Quantile Factor Models*

Liang Chen[1], Juan J. Dolado[2], and Jesús Gonzalo[3]

[1] *Department of Economics, Nuffield College and INET, University of Oxford*
[2] *Department of Economics, European University Institute*
[3] *Department of Economics, Universidad Carlos III de Madrid*

January 2016

Abstract

In this paper we introduce a novel concept: *Quantile Factor Models* (QFM), where a few unobserved common factors may affect all parts of the distributions of many observed variables in a panel dataset of dimension $N \times T$. When the factors affecting the quantiles also affect the means of the observed variables, a simple two-step procedure is proposed to estimate the common factors and the quantile factor loadings. Conditions on N and T ensuring uniform consistency and weak convergence of the entire quantile factor loadings processes differ from standard conditions in factor-augmented regressions with smooth object functions. Based on these results, we show how to make inference on the quantile factor loadings in a location-scale shift factor model. When factors affecting the quantiles differ from those affecting the means of the observed variables, we propose an iterative procedure to estimate both factors and factor loadings at a given quantile. Simulation results confirm a satisfactory performance of our estimators in small to moderate sample sizes. In particular, it is shown that the iterative procedure can consistently estimate common factors that cannot be captured by PC estimators. Empirical applications of our methods to stocks and mutual fund returns are considered.

Keywords: Factor models, quantile regression, generated regressors, incidental parameters.

JEL codes: C31, C38.

*We are grateful to Dante Amengual, Manuel Arellano, Xu Cheng, Alfred Galichon, Jerry Hausman, Enrique Sentana and Guillermo Tellechea for helpful discussions and insightful comments. Financial support from the Open Society Foundations, the Oxford Martin School, and the Spanish Ministerio de Economía y Competitividad is gratefully acknowledged.

1 Introduction

The last two decades have seen a rapid and fruitful progress in the theoretical development of large dimensional factor models, which are now broadly applied in macroeconomic forecasting and modeling; see Bai and Ng (2008b) and Stock and Watson (2011) for reviews of recent developments. The primary advantage of factor models is that they provide a parsimonious and flexible way of characterizing the co-movement of many observed variables through a small number of unobserved factors. Our goal in this paper is to introduce a novel concept in this setup: *quantile factor models* (QFMs hereafter), as well as to analyze estimation and inference of such models.

The main feature of QFMs is that for a panel of N observed variables, there exists a vector of $r(<< N)$ common factors F_t such that, conditional on F_t, the quantiles of these observed variables are linear in F_t. The coefficients in these linear functions for the τth quantile (where $0 < \tau < 1$), called the *quantile factor loadings* at τ, are allowed to be different for all variables, so that they are analogues to factor loadings in standard factor models. Moreover, for each individual variable, the loadings can be different at different quantiles, allowing the common factors to have heterogeneous effects on different parts of the conditional distributions of the observed variables. The quantile factor loadings at different quantiles, called the *quantile factor loadings processes*, can be viewed as functions of τ, and they are the main objects of interest.

To estimate the common factors and the quantile factor loading processes, we start by proposing a simple two-step procedure that is easy to implement in practice. In the first step, the common factors are estimated using principal components analysis (PCA hereafter); in the second step, the quantile factor loadings at various τs are estimated using quantile regressions (QR hereafter), where the unobserved factors are replaced by their estimates in the first step. Uniform consistency and weak convergence of the estimated quantile factor loading processes are established under general assumptions. In particular, we show that, among other conditions, if $T^{5/4}/N \to 0$ as $N, T \to \infty$ jointly, the distributional effects of estimating the common factors can be asymptotically ignored in the second step.

The asymptotic distributions of the entire factor loadings process can be used to test hypothesis of very general forms. In this paper, we show how to use these results to make inference in a location-scale shift factor model. When the null hypotheses involves unknown parameters that need to be estimated, we follow Koenker and Xiao (2002) in using the *Khmaladez martingale transformation* to solve the so-called *Durbin's problem*.

While the two-step procedure works for a large class of QFMs, it may fail when there are factors that affect the quantiles but do not affect the means of the observed variables. To solve this problem, an iterative procedure is proposed to estimate the common factors and the factor loadings at a given quantile, while the consistency of the such estimators is proven for a smoothed version of the iterative procedure.

Our paper is related to a burgeoning literature on QRs which is summarized by Koenker (2005). In particular, there is a growing number of studies on the intersection of QRs and panel data models; cf. Koenker (2004), Abrevaya and Dahl (2008), Graham et al. (2009), Lamarche (2010), Canay (2011), Rosen (2012), Kato et al. (2012) and Harding and Lamarche (2014), among others. Our model is significantly different from all these quantile panel data models since our regressors (the common factors) are not observable. Additionally, our paper is also related to the problem of QRs with generated regressors, see e.g. Xiao and Koenker (2009).

This paper also contributes to the rapidly growing literature on factor models. To the best of our knowledge, this is the first paper that proposes the concept of QFM. Ando and Tsay (2011) study a factor-augmented QR where the factors are estimated by PC from a standard factor model. Their result can be viewed as an extension of factor-augmented regressions (see Bai and Ng 2006) and factor-augmented extremum estimators (see Bai and Ng 2008a) to factor-augmented QRs. Although our model is obviously different from the settings of all these papers, our results are related to the factor-augmented QRs since, in the second step of our estimation procedure, the true factors are also replaced by the estimated factors. However, we show that due to the non-smoothness of the object function, one needs $T^{5/4}/N \to 0$ for the estimated factors to be treated as known. This is in contrast to the condition $T^{5/8}/N \to 0$ obtained by Bai and Ng (2008a) for smooth object functions.

The rest of the paper is organized as follows: In Section 2, the QFMs are defined, several examples are given, and the two-step estimator is proposed. Section 3 present the main asymptotic results for the estimated quantile factor loadings processes. Second 4 proposes an iteration procedure when the two-step procedure fails. Section 5 contains some simulation results, and Section 6 considers applications of our methods to stock and mutual fund returns, where we show how to make inference for the entire quantile factor loading process based on the asymptotic distributions established in Section 2. Finally, Second 7 concludes and suggests several directions for future research. Proofs of the main results are collected in the appendix.

2 The Model and the Estimator

2.1 Quantile Factor Models

Suppose there is a panel of observable random variables $\{X_{it}\}$ generated by

$$X_{it} = \lambda_i'(U_{it})F_t, \text{ where } U \perp F, \text{ and } U_{it} \sim U[0,1] \tag{1}$$

for $i = 1, \ldots, N$ and $t = 1, \ldots, T$. The common factors F_t is a $r \times 1$ vector of unobservable random variables, with $F_t \in \mathcal{F} \subset \mathbb{R}^r$ for all t. Let $\mathcal{T}$ denote a closed subinterval of $(0,1)$, and suppose that $\lambda_i(\tau) \in \mathcal{A} \subset \mathbb{R}^r$ for all i and $\tau \in \mathcal{T}$. If we further assume the mapping $\tau \mapsto \lambda_i'(\tau)f$ to be strictly incresing for all i and any $f \in \mathcal{F}$, then $\lambda_i'(\tau)F_t$ is the τth quantile of X_{it} conditional on F_t since:

$$P\big[X_{it} \leq \lambda_i'(\tau)F_t|F_t\big] = P\big[\lambda_i'(U_{it})F_t \leq \lambda_i'(\tau)F_t|F_t\big] = P[U_{it} \leq \tau] = \tau.$$

In other words, model (1) implies

$$Q_{X_{it}}[\tau|F_t] = \lambda_i'(\tau)F_t \text{ for all } \tau \in \mathcal{T}. \tag{2}$$

Therefore, conditional on F_t, the quantiles of X_{it} have a factor model structure. As a result, we call (1) a *Quantile Factor Model*, and $\Lambda(\tau) = (\lambda_1(\tau), \ldots, \lambda_N(\tau))'$ are called the *quantile factor loadings at* τ. It seems that representations (1) and (2) are equivalent, but Example 3 (to be shown below) provides a counterexample showing that (1) is in fact more restrictive than (2).

Similar representation for conditional quantiles can be found in Chernozhukov and Hansen (2006, 2008), Canay (2011), and many other papers. It also has an interesting random coefficient interpretation (see Koenker 2005) as we can define $\tilde{\lambda}_{it} = \lambda_i(U_{it})$ as random coefficients. Moreover, since the dependence between the elements of F_t is left unrestricted, the factors can include different transformations of the same variable, and thus model (1) can approximate any nonlinear conditional quantile functions arbitrarily well by increasing the number of factors. In this sense, the linearity of the quantile factor model (1) is not as restrictive as it looks.

2.2 Examples

Example 1. ***Location shift model.*** *$X_{it} = \alpha_i f_t + \epsilon_{it}$, where $\{\epsilon_{it}\}$ are i.i.d errors with cumulative distribution function (CDF) F_ϵ. This is a standard factor model and it can be equivalently written as $X_{it} = \alpha_i f_t + Q_\epsilon(U_{it})$, where $Q_\epsilon(\tau) = \mathsf{F}_\epsilon^{-1}(\tau) = \inf\{c : \mathsf{F}_\epsilon(c) \leq \tau\}$ is assumed to be uniquely defined for each $\tau \in (0,1)$, and $\{U_{it}\}$ are i.i.d and uniformly distributed over $[0,1]$. Thus, this model is can be expressed as model (1) by defining $\lambda_i(U_{it}) = [Q_\epsilon(U_{it}), \alpha_i]'$ and $F_t = [1, f_t]'$.*

Example 2. ***Location-scale shift model (same sign-restricted factor).*** *$X_{it} = \alpha_i f_t + f_t \epsilon_{it}$, where $f_t \geq 0$ for all t and $\{\epsilon_{it}\}$ are defined as in Example 1. The model can be written as in (1) by defining $\lambda_i(U_{it}) = Q_\epsilon(U_{it}) + \alpha_i$ and $F_t = f_t$.*

Example 3. ***Location-scale shift model (sign-unrestricted factor).*** *$X_{it} = \alpha_i f_t + f_t \epsilon_{it}$, where $\{\epsilon_{it}\}$ are defined as in Example 1 and the sign of f_t is unrestricted. As in Example 2, this model can be written as $X_{it} = (Q_\epsilon(U_{it}) + \alpha_i)f_t$. When $f_t \geq 0$, the conditional τth quantile* *of X_{it} given f_t is $(Q_\epsilon(\tau)+\alpha_i)f_t$; when $f_t<0$, the conditional τth quantile of X_{it} given f_t is $(Q_\epsilon(1-\tau)+\alpha_i)f_t$. Therefore, this model cannot be nested by model (1) since the quantile factor loadings depend on the signs of the factors.*

