ is a random permutation in Π_N, independent of $Y_1, ..., Y_N, X_{12}, ..., X_{N2}$.[9]

The estimator $\widehat{X}_1$ minimizes the Wasserstein distance between the empirical distributions of $Z_i = X_{\sigma(i),1} + X_{i2}$, $i = 1, ...N$, and Y_i, $i = 1, ..., N$. The Wasserstein distance is defined as:

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

The values $Z_i = X_{\sigma(i),1} + X_{i2}$, $i = 1, ..., N$, are draws from $X_1 + X_2$; that is, predicted values from the model. Hence $\widehat{X}_1$ minimizes the Wasserstein distance between the empirical distribution of the data and the empirical distribution of model predictions.

Since Y_i and Z_i are scalar, by the Hardy Littlewood and Polya rearrangement inequality the solution to (3) is to sort Y_i's and Z_i's in the same order. That is, letting $\widehat{\pi}$ denote the minimum argument in (3), $\widehat{\pi}(i)$ is the *rank* of Z_i:

$$\widehat{\pi}(i) = \text{Rank}(Z_i) \equiv N \widehat{F}_Z(Z_i).$$

Remark 1: averaging. The estimates $\widehat{X}_{i1}$, $i = 1, ..., N$, depend on the permutation σ. A simple way to reduce the dependence on this random 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 $i = 1, ..., M$. For fixed M, such averages will be consistent as N tends to infinity under similar conditions as the baseline estimator.

Remark 2: draws from the model. Given X_{i1}'s and X_{i2}'s, predicted values from the model could be generated in other ways. For example, one could set $Z_i = X_{i1} + \widetilde{X}_{i2}$, where $\widetilde{X}_{12}, ..., \widetilde{X}_{N2}$ is a random sample from $X_{12}, ..., X_{N2}$ drawn with replacement. Alternatively, one could generate $R > 1$ predictions per observation i, $Z_{i,r} = X_{\sigma(i,r),1} + X_{i2}$, for $r = 1, ..., R$.

[9] Random permutations are uniform draws on Π_N. Simple algorithms exist to generate random permutations (e.g., Knuth, 1997).

In the latter case, π and σ would map $\{1, ..., NR\}$ to $\{1, ..., N\}$, so increasing R is associated with an increase in computational cost.

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.[10] We seek to compute pseudo-observations $\widehat{X}_{11}, ..., \widehat{X}_{N1}$, ..., $\widehat{X}_{1K}, ..., \widehat{X}_{NK}$, which 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.

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

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

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

Note that $Z_{it} = \sum_{k=1}^K a_{tk} X_{\sigma_k(i),k}$, $i = 1, ..., N$, are predicted values from the factor model. Hence, as before, the vector $\widehat{X}$ minimizes the Wasserstein distance between the empirical distribution of the data $(Y_{i1}, ..., Y_{iT})$, and the one of model predictions $(Z_{i1}, ..., Z_{iT})$. A difference with the scalar deconvolution model is that, 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 as the solution to a linear program. We will take advantage of this in our estimation algorithm.

3.3 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 $\max_{i=1,...,N} |\widehat{X}_{ik} - F^{-1}_{X_k}(\frac{i}{N+1})|$ tends to zero in probability asymptotically. This provides uniformly consistent estimators

[10] The mean of X is non-identifiable when $T < K$. 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 fixed-effects model assuming that $\mathbb{E}(X_1) = 0$ suffices for identification. Our algorithm can easily be adapted to such cases.

of the quantile functions of the latent variables. These estimators can in turn be used for density estimation, under a slight modification of the parameter space $\mathcal{X}_N$. To proceed, 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: $(N+1)^2 |X_{i+2,k} - 2X_{i+1,k} + X_{i,k}| \leq \overline{C}_N$, 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 standard conditions on the kernel κ and bandwidth b.

Expectations. Our estimator delivers simple consistent estimators of unconditional expectations. For example, for any Lipschitz function h, the expectation $\mathbb{E}(h(X_k))$ can be consistently estimated as $\frac{1}{N} \sum_{i=1}^{N} h\left(\widehat{X}_{ik}\right)$. Likewise, for all t, $\mathbb{E}(h(X_k, Y_t))$ is consistently estimated as $\frac{1}{N} \sum_{i=1}^{N} h\left(\widehat{X}_{\sigma_k(i),k}, \sum_{\ell=1}^{K} a_{t\ell} \widehat{X}_{\sigma_\ell(i),\ell}\right)$, for independent random permutations $\sigma_1, ..., \sigma_K$ in Π_N.

Conditional expectations are of particular interest in prediction problems. Given the $\widehat{X}_{ik}$'s and the $\widehat{f}_{X_k}$'s, a consistent estimator of the conditional expectation $\mathbb{E}(X_k \,|\, Y = y)$ 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). We can 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 $\mathbb{E}(X_k \,|\, Y = y)$ is then:

$$\widehat{\mathbb{E}}(X_k \,|\, Y = y) = \frac{\sum_{i=1}^{N} \widehat{f}_{X^{B_k}}\left(\widehat{X}^{B_k}_{\sigma(i)}\right) \widehat{f}_{X^{C_k}}\left(C_k^{-1}\left[y - B_k \widehat{X}^{B_k}_{\sigma(i)}\right]\right) \widehat{X}_{\sigma_k(i),k}}{\sum_{i=1}^{N} \widehat{f}_{X^{B_k}}\left(\widehat{X}^{B_k}_{\sigma(i)}\right) \widehat{f}_{X^{C_k}}\left(C_k^{-1}\left[y - B_k \widehat{X}^{B_k}_{\sigma(i)}\right]\right)}. \tag{6}$$

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

$$\widehat{\mathbb{E}}(X_1 \,|\, Y = y) = \frac{\sum_{i=1}^{N} \prod_{t=1}^{T} \widehat{f}_{X_{t+1}}\left(y_t - \widehat{X}_{\sigma_1(i),1}\right) \widehat{X}_{\sigma_1(i),1}}{\sum_{i=1}^{N} \prod_{t=1}^{T} \widehat{f}_{X_{t+1}}\left(y_t - \widehat{X}_{i1}\right)} = \frac{\sum_{i=1}^{N} \prod_{t=1}^{T} \widehat{f}_{X_{t+1}}\left(y_t - \widehat{X}_{i1}\right) \widehat{X}_{i1}}{\sum_{i=1}^{N} \prod_{t=1}^{T} \widehat{f}_{X_{t+1}}\left(y_t - \widehat{X}_{i1}\right)}. \tag{7}$$

More generally, the densities $\widehat{f}_{X^{B_k}}$ and $\widehat{f}_{X^{C_k}}$ in (6) are products of marginal densities of individual latent factors.

Remark 3: constrained prediction. In the present setting, an alternative to the usual prediction problem consists in minimizing expected square loss subject to the constraint that the *cross-sectional distribution* of the predicted values coincide with the population distribution of the latent variable. The resulting *constrained* optimal predictor can be estimated as: $\widetilde{X}_{ik} = \widehat{X}_{\pi^*(i),k}$, $i = 1, ..., N$, where the $\widetilde{X}_i$'s are equal to the $\widehat{X}_j$'s sorted in the same order as the $\widehat{\mathbb{E}}(X_k \,|\, Y = Y_i)$'s; that is: $\pi^* = \text{argmin}_{\pi \in \Pi_N} \sum_{i=1}^N \left(\widehat{\mathbb{E}}(X_k \,|\, Y = Y_i) - \widehat{X}_{\pi(i)} \right)^2$.[11] We leave the characterization of the properties of such constrained predictors to future work.

4 Computation

The optimization problems in (2) and (4) are mixed integer quadratic programs. The *relaxed* problem obtained when π is not required to have $\{0, 1\}$ elements is a convex quadratic program. This facilitates the implementation of exact solution algorithms based on branch and cuts and other efficient enumerative techniques. Yet, although the literature on mixed integer programming has made substantial progress in the past decades (e.g., Bliek, Bonami and Lodi, 2014), such exact algorithms are currently limited in the dimensions they can allow for, making their use in empirical applications often impractical. Here we describe a simple heuristic algorithm to minimize (2) and (4).

4.1 Algorithm

The algorithm we propose is based on the observation that, for given $X_1, ..., X_N$ values, problem (4) is a *linear assignment* (or *discrete optimal transport*) problem, hence it can be solved by any linear programming routine. In turn, given π, problem (4) is a simple least squares problem subject to linear restrictions. 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.

[11] In a similar spirit, one can construct a matching-based alternative to $\widehat{\mathbb{E}}(X_k \,|\, Y = Y_i)$ as: $\frac{1}{M} \sum_{j=1}^N \sum_{m=1}^M \mathbf{1}\{\widehat{\pi}^{(m)}(j) = i\} \widehat{X}_{\sigma_k^{(m)}(j),k}$, where $\sigma_k^{(m)}$, $m = 1, ..., M$, are independent random permutations in Π_N, and $\widehat{\pi}^{(m)} = \text{argmin}_{\pi \in \Pi_N} \sum_{i=1}^N \sum_{t=1}^T \left(Y_{\pi(i),t} - \sum_{k=1}^K a_{tk} X_{\sigma_k(i),k} \right)^2$.

Algorithm 1

- *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:*[12]

$$\widehat{\pi}^{(s+1)} = \underset{\pi \in \Pi_N}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\sum_{k=1}^{K} a_{tk} \widehat{X}_{\sigma_k(i),k}^{(s)} \right) Y_{\pi(i),t}. \tag{8}$$

- *(Update step) Compute:*

$$\widehat{X}^{(s+1)} = \underset{X \in \mathcal{X}_N}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{\widehat{\pi}^{(s+1)}(i),t} - \sum_{k=1}^{K} a_{tk} X_{\sigma_k(i),k} \right)^2. \tag{9}$$

Both steps in the algorithm are straightforward to implement. The matching step (8) can be computed by a linear programming routine, due to the fact that the linear programming relaxation of a discrete optimal transport problem has integer-valued solutions.[13] Formally, $\widehat{\pi}^{(s+1)}$ in (8) is a solution to the following *linear program*:

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

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

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

$$(\widehat{X}, \widehat{P}) = \underset{X \in \mathcal{X}_N, P \in \mathcal{P}_\mathcal{N}}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} \left\{ \left(\sum_{k=1}^{K} a_{tk} X_{\sigma_k(i),k} \right)^2 - 2 \left(\sum_{k=1}^{K} a_{tk} X_{\sigma_k(i),k} \right) \left(\sum_{j=1}^{N} P_{j\ell} Y_{jt} \right) \right\}. \tag{10}$$

However, (10) is not convex in general. Our estimation algorithm may be interpreted as a heuristic method to solve this non-convex quadratic program.

[12] Notice that, since π is a permutation, the quadratic term in $\sum_{i=1}^{N} \sum_{t=1}^{T} Y_{\pi(i),t}^2 = \sum_{i=1}^{N} \sum_{t=1}^{T} Y_{it}^2$ does not depend on π.

[13] See for example Chapter 3 in Galichon (2016) for a survey of discrete Monge-Kantorovitch problems, and Conforti, Cornuejols and Zambelli (2014) for an extensive discussion of integer programming problems and perfect formulations.

Once a solution to the algorithm has been reached it is possible to refine it using a local search improvement. 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 in Section 6.

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:

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

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} is not restricted, and neither are the values and increments of the X_{i1}. Formally, the sub-steps of the Mallows algorithm are the following:

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

In Section 6 we will compare the performance of our approach with Mallows' stochastic algorithm on simulated data. Note that the methods introduced in this paper naturally deliver counterparts to the Mallows algorithm for other models beyond nonparametric deconvolution, such as general linear independent factor models. However, consistency properties of the Mallows estimator are currently unknown.

[14]Strictly speaking, Mallows (2007) redefined $\widehat{X}^{(s+1)}_{i1} \equiv \widehat{X}^{(s+1)}_{\sigma^{(s)}(i),1}$ for all $i = 1, ..., N$ at the end of step s, and then applied 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 $\{1, ..., N\}$.

5 Consistency analysis

In this section we provide conditions under which the estimators introduced in Section 3 are consistent. We start with the scalar nonparametric deconvolution model.

5.1 Nonparametric deconvolution

Let us denote the quantile function of X as:

$$F_X^{-1}(\tau) = \inf\{x \in \text{Supp}(X) \,:\, F_X(x) \geq \tau\}, \text{ for all } \tau \in (0,1).$$

Let us define the following two Sobolev sup-norms of a function $H : (0,1) \to \mathbb{R}$:

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

where $\nabla^k H$ denotes the k-th derivative of H (when it exists). We denote $\nabla = \nabla^1$ for the first derivative. The parameter space $\mathcal{H}$ of $H = F_X^{-1}$ is taken to be the $\|\cdot\|_\infty$-closure of the set of continuously differentiable functions H that belong to a $\|\cdot\|$-ball with derivatives bounded from below by a positive constant; see Assumption 1 (ii) below for a formal definition.

To a solution $\widehat{X}_1$ to (2)[15] we will associate an interpolating function $\widehat{H}$ in $\mathcal{H}$ such that $\widehat{H}\left(\frac{i}{N+1}\right) = \widehat{X}_{i1}$ for all i. We are then going to show that $\|\widehat{H} - F_{X_1}^{-1}\|_\infty = o_p(1)$. This result will be obtained as an application of the consistency of sieve extremum estimators (e.g., Chen, 2007).

We make the following assumptions.

Assumption 1

(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.

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

(iii) (Identification) The characteristic function of X_2 does not vanish on the real line.

(iv) (Penalization) $\overline{C}_{N-1} \leq \overline{C}_N < \overline{C} - \underline{C}/(N+1)$ and $\underline{C}_{N-1} \geq \underline{C}_N > \underline{C}$ for all N. Moreover, $\overline{C}_N \xrightarrow{p} \overline{C}$ and $\underline{C}_N \xrightarrow{p} \underline{C}$ as $N \to \infty$.

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

[15] In fact, it is not necessary for $\widehat{X}_1$ to be an exact minimizer of (2). As shown in the proof, it suffices that the value of the objective function at $\widehat{X}_1$ be in an ϵ_N-neighborhood of the global minimum, for ϵ_N tending to zero as N tends to infinity.