Example 4. *Location-scale shift model (different factors).* *$X_{it}=\alpha_i f_t+g_t\epsilon_{it}$, where $\{\epsilon_{it}\}$ are defined as in Example 1 and $g_t>0$. In this case, X_{it} has an equivalent representation in form of (1) with $\lambda_i(U_{it})=[\alpha_i,Q_\epsilon(U_{it})]'$ and $F_t=[f_t,g_t]'$.*

Example 5. *Nonlinear factor model.* *The model $X_{it}=\lambda_i\cdot f_t\cdot e^{\epsilon_{it}}$, where $\lambda_i>0, f_t>0$, can be written as a special case of model (1) with $F_t=f_t$ and $\lambda_i(U_{it})=\lambda_i\cdot e^{Q_\epsilon(U_{it})}$. Note that taking logarithm on both sides we get $\log X_{it}=\log\lambda_i+\log f_t+\epsilon_t$, a linear factor model for $\log X_{it}$ where the new factor $g_t=\log f_t$ can be easily estimated. However, it is not possible to estimate the original factors and factor loading processes from this transformed linear model.*

The five examples above represent some but not all of the possible cases where an approximate factor model (AFM, as defined in Chamberlain and Rothschild 1983) implies a QFM, and they highlight some important points in our specification of QFMs. First, it is crucial that the mapping $\tau\mapsto\lambda_i'(\tau)f$ is monotone for all possible values of F_t in $\mathcal{F}$. In a simple linear model like (1), this may require the domain of F_t to be restricted, such as in Example 2. Second, the factors and the number of factors in the QFM may be different from those in the approximate factor model, such as in Examples 1 and 4. For instance, in Example 4, if $\mathbb{E}(\epsilon_{it})=0$ and $f_t\neq g_t$, then there is only one factor in the approximate factor model: f_t, but there are two factors in the QFM: f_t and g_t. The implication of such differences for the estimation of the quantile factor loadings will be discussed in detail in the next section.

Example 3 is an interesting example since it shows that representation (2) is more general than model (1). To see this, note that in Example 3 we have

$$Q_{X_{it}}[\tau|f_t]=(Q_\epsilon(\tau)+\alpha_i)f_t\cdot\mathbf{1}\{f_t\geq 0\}+(Q_\epsilon(1-\tau)+\alpha_i)f_t\cdot\mathbf{1}\{f_t<0\}, \tag{3}$$

which is a special case of (2) by setting

$$\lambda_i(\tau)=[Q_\epsilon(\tau)+\alpha_i,Q_\epsilon(1-\tau)+\alpha_i]' \text{ and } F_t=[f_t\cdot\mathbf{1}\{f_t\geq 0\},f_t\cdot\mathbf{1}\{f_t<0\}]'.$$

Moreover, it is easy to see that for any uniformly distributed random variables $U_{1,it}$ and $U_{2,it}$, the model

$$X_{it}=(Q_\epsilon(U_{1,it})+\alpha_i)f_t\cdot\mathbf{1}\{f_t\geq 0\}+(Q_\epsilon(U_{2,it})+\alpha_i)f_t\cdot\mathbf{1}\{f_t<0\} \tag{4}$$

gives conditional quantiles of form (3). Example 3 is a special case of model (4) with $U_{1,it}=U_{2,it}=U_{it}$. Therefore, for Example 3 the conditional quantiles has the form of (2), but it is impossible to write it in the form of model (1) since the mapping $\tau\mapsto\lambda_i'(\tau)f$ is not monotone.

Interestingly, by choosing $U_{1,it} = U_{it}$ and $U_{2,it} = 1 - U_{it}$ in model (4) we get a strictly increasing mapping and the model becomes a special case of model (1). In particular, when the distribution of U_{it} is symmetric around 0, it is easy to see that the model reduces to $X_{it} = \alpha_i f_t + |f_t|\epsilon_{it}$, a special case of Example 4.

2.3 The Estimator

Note that we can also write model (1) as:

$$X_{it} = \lambda_i'(\tau)F_t + [\lambda_i(U_{it}) - \lambda_i(\tau)]'F_t = \lambda_i'(\tau)F_t + v_{it}, \tag{5}$$

where $v_{it} = [\lambda_i(U_{it}) - \lambda_i(\tau)]'F_t$ and $P[v_{it} \leq 0|F_t] = \tau$. The main objects of interest are the common factors, and the quantile factor loadings at all $\tau \in \mathcal{T}$. If F_t can be observed, using standard QR of X_{it} on F_t leads to consistent and asymptotically normally distributed estimators of $\lambda_i(\tau)$ for each i and $\tau \in \mathcal{T}$. However, since F_t are not observable, a feasible procedure is to estimate the factors first, and then run QR of X_{it} on the estimated factors, $\hat{F}_t$.

Define $\lambda_i = \mathbb{E}[\lambda_i(U_{it})]$, then model (1) can also be rewritten as:

$$X_{it} = \lambda_i' F_t + [\lambda_i(U_{it}) - \lambda_i]'F_t = \lambda_i' F_t + e_{it}, \tag{6}$$

where $e_{it} = [\lambda_i(U_{it}) - \lambda_i]'F_t$, and $\mathbb{E}[e_{it}|F_t] = 0$. Thus, if λ_i, F_t and e_{it} satisfy some assumptions (see Assumption 1 below), (6) can be viewed as an AFM, and the factors can be consistently estimated by PCA as in Stock and Watson (2002) and Bai (2003). This leads us to the following two-step estimation procedure for the common factors and the quantile factor loadings at various τs:[1]

1. First, obtain the estimated factors $\hat{F}$. For example, following Bai (2003), one can use PCA where $\hat{F} = (\hat{F}_1, \ldots, \hat{F}_T)'$ are the r eigenvectors (multiplied by $\sqrt{T}$) of XX' associated with the r largest eigenvalues, where $X = \{X_{it}\}'$ is a $T \times N$ matrix collecting all the observable variables.

2. For $i = 1, \ldots, N$ and each $\tau \in \mathcal{T}$, the QR estimator $\hat{\lambda}_i(\tau)$ is then defined as:

$$\hat{\lambda}_i(\tau) = \arg\min_{\lambda \in \mathcal{A}} T^{-1} \sum_{t=1}^{T} \rho_\tau(X_{it} - \lambda' \hat{F}_t) \tag{7}$$

where $\rho_\tau(u) = u(\tau - \mathbf{1}\{u < 0\})$, i.e., the so-called check function.

Both steps of our estimation procedure can be easily implemented in standard econometric packages, providing in this way a very convenient tool for the practitioners. Furthermore, an

[1] As discussed in Section 4 below, this two-step procedure is not useful if the data are generated by location and scale-shift models as in Example 4 above.

observation of independent interest is that when the errors e_{it} in model (6) have symmetric distributions around zero, our second step at $\tau = 0.5$ can be viewed as a median regression for estimating the factor loadings in an AFM while the estimated factor loadings in Bai (2003) are obtained by OLS regressions of X_{it} on $\hat{F}_t$.

A generic problem of factor analysis is the indeterminacy of the factors and factor loadings up to to a rotation, which also pertains to the QFMs defined above. To see this, note that for any invertible $r \times r$ matrix A, model (1) is observationally equivalent to $(\lambda_i'(U_{it})A^{-1})(AF_t)$. Therefore, the factors and the quantile factor loadings can only be identified up to a rotation, unless r^2 restrictions are imposed to pin down a unique rotation matrix. The PCA estimators defined above implicitly adopt the normalization that $T^{-1}\sum_{t=1}^{T} F_t F_t' = I_r$ and $N^{-1}\sum_{i=1}^{N} \lambda_i \lambda_i'$ is orthogonal, which is equivalent to a specific choice of A. Bai and Ng (2013) also consider other normalizations that uniquely determine the rotation matrix. However, unless justified by some economic theories, it is hard to argue that any specific rotation is better than the others. For example, the restriction PC2 in Bai and Ng (2013) assumes that $T^{-1}\sum_{t=1}^{T} F_t F_t' = I_r$ and $[\lambda_1, \ldots, \lambda_r]'$ is a lower triangular matrix, while their restriction PC3 assumes $[\lambda_1, \ldots, \lambda_r]' = I_r$. The two restrictions imply two different rotation matrices, but one has to resort to economic theories to determine which one is more appropriate. In this paper we do not consider the problem of imposing identification restrictions. Therefore, our main results in the next section are stated for a (possibly random) rotation of $\lambda_i(\tau)$, because the PC estimators of the factors are only consistent for a rotation of F_t. However, it is straightforward to extend our results to estimators with identification restrictions in the manner of Bai and Ng (2013).

3 Asymptotic Results

3.1 Consistency

To establish the uniform consistency of the estimated quantile factor loadings, we impose the following assumptions for each $i = 1, \ldots, N$:

Assumption 1. *Suppose that the observed data* $\{X_{it}\}$ *are generated by model* (1) *and*

(i) The sequence $\{F_t\}$ *is strictly stationary and m-dependent with* $\mathbb{E}\|F_t\|^4 < \infty$*, and* $\Sigma_F = \mathbb{E}(F_t F_t') > 0$.

(ii) The random variables $\{U_{it}\}$ *are uniformly distributed over* $[0, 1]$ *and independent across i and t, and* U_{it} *is independent of* F_t *for all* i, t.

(iii) There is a compact set $\mathcal{A} \subset \mathbb{R}^r$ *such that* $\lambda_i(\tau) \in \mathcal{A}$ *for all i and* $\tau \in \mathcal{T}$*, and there is a* $\Sigma_\Lambda > 0$ *such that* $\|N^{-1}\sum_{i=1}^{N} \lambda_i \lambda_i' - \Sigma_\Lambda\| \to 0$ *as* $N \to \infty$.

(iv) The eigenvalues of $\Sigma_F \Sigma_\Lambda$ *are distinct.*

(v) The conditional density $\mathsf{f}_X(x|F_t = f)$ *exists, and is bounded and uniformly continuous in*

x for all $f \in \mathcal{F}$; $J(\lambda_i(\tau)) = \mathbb{E}[\mathsf{f}_X(\lambda_i(\tau)'F_t|F_t)F_tF_t']$ is positive definite for all τ.

Define $H_{NT} = (\Lambda'\Lambda/N)(F'\hat{F}/T)V_{NT}^{-1}$, where $\Lambda' = [\lambda_1, \ldots, \lambda_N]$, $F' = [F_1, \ldots, F_T]$, $\hat{F}' = [\hat{F}_1, \ldots, \hat{F}_T]$, and V_{NT} is a $r \times r$ diagonal matrix with the eigenvalues of $(NT)^{-1}XX'$ in decreasing order. Further, define $H_0 = \Sigma_\Lambda^{1/2}\Upsilon V^{-1/2}$, where V is a diagonal matrix with the eigenvalues of $\Sigma_\Lambda^{1/2}\Sigma_F\Sigma_\Lambda^{1/2}$ in decreasing order, and Υ is a matrix of corresponding eigenvectors. It can be shown that:

Theorem 1 (Uniform Consistency). *Under Assumption 1, $\sup_{\tau \in \mathcal{T}} \|\hat{\lambda}_i(\tau) - H_{NT}^{-1}\lambda_i(\tau)\| = o_P(1)$ and $\sup_{\tau \in \mathcal{T}} \|\hat{\lambda}_i(\tau) - H_0^{-1}\lambda_i(\tau)\| = o_P(1)$ for all $i = 1, \ldots, N$.*

Remark 1: The proof of Theorem 1 consists of two steps. In the first step, it is shown that $T^{-1}\sum_{t=1}^T \rho_\tau(X_{it} - \lambda'\hat{F}_t)$ converges to $\mathbb{E}[\rho_\tau(X_{it} - H_0'F_t)]$ uniformly in τ and λ. In the second step, given that $H_0^{-1}\lambda_i(\tau)$ is the unique minimizer of $\mathbb{E}[\rho_\tau(X_{it} - H_0'F_t)]$ by Assumption 1(v) and that $\hat{\lambda}_i(\tau)$ is the minimizer of $T^{-1}\sum_{t=1}^T \rho_\tau(X_{it} - \lambda'\hat{F}_t)$ by definition, the uniform consistency of $\hat{\lambda}_i(\tau)$ for $H_0^{-1}\lambda_i(\tau)$ follows from Lemma B.1 of Chernozhukov and Hansen (2006), which is a generalization of the consistency of M-estimators to estimated processes. A key result to show the uniform convergence of $T^{-1}\sum_{t=1}^T \rho_\tau(X_{it} - \lambda'\hat{F}_t)$ to $\mathbb{E}[\rho_\tau(X_{it} - H_0'F_t)]$ and to prove Theorem 2 below is the following consistency result for the estimated factors: $T^{-1}\sum_{t=1}^T \|\hat{F}_t - H_{NT}'F_t\|^2 = o_P(1)$. This result becomes a direct consequence of Theorem 1 in Bai and Ng (2002) if one can show that the factors, factor loadings λ_i and the error terms e_{it} in Model (6) all satisfy Assumptions A to D in their paper. However, in our setting, the error terms $e_{it} = (\lambda_i(U_{it}) - \lambda_i)'F_t$ do not satisfy Assumption C.5 of Bai and Ng (2002), which requires

$$\mathbb{E}\left|N^{-1/2}\sum_{i=1}^N [e_{it}e_{is} - \mathbb{E}(e_{it}e_{is})]\right|^4 < \infty \text{ for all } t, s. \tag{8}$$

To see this, consider the simple case $r = 1$. Define $u_{it} = \lambda_i(U_{it}) - \lambda_i$ so that in our model $e_{it} = u_{it}F_t$. When $t = s$, we have

$$\mathbb{E}\left|N^{-1/2}\sum_{i=1}^N [e_{it}e_{is} - \mathbb{E}(e_{it}e_{is})]\right|^2 = N^{-1}\sum_{i=1}^N\sum_{j=1}^N \Big(\mathbb{E}[e_{it}^2e_{jt}^2] - \mathbb{E}[e_{it}^2]\mathbb{E}[e_{jt}^2]\Big).$$

Since in our setup, $\mathbb{E}[e_{it}^2e_{jt}^2] - \mathbb{E}[e_{it}^2]\mathbb{E}[e_{jt}^2] = \mathbb{E}[u_{it}^2]\mathbb{E}[u_{jt}^2]\big(\mathbb{E}[F_t^4] - (\mathbb{E}[F_t^2])^2\big) \neq 0$ for any i, j unless F_t^2 is a constant, the above expression can not be bounded, and thus Assumption C5 of Bai and Ng (2002) is violated. As shown in the Appendix, by imposing the stronger condition $\mathbb{E}\|F_t\|^4 < \infty$[2], we prove that Theorem 1 of Bai and Ng (2002) still holds even when their Assumption C.5 is not satisfied in our model. ∎

[2]Assumption A of Bai and Ng (2002) does require $\mathbb{E}\|F_t\|^4 < \infty$, which is only needed to prove Theorem 2 in their paper. To prove their Theorem 1, $\mathbb{E}\|F_t\|^2 < \infty$ is sufficient.

3.2 Weak Convergence

To establish the limiting distribution of the estimated quantile factor loading processes, we impose the following extra assumptions:

Assumption 2. *(i)* $\mathbb{E}\|F_t\|^8 < \infty$*; (ii)* $T^{5/4}/N \to 0$ *as* $N, T \to \infty$*; (iii) For each* $i \leq N$*, the eigenvalues of* $J_{H_0}(\lambda_i(\tau)) = H_0' J(\lambda_i(\tau)) H_0$ *are bounded below by a constant* $\rho^* > 0$ *uniformly in* τ*.*

Define $\varphi_\tau(u) = \mathbf{1}\{u < 0\} - \tau$, and let $\mathcal{B}_r$ be a vector of r independent standard Brownian Bridges, then:

Theorem 2 (Weak Convergence)**.** *Under Assumptions 1 and 2, it holds that, for each* i*,*

$$J_{H_0}(\lambda_i(\cdot)) \cdot \sqrt{T}[\hat{\lambda}_i(\cdot) - H_{NT}^{-1}\lambda_i(\cdot)] = -\mathbb{V}_{iT}(\cdot) + o_P(1) \text{ in } \ell^\infty(\mathcal{T}),$$

where $\mathbb{V}_{iT}(\cdot) = T^{-1/2}\sum_{t=1}^{T} \varphi_\tau(X_t - \lambda_i(\cdot)' F_t) H_0' F_t$ *converges weakly to* $\mathcal{B}_r(\cdot)$ *in* $\ell^\infty(\mathcal{T})$*.*

Remark 2: Bai and Ng (2008a) show that, for extremum estimators with twice continuously differentiable object functions, the estimated factors can be treated as known when they are regressors, if (among other conditions) $T^{5/8}/N \to 0$. In contrast, the estimation-effects-free property of our estimators requires a much larger N compared to T, i.e., $T^{5/4}/N \to 0$. This difference is mainly due to the fact that our object function is not smooth, and thus a necessary condition for the estimated factors to have no distributional effects is

$$\sqrt{T} \cdot \max_{1\leq t\leq T} \|\hat{F}_t - H_{NT}' F_t\| = o_P(1). \tag{9}$$

While in Bai and Ng (2008a), due to the smoothness of their object function, it is enough to have

$$\big(O_P(1) + O_P(\sqrt{T}/\sqrt{N})\big) \cdot \max_{1\leq t\leq T} \|\hat{F}_t - H_{NT}' F_t\| = o_P(1).$$

In the Appendix, we establish the following uniform convergence rate for the estimated factors :

$$\max_{1\leq t\leq T} \|\hat{F}_t - H_{NT}' F_t\| = O_P(T^{-5/8}) + O_P(T^{1/8}/\sqrt{N}), \tag{10}$$

illustrating that the required condition $T^{5/4}/N \to 0$ is therefore a consequence of (9) and (10). ■

Remark 3: Suppose that $\mathcal{T} = [\epsilon, 1-\epsilon]$ for some $\epsilon > 0$. For small values of ϵ, Theorem 2 may not provide a good approximation for the finite sample distributions of the estimators. Usually, the Gaussian approximation performance well for $\epsilon > 30/T$ (e.g., when $T = 200$, $\epsilon > 0.15$), and for more extreme quantiles the small sample distributions are better approximated by the asymptotic distributions of extremal conditional quantiles (see Chernozhukov 2005). ■

The above-mentioned asymptotic theory involves a random rotation of the original quantile factor loadings. As discussed in the end of Section 2.1, this random rotation matrix H_{NT}^{-1} depends on the factor and factor loadings in (6). As a result, it is not possible to make any inference about the individual elements of the quantile factor loadings unless some identification restrictions are imposed. Suppose we consider the following widely adopted (also called restriction PC1 in Bai and Ng 2013) restrictions in factor analysis:

$$T^{-1}\sum_{t=1}^{T} F_t F_t' = I_r \text{ and } N^{-1}\sum_{i=1}^{N} \lambda_i\lambda_i' \text{ is diagonal,} \tag{11}$$

then the representation in Theorem 2 still holds if we replace H_{NT}^{-1} by I_r. Formally, we have:

Corollary 1. *Under Assumptions 1 and 2 , the following representation holds for each i if the restrictions in (11) are satisfied for large N and T:*

$$J_{H_0}(\lambda_i(\cdot)) \cdot \sqrt{T}[\hat{\lambda}_i(\cdot) - \lambda_i(\cdot)] = T^{-1/2}\sum_{t=1}^{T} \varphi_\tau(X_t - \lambda_i(\cdot)'F_t)H_0'F_t + o_P(1) \text{ in } \ell^\infty(\mathcal{T}). \tag{12}$$

The result above follows directly from Theorem 2 by noting that, as proven in Bai and Ng (2013), $H_{NT}^{-1} - I_r = O_P(\min[N,T]^{-1})$ under restrictions (11).

Theorem 2 also allows us to construct confidence band and make inference for the entire quantile processes if uniform (in τ) consistent estimators of $J_{H_0}(\lambda_i(\tau))$ are available. Following Powell (1986) we consider the the following estimator

$$\hat{J}(\hat{\lambda}_i(\tau)) = \frac{1}{2h_T \cdot T}\sum_{t=1}^{T}\Big\{\mathbf{1}\{|X_{it} - \hat{\lambda}_i(\tau)'\hat{F}_t| \le h_T\}\hat{F}_t\hat{F}_t'\Big\}, \tag{13}$$

and impose the following assumptions:

Assumption 3. *The bandwidth parameter h_T satisfies: $h_T \to 0$ and $h_T \cdot T^{1/2} \to \infty$ and $\|H_{NT} - H_0\|/h_T = o_P(1)$.*

The following result shows that the weak convergence still holds by replacing $J_{H_0}(\lambda_i(\tau))$ with its estimate.