Though convenient for the derivations, the compact supports assumption (i) is strong, and so is the sup-norm definition of the parameter space in (ii). In particular, both (i) and (ii) restrict the tail behavior of $F_{X_1}^{-1}$ and its derivative. The conditions could be weakened by working with weighted norms, at the cost of achieving a weaker consistency result. (ii) ensures that $\mathcal{H}$ is compact with respect to $\|\cdot\|_\infty$. This type of construction was pioneered by Gallant and Nychka (1987). Compactness can be preserved when sup norms are replaced by weighted Sobolev sup-norms (e.g., using polynomial or exponential weights); see for example Theorem 7 in Freyberger and Masten (2015). The simulation experiments reported below suggest that the estimator continues to perform well when supports are unbounded.

(iii) is an identification condition which is commonly assumed in nonparametric deconvolution. It can be relaxed to some extent by allowing for the presence of isolated zeros (Evdokimov and White, 2012). The constants $\underline{C}_N$ and $\overline{C}_N$ appearing in (iv) ensure that the $\widehat{X}_{i1}$ values are bounded and of bounded variation. We will document the sensitivity of our estimator to those constants in the simulation section.

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

Theorem 1 *Let Assumption 1 hold. Then, as N tends to infinity:*

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

5.2 Nonparametric factor models

Consider next the linear independent factor model $Y = AX$, where $X = (X_1, ..., X_K)'$, with the X_k's mutually independent, and A is a known $T \times K$ matrix with generic element a_{tk}. We make the following assumptions.

Assumption 2

(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.

(ii) (Parameter space) $\mathcal{H}_K = \left\{(H_1, ..., H_K) : H_k \in \mathcal{H}$ and $\sum_{i=1}^N H_k\left(\frac{i}{N+1}\right) = 0$ for all $k\right\}$, where $\mathcal{H}$ is defined in Assumption 1 (ii).

(iii) (Identification) The characteristic function of X_k does not vanish on the real line for all k, and the vectors $\operatorname{vec} A_k A_k'$, *$k = 1, ..., K$, are linearly independent.*

(*iv*) *(Penalization) As in Assumption 1* (*iv*).

(*v*) *(Sampling)* $(Y_{i1}, ..., Y_{iT})$ *are i.i.d.*

These conditions are similar to the scalar deconvolution case. (*iii*) is a sufficient condition for the distributions of latent variables X_k to be nonparametrically identified (e.g., Székely and Rao, 2000, Bonhomme and Robin, 2010). We have the following consistency result.

Theorem 2 *Let Assumption 2 hold. Then, as N tends to infinity:*

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

Densities and expectations. Under slightly stronger assumptions, Theorem 2 can be strengthened to obtain consistent estimators of both $F^{-1}_{X_k}$ and its derivative. This will deliver a smoother estimator of $F^{-1}_{X_k}$. The estimators of $F^{-1}_{X_k}$ and its derivative can then be used for density estimation. To see this, let us denote as $\mathcal{X}^{(2)}_N$ 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. Likewise, denote as $\mathcal{H}^{(2)}_K$ the set of functions $(H_1, ..., H_K) \in \mathcal{H}_K$ which additionally satisfy $|\nabla^2 H_k| \leq \overline{C}$ for all k. Modifying Assumption 2 to accommodate these two differences, and modifying the proof of Theorem 2 accordingly, we obtain that:

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

We then have the following result.[16,17]

Corollary 1 *Let b in (5) be such that* $b \to 0$ *and* $Nb \to \infty$ *as N tends to infinity. Let* κ *be a Lipschitz kernel which integrates to one and has finite first moments. Then, under the modifications of Assumption 2 described in the previous paragraph, we have:*

$$\sup_{x\in\mathbb{R}} \left|\widehat{f}_{X_k}(x) - f_{X_k}(x)\right| = o_p(1), \qquad \textit{for all } k = 1, ..., K. \tag{11}$$

Given Corollary 1, it can readily be checked that conditional expectations estimators given by (6) and (7) are consistent in sup norm for their population counterparts.

[16] Consistency also holds when a uniform kernel is used, although the proof is omitted for brevity.

[17] An alternative density estimator, which can be shown to be uniformly consistent for f_{X_k} under the same conditions, is: $\widetilde{f}_{X_k}(x) = 1/\nabla\widehat{H}_k(\widehat{H}_k^{-1}(x))$.

6 Numerical experiments

In this section we illustrate the finite-sample performance of our estimator on simulated data. We consider two models in turn: the scalar nonparametric deconvolution model, and the fixed-effects model.

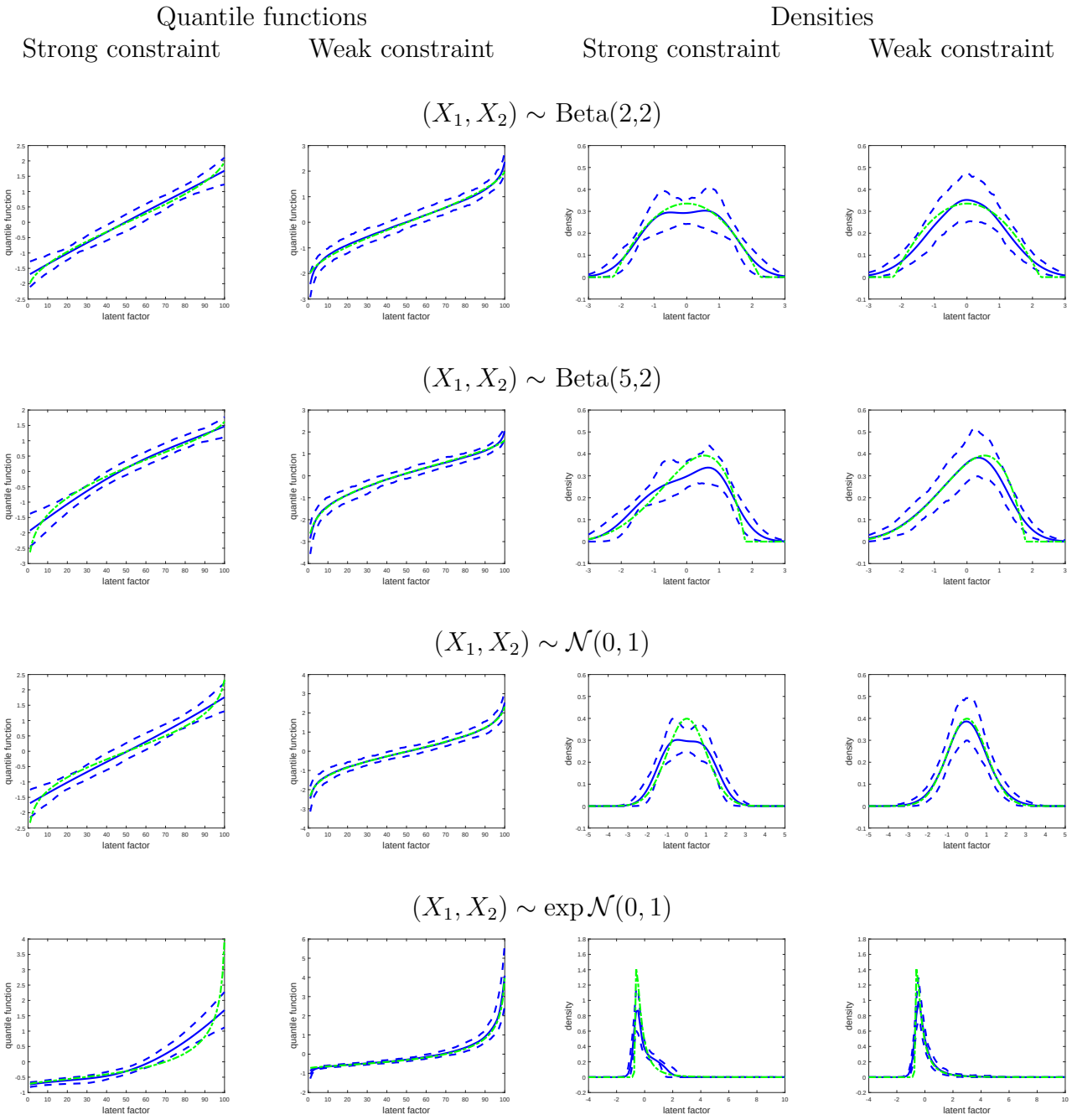
6.1 Nonparametric deconvolution

We start with the deconvolution model $Y = X_1 + X_2$, where X_1 and X_2 are scalar, independent, and follow identical distributions. We consider four specifications: Beta$(2, 2)$, Beta$(5, 2)$, normal, and log-normal, all standardized so that X_1 and X_2 have mean zero and variance one. To restrict the maximum values of $\widehat{X}_{i1}$, its increments and its second-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 (2) 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 $i = 1, ..., N$.

The first two columns in Figure 2 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 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 $N = 100$. Even for such a small sample size, our nonparametric estimator performs well, although there is some evidence of bias when imposing a stronger constraint on the parameters (first column). Estimates under weak constraint are virtually unbiased and quite precise. On 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 larger biases in the strong constraint case the results reproduce the shape of the unknown densities rather well.

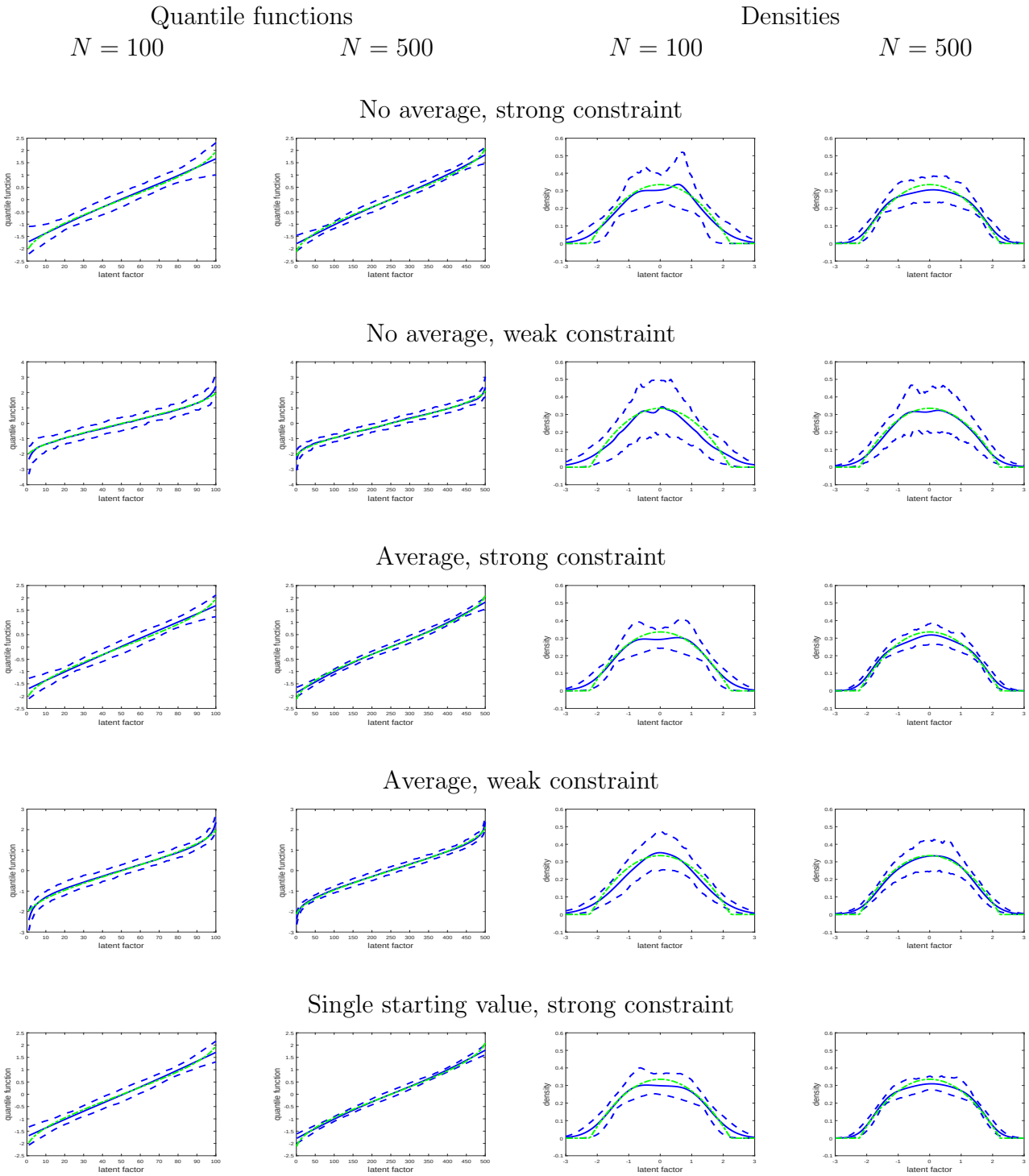
In Figure 3 we report additional results for the Beta$(2, 2)$ specification, for $N = 100$ (columns 1 and 3) and $N = 500$ (columns 2 and 4). In the first two rows we report the results based on a single σ draw per estimate (i.e., $M = 1$), whereas in the next two rows we show the results for the estimator averaged over $M = 10$ different σ draws. While we see that averaging seems to slightly increase the precision of estimated quantile functions and

Figure 2: Monte Carlo results, deconvolution model, $N = 100$



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

Figure 3: Monte Carlo results, deconvolution model, Beta(2,2), $N = 100, 500$



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

densities, the results based on one σ draw are comparable to the ones based on 10 draws. In the last row of Figure 3 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 very little affected, suggesting that the impact of starting values on the performance of the estimator is moderate.

In Table 1 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 1 we report an "implied 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 $N^{-\frac{7}{10}}$. 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, characteristic functions converge at the quadratic rate, so the corresponding optimal rate is $N^{-\frac{4}{7}}$.