Theorem 3. *Under Assumptions 1 to 3, it holds that $\sup_{\tau\in\mathcal{T}} \big\|\hat{J}(\hat{\lambda}_i(\tau)) - J_{H_0}(\lambda_i(\tau))\big\| = o_P(1)$, and thus for each $i \le N$, $\hat{J}(\hat{\lambda}_i(\cdot)) \cdot \sqrt{T}[\hat{\lambda}_i(\cdot) - H_{NT}^{-1}\lambda_i(\cdot)] \Rightarrow \mathcal{B}_r(\cdot)$ in $\ell^\infty(\mathcal{T})$.*

By Theorem 3, we can construct confidence bands for $H_{NT}^{-1}\lambda_i(\tau)$. For example, when $r = 1$, the α level confidence band is $\hat{\lambda}_i(\tau) \pm T^{-1/2}\hat{J}(\hat{\lambda}_i(\tau))^{-1}C_\alpha$, where C_α is the αth quantile of

$\sup_{\tau\in\mathcal{T}}|\mathcal{B}(\tau)|$. Theorem 3 also implies that for each $i \le N$ and each $\tau \in \mathcal{T}$,

$$[\tau(1-\tau)]^{-1/2} \cdot \hat{J}(\hat{\lambda}_i(\tau)) \cdot \sqrt{T}[\hat{\lambda}_i(\tau) - H_{NT}^{-1}\lambda_i(\tau)] \rightsquigarrow \mathcal{N}(0, I_r).$$

3.3 Discussions

3.3.1 Misspecification

Note that Assumption 1(iii) excludes examples 1 and 4, since in these cases we have $\lambda_i' = \mathbb{E}[\alpha_i, Q_\epsilon(U_{it})] = [\alpha_i, 0]$, so Σ_Λ will have reduced rank. As we discussed above, for these 2 examples, there is only 1 factor f_t in the AFM but there are 2 factors in the QFM ($F_t = [1, f_t]'$ in example 1, and $F_t = [f_t, g_t]'$ in example 4). Therefore, in the first step we can only consistently (up to a scale) estimate f_t, and in the second step the QR of X_{it} on $\hat{f}_t$ will not consistently estimate the quantile factor loadings due to omitted regressors.

While the general effects of omitted regressors in quantile regressions have been studied by Angrist et al. (2006a), in this paper, we are more interested in analyzing how to estimate the quantile factor loadings if the estimated factors in the first step are only consistent for a subspace of the factors in the QFM. Consider the following general location-scale model:

$$X_{it} = \lambda_i' F_t + g_i(F_t)\epsilon_{it},$$

where the first element of F_t is 1, $g_i(F_t) > 0$ with probability 1, and $\{\epsilon_{it}\}$ are as defined in the examples. If the functions g_i are known, our method should still work if the regressors in the second step are set as $\hat{F}_t$ and $g_i(\hat{F}_t)$. When g_i are unknown nonlinear functions, our two-step method generally doesn't work.

When g_i is unknown but is linear up to a vector of unknown parameters, we have

$$X_{it} = \lambda_i' F_t + (\gamma_i' F_t)\epsilon_{it},$$

where $\gamma_i' F_t > 0$ with probability 1. The omitted variable problem arises if there is a $k \le r$ such that: $\lambda_{ik} = 0$ but $\gamma_{ik} \ne 0$ for all i, because in this case the kth factor appears in the QFM but not in the AFM. Note that our method still works if for some k: $\lambda_{ik} \ne 0$ but $\gamma_{ik} = 0$ for all i, because in this case the kth factor appears in the AFM but not in the QFM.

Example 1 belongs to the first case, where the omitted regressors is a constant factor. In general, when the only omitted regressor is the constant factor, we can use $[1, \hat{F}_t']'$ as the regressors in the second step, and as we show in Section 6, it is similar to Theorem 2 to derive the limiting distributions of the estimated quantile factor loading processes. In particular, in Section 6 we show show to test the hypothesis that the factors only have location-shift effect, i.e., the quantile factor loading processes are constant functions of τ.

3.3.2 Cross-sectional quantiles

In a working paper version of Gouriéroux and Jasiak (2008) (GJ hereafter), the author also discussed the concept of quantile factor models, focusing mainly on the cross-sectional quantiles of the observed variables. To compare our models with theirs, consider the following model:

$$X_{it} = \alpha_i f_t + g_t \epsilon_{it}.$$

Instead of treating α_i as fixed parameters, GJ assume that α_i are i.i.d random variables with

$$\begin{pmatrix} \alpha_i \\ \epsilon_{it} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\alpha \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \\ & \sigma_\epsilon^2 \end{pmatrix} \right),$$

and treat f_t and g_t as fixed parameters (i.e., everything is conditional on them). As a result, X_{it} are i.i.d across N for each t, and

$$P[X_{it} \leq x] = P[\alpha_i f_t + g_t \epsilon_{it} \leq x] = \Phi\left(\frac{x - \mu_\alpha f_t}{\sqrt{\sigma_\alpha^2 f_t^2 + \sigma_\epsilon^2 g_t^2}} \right)$$

where Φ is the CDF of the standard normal distribution. Thus, we have the cross-sectional quantile of X_{it} at time t as:

$$Q_{X_t}(\tau) = \Phi^{-1}(\tau) \sqrt{\sigma_\alpha^2 f_t^2 + \sigma_\epsilon^2 g_t^2} + \mu_\alpha f_t,$$

while, in general, they assume that

$$Q_{X_t}(\tau) = \beta(\tau)' G_t \tag{14}$$

for some unknown factors G_t. In the our specific example, $\beta(\tau) = [\Phi^{-1}(\tau), \mu_\alpha]'$ and $G_t = [\sqrt{\sigma_\alpha^2 f_t^2 + \sigma_\epsilon^2 g_t^2}, f_t]$.

By treating α_i as i.i.d and treating f_t and g_t as fixed parameters, $X_{1t}, \ldots, X_{Nt}$ are i.i.d for each t, then it becomes possible to consistently estimate $Q_{X_t}(\tau)$. Furthermore, given a consistent estimator $\hat{Q}_{X_t}(\tau)$, it is easy to get a consistent estimator $\hat{G}_t$ for the space of G_t, using equation (14).two-step

However, compared to ours, their approach has two main limitations. On the one hand, it is impossible to consistently estimate the space of $[g_t, f_t]$, which is the true object of interest, because the factors $G_t = [\sqrt{\sigma_\alpha^2 f_t^2 + \sigma_\epsilon^2 g_t^2}, f_t]$ for the cross-sectional quantiles do not span the linear space of $[g_t, f_t]$. So, even if G_t could be consistently estimated, the space of $[g_t, f_t]$ cannot. Notice that this is the case under the aforementioned very strong distributional assumptions, and the relationship between G_t and $[g_t, f_t]$ could even be more complicated under other distributional assumptions. On the other hand, it is impossible to estimate the loadings α_i, or the quantile factors loadings defined as in our paper: $[\alpha_i, Q_\epsilon(\tau)]$.

4 Extensions

4.1 A Solution When the Two-Step Approach Fails

The two-step approach relies on the assumption that a QFM can be transformed into an AFM, from which the factors can be extracted as the PC estimators. One key restriction is Assumption 1(iii), which requires that the factors shifting the quantiles of X should also shift the means of X. In Example 4, $X_{it} = \alpha_i f_t + g_t \epsilon_{it}$, but $\lambda_i = \mathbb{E}[\lambda_i(U_{it})] = [\alpha_i, \mathbb{E}[Q_\epsilon(U_{it})]]' = [\alpha_i, 0]'$, thus Assumption 1(iii) is violated. As a result, the factor g_t, which shifts the quantiles but not the means of X, can not be recovered from the first step PC estimators. In general, the major limitation of our two-step approach is that the first step cannot consistently estimate the factors that only shift the quantiles but not the means.

However, note that in Example 4, if we further assume that either g_t is a function of f_t or g_t is independent of f_t and ϵ_{it}, and that the median of ϵ_{it} is 0, we have for each $\tau \in (0,1)$ and $\tau \neq 0.5$,

$$Q_{X_{it}}(\tau|F_t) = \alpha_i f_t + Q_\epsilon(\tau) g_t$$

where $F_t = [f_t, g_t]'$, and the loadings $\lambda_i(\tau) = [\alpha_i, Q_\epsilon(\tau)]'$ satisfy Assumption 1(iii) if α_i have enough cross-sectional variations. Even though in the AFM form of the model, the factor g_t plays no special role since it does not shift the means of X, the above expression implies that the quantiles of X across individuals at each τ are informative about the factor g_t. Now consider the following general step up for a given $\tau \in (0,1)$:

$$X_{it} = \lambda_i(\tau)' F_t(\tau) + U_{it}, \tag{15}$$

where the errors U_{it} satisfy

$$P[U_{it} \leq 0 | F_t(\tau)] = \tau,$$

and $\lambda_i(\tau)$ and $F_t(\tau)$ are $r(\tau) \times 1$ vector of factor loadings and factors at quantile τ. As in model (1), we allow the factor loadings to be different not only across i but also across τ, but the main difference is that in model (15) the factors and the number of factors are also allowed to differ across τ. This new set up includes Example 4 since, when $\tau \neq 0.5$, we have $r(\tau) = 2$ and $F_t(\tau) = [f_t, g_t]'$, and when $\tau = 0.5$, we have $r(\tau) = 1$ and $F_t(\tau) = f_t$. In the two-step approach, the first step focuses on the case $\tau = 0.5$ (since the mean and the median of ϵ_{it} are both 0), where the factor g_t is treated as part of the idiosyncratic errors. In order to recover the factor g_t, we need to consider the quantiles of X at other τs, and go beyond the AFM and the standard

PC estimators for the factors.