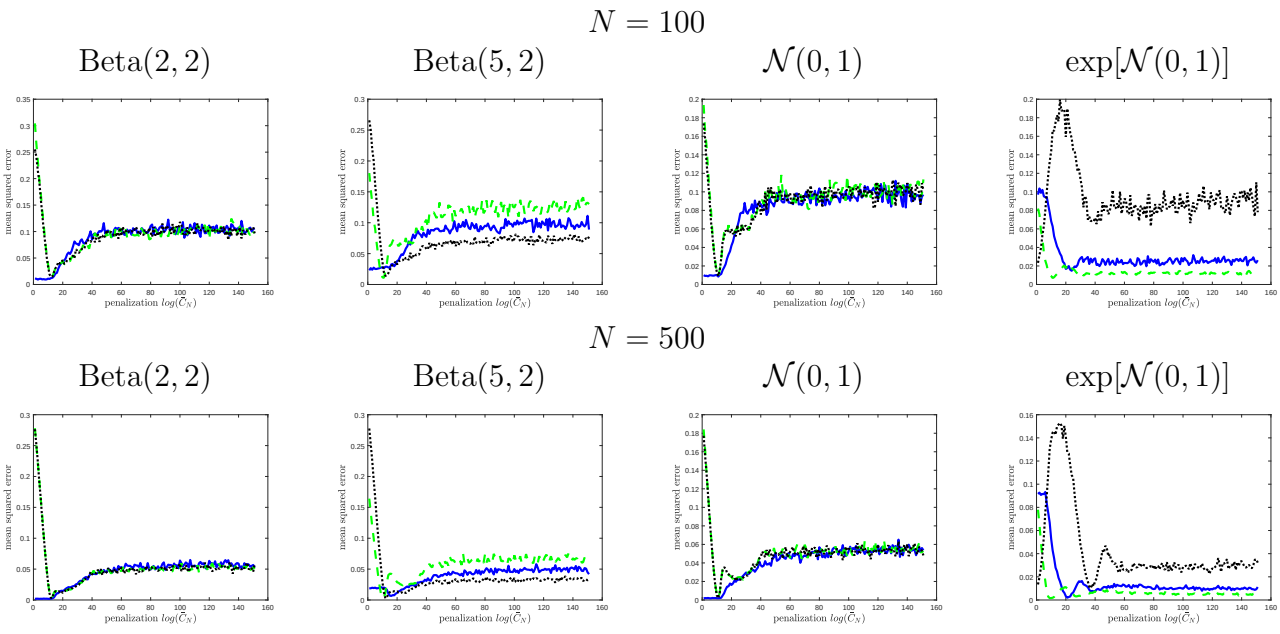
Next, we assess the impact of the penalization parameters $\overline{C}_N$ and $\underline{C}_N$ on the mean squared error of quantile estimates, at the median and 25% and 75% percentiles. In Figure 4 we show the results for the four specifications, when varying the logarithm of $\overline{C}_N$ between 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 usual regularization methods for ill-posed inverse problems such as Tikhonov regularization or spectral cut-off (e.g., Carrasco, Florens and Renault, 2007), for which decreasing the amount of penalization typically causes large increases in variance. The high 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 optimal choice of the penalization parameters, which would be very interesting to study

Table 1: 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 4: Monte Carlo simulation, mean squared error of estimated quantiles of X_1 as a function of the penalization parameter

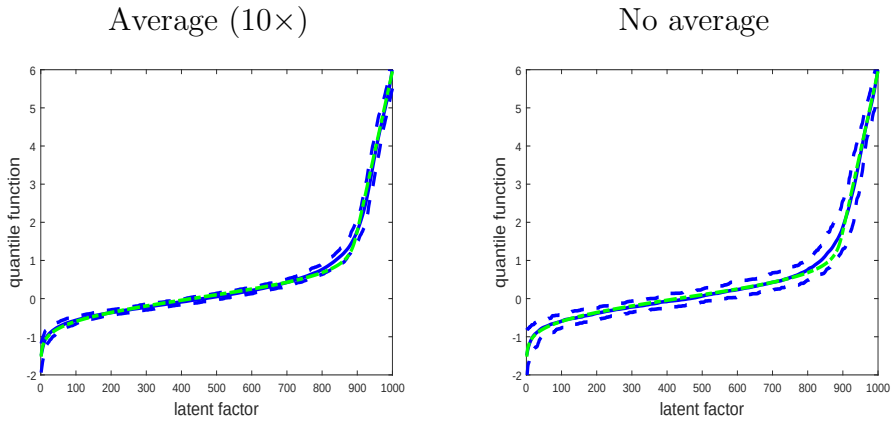


Notes: Simulated data from the deconvolution model $Y = X_1 + X_2$. *Log of penalization* $\overline{C}_N$ *(x-axis) against mean squared error (y-axis).* $\underline{C}_N$ *is set to* $\overline{C}_N^{-1}$. *Solid corresponds to the median, dashed to the 25% quantile, dotted to the 75% quantile. No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

in future work.

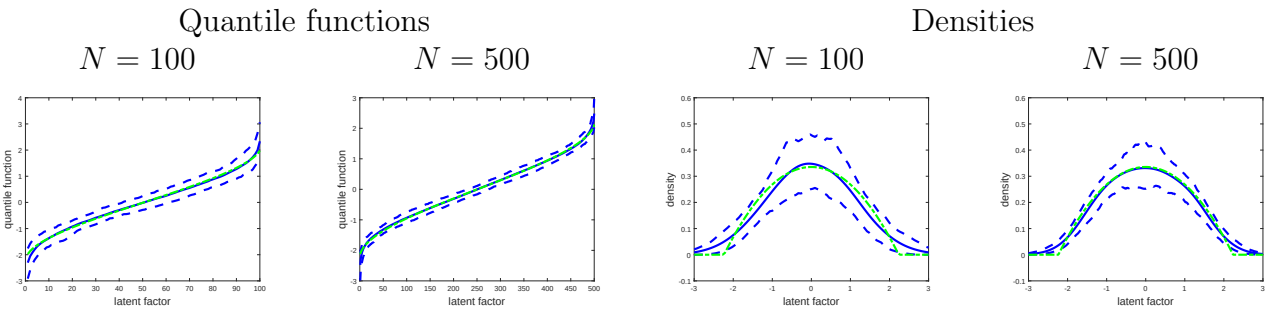
We next consider a data generating process (DGP) which has been previously used to assess the finite-sample behavior of several estimators in the nonparametric deconvolution model. This DGP was used in Koenker (2016), and it is a slight variation of a DGP introduced by Efron (2016). Let $Y = X_1 + X_2$, 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 reports that the Stefanski and Carroll (1990) characteristic-function based estimator does quite poorly on this DGP, distribution functions estimated on a sample of 1000 observations showing wide oscillations. In Figure 5 we apply our estimator to this DGP, and report the results of 100 simulations. On the left graph we show quantile function estimates averaged 10 times, whereas on 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

Figure 5: Monte Carlo results, deconvolution model, Efron-Koenker specification, $N = 1000$



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

Figure 6: Monte Carlo results, deconvolution model, Beta(2,2), Mallows' (2007) algorithm



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

Table 2: Monte Carlo simulation, mean integrated squared and absolute errors of density estimators in the fixed-effects model, results for X_1

MISE	MIAE	MISE	MIAE	MISE	MIAE
$(X_1, X_2, X_3) \sim$ Beta(2,2)					
Strong constraint		Weak constraint		Fourier	
0.0036	0.0654	0.0035	0.0631	0.0123	0.2274
$(X_1, X_2, X_3) \sim$ Beta(5,2)					
Strong constraint		Weak constraint		Fourier	
0.0050	0.0750	0.0042	0.0677	0.0249	0.2979
$(X_1, X_2, X_3) \sim \mathcal{N}(0,1)$					
Strong constraint		Weak constraint		Fourier	
0.0056	0.0796	0.0040	0.0674	0.0122	0.2372
$(X_1, X_2, X_3) \sim \exp[\mathcal{N}(0,1)]$					
Strong constraint		Weak constraint		Fourier	
0.1003	0.2415	0.0536	0.1492	0.3344	0.8613

Notes: Mean integrated squared and absolute errors across 100 simulations from the fixed-effects model. $N = 100$, $T = 2$. "Fourier" is the characteristic-function based estimator of Bonhomme and Robin (2010). Results for the first factor X_1.

of the parametric estimator analyzed in Efron (2016).

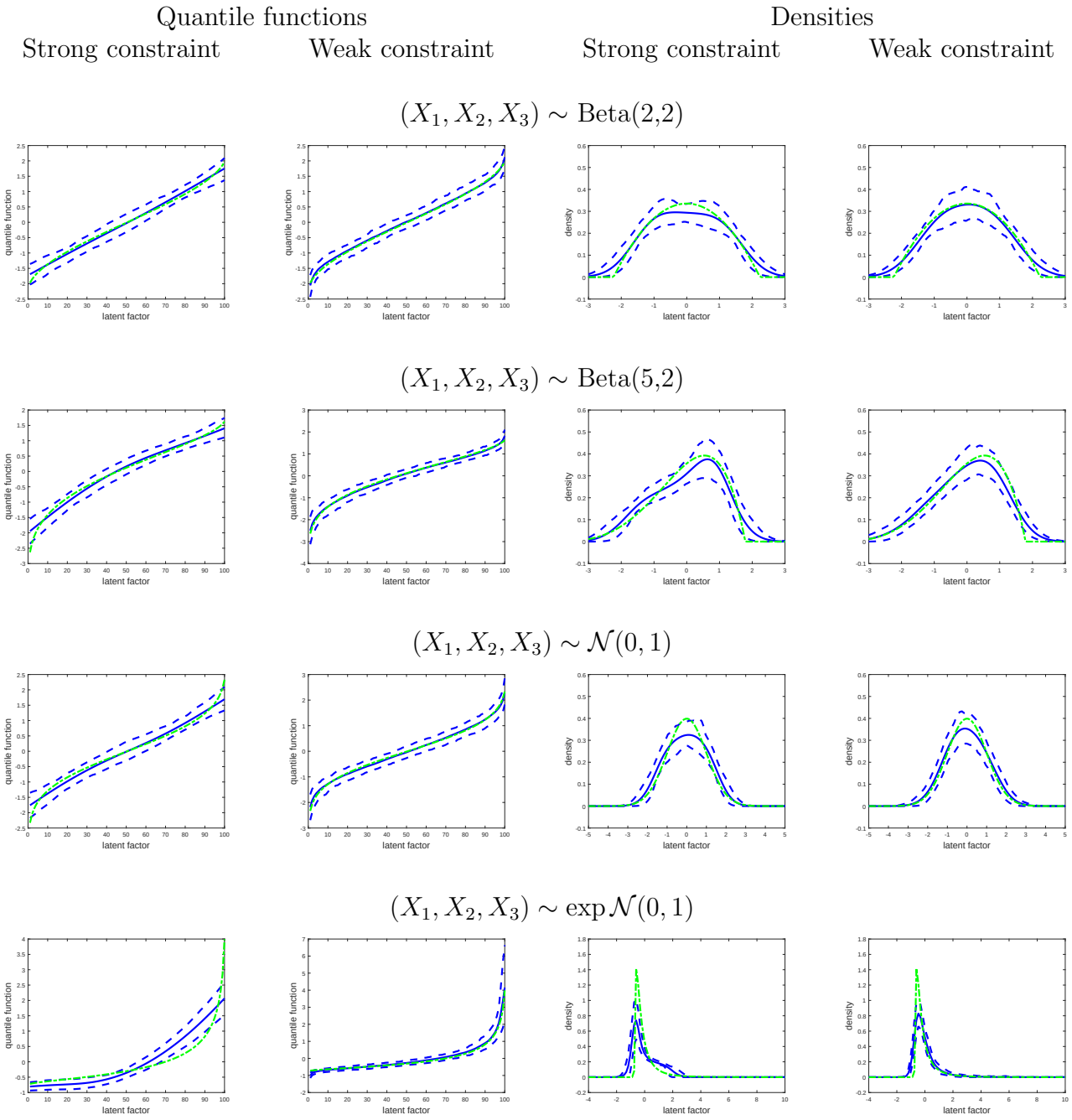
Lastly, in Figure 6 we report simulation results for Mallows' (2007) stochastic estimator, in the case of the Beta$(2, 2)$ specification. As we pointed out in Section 4, this algorithm is closely related to ours but it differs from it since new random permutations are re-drawn in every step. We draw 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 with Figure 3.

6.2 Fixed-effects model

We next turn to the fixed-effects model $Y_1 = X_1 + X_2$, $Y_2 = X_1 + X_3$, where X_1, X_2, X_3 are independent of each other and have identical distributions. As before we consider four specifications for the distribution of X_k. In Figure 7 we report the estimates of the quantile function and density of the first factor X_1. Results corresponding to X_2 and X_3 are similar, and can be found in Appendix B. The sample size is $N = 100$. The estimates are comparable to the ones in Figure 2.

In Table 2 we report the mean integrated squared and absolute errors (MISE and MIAE,

Figure 7: Monte Carlo results for X_1 in the fixed-effects model, $N = 100$, $T = 2$



Notes: Simulated data from the fixed-effects model, results for the first factor X_1. Solid is the mean across simulations, dashed are 10 and 90 percent pointwise quantiles, and dashed-dotted is the true quantile function or density. 100 *simulations. 10 averages over permutation draws.*

respectively) of our density estimators, for the four distributional specifications and $N = 100$. We see that the version of the estimator under weak constraint performs better. Moreover, interestingly, as shown by the last two columns of Table 2 our estimator outperforms characteristic-function based density estimators.[18] Inspection of the estimates suggests that the differences are mainly driven by estimates of the tails of the densities. Unlike our estimator, characteristic-function based ones do not guarantee that estimated densities be non-negative, and their values tend to oscillate in the left and right tails.

7 Empirical illustrations

In this section we present two illustrations of our method. We first estimate distributions of earnings shocks, based on a subsample from the PSID for the years 1978 to 1987 taken from Bonhomme and Robin (2010). We then estimate distributions of school quality, based on data from Madrid for 2005-2015.

7.1 Permanent-transitory earnings dynamics

Following Bonhomme and Robin (2010) we estimate a simple permanent-transitory model where log-earnings, net of the effect of some covariates, is the sum of a random walk η_{it} and an independent innovation ε_{it}. In first differences we have, denoting log-earnings growth as $\Delta Y_{it} = Y_{it} - Y_{i,t-1}$:

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

which is a linear factor model with $2T - 1$ independent factors.[19]

We use the same sample selection as in Bonhomme and Robin, focusing on a balanced panel of 624 employed male workers. Log-earnings growth ΔY_{it} is net of education, race,

[18] The results in the "Fourier" column are based on the characteristic-function based generalized deconvolution estimator of Bonhomme and Robin (2010). We use their recommended choice to set the regularization parameter in each replication.

[19] Indeed we have:

$$\underbrace{\begin{pmatrix} \Delta Y_1 \\ \Delta Y_2 \\ \Delta Y_3 \\ ... \\ \Delta Y_T \end{pmatrix}}_{=Y} = \underbrace{\begin{pmatrix} 1 & 0 & ... & 0 & 1 & 0 & ... & 0 \\ 0 & 1 & ... & 0 & -1 & 1 & ... & 0 \\ 0 & 0 & ... & 0 & 0 & -1 & ... & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & ... & 1 & 0 & 0 & ... & -1 \end{pmatrix}}_{=A} \underbrace{\begin{pmatrix} v_1 - \varepsilon_0 \\ v_2 \\ ... \\ v_T + \varepsilon_T \\ \varepsilon_1 \\ \varepsilon_2 \\ ... \\ \varepsilon_{T-1} \end{pmatrix}}_{=X}.$$

geographic and year dummies, and a quadratic polynomial in age. We estimate the quantile functions of permanent shocks v_{it} and transitory shocks ε_{it} for different years t using our matching estimator.

In Figures 8 and 9 we show the estimated quantile functions of permanent and transitory shocks, respectively. We report average estimates based on $M = 10$ σ draws, and use 10 different starting values in the algorithm. The estimates in the graphs are based on $(\underline{C}_N, \overline{C}_N) = (.1, 10)$ (strong constraint). The dotted line shows a fitted Gaussian quantile function. In dotted lines we show 10%-90% bootstrap confidence bands.[20] In Figures B3 and B4 in Appendix B we compare the estimated quantile functions under strong and weak constraints. We see that both permanent and transitory shocks are far from being normally distributed. This confirms the findings of strong non-Gaussianity found in Horowitz and Markatou (1996), Geweke and Keane (2000), and Bonhomme and Robin (2010), among others.

Next, in Figures 10 and 11 we show density estimates for permanent and transitory shocks. The results obtained under a stronger penalization (strong constraint) are shown in solid lines, whereas the results under a weaker penalization are in dashed lines. Density estimates confirm the evidence of non-Gaussianity and suggest the presence of excess kurtosis in permanent and transitory shocks. Moreover, while the effect of the penalization on the density estimates is stronger in the tails, it does not affect much their central parts.

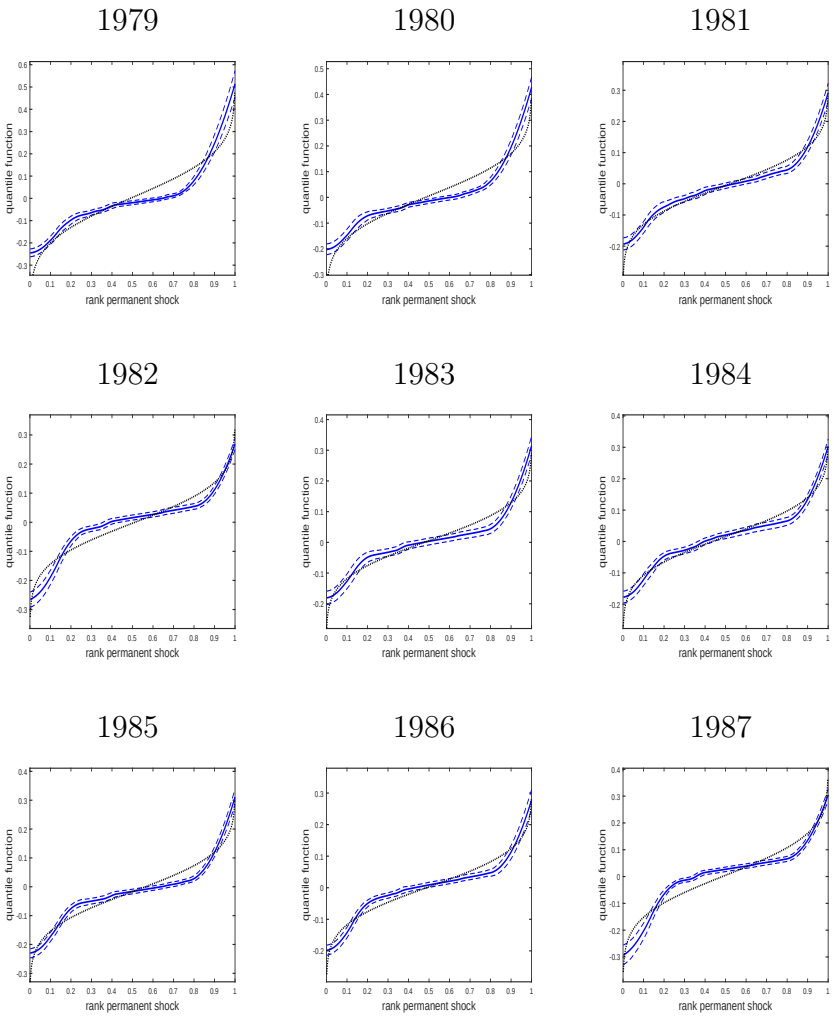
Lastly, in Figures B5 and B6 in Appendix B we show how the model fits distributions of log-earnings growth $Y_{it} - Y_{i,t-s}$ at various horizons s, and the distribution of year-to-year growth $Y_{it} - Y_{i,t-1}$ for different years. We simulate the model 200 times for every individual in the sample and report the resulting measures of fit. We see that the model produces a good fit both at different horizons and over time.

7.2 School effects in the Spanish region of Madrid

In this second illustration we focus on a standardized exam administered each year in all primary schools from the Madrid area to sixth-grade students. This exam, called the CDI ("prueba de Conocimientos y Destrezas Indispensables") is compulsory for all schools. It has no academic consequences for students (Anghel, Cabrales and Carro, 2016). We focus on average maths test scores in every school in Madrid capital, which represents 481 schools

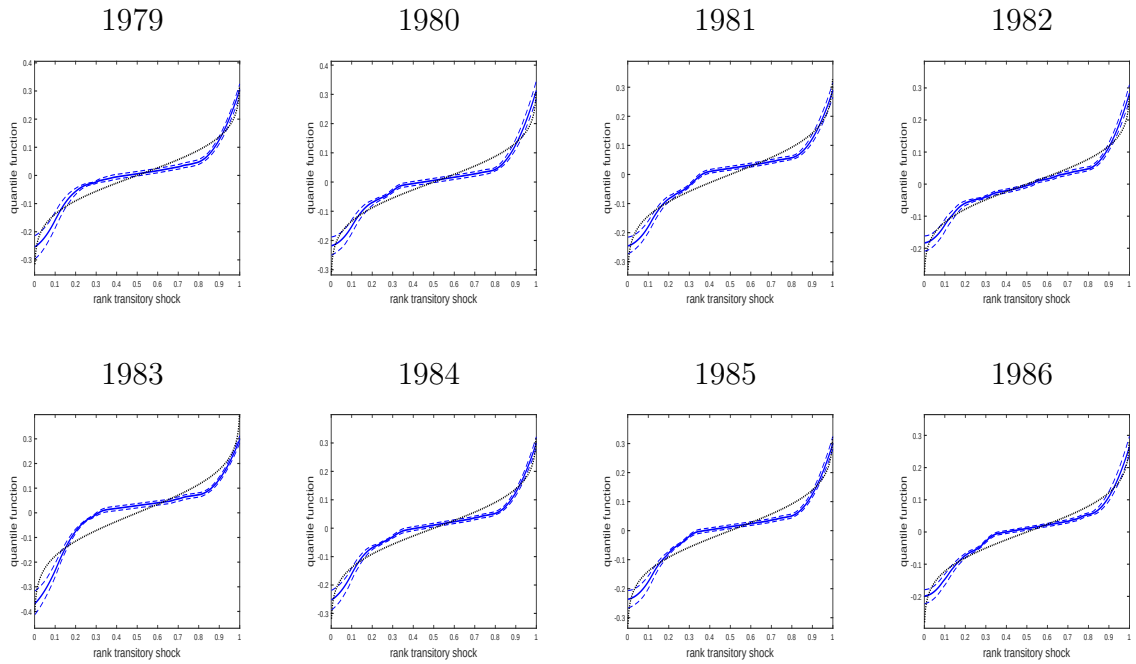
[20] Note that our theory does not provide asymptotic guarantees on the validity of the bootstrap.

Figure 8: Estimated quantile functions of permanent shocks in different years



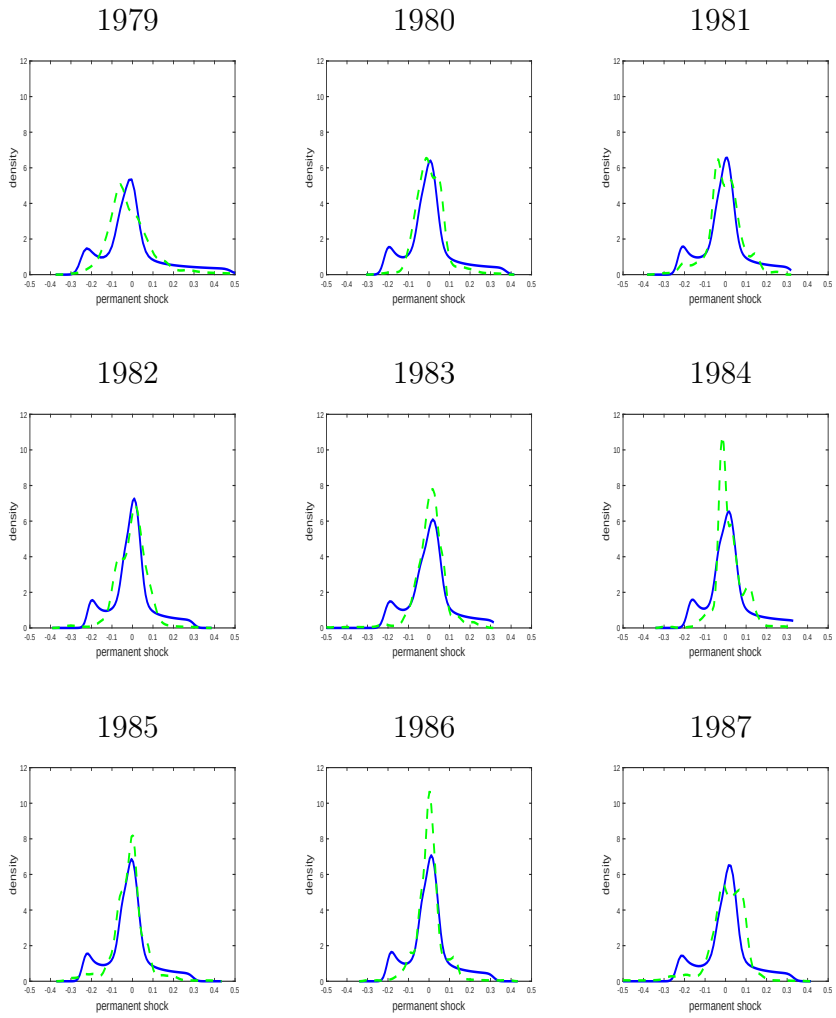
Notes: PSID, 1978-1987. Permanent shock in every year (note: the first and last years are a combination of permanent and transitory shocks). Sample selection and construction of log-earnings growth residuals as in Bonhomme and Robin (2010). Model estimation: strong constraint, 10 averages over permutation draws. Point estimates in solid, 10 and 90 pointwise bootstrap confidence bands in dashed (100 replications), normal quantile function in dotted.

Figure 9: Estimated quantile functions of transitory shocks in different years



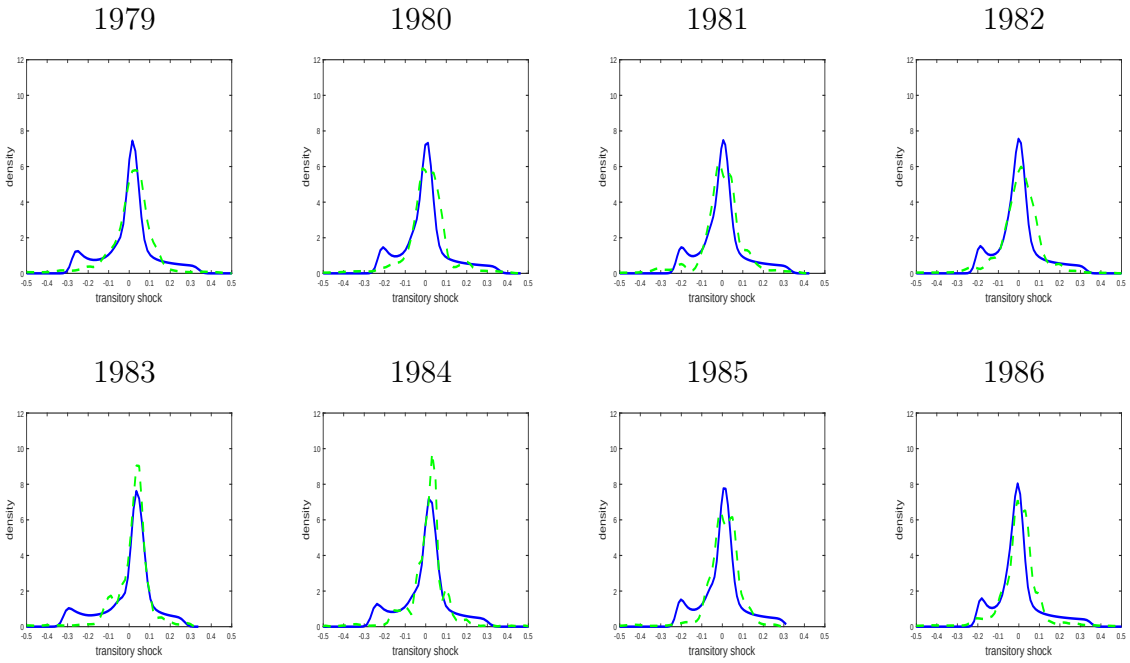
Notes: PSID, 1978-1987. Transitory shock in every year. Sample selection and construction of log-earnings growth residuals as in Bonhomme and Robin (2010). Model estimation: strong constraint, 10 averages over permutation draws. Point estimates in solid, 10 and 90 pointwise bootstrap confidence bands in dashed (100 replications), normal quantile function in dotted.