Just as the standard QR replaces the least-square object function by the check function, we consider the following object function where the check function replaces the least-square objection function of the PC estimators:

$$S(F,\Lambda,\tau)=\sum_{i=1}^{N}\sum_{t=1}^{T}\rho_\tau(X_{it}-\lambda_i'F_t),$$

The estimated factors $\hat{F}(\tau,k)=[\hat{F}_1(\tau,k),\ldots,\hat{F}_T(\tau,k)]'$ and estimated factor loadings $\hat{\Lambda}(\tau,k)=[\hat{\Lambda}_1(\tau,k),\ldots,\hat{\Lambda}_N(\tau,k)]'$ at quantile τ are defined as

$$[\hat{F}(\tau,k),\hat{\Lambda}(\tau,k)]=\underset{F\in\mathbb{R}^{T\times k},\Lambda\in\mathbb{R}^{N\times k}}{\arg\min}\ S(F,\Lambda,\tau), \tag{16}$$

where k is a predetermined positive integer. Unlike the PC estimators, the optimal factors for problem (16) with given loadings do not have a closed form expression. The analysis of the asymptotic properties of $\hat{F}(\tau,k)$ and $\hat{\Lambda}(\tau,k)$ is particularly challenging even when the true number of factors $r(\tau)$ is known, mainly due to the non-smoothness of the object function and the increasing dimension of the parameters.

There are some recent studies related to problem (16) in the literature of panel data models. Fernández-Val and Weidner (2015) and Chen et al. (2014) consider bias-corrected fixed-effects estimators for nonlinear panel data models with both individual and time effects. Similar to our QFM, their models contain $N+T$ incidental parameters, but their object functions are assumed to be smooth and strictly concave. Kato and Galvao (2011) study quantile regressions for panel data models where they replace the check functions by some smooth functions, but in their model there are only N incidental parameters. Problem (16) features both a non-smooth object function and $(N+T)*k$ incidental parameters. In the next subsection, along the lines of Kato and Galvao (2011), we show how to overcome the first difficulty by smoothing the object function $S(F,\Lambda,\tau)$, as well as provide a consistency result for the estimated factors and factor loadings. In the rest of this subsection, we describe a simple and fast computational algorithm for problem (16).

Starting with any $T\times k$ matrix $\hat{F}^{(1)}$ (for notational simplicity we omit the dependence of the estimated factors and loadings on τ and k), the estimated factors in problem (16) can be obtained using the following iterative procedure:

1. Given $\hat{F}^{(m)}=[\hat{F}_1^{(m)},\ldots,\hat{F}_T^{(m)}]$, using QR of X_{it} on $\hat{F}^{(m)}$ to estimate $\hat{\Lambda}_i^{(m+1)}$ for $i=1,\ldots,N$.
2. Given $\hat{\Lambda}^{(m+1)}=[\hat{\Lambda}_1^{(m+1)},\ldots,\hat{\Lambda}_N^{(m+1)}]$, using QR of X_{it} on $\hat{\Lambda}^{(m+1)}$ to estimate $\hat{F}_t^{(m+1)}$ for $t=1,\ldots,T$.
3. Repeat Steps 1 and 2 until $\hat{F}^{(k)}$ and $\hat{F}^{(k+1)}$ are close enough.

4.2 A Smoothed Version of the Solution

The non-smoothness of the object function in problem (16) makes it difficult to analyse the asymptotic property of the estimators. To overcome this difficulty, we use the idea of Horowitz (1998) to smooth the object function. In particular, let $K(u)$ be a kernel function, and define

$$G(u) = 1 - \int_{-\infty}^{u} K(s)ds,$$

then the indicator function $\mathbf{1}\{u \leq 0\}$ can be approximated by $G_{c_{NT}}(u) = G(u/c_{NT})$, where c_{NT} is a sequence of positive numbers that goes to 0 as N and T get large. Define the object function as follows:

$$S^*(F, \Lambda, \tau) = \sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it} - \lambda_i' F_t)[\tau - G_{c_{NT}}(X_{it} - \lambda_i' F_t)]$$

Following Bai (2009), we estimate the realizations of the quantile factors by treating them as fixed parameters. So, from now on, let F^0 denote a particular realization of the the random quantile factors.[3] Let $\mathcal{F}$ and $\mathcal{A}$ be subsets of $\mathbb{R}$ such that $F_{tj}^0(\tau) \in \mathcal{F}$ and $\lambda_{ij}^0(\tau) \in \mathcal{A}$ for $j = 1, \ldots, r(\tau)$. Then estimators are defined as

$$[\hat{F}(\tau, k), \hat{\Lambda}(\tau, k)] = \underset{F_t \in \mathcal{F}^k, \lambda_i \in \mathcal{A}^k}{\arg\min}\ S^*(F, \Lambda, \tau). \tag{17}$$

Similar to the PC estimator, we impose the following restrictions to avoid rotational indeterminacy:

$$\hat{F}(\tau, k)'\hat{F}(\tau, k)/T = I_k \quad \text{and} \quad \hat{\Lambda}(\tau, k)'\hat{\Lambda}(\tau, k)/N \text{ is diagonal.}$$

It should be noted that Theorem 4 also holds for other identification restrictions. Define $l(z) = z[\tau - G_{c_{NT}}(z)]$, $\partial_z l(z) = \partial l(z)/\partial z$, and $\partial_{z^2} l(z) = \partial^2 l(z)/\partial z^2$. We impose the following assumptions:

Assumption 4. *For a given $\tau \in (0, 1)$,*

(i) There exists a sequence of $r(\tau) \times 1$ factors $\{F_t^0(\tau)\}_{t=1}^T$, and a sequence of $r(\tau) \times 1$ non-random factor loadings $\{\lambda_i^0(\tau)\}_{i=1}^N$, such that $X_{it} = \lambda_i^0(\tau)' F_t^0(\tau) + U_{it}$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T$, and $P[U_{it} \leq 0] = \tau$. Moreover, there exists positive definite matrices $\Sigma_\Lambda(\tau)$ and $\Sigma_F(\tau)$ such that $N^{-1}\sum_{i=1}^N \lambda_i^0(\tau)\lambda_i^0(\tau)' \to \Sigma_\Lambda(\tau)$ and $T^{-1}\sum_{t=1}^T F_t^0(\tau)F_t^0(\tau)' \to \Sigma_F(\tau)$.

(ii) The errors $\{U_{it}\}$ are i.i.d across i, and independent across t. Their density functions $\{\mathsf{f}_{Ut}\}_{t=1}^T$ and the first derivatives $\{\mathsf{f}_{Ut}^{(1)}\}_{t=1}^T$ exist, and there is a finite constant M such $|\mathsf{f}_{Ut}^{(1)}(\cdot)| < M$ and $\mathbb{E}[\partial_z l(U_{it})]^8 < M$ for all t.

[3] Alternatively, we can treat F^0 as random variables but make all assumptions and conclusions conditional on F^0.

(iii) K *is differentiable and* $\int_{-\infty}^{\infty} K(s)ds = 1$, $\int_{-\infty}^{\infty} s^j K(s)ds = 0$ *for* $j = 1, \ldots, d-1$ *and* $C_K := \int_{-\infty}^{\infty} s^d K(s)ds \leq \infty$. $c_{NT} \to 0$ *as* $N, T \to \infty$.

(iv) There exist a sequence of positive constants $\{b_{NT}^*\}$, *such that* $\min_{i \leq N, t \leq T} \partial_{z^2} l(Z_{it}) \geq b_{NT}^*$ *for any* $Z_{it} = X_{it} - \lambda_i(\tau)' F_t(\tau)$ *with* $F_t \in \mathcal{F}^k, \lambda_i \in \mathcal{A}^k$.

(v) $\max\{T^{-5/8}, T^{-1/2}N^{-1/8}, T^{-1/8}N^{-1/4}, c_{NT}^d\}/b_{NT}^* \to 0$ *as* $N, T \to \infty$.

Assumption 4(i) defines the true quantile factors, factor loadings, and the number of factors at $\tau \in (0,1)$. It also requires the quantile factors to be strong. Part (ii) allow us to consider models like Example 4, and it can be relaxed to allow serial dependence of the errors (i.e., α-mixing as in Fernández-Val and Weidner 2015 for each i). Part (iii) requires K to be a dth order kernel, and c_{NT} to be a sequence of positive numbers going to zero. Parts (iv) and (v) impose implicit restrictions on the kernel function and the bandwidth parameter, because b_{NT}^* depends on $l(Z_{it})$ and therefore on c_{NT}. Fernández-Val and Weidner (2015) has a similar assumption with $b_{NT}^* = b^*$ for all N, T. Since we are using the function $l(z)$ to approximate the check function whose second derivatives are 0 (except at $z = 0$), it is impossible to bound $\min_{i \leq N, t \leq T} \partial_{z^2} l(Z_{it})$ below by a constant positive number. Instead, we allow the lower bound to be a positive number depending on N and T.

Let $P_C = C(C'C)^{-1}C'$ denotes the project matrix for C. Let $P_{\hat{F}(\tau)} = P_{\hat{F}(\tau, r(\tau))}$ and $P_{\hat{\Lambda}(\tau)} = P_{\hat{\Lambda}(\tau, r(\tau))}$. Then we can show that

Theorem 4. *Under Assumption 4,*

$$\left\| P_{\hat{F}(\tau)} - P_{F^0(\tau)} \right\|^2 = o_P(1), \text{ and } \left\| P_{\hat{\Lambda}(\tau)} - P_{\Lambda^0(\tau)} \right\|^2 = o_P(1).$$

Theorem 4 implies that if we know the true number of quantile factors at τ, the estimated factors given by (17) span the same space of the true quantile factors in the sense that, if one regress a $T \times 1$ vector Y with $\|Y/\sqrt{T}\| = O_P(1)$ on the estimated factors and the true factors separately, and let $\hat{Y}_{\hat{F}}$ and $\hat{F}_{F^0}$ denote the two sets of fitted values, then we have

$$\frac{1}{T} \sum_{t=1}^{T} (\hat{Y}_{\hat{F},t} - \hat{Y}_{F^0,t})^2 = o_P(1).$$

Moreover, an important intermediate result in proving Theorem 4 is

$$(NT)^{-1/2} \|\hat{F}(\tau, k)\hat{\Lambda}(\tau, k)' - F^0 \Lambda^{0'}\| = o_P(1) \text{ for } k \geq r,$$

which implies that the common components of the quantiles of X can be consistently estimated on average as long as $k \geq r$.

The computation of problem (17) can be implemented using the similar iteration procedure for problem (16), the only difference is that in Steps (1) and (2) we need to use nonlinear maximization instead of QR.

5 Simulations

5.1 The Estimation of Quantile Factor Loadings

To evaluate the finite sample performance of our two-step estimator, we consider the following data generating processes (DGP) with 1 common factor:

$$X_{it} = \lambda_i F_t + F_t \epsilon_{it},$$

where λ_i and ϵ_{it} are drawn independently from $\mathcal{N}(0,1)$. F_t is generated by $F_t = e^{\sigma Z_t}$, where Z_t are independent standard normal variables, and $\sigma = 0.7$ such that $\mathbb{E}(X) \approx 1.28$ and $Var(X) \approx 1$. This DGP implies a linear quantile factor model of form (1) with $\lambda_i(\tau) = \lambda_i + \Phi^{-1}(\tau)$, where $\Phi(\cdot)$ is the CDF of standard normal distribution. The histograms of $[\tau(1-\tau)]^{-1/2} \cdot \hat{J}(\hat{\lambda}_1(\tau)) \cdot \sqrt{T}[\hat{\lambda}_1(\tau) - H_{NT}^{-1}\lambda_1(\tau)]$ from 5000 simulations are plotted together with the density function of $\mathcal{N}(0,1)$. We consider the sample sizes with $T = 100, 200$ and $N = 100, 200, 500, 1000$, and the estimates at quantiles $\tau = 0.1, 0.25, 0.5, 0.75, 0.9$. The results are reported in Figures 1 to 5. It can be observed that, as expected, the histograms of the constructed statistics are close to the density functions of standarized normal variables as N and T get large. It should be noticed that the approximations are more accurate for the quintiles at the middle than in the tails. Moreover, as a common problem in nonparametric density estimations, the bandwidth parameter h_T has a significant effect on the distributions of the statistics. In our simulations we simply set $h_T = T^{-1/3}$, so there should be enough room for improvements if one allows h_T to be data-dependent.

5.2 PC v.s. Iteration Procedures

To illustrate the advantage of iterative procedures compared to PC estimators, we use simulations where the data sets are generated as in the location-scale model of Example 4: $X_{it} = \alpha_i f_t + g_t \epsilon_{it}$, where $f_t \sim i.i.d\ \mathcal{N}(0,1)$, $g_t = e^{h_t}$ with $h_t \sim i.i.d\ \mathcal{N}(0, 0.5)$ and $\epsilon_{it} \sim i.i.d\ \mathcal{N}(0,1)$. Ideally, we would expect the iterative procedures to capture the two factors f_t and g_t at $\tau \neq 0.5$, while the PC estimators would only extract f_t. Table 1 reports the values of R^2 of regressing f_t and g_t on the two estimated factors $\hat{F}_{PC}$, $\hat{F}_{QR}$ and $\hat{F}_{SQR}$ using PC (columns 2 and 3), and the iterative approaches (un-smoothed and smoothed version, columns 4 to 7) respectively. It is evident that the factor g_t, which only shifts the scales but not the means of X, is captured by the iterative procedures at $\tau = 0.25, 0.75$ but not by the PC estimators. Also note that, as discussed above, there is only one quantile factor f_t at $\tau = 0.5$ due to the symmetry of the distribution, so the iterative procedures are unable to recover the factor g_t. In particular, we can observe that the smoothed version of the proposed iterative procedure given the symmetry of the error term performs as well as the un-smoothed version in recovering the quantile factor g_t.

Table 1: PC v.s. Iteration Procedure

	$f,\hat{F}_{PC}$	$g,\hat{F}_{PC}$	$f,\hat{F}_{QR}$	$g,\hat{F}_{QR}$	$f,\hat{F}_{SQR}$	$g,\hat{F}_{SQR}$
$N,T=20,\tau=0.25$	.8940	.2705	.8350	.8370	.8868	.8251
$N,T=50,\tau=0.25$	.9596	.1662	.9363	.9287	.9399	.9242
$N,T=100,\tau=0.25$	.9802	.1058	.9621	.9626	.9617	.9604
$N,T=20,\tau=0.5$	.8940	.2705	.9060	.2064	.9152	.2511
$N,T=50,\tau=0.5$	.9596	.1662	.9540	.1535	.9536	.1362
$N,T=100,\tau=0.5$	.9802	.1058	.9716	.1065	.9713	.0967
$N,T=20,\tau=0.75$	.8940	.2705	.7592	.8347	.8684	.8338
$N,T=50,\tau=0.75$	.9596	.1662	.9354	.9314	.9402	.9279
$N,T=100,\tau=0.75$	.9802	.1058	.9638	.9624	.9639	.9601

6 Applications

In this section we consider the applications of our methods to three datasets. The first dataset consists of monthly returns of all US common stocks from 1980 to 2014, while the second one includes monthly returns of all US mutual funds from 2000 to 2014. Both datesets are obtained from the Center for Research in Security Prices (CRSP). We eliminate all the stocks and mutual funds with missing values in those periods, and the resulting datasets have dimensions $N=475, T=420$ for the stocks and $N=2419, T=180$ for the mutual funds. The third dataset contains the excess returns of the well known 100 portfolios ($N=100$) constructed by Fama and French from 1985 to 2012 ($T=324$).

6.1 Applying the two-step approach

We first apply the eigenvalue-ratio estimator of Ahn and Horenstein (2013) for the number of factors and find that $\hat{r}=1$ for both stock returns and mutual fund returns and $\hat{r}=3$ for the Fama-French (FF) portfolios. The last result confirms Fama French (1993)'s well-known result that a large proportion of the variance in the portfolio returns can be explained by three common factors.

Notice that, since the estimated factors for the stock and mutual fund returns are obviously time varying, and that the estimated factors for the FF portfolios do not contain a constant factor, in the second step QR we include a constant term along the regressors. The estimated N

quantile factor loadings processes for the constant and the factors are plotted in Figures 6 for the stock returns, in Figure 7 for the mutual fund returns, and in Figure 8 for FF portfolios. As can be inspected, for both datasets, factor loadings for the constant term change across quantiles, but the factor loadings for the estimated factors seem to be stable, implying a standard factor model as Example 1, i.e., the common factors have only location-shift effect in the two datasets. However, to formally justify this hypothesis, we need to implement a test for the constancy of the quantile factor loading process for each of the returns in all three datasets.

6.2 Testing for constancy of QFM loadings

For simplicity, we focus on the case where the number of PC factors is 1, as in the first two datasets, though the following results can be easily generalized to models with more than 1 factors. The aforementioned estimation results lead us to consider the following location-scale shift model:

$$X_{it} = \lambda_i F_t + (1 + \gamma_i F_t)\epsilon_{it}.$$

With $\mathbb{E}[\epsilon_{it}|F_t] = 0$, this is a conditional mean factor model with one common factor. In this model, the hypothesis of constant factor loadings across quantiles is equivalent to

$$H_0 : \gamma_i = 0.$$

Letting $\hat{\lambda}_i(\cdot)$ be the estimated factor loading process for individual i, then it is natural to consider the process $\hat{\lambda}_i(\cdot) - \lambda_i$. However, since λ_i is unknown, we should base our test on the process $\hat{\lambda}_i(\cdot) - \hat{\lambda}_i(0.5)$, where $\hat{\lambda}_i(0.5)$ can be replaced by any consistent estimator of λ_i under H_0. Let us assume that:

Assumption 5. *(i) The sequence $\{F_t\}$ is strictly stationary and m-dependent with $\mathbb{E}|F_t|^8 < \infty$, and $1 + \gamma_i f > 0$ for all i and all f in the support of F_t; (ii) $N^{-1}\sum_{i=1}^{N} \lambda_i^2 \to \sigma_\lambda^2 > 0$ as $N \to \infty$; (iii) The errors $\{\epsilon_{it}\}$ are i.i.d with CDF F_ϵ and they are independent of the factor. The quantile function $Q_\epsilon(\tau) = \inf\{c : \mathsf{F}_\epsilon(c) \leq \tau\}$ is well defined, and the density function f_ϵ is bounded and uniformly continuous.*

Let $\hat{F}_t$ be the PC estimator of F_t, and define

$$\hat{\theta}_i(\tau) = [\hat{\alpha}_i(\tau), \hat{\lambda}_i(\tau)]' = \arg\min_{\theta \in \mathbb{R}^2} \sum_{t=1}^{T} \rho_\tau(X_{it} - \theta'(1, \hat{F}_t)) \text{ for all } \tau \in \mathcal{T},$$

then as in the proof of Theorem 2, we can show that under H_0 and the above assumption

$$\mathsf{f}_\epsilon(Q_\epsilon(\cdot)) \cdot (1 - h_0^2(\mathbb{E}F_t)^2)^{1/2} \cdot \sqrt{T}(\hat{\lambda}_i(\cdot) - h^{-1}\lambda_i(\cdot)) \Rightarrow \mathcal{B}(\cdot) \text{ in } \ell^\infty(\mathcal{T}),$$

where $\lambda_i(\tau) = \lambda_i + \gamma_i Q_\epsilon(\tau)$, $h = (N^{-1}\sum_{i=1}^N \lambda_i^2)(T^{-1}\sum_{t=1}^T F_t \hat{F}_t)/v$, $h_0 = (\mathbb{E}[F_t^2])^{-1/2}$, and v is the largest eigenvalue of $(NT)^{-1}XX'$. It then follows that under H_0

$$\begin{aligned}\hat{v}_T(\cdot) &= \mathsf{f}_\epsilon(Q_\epsilon(\cdot)) \cdot (1 - h_0^2(\mathbb{E}F_t)^2)^{1/2} \cdot \sqrt{T}(\hat{\lambda}_i(\cdot) - \hat{\lambda}_i(1/2)) \\ &= \mathsf{f}_\epsilon(Q_\epsilon(\cdot)) \cdot (1 - h_0^2(\mathbb{E}F_t)^2)^{1/2} \cdot \sqrt{T}(\hat{\lambda}_i(\cdot) - h^{-1}\lambda_i) + \mathsf{f}_\epsilon(Q_\epsilon(\cdot)) \cdot O_P(1), \quad (18)\end{aligned}$$

where the first term on the right converges weakly to a Brownian bridge, and the second term depends on the distribution of ϵ_{it}, which is usually unknown. Note that the second term, which makes the standard Kolmogorov-Smirnov (KS) test $\sup_{\tau\in\mathcal{T}} |\hat{v}_T(\tau)|$ invalid, is due to the estimation of the unknown parameter λ_i, which is known in the literature as Durbin's problem (See Durbin (1973) and Koenker and Xiao 2002). Following Koenker and Xiao (2002), we use the Khmaladze transformation to purge the estimation effects and get a nuisance-parameter free test.