Figure 10: Estimated density functions of permanent shocks in different years, weak constraints (dashed) and strong constraints (solid)



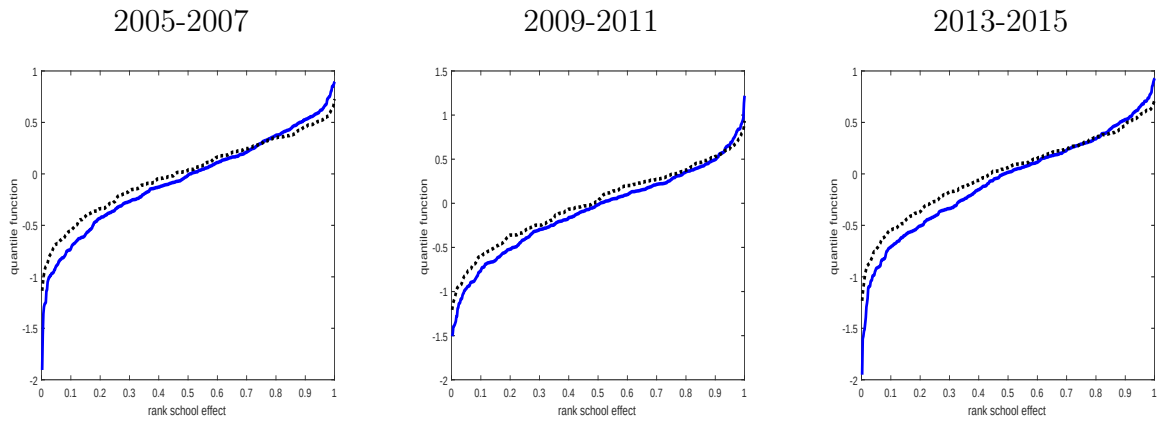
Notes: PSID, 1978-1987. Permanent shock in every year (note: the first and last years are a combination of permanent and transitory shocks). Sample selection and construction of log-earnings growth residuals as in Bonhomme and Robin (2010). Model estimation: strong (solid line) and weak (dashed line) constraint, 10 averages over permutation draws.

Figure 11: Estimated density functions of transitory shocks in different years, weak constraints (dashed) and strong constraints (solid)



Notes: PSID, 1978-1987. Transitory shock in every year. Sample selection and construction of log-earnings growth residuals as in Bonhomme and Robin (2010). Model estimation: strong (solid line) and weak (dashed line) constraint, 10 averages over permutation draws.

Figure 12: Quantile functions of school effects



Notes: Administrative data from the Spanish region of Madrid. Solid is the quantile function of time-averaged test scores $\overline{Y}_i$, dashed is the estimated quantile function of η_i.

from 2005 to 2015. Here we focus on overall school performance, without attempting to separate student composition from the causal effect of the school.

The covariance structure of school averages of test scores (given in Table B1 in Appendix B) suggests that a fixed-effects model with uncorrelated additive errors provides a good approximation. We find there is substantial year-to-year fluctuations in test scores. Estimating a simple stationary covariance structure on school averages gives a variance of the fixed effect of .166, and a variance of the transitory shock of .141. In other words, close to half of the variation in test scores across schools is of a transitory nature.

We then estimate the fixed-effects model:

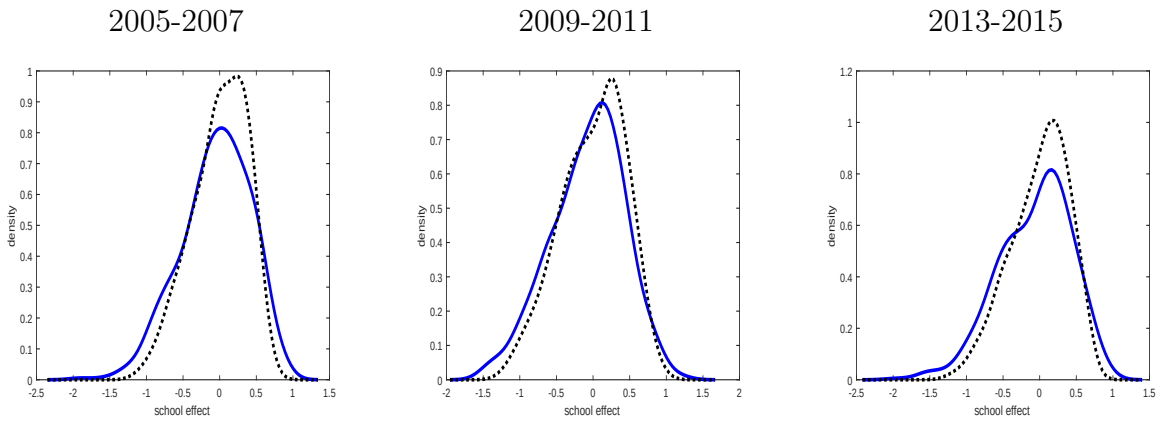
$$Y_{it} = \eta_i + \varepsilon_{it},$$

where Y_{it} is the average CDI test score in mathematics in school i in year t, and $\eta_i, \varepsilon_{i1}, ..., \varepsilon_{iT}$ are modeled as independent of each other with unspecified distributional forms. By estimating the distribution of η_i we aim to provide a "nonparametric shrinkage" of the time averages $\overline{Y}_i = (1/T)\sum_{t=1}^{T} Y_{it}$, so as to document the distribution of permanent school quality, net of transitory fluctuations.

We estimate the fixed-effects model on three pairs of years: $2005-2007$, $2009-2011$, and $2013-2015$. In Figures 12 and 13 we report the estimated quantile functions and densities of η_i (in dashed lines) together with the quantile functions and densities of school test scores averaged over each of the three periods $\overline{Y}_i$ (in solid lines).[21] We see that, compared to

[21] In Figure B7 in Appendix B we show the fit of the model for marginal quantile functions of school-

Figure 13: Densities of school effects



Notes: Administrative data from the Spanish region of Madrid. Solid is the density of time-averaged test scores $\overline{Y}_i$, dashed is the estimated density of η_i.

the distributions of $\overline{Y}_i$, the distributions of permanent school quality η_i are less dispersed, consistently with the idea that η_i's are net of transitory variation. At the same time, the noise correction is asymmetric, making a larger difference at the bottom of the distribution than at the top. Such asymmetric effects could not be captured using conventional Gaussian empirical Bayes methods.

8 Extension: finite mixture models

In this section we outline an extension of our matching approach to the following finite mixture model with G groups, for a T-dimensional outcome Y:

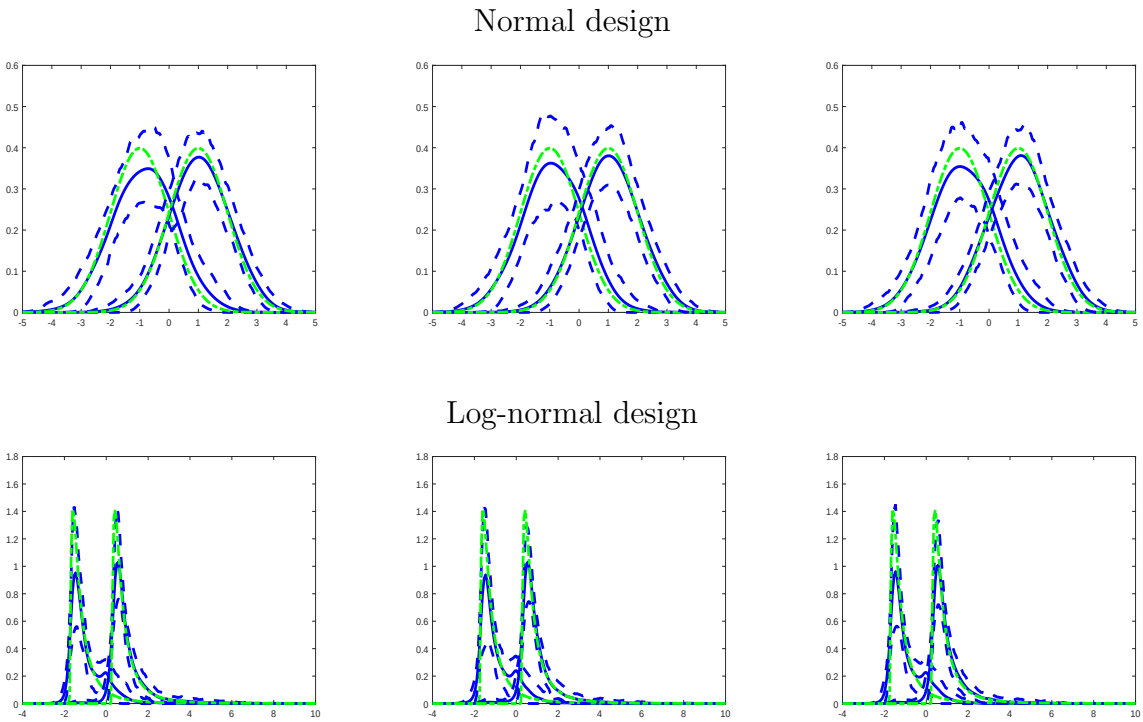
$$Y_t = \sum_{g=1}^{K} Z_g X_{gt}, \quad g = 1, ..., G, \tag{12}$$

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 all mutually independent. The nonparametric version of model (12) has been extensively analyzed in the literature (e.g., Hall and Zhou, 2003, Hu, 2008, Allmann, Matias and Rhodes, 2009, Bonhomme, Jochmans and Robin, 2016a, 2016b).

To construct a matching estimator in model (12) we first note that, by the threshold crossing representation, there exist a parameter vector $\mu = (\mu_1, ..., \mu_{G-1})$ and a standard

averaged test scores for 2005, 2009 and 2013. The model produces an excellent fit in this dimension.

Figure 14: Monte Carlo results, finite mixture model, $G = 2$



Notes: Simulated data from a finite mixture model with $G = 2$ components. Solid is the mean across simulations, dashed are 10 and 90 percent pointwise quantiles, and dashed-dotted is the true density. 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.

uniform random variable V such that $Z_g = Z_g(V, \mu)$, where $Z_1(V, \mu) = 1$ if and only if $V \leq \mu_1$, $Z_g(V, \mu) = 1$ if and only if $\mu_{g-1} < V \leq \mu_g$ for $g = 2, ..., G-1$, and $Z_G(V, \mu) = 1$ if and only if $\mu_{G-1} < V$. We denote as $\mathcal{M}_{G-1}$ 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 then define the following estimator:

$$(\widehat{X}, \widehat{\mu}) = \underset{X \in \mathcal{X}_N, \mu \in \mathcal{M}_{G-1}}{\operatorname{argmin}} \left\{ \min_{\pi \in \Pi_N} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{\pi(i),t} - \sum_{g=1}^{G} Z_g(V_i, \mu) X_{\sigma_{gt}(i),gt} \right)^2 \right\}, \tag{13}$$

where $V_1, ..., V_N$ are standard uniform draws, and σ_{gt} are random permutations in Π_N for all $g = 1, ..., G$, $t = 1, ..., T$.

For given μ, we propose to use an algorithm analogous to the one described in Section 4 to compute $\widehat{X}$. The outer minimization with respect to μ can be done using simulated annealing or other methods to minimize non-differentiable objective functions. When G is

small (as in the simulation exercise below) grid search is a viable option. Consistency of the estimator can be studied using the same techniques as in Section 5.

In Figure 14 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 the components are ordered by increasing means.[22] The results are encouraging, and suggest that matching estimators can perform well in nonparametric finite mixture models too.

9 Conclusion

In this paper we have proposed an approach to nonparametrically estimate models with 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 excellent performance of our nonparametric estimator in small samples, in particular compared to characteristic-function based estimators. Substantial progress on computation might be possible by leveraging recent advances on regularized optimal transport (e.g., Cuturi, 2013). An important question for future work will be to characterize rates of convergence and asymptotically valid confidence sets for our estimator.

Lastly, although we have focused on linear independent factor models, our approach could be generalized to other nonparametric or semiparametric models with latent variables such as finite mixture models, which we have briefly analyzed, and more generally finite mixtures of linear independent factor models (or "mixtures of factor analyzers"); see Ghahramani and Hinton (1997) and McLachlan, Peel and Bean (2003). Our approach can also be extended to estimate linear random coefficients models, such as:

$$Y = X_1 + \sum_{k=2}^{K} W_k X_k, \tag{14}$$

where $(W_2, ..., W_K)$ is independent of $(X_1, ..., X_K)$, the scalar outcome Y and the covariates $W_2, ..., W_K$ are observed, and $X_1, ..., X_K$ are latent.[23] To construct a matching estimator

[22] We use a version of (13) with multiple draws $\sigma_{gt}(i, r)$ for all i, with $R = 10$ simulations by observation. We use 3 starting values in every inner loop, and perform an outer loop for 10 equidistant values of the first group's probability.

[23] Model (14) has been extensively studied. See for example Beran and Hall (1992), Beran and Millar (1994), Beran, Feuerverger and Hall (1996), and Hoderlein, Klemela and Mammen (2010).

in this case, we augment (14) 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 parameters (that is, the joint distributions of $(X_1, ..., X_K)$ and $(V_2, ..., V_K)$) are estimated by minimizing the Euclidean distance between the model's predictions of Y, W observations, and their matched values in the data. A similar approach could be used in binary choice models with random coefficients.[24]

[24]See Ichimura and Thompson (1998) and Gautier and Kitamura (2013).

References

[1] Allman, E. S., C. Matias, and J. A. Rhodes (2009): "Identifiability of Parameters in Latent Structure Models with Many Observed Variables," *Annals of Statistics*, 3099–3132.

[2] Anghel, B., A. Cabrales, A., and J. M. Carro (2016): "Evaluating a Bilingual Education Program in Spain: The Impact Beyond Foreign Language Learning," *Economic Inquiry*, 54(2), 1202–1223.

[3] Arellano, M., R. Blundell, and S. Bonhomme (2017): "Earnings and Consumption Dynamics: A Nonlinear Panel data Framework," *Econometrica*, 85(3), 693–734.

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

[5] Arellano, M., and S. Bonhomme (2016): "Nonlinear Panel Data Estimation via Quantile Regressions," *Econometrics Journal*, 19, C61-C94.

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

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

[8] Beran, R., and P. W. Millar (1994): "Minimum Distance Estimation in Random Coefficient Regression Models," *Annals of Statistics*, 1976–1992.

[9] 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.

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

[11] Bonhomme, S., K. Jochmans, and J.M. Robin (2016a): "Nonparametric Estimation of Finite Mixtures from Repeated Measurements," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1), 211–229.

[12] Bonhomme, S., K. Jochmans, and J.M. Robin (2016b): "Estimating Multivariate Latent-Structure Models," *Annals of Statistics*, 44(2), 540–563.

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

[14] 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.

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

[16] Carneiro, P., K. T. Hansen, and J. J. Heckman (2003): "Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice," *International Economic Review*, 44(2), 361–422.

[17] Carrasco, M., and J.P. Florens (2011): "Spectral Method for Deconvolving a Density," *Econometric Theory*.

[18] Carrasco, M., J.P. Florens, and E. Renault (2007): "Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization," *Handbook of Econometrics*, vol. 6, 5633–5751.

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

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

[21] Chen, X. (2007): "Sieve Methods in Econometrics," *Handbook of Econometrics*.