To do so, let us define

$$g(\tau) = [\tau, \mathsf{f}_\epsilon[Q_\epsilon(\tau)]]' \quad \dot{g}(\tau) = dg(\tau)/d\tau \quad C(\tau) = \int_\tau^1 \dot{g}(s)\dot{g}(s)'ds,$$

and assume that

Assumption 6. *(i)* $\int_0^1 \left|(\dot{\mathsf{f}}_\epsilon/\mathsf{f}_\epsilon)(Q_\epsilon(\tau))\right|^2 d\tau < \infty$*; (ii) the function* $(\dot{\mathsf{f}}_\epsilon/\mathsf{f}_\epsilon)(Q_\epsilon(\tau))$ *is not a constant in the neighbourhood of* 1*.*

Consider the following transformed process

$$\tilde{v}_T(\tau) = \Phi_g(\hat{v}_T) = \hat{v}_T(\tau) - \int_0^\tau \left(\dot{g}'(s)C^{-1}(s)\int_s^1 \dot{g}(t)d\hat{v}_T(t)\right)ds. \quad (19)$$

Essentially, the linear operator Φ_g projects out the functions belonging to the space of $g(\tau)$. Formally, we have:

Proposition 1 (Koenker and Xiao (2002))**.** *Under Assumptions 5 and 6, we have* $\tilde{v}_T(\cdot) = \Phi_g(\hat{v}_T)(\cdot) \Rightarrow \mathcal{W}(\cdot)$ *in* $\ell^\infty(\mathcal{T})$ *under* H_0 *as* $N, T \to \infty$*, where* $\mathcal{W}$ *denotes the Brownian motion process. The results still holds if the function* $\dot{g}$ *is replaced by an estimator* $\dot{g}_T$ *satisfying* $\sup_{\tau\in\mathcal{T}} \|\dot{g}_T(\tau) - \dot{g}(\tau)\| = o_P(1)$*, and* $h_0^2(\mathbb{E}F_t)^2$ *is replaced by* $(T^{-1}\sum_{t=1}^T \hat{F}_t)^2$*.*

To give a computationally feasible formula for the test statistics, let $\tau_1 = a_0 < a_1 < a_2 < \cdots < a_m < a_{m+1} = \tau_2$ be a partition of $\mathcal{T}$, and define

$$C_k = \sum_{j=k}^m \dot{g}(a_j)\dot{g}(a_j)'(a_{j+1} - a_j), \quad D_k = \sum_{j=k}^m \dot{g}(a_j)(\hat{v}_T(a_{j+1}) - \hat{v}_T(a_j))' \quad \text{for } k = 0, \ldots, m-1,$$

$$\tilde{v}_T(a_h)' = \hat{v}_T(a_h)' - \sum_{k=0}^{h} \dot{g}(a_k)' C_k^{-1} D_k(a_{k+1} - a_k) \quad \text{for } h = 0, \ldots, m-1,$$

and we consider the following KS test statistics:

$$\sup_{0 \le j \le m} \frac{\|\tilde{v}_T(a_j) - \tilde{v}_T(a_0)\|}{\sqrt{\tau_2 - \tau_1}}.$$

In practice, we replace $\dot{g}(\tau)$ by an uniform consistent estimator and replace $h_0^2(\mathbb{E}F_t)^2$ by $(T^{-1}\sum_{t=1}^{T}\hat{F}_t)^2$. In light of Proposition 1, the test statistics converges in distribution to $\sup_{\tau\in[0,1]}|\mathcal{W}(\tau)|$ by the continuous mapping theorem. The numbers of variables (percentages in bracket) for which H_0 cannot be rejected are reported in Table 2 for all three datasets, using critical values at 1%, 5% and 10% significance levels. As can be observed, the null hypothesis of a simple location-shift model cannot be rejected for almost 80% of the returns in all three datasets, whereas for the remaining 20% there is evidence of a location-scale model with a single factor.

Table 2: Testing for constant quantile factor loading processes.

Critical Values	10%(1.94)	5%(2.22)	1%(2.80)
Common Stocks	303(70.63%)	341(79.49%)	382(89.04%)
Mutual Funds	1894(80.05%)	1994(84.28%)	2086(88.17%)
FF portfolios	266(82.10%)	287(88.58%)	295(91.04%)

6.3 Iteration Approach and Model Checks

As discussed in Section 3.3, the two-step approach only works when the first-step PC estimators of the factors are consistent for the space of all quantile factors. It there were extra quantile factors that cannot be consistently estimated by PC estimators, then the estimated quantile factor loading process may not be consistent due to omitted factors (see Angrist et al 2006 for more details), hence the tests based on such estimates may be misleading. In the special case of location-shift factor models such as in Example 1, the omitted factor is a constant, and the problem can be easily solved by adding a constant in the regressors in the second-step QR, as we did in Section 3.1. However, when the missing factor is not a constant, there is no easy solution and we have to rely on the iterative approach discussed in Section 4 to recover all the quantile factors.

More importantly, in the spirit of Hausman's test, the iterative approach provides a simple heuristic way of testing whether the two-step approach works by checking whether the PC estimators misses any factors other than the constant factor. To simplify the discussions, let $\hat{F}_{PC}$ and $\hat{F}_{QR}(\tau)$ denote the estimated factors using PC and the iterative approach at τ respectively.

First, note that if Assumptions 1 and 4 are reasonable approximations of the true model, $\hat{F}_{PC}$ should be close to the space of the location-shift factors (or mean factors), which should also be close to a subspace of $\hat{F}_{QR}(\tau)$, because the quantile factors may also include some other factors (such as the constant factor or scale-shift factors) except the location-shift factors. Therefore, running linear regressions of $\hat{F}_{PC}$ on $\hat{F}_{QR}(\tau)$ should yield a R^2 close to 1. Second, $\hat{F}_{QR}(\tau)$ should be close to the space of all quantile factors, including the factors that cannot be captured by $\hat{F}_{PC}$. Therefore, if the factor missed by $\hat{F}_{PC}$ is a constant factor, running linear regressions of a constant on $\hat{F}_{QR}(\tau)$ will result in R^2 close to 1 for most τs (for instance in Example 4 there is no constant factor at $\tau = 0.5$).

In sum, our model checks can be based on the two R^2 mentioned above. On the one hand, if the first R^2 is small for many τs, the validity of Assumptions 1 and 4 should be questioned. On the other hand, if the first R^2 is close to 1 for all τs, but the second R^2 is low for many τs, then there is evidence for the existence of extra quantile factors that are not constant. [4]

To provide an empirical illustration of the discussion above, we first generate a simple location-shift factor model as in Example 1 with $N = T = 200$, and plot the two R^2 across τs in the upper left panel of Figure 9. Recall that in Example 1, the PC estimator consistently estimates the common factors f_t, but there is an extra quantile factor – the constant, for all τs except $\tau = 0.5$. As predicted, the first R^2 (red) is close to 1 for all $\tau = 0.1, \ldots, 0.9$, and the second R^2 (blue) is also close to 1 except at $\tau = 0.45, 0.5, 0.55$. Although $Q_\epsilon(\tau)$, the factor loadings for the constant factor at τ, are different from zero at $\tau = 0.45, 0.55$, $|Q_\epsilon(\tau)| = 0.1257$ are small compared with the factor loadings of f_t, so that $\hat{F}_{QR}$ can only span the space of f_t in finite samples.

Using the two sets of R^2 calculated from the simulated dataset as the benchmark, we can get such R^2 for each dataset and have an idea of how the two-step approach works, i.e., whether there exists extra quantile factors of our financial returns datasets that cannot be captured by PC estimators. When using the iterative approach to estimate the quantile factors, we set $r = 2$ for the stock and mutual fund returns and $r = 4$ for FF portfolio returns. The two sets of R^2 for each dataset are plotted in Figure 9. The first set of R^2 (red) for the mutual fund returns and FF portfolio returns are both close to 1 for all τs, indicating that Assumptions 1 and 4 are satisfied in both datasets indicating that both $\hat{F}_{PC}$ and $\hat{F}_{QR}$ are consistent for the space of the location-shift factors. As for the stock returns, the first set of R^2 has an inverted U shape, with very low values at tail quantiles, which implies that either Assumption 1 or Assumption 4 is not satisfied such that neither $\hat{F}_{PC}$ nor $\hat{F}_{QR}$ is close the true factor space. The second set of R^2 (blue) for the mutual fund returns has a similar pattern as in the benchmark models, with

[4] It could also be the case that all the quantile factors are also location-shift factors which can be consistently estimated by $\hat{F}_{PC}$. But this case is unlikely in our applications because it implies that in the two-step approach the estimated quantile factor loading process for the constant are all 0, which is obviously different from what is shown in Figures 6 to 8.

values close to 1 at tail quantiles and very low values around $\tau = 0.5$. Finally, when considering the FF portfolio returns the second set of R^2 also has a similar pattern as in the benchmark model but with all the values being far from 1.

Overall, the evidence discussed above indicates that the extra factor for the mutual fund returns is just the constant factor, so that the two-step approach works well for this dataset, while there might exist extra non-constant factors for the FF portfolio returns and that such factors cannot be consistently estimated by PC.

7 Conclusions

In this paper we propose the concept of QFM, and consider a two-step procedure to estimate the common factors and the quantile factor loading processes. At a first pass, a useful weak convergence result for the entire estimated quantile factor loading processes is obtained. This result provides the basis for testing various interesting hypotheses about the effects of the common factors on the distributions of the observed variables, as illustrated in the applications.

Yet, when there exists common factors that affect the quantiles but not the means, the two-step procedure may result in inconsistent estimators due to omitted variables, because the PC estimators in the first step cannot capture all the relevant factors for the second-step QR. To solve this problem, we propose an iterative procedure that can successfully extract not only the mean factors but also the quantile factors. Consistency of these estimators is proven for a smoothed version of the iteration procedure.

There still remains several important questions which deserve further research. Firstly, when the two-step procedure works, the number of quantile factors could be consistently estimated using many existing methods. Yet, it is important to have a consistent estimator for the number of quantile factors when the two-step procedure fails (as in Example 4). Second, while our iterative procedure can recover factors that cannot be captured by PC estimators, it is interesting to see how these extra quantile factors can improve macro forecasts compared to current practices based exclusively on factors estimated by PC. Finally, a challenging but interesting problem is to derive the asymptotic distributions of the estimated factors stemming from the iterative procedure.