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

[23] Chernozhukov, V., A. Galichon, M. Hallin, and M. Henry (2017): "Monge?Kantorovich Depth, Quantiles, Ranks and Signs," *Annals of Statistics*, 45(1), 223–256.

[24] Conforti, M., G. Cornuéjols, and G. Zambelli (2014): *Integer programming*. Vol. 271. Berlin: Springer.

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

[26] Cunha, F., J. J. Heckman, and S. M. Schennach (2010): "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica*, 78(3), 883–931.

[27] Cuturi, M. (2013): "Sinkhorn Distances: Lightspeed Computation of Optimal Transport," in *Adv. in Neural Information Processing Systems*, 2292–2300.

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

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

[30] Efron, B., and T. Hastie (2016): *Computer Age Statistical Inference.* Vol. 5. Cambridge University Press.

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

[34] Fournier, N., and A. Guillin (2015): "On the Rate of Convergence in Wasserstein Distance of the Empirical Measure," *Probability Theory and Related Fields*, 162(3–4), 707–738.

[35] Freyberger, J., and M. Masten (2015): "Compactness of Infinite Dimensional Parameter Spaces," Cemmap working paper No. CWP01/16.

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

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

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

[39] Gautier, E., and Y. Kitamura (2013): "Nonparametric Estimation in Random Coefficients Binary Choice Models," *Econometrica*, 81(2), 581–607.

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

[41] 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.

[42] Ghahramani, Z., and G. E. Hinton (1996): "The EM Algorithm for Mixtures of Factor Analyzers," Vol. 60, Technical Report CRG-TR-96-1, University of Toronto.

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

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

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

[46] 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.

[47] 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.

[48] Hoderlein, S., J. Klemelä, and E. Mammen (2010): "Reconsidering the Random Coefficient Model," *Econometric Theory*, 26(3), 804–837.

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

[50] Hu, Y. (2008): "Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables: A General Solution," *Journal of Econometrics*, 144(1), 27–61.

[51] Ichimura, H., and T. S. Thompson (1998): "Maximum Likelihood Estimation of a Binary Choice Model with Random Coefficients of Unknown Distribution," Journal of Econometrics, 86(2), 269–295.

[52] Kasahara, H., and K. Shimotsu (2009): "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices," *Econometrica*, 77(1), 135–175.

[53] Knuth, D. E. (1997): *The Art of Computer Programming*, third edition, Volume 2: Seminumerical Algorithms, Addison-Wesley.

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

[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] McLachlan, G. J., D. Peel, and R. W. Bean (2003): "Modelling High-Dimensional Data by Mixtures of Factor Analyzers," *Computational Statistics & Data Analysis*, 41(3), 379–388.

[59] Meghir, C., and L. Pistaferri, (2011): "Earnings, Consumption and Life Cycle Choices," *Handbook of Labor Economics*, Elsevier.

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

[61] 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.

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

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

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

[65] Van der Vaart, A. W., and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer.

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

[67] Villani, C. (2008): *Optimal Transport: Old and New.* Vol. 338. Springer Science & Business Media.

[68] Wu, X., and J. M. Perloff (2006): "Information-Theoretic Deconvolution Approximation of Treatment Effect Distribution," unpublished manuscript.

APPENDIX

A Proofs

A.1 Proof of Theorem 1

Define the empirical objective function, for any 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 $\widehat{F}_Y^{-1}(\tau) = \inf\{y \in \text{Supp}(Y) : \widehat{F}_Y(y) \geq \tau\}$, and $\widehat{\text{Rank}}(Z_i) = N\widehat{F}_Z(Z_i)$. The second equality follows from Hardy, Littlewood and Polya's rearrangement inequality. 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, where V is standard uniform and $X_2 \sim F_{X_2}$.

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

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

Consider a sieve estimator $\widehat{H}$ such that:

$$\widehat{Q}(\widehat{H}) \leq \min_{H\in\mathcal{H}_N} \widehat{Q}(H) + \epsilon_N,$$

where ϵ_N tends to zero as N tends to infinity. Let $\widetilde{X}_i = \widehat{H}\left(\frac{i}{N+1}\right)$ for all i. We have:

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

To see that (A1) holds, note that $\widetilde{X} \in \mathcal{X}_N$ (by the definition of $\mathcal{H}_N$), and that, for all $X \in \mathcal{X}_\mathcal{N}$, there exists an $H \in \mathcal{H}$ such that $H\left(\frac{i}{N+1}\right) = X_i$ for all i.[25]

Let $H_0 = F_{X_1}^{-1}$. To show Theorem 1 it is thus sufficient to show that $\|\widehat{H} - H_0\|_\infty = o_p(1)$. This will follow from verifying conditions (3.1"), (3.2), (3.4), and (3.5(i)) in Chen (2007).

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

$\mathcal{H}$ is compact under $\|\cdot\|_\infty$ and $Q(H)$ is upper semicontinuous on $\mathcal{H}$. Compactness holds as indicated in the text. (3.4) follows since $\mathcal{H}_N$ is a closed subset of $\mathcal{H}$. To show that $Q(H)$ is continuous on $\mathcal{H}$ under $\|\cdot\|_\infty$, let H_1, H_2 in $\mathcal{H}$. By Assumption 1 (i), F_Y^{-1} and F_{X_2} are Lipschitz. It follows that, for some constant $\widetilde{C}$, $|Q(H_2) - Q(H_1)| \leq \widetilde{C}\|H_2 - H_1\|_\infty$. This implies continuity of Q. This shows (3.1") in Chen (2007).

$\mathcal{H}_N \subset \mathcal{H}_{N+1} \subset \mathcal{H}$ for all N, and there exists a sequence $H_N \in \mathcal{H}_N$ such that $\|H_N - H_0\|_\infty = o_p(1)$. If H_0 is linear with slope $\underline{C}$, take H_N linear too, with slope $\underline{C}_N$. Assume from now on that H_0 is not linear with slope $\underline{C}$. Then there is an $\epsilon > 0$ such that $H_0(1) - H_0(0) > \underline{C} + \epsilon$. Let G_0 be linear with $G_0(0) = H_0(0) + \delta$ and $G_0(1) = H_0(1) - \delta$, for $0 < \delta < (H_0(1) - H_0(0) - (\underline{C} + \epsilon))/2$. For an increasing sequence λ_N which tends to one as N tends to infinity, let $H_N = \lambda_N H_0 + (1 - \lambda_N)G_0$. Taking λ_N such that $(1 - \lambda_N) \geq \max\left\{\frac{\overline{C} - \overline{C}_N}{\delta}, \frac{\underline{C} - \underline{C}_N}{\epsilon}\right\}$, we have $|H_N| \leq \overline{C}_N$ and $\underline{C}_N \leq \nabla H_N \leq \overline{C}_N$, hence $H_N \in \mathcal{H}_N$.[26] Moreover:

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

This shows (3.2) in Chen (2007).

$Q(H)$ is uniquely minimized at H_0 on $\mathcal{H}$, and $Q(H_0) < \infty$. 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:

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

Since the left-hand side in this equation is distributed as F_Y, it thus follows that, almost surely:

$$F_{H(V)+X_2}\left(H(V) + X_2\right) = F_Y\left(H(V) + X_2\right).$$

It follows that $F_{H(V)+X_2} = F_Y$ almost everywhere on the real line. Since Y and X_2 have densities f_Y and f_{X_2}, this also implies that, y-almost everywhere:

$$f_Y(y) = \int_0^1 f_{X_2}(y - H(\tau))d\tau.$$

[26] Note that, by the mean value theorem we have, for all continuously differentiable H and all i:

$$\inf_{\tau \in [i/(N+1),(i+1)/(N+1)]} \nabla H(\tau) \leq (N+1)\left(H\left(\frac{i+1}{N+1}\right) - H\left(\frac{i}{N+1}\right)\right) \leq \sup_{\tau \in [i/(N+1),(i+1)/(N+1)]} \nabla H(\tau).$$

Now, since $H \in \mathcal{H}$, the function $f_{\widetilde{X}}(x) \equiv 1/\nabla H(H^{-1}(x))$ is well-defined, continuous and bounded. We then have by a change of variables:

$$f_Y(y) = \int_0^1 f_{X_2}(y-x) f_{\widetilde{X}}(x) dx.$$

Taking Fourier transforms in this equation yields, denoting as Ψ_Z the characteristic function of any random variable Z:

$$\Psi_Y(s) = \Psi_{X_1}(s)\Psi_{X_2}(s) = \Psi_{\widetilde{X}}(s)\Psi_{X_2}(s), \text{ for all } s \in \mathbb{R}.$$

As Ψ_{X_2} is non-vanishing we thus have $\Psi_{X_1} = \Psi_{\widetilde{X}}$. It follows that $f_{X_1} = f_{\widetilde{X}}$, hence that $H = H_0$. This shows (3.1"(ii)) in Chen (2007).

$\mathbf{plim}_{N\to\infty}$ $\mathbf{sup}_{H\in\mathcal{H}}$ $|\widehat{Q}(H) - Q(H)| = 0$. First, notice that since $\mathcal{H}$ consists of Lipschitz functions its ϵ-bracketing entropy is finite for any $\epsilon > 0$ (e.g., Corollary 2.7.2 in van der Vaart and Wellner, 1996). Hence $\mathcal{H}$ is Glivenko Cantelli for the $\|\cdot\|_\infty$ norm.

Let:

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

Notice that, for all $H \in \mathcal{H}$:[27]

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^N G_H\left(\frac{\sigma(i)}{N+1}, X_{i2}\right) &= \frac{1}{N}\sum_{i=1}^N G_H\left(\frac{i}{N+1}, X_{\sigma^{-1}(i),2}\right) \\ &= \int_0^1 \mathbb{E}\left(G_H(\tau, X_2)\right) d\tau + o_p(1) = Q(H) + o_p(1). \end{aligned} \tag{A2}$$

Moreover, since $H \mapsto G_H$ is Lipschitz on $\mathcal{H}$,[28], and $\mathcal{H}$ is Glivenko Cantelli, the set of functions $\{G_H : H \in \mathcal{H}\}$ is also Glivenko Cantelli. Hence:

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

Next, we are going to show that:

$$\begin{aligned} \sup_{H\in\mathcal{H}} \left| \frac{1}{N}\sum_{i=1}^N \frac{1}{N} \widehat{\text{Rank}}\left(H\left(\frac{\sigma(i)}{N+1}\right) + X_{i2} \right) - \int_0^1 F_{X_2}\left(H\left(\frac{\sigma(i)}{N+1}\right) + X_{i2} - H(\tau) \right) d\tau \right| \\ = o_p(1). \end{aligned} \tag{A3}$$

[27]Recall that σ is a random permutation of $\{1, ..., N\}$. σ is thus a shorthand for σ_N.

[28]This follows from the fact that f_Y is bounded away from zero and f_{X_2} is bounded away from infinity.

From (A3) and the fact that F_Y^{-1} is Lipschitz we will then have:

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

To show (A3) we are going to show that:

$$\sup_{H\in\mathcal{H},\,a\in\mathbb{R}}\left|\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left\{H\left(\frac{\sigma(i)}{N+1}\right)+X_{i2}\leq a\right\}-\int_0^1 F_{X_2}\left(a-H(\tau)\right)d\tau\right| = o_p(1). \qquad \text{(A4)}$$

Pointwise convergence in (A4) is readily verified (similarly as in (A2)). Uniform convergence follows provided we can show that $\mathcal{G}=\{g_{H,a}:H\in\mathcal{H},a\in\mathbb{R}\}$ is Glivenko Cantelli, where $g_{H,a}(v,u)=\mathbf{1}\{H(v)+u\leq a\}$. We are going to show this using a bracketing technique from empirical process theory. Fix an $\epsilon>0$. Since $\mathcal{H}$ has finite ϵ-bracketing entropy 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 $\|H_j-H_{j-1}\|_\infty<\epsilon$ for all j. Moreover, there exists a set of scalars a_k, $k=1,...,K$, such that the real line is covered by the intervals $[a_k,a_{k+1}]$, and $F_{X_2}(a_{k+1})-F_{X_2}(a_k)<\epsilon$ for all k. Since X_2 has bounded support we can assume without loss of generality that $a_{k+1}-a_k<\epsilon$. 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). Since $\int_0^1 F_{X_2}(a_{k+1}-H_j(\tau))d\tau-\int_0^1 F_{X_2}(a_k-H_{j+1}(\tau))d\tau<\widetilde{C}\epsilon$, where $\widetilde{C}>0$ is finite as f_{X_2} is bounded away from infinity, $\mathcal{G}$ is Glivenko Cantelli and (A4) has been shown.

Lastly, since f_Y is bounded away from zero and infinity and differentiable, the empirical quantile function of Y is such that (e.g., Corollary 1.4.1 in Csörgö, 1983):

$$\sup_{\tau\in(0,1)}\left|\widehat{F}_Y^{-1}(\tau)-F_Y^{-1}(\tau)\right| = o_p(1).$$

Hence:

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

This shows (3.5(i)) in Chen (2007) and ends the proof of Theorem 1.

A.2 Proof of Theorem 2

Define the empirical objective function as, for any $H = (H_1, ..., H_K)$:

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

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 $\| \cdot \|$ is the Euclidean norm on $\mathbb{R}^T$. 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 $\widetilde{\mu}_{AH}$. See Chapter 7 in Villani (2003) for an analysis of some of the main properties of Wasserstein distances.

Likewise, let us define the population counterpart to $\widehat{Q}$, for any $H \in \mathcal{H}_K$, 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, or *couplings*, 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}. In this case $Q(H)^{\frac{1}{2}} = W_2 \left(\mu_Y, \mu_{AH} \right)$ is the Wasserstein distance between the two population marginals.

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

***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:

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

for some constant $\widetilde{C} > 0$. Hence, since W_2 satisfies the triangular inequality (see Theorem 7.3 in Villani, 2003):

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

Next, we use that, 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 the Kantorovich-Rubinstein distance, W_1 can be equivalently written as (see Theorem 1.14 in Villani, 2003):

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

where φ are 1-Lipschitz functions on $\mathbb{R}^T$; that is, such that $|\varphi(y_2) - \varphi(y_1)| \leq \|y_2 - y_1\|$ for all $(y_1, y_2) \in \mathbb{R}^T \times \mathbb{R}^T$.

Hence:

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

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

$Q(H)$ is uniquely minimized at H_0 on $\mathcal{H}_K$. Let H be such that $Q(H) = 0$. Then $W_2(\mu_Y, \mu_{AH}) = 0$. By Theorem 7.3 in Villani (2003) this implies that $\mu_Y = \mu_{AH}$. Hence the cdfs of $Y = \sum_{k=1}^{K} A_k H_{0k}(V_k)$ and $\sum_{k=1}^{K} A_k H_k(V_k)$ are equal. By Assumption 2 (iii), it follows from the identification result in Bonhomme and Robin (2010) that $H_k = H_{0k}$ for all k.

$\text{plim}_{N\to\infty} \sup_{H\in\mathcal{H}_K} |\widehat{Q}(H) - Q(H)| = 0$. Using similar arguments as for the continuity of $Q(H)$, we have:

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

where we have used again the triangular inequality.

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

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

where the last equality follows from the set of 1-Lipschitz functions φ being Glivenko-Cantelli.[29]

Next, we have:

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

where the last equality follows from the fact that the following set is Glivenko-Cantelli:

$$\left\{\varphi\circ\left(\sum_{k=1}^{K} A_k H_k\right) : \varphi \text{ is 1-Lipschitz, } H=(H_1,...,H_K)\in\mathcal{H}_K\right\}.$$

This concludes the proof of Theorem 2.

A.3 Proof of Corollary 1

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, where $\mathcal{H}_N^{(2)}$ is the set of functions in $\mathcal{H}^{(2)}$ such that $\left\{H_k\left(\frac{i}{N+1}\right) : i=1,...,N\right\}$ belongs to $\mathcal{X}_N^{(2)}$. We have:

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

where $C>0,\widetilde{C}>0$ are constants, and we have used that κ is Lipschitz, $\nabla\widehat{H}_k$ is uniformly bounded, and $Nb\to\infty$.

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

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

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

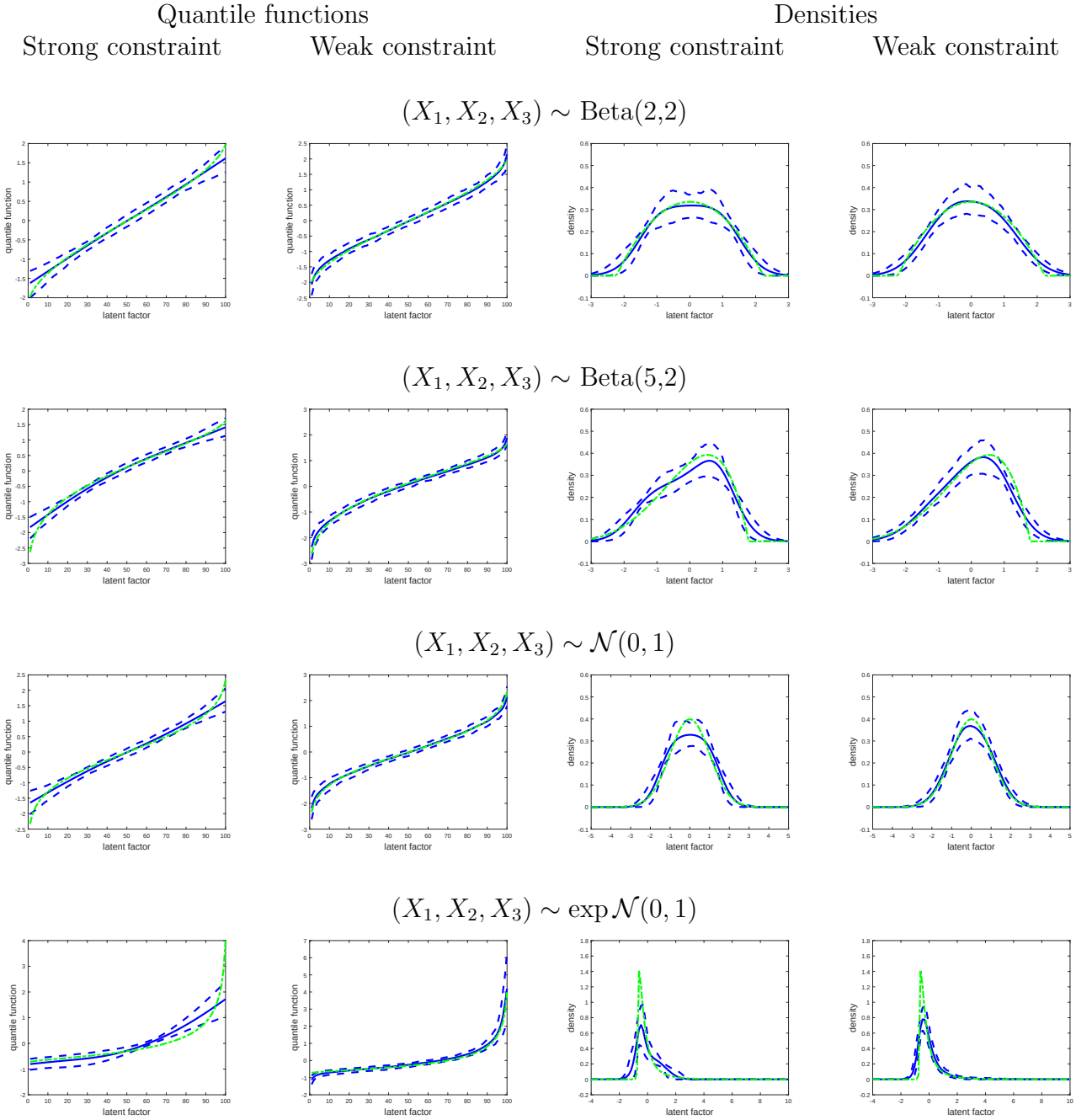
[29]Non-asymptotic bounds and asymptotic rates results are available for $W_2\left(\mu_Y,\widehat{\mu}_Y\right)$ in the literature (e.g., Fournier and Guillin, 2015).

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

This shows Corollary 1.

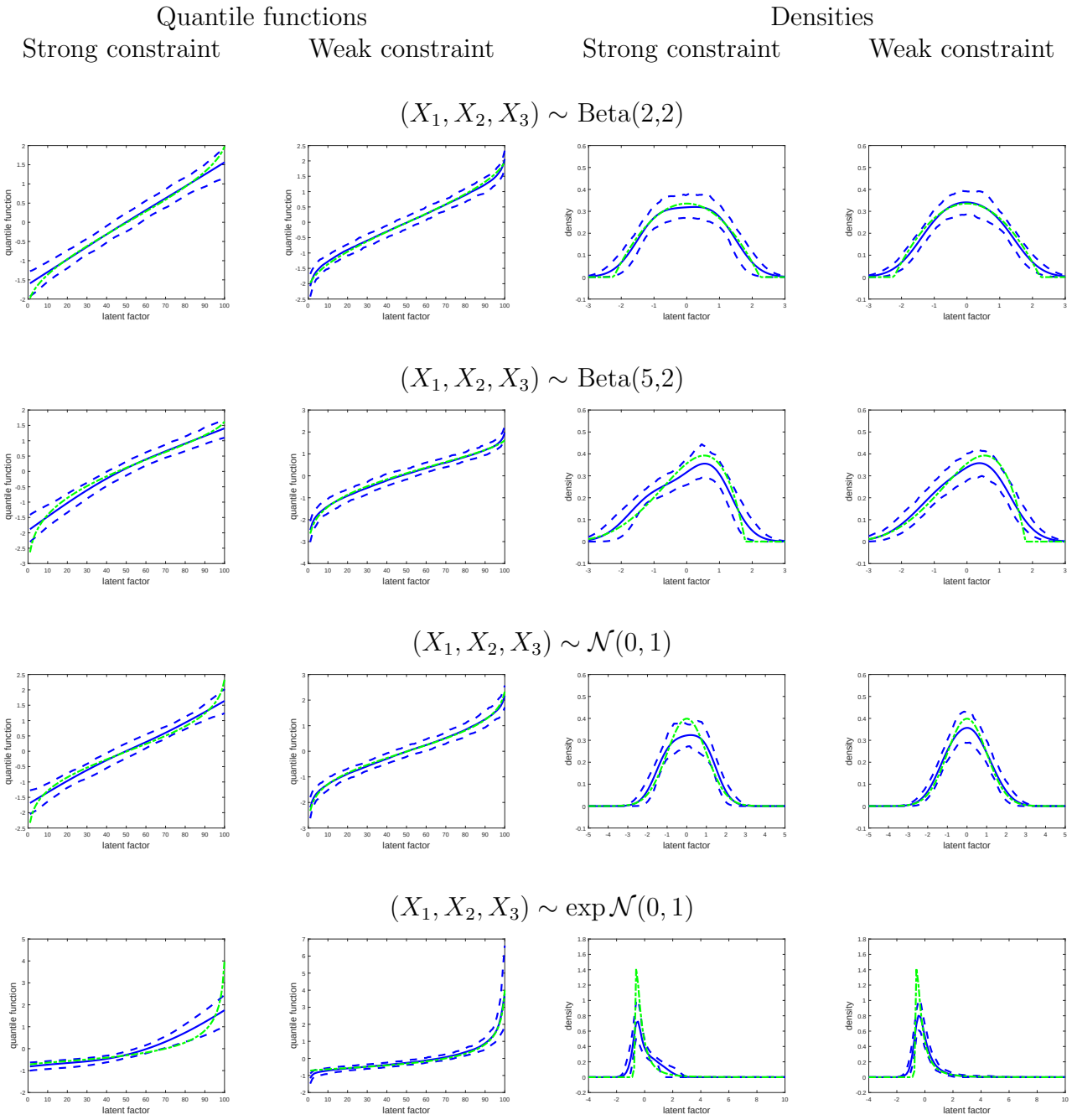
B Additional results

Figure B1: Monte Carlo results for X_2 in the fixed-effects model, $N = 100$, $T = 2$



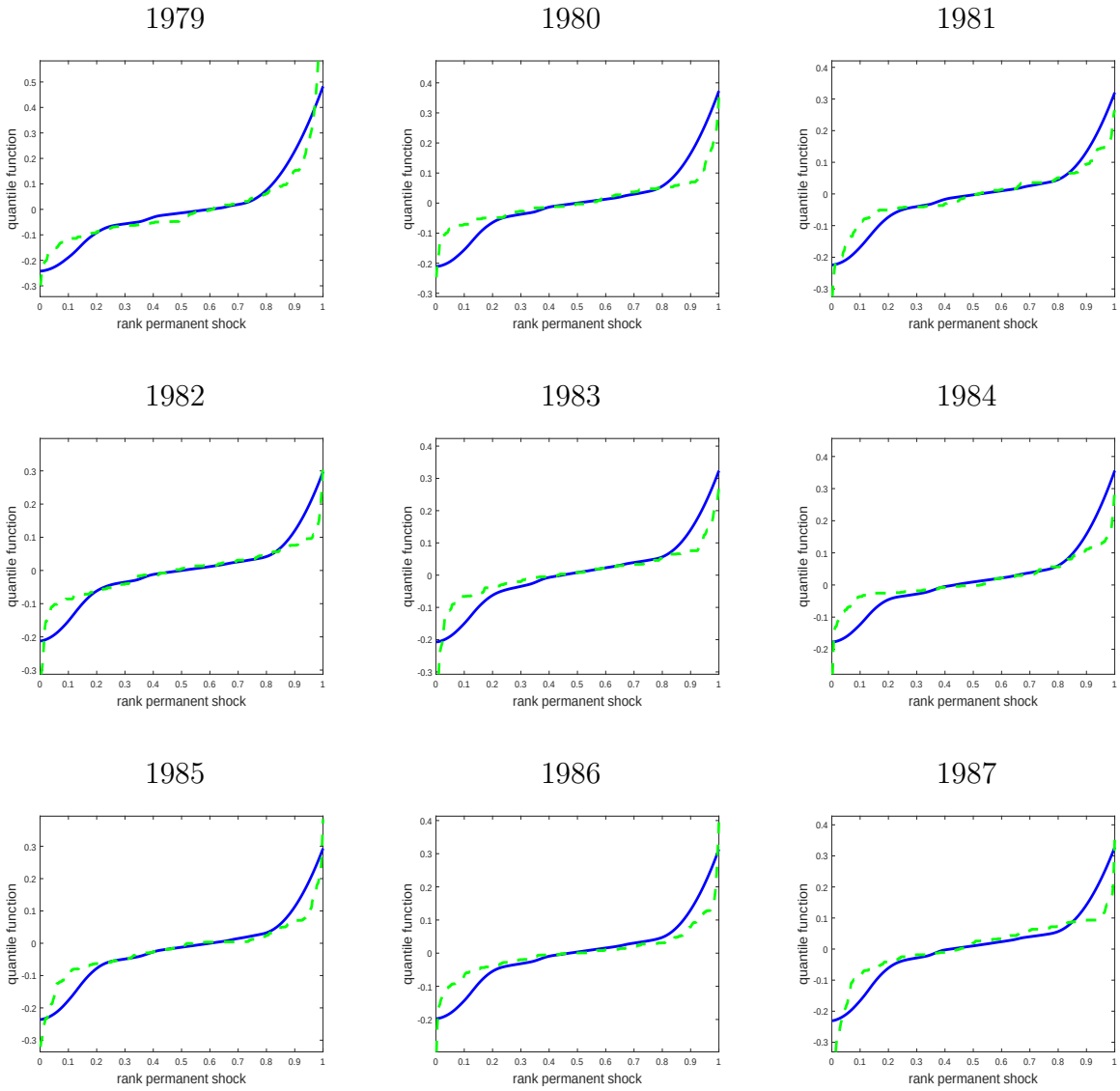
Notes: Simulated data from the fixed-effects model, results for the second factor X_2. Solid is the mean across simulations, dashed are 10 and 90 percent pointwise quantiles, and dashed-dotted is the true quantile function or density. 100 *simulations. 10 averages over permutation draws.*

Figure B2: Monte Carlo results for X_3 in the fixed-effects model, $N = 100$, $T = 2$



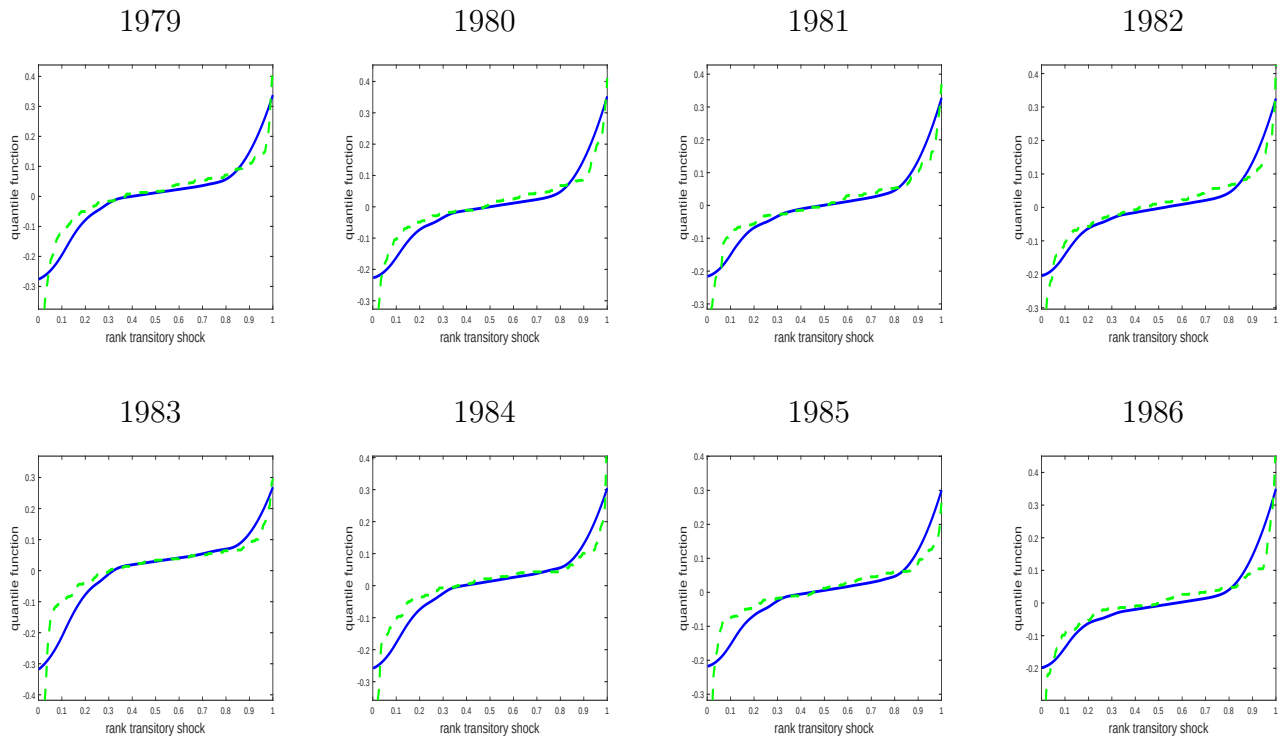
Notes: Simulated data from the fixed-effects model, results for the third factor X_3. Solid is the mean across simulations, dashed are 10 and 90 percent pointwise quantiles, and dashed-dotted is the true quantile function or density. 100 *simulations. 10 averages over permutation draws.*

Figure B3: Estimated quantile functions of permanent shocks in different years, weak constraints (dashed) and strong constraints (solid)



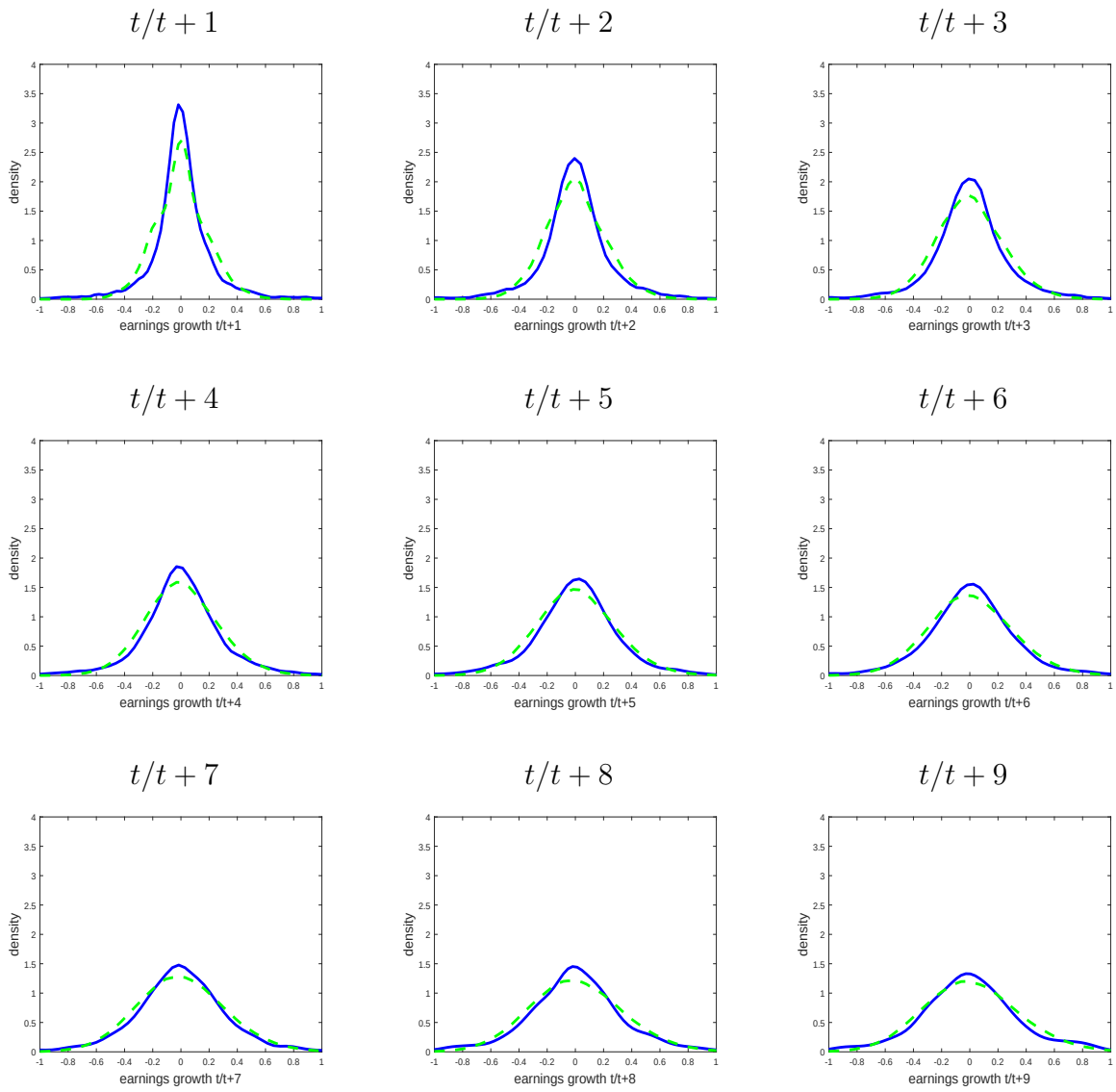
Notes: PSID, 1978-1987. Permanent shock in every year (note: the first and last years are a combination of permanent and transitory shocks). Sample selection and construction of log-earnings growth residuals as in Bonhomme and Robin (2010). Model estimation: strong (solid line) and weak (dashed line) constraint, 10 averages over permutation draws.

Figure B4: Estimated quantile functions of transitory shocks in different years, weak constraints (dashed) and strong constraints (solid)



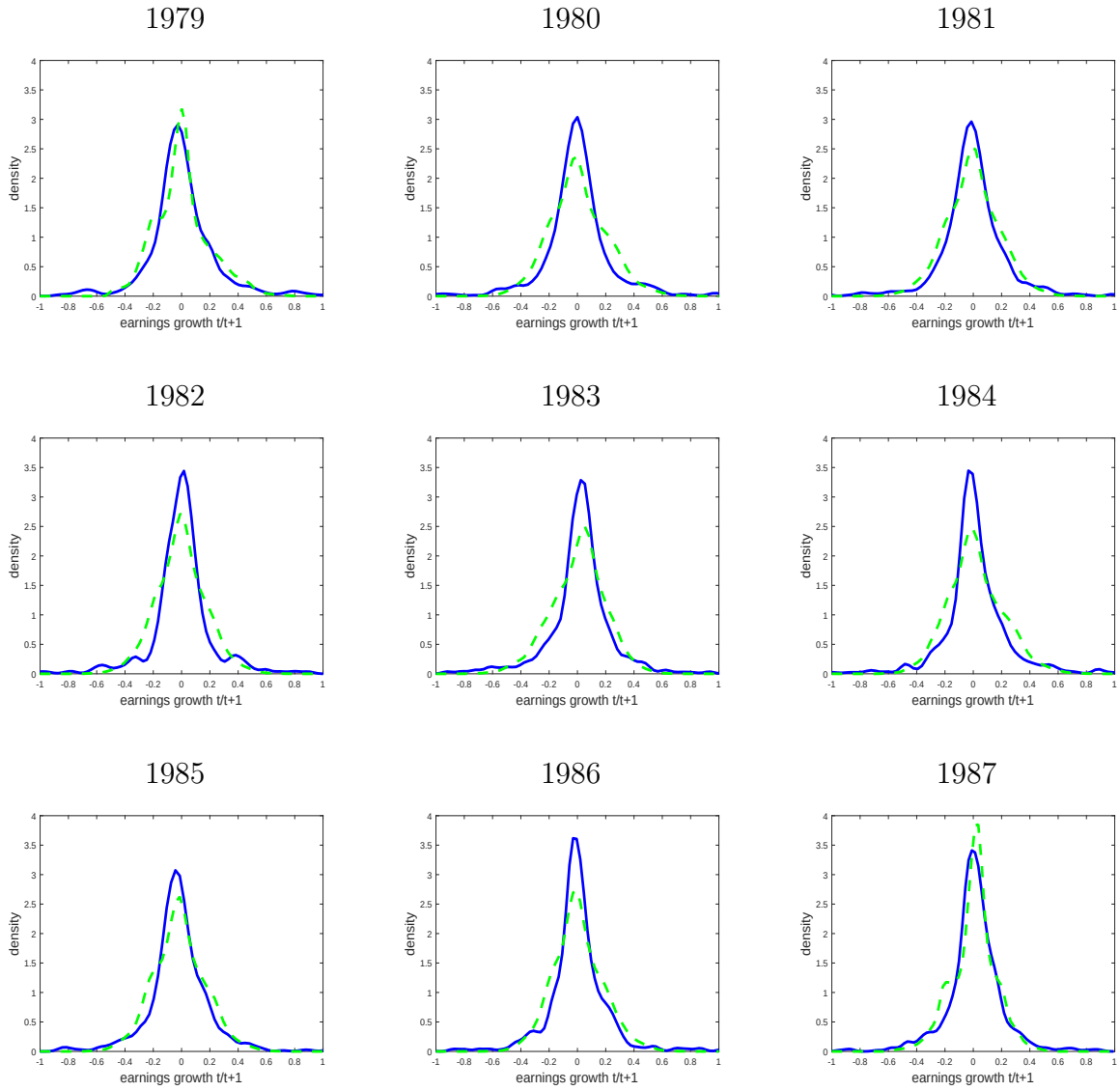
Notes: PSID, 1978-1987. Transitory shock in every year. Sample selection and construction of log-earnings growth residuals as in Bonhomme and Robin (2010). Model estimation: strong (solid line) and weak (dashed line) constraint, 10 averages over permutation draws.

Figure B5: Densities of earnings growth residuals at various horizons, data (solid) and model (dashed)



Notes: PSID, 1978-1987. Results pooled over all years. Sample selection and construction of log-earnings growth residuals as in Bonhomme and Robin (2010). Model estimation: strong constraint, 10 averages over permutation draws. Model simulations: 200 simulations per individual observation.

Figure B6: Densities of earnings growth residuals in different years, data (solid) and model (dashed)



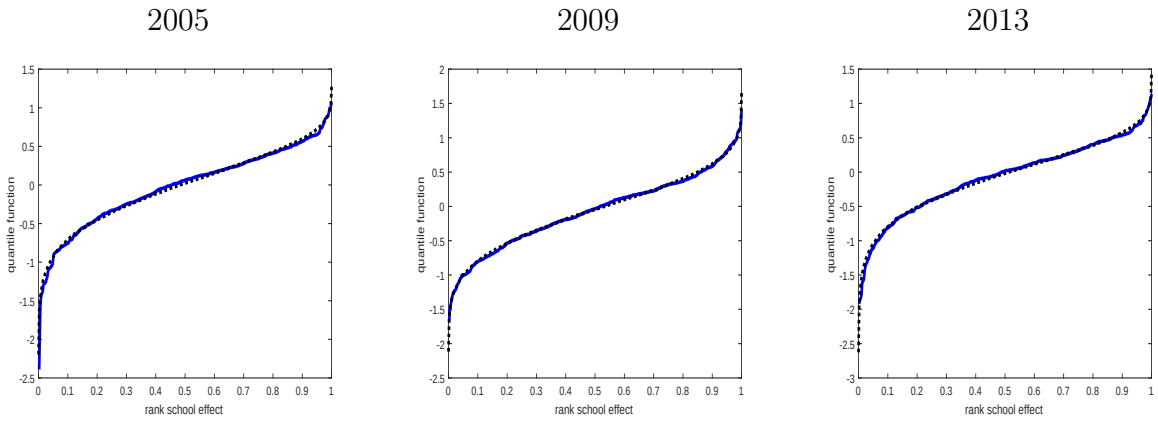
Notes: PSID, 1978-1987. Log-earnings $t/t+1$ growth residuals in every year. Sample selection and construction of log-earnings growth residuals as in Bonhomme and Robin (2010). Model estimation: strong constraint, 10 averages over permutation draws. Model simulations: 200 simulations per individual observation.

Table B1: Covariance matrix of school averages of test scores

	2005	2006	2007	2008	2009	2010	2011	2013	2015
2005	.2955								
2006	.1408	.2986							
2007	.1639	.1877	.3289						
2008	.1467	.1920	.2030	.3055					
2009	.1707	.1692	.2016	.1911	.3070				
2010	.1441	.165	.1878	.1961	.2130	.3054			
2011	.1435	.1611	.1972	.1712	.2078	.2014	.3099		
2013	.1316	.1370	.1683	.1447	.1794	.1750	.1969	.3051	
2015	.0960	.1291	.1387	.1286	.1606	.1602	.1820	.1731	.3286

Notes: Administrative data from the Spanish region of Madrid. The test was not administered in 2012 and 2014.

Figure B7: Model fit to quantile functions of school averages of test scores



Notes: Administrative data from the Spanish region of Madrid. Quantile function of Y_{it} in year t. Data in solid, model in dashed. Model simulations, 200 simulations per observation.