A Proof of the Theorems

A.1 Proof of Theorem 1

Lemma 1. *Define $C_{NT} = \min[N, T]$, the following results hold under Assumption 1:*
(i) $T^{-1}\sum_{t=1}^{T}\|\hat{F}_t - H_{NT}'F_t\|^2 = O_P(C_{NT}^{-1})$.
(ii) $\|H_{NT} - H_0\| = o_P(1)$.

Proof. Since the errors $\{e_{it}\}$ defined in (6) are uncorrelated across i and t and $E\|F_t\|^4 < \infty$, Assumptions C1 to C4 of Bai and Ng (2002) are trivially satisfied (Note that we don't need the $\mathbb{E}|e_{it}|^8 < \infty$, which are only required to consistently estimate the number of factors). Moreover, By Assumption 1(i) we also have $T^{-1}F_tF_t' - \Sigma_F = o_P(1)$ by applying the law of large numbers, thus Assumptions A and B of Bai and Ng (2002) are also satisfied by our Assumptions 1(i) and (iii).

However, Assumption C5 of Bai and Ng (2002) is not satisfied by our model. Note that this assumption is only needed in proving the following result (we adopt the same notation for convenience)[5]:

$$T^{-1}\sum_{t=1}^{T} b_t = O_P(1/N) \text{ where } b_t = T^{-2}\left\|\sum_{s=1}^{T}\hat{F}_s\zeta_{st}\right\|^2 \text{ and } \zeta_{st} = N^{-1}\sum_{i=1}^{N}\big(e_{it}e_{is} - \mathbb{E}[e_{it}e_{is}]\big).$$

Next, we show that under our Assumption 1, $T^{-1}\sum_{t=1}^{T} b_t = O_P(C_{NT}^{-1})$ and thus part (i) of the desired result, which is Theorem 1 of Bai and Ng (2002), still holds. Note that

$$\sum_{t=1}^{T} b_t = T^{-2}\sum_{t=1}^{T}\left\|\sum_{s=1}^{T}\hat{F}_s\zeta_{st}\right\|^2 \le \left(\frac{1}{T}\sum_{s=1}^{T}\|\hat{F}\|_s^2\right)\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\zeta_{st}^2 = r\cdot\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\zeta_{st}^2,$$

and

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\zeta_{st}^2 = \frac{1}{N}\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1,s\neq t}^{T}\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\big(e_{it}e_{is} - \mathbb{E}[e_{it}e_{is}]\big)\right]^2 + \frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{N}\sum_{i=1}^{N}\big(e_{it}^2 - \mathbb{E}[e_{it}^2]\big)\right]^2 \quad (20)$$

Note that under our assumptions $\mathbb{E}[e_{it}e_{is}e_{jt}e_{js}] - \mathbb{E}[e_{it}e_{is}]\mathbb{E}[e_{jt}e_{js}] = 0$ for any $i \neq j$ and $t \neq s$. Hence, for all $t \neq s$,

$$\mathbb{E}\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\big(e_{it}e_{is} - \mathbb{E}[e_{it}e_{is}]\big)\right]^2 \le \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[e_{it}^2e_{is}^2] \le \max_{1\le i\le N}\mathbb{E}[e_{it}^4] \le \mathbb{E}\|F_t\|^4\cdot\max_{1\le i\le N}\mathbb{E}\|u_{it}\|^4 < \infty$$

since $\mathbb{E}\|F_t\|^4 < \infty$ by Assumption 1(i), and by Assumption 1(iii) there exists a $M < \infty$ such that $\mathbb{E}\|u_{it}\|^p < M$ for all i and for any finite $p > 0$. So the first part on the right hand side of (20) is

[5] In Bai and Ng (2002) the authors consider the estimator $\tilde{F}_t = V_{NT}^{-1}\hat{F}_t$, which does not affect our results because $V_{NT} \to_p V$ and thus $\|V_{NT}\| = O_P(1)$.

$O_P(T/N)$. And when $t = s$,

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(e_{it}^2 - \mathbb{E}[e_{it}^2]\right)\right]^2 \le \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}[e_{it}^2 e_{jt}^2] \le \mathbb{E}\|F_t\|^4 \cdot \max_{1\le i\le N}\mathbb{E}\|u_{it}\|^4 < \infty$$

for all t as shown above. Then the second part on the right hand side of (20) is $O_P(1)$. In sum, we have $T^{-1}\sum_{t=1}^{T} b_t = O_P(1/N) + O_P(1/T) = O_P(C_{NT}^{-1})$, and the other parts of the proof is similar to those in Bai and Ng (2002). Part (ii) Follows directly from Proposition 1 of Bai (2003) and Assumption 1(iv).

Another direct consequence of (i) is that

$$\max_{1\le t\le T}\left\|T^{-1}\sum_{s=1}^{T}\hat{F}_s\zeta_{st}\right\| = \sqrt{\max_{1\le t\le T} b_t} \le \sqrt{\sum_{t=1}^{T} b_t} = O_P(\sqrt{T}/\sqrt{N}) + O_P(1).$$

By a proof similar to that of Proposition 2 in Bai (2003), we can obtain the following useful result:

$$\max_{1\le t\le T}\|\hat{F}_t - H_{NT}'F_t\| = O_P(T^{-1/2}C_{NT}^{-1/2}) + O_P(\alpha_T/T) + O_P(\sqrt{T}/\sqrt{N}) + O_P(1), \tag{21}$$

where $\max_{1\le t\le T}\|F_t\| = O_P(\alpha_T)$. □

Proof of Theorem 1: Define $\mathcal{D} = \{D \in \mathbb{R}^{r\times r} : D > 0 \text{ and } \|D\| < \infty\}$, $\mathbb{Q}_\infty(\tau,\lambda) = \mathbb{E}[\rho_\tau(X_{it} - \lambda' F_t)]$ and $\varphi_\tau(u) = \mathbf{1}\{u < 0\} - \tau$. Under Assumption 1(vi) we have for each τ in $\mathcal{T}$, $\partial\mathbb{Q}_\infty(\tau,\lambda)/\partial\lambda = \mathbb{E}[\varphi_\tau(X_{it} - \lambda' F_t)F_t]$ and $\partial\mathbb{Q}_\infty(\tau,\lambda)/\partial\lambda\partial\lambda' = J(\lambda)$. From (2.1) we have $\partial\mathbb{Q}_\infty(\tau,\lambda_i(\tau))/\partial\lambda = \mathbb{E}[\varphi_\tau(X_{it} - \lambda_i(\tau)'F_t)F_t] = 0$, thus by Assumption 1(v) $\lambda_i(\tau)$ uniquely minimizes $\mathbb{Q}_\infty(\tau,\lambda)$ uniformly over $\mathcal{T}$. It then follows that $D^{-1}\lambda_i(\tau)$ uniquely minimizes $\mathbb{Q}_{\infty,D}(\tau,\lambda) = \mathbb{E}[\rho_\tau(X_{it} - \lambda' D' F_t)]$ uniformly over $\mathcal{T}$ for any $D \in \mathcal{D}$.

Define $\mathbb{Q}_{T,D}(\tau,\lambda) = T^{-1}\sum_{t=1}^{T}\rho_\tau(X_{it} - \lambda' D' F_t)$. Notice that the function $(\tau,\lambda) \mapsto \rho_\tau(x - \lambda' f)$ is continuous for each $x \in \mathcal{X}$ and $f \in \mathcal{F}$, and $|\rho_\tau(X_{it} - \lambda' H_0' F_t)| \le C \cdot \sup_{\lambda\in\mathcal{A}}\|\lambda\| \cdot \|H_0\| \cdot \|F_t\|$ for some constant $C < \infty$ for all $(\tau,\lambda) \in \mathcal{T}\times\mathcal{A}$. Since $\mathbb{E}\|F_t\| < \infty$ and $\mathcal{A}$ is compact by Assumption 1, it follows that

$$\sup_{(\tau,\lambda)\in\mathcal{T}\times\mathcal{A}}\|\mathbb{Q}_{T,H_0}(\tau,\lambda) - \mathbb{Q}_{\infty,H_0}(\tau,\lambda)\| = o_P(1) \tag{22}$$

by invoking Lemma 2.4 of Newey and MaFaden (1994).

Define $\hat{\mathbb{Q}}_T(\tau,\lambda) = T^{-1}\sum_{t=1}^{T}\rho_\tau(X_{it} - \lambda'\hat{F}_t)$. By definition, $\hat{\lambda}_i(\tau)$ is the minimizer of $\hat{\mathbb{Q}}_T(\tau,\lambda)$ over $\mathcal{A}$ for each τ. Note that $\rho_\tau(u - v) - \rho_\tau(u) = v\psi_\tau(u) + \int_0^v(\mathbf{1}\{u < s\} - \mathbf{1}\{u < 0\})ds$, so[6]

$$|\hat{\mathbb{Q}}_T(\tau,\lambda) - \mathbb{Q}_{T,H_0}(\tau,\lambda)| \le C\cdot\|\lambda\|\cdot T^{-1}\sum_{t=1}^{T}\|\hat{F}_t - H_0'F_t\| \le C\cdot\|\lambda\|\cdot\sqrt{T^{-1}\sum_{t=1}^{T}\|\hat{F}_t - H_0'F_t\|^2}$$

for some constant $C > 0$. By Lemma 1 we have $T^{-1}\sum_{t=1}^{T}\|\hat{F}_t - H_0'F_t\|^2 \le T^{-1}\sum_{t=1}^{T}\|\hat{F}_t - H_{NT}'F_t\|^2 + \|H_{NT} - H_0\|^2 \cdot T^{-1}\sum_{t=1}^{T}\|F_t\|^2 = o_P(1)$, it then follows that $\sup_{(\tau,\lambda)\in\mathcal{T}\times\mathcal{A}}|\hat{\mathbb{Q}}_T(\tau,\lambda) - \mathbb{Q}_{T,H_0}(\tau,\lambda)| = o_P(1)$. The latter result together with (22) imply: $\sup_{(\tau,\lambda)\in\mathcal{T}\times\mathcal{A}}|\hat{\mathbb{Q}}_T(\tau,\lambda) - \mathbb{Q}_{\infty,H_0}(\tau,\lambda)| = o_P(1)$. Since $\hat{\lambda}_i(\tau)$

[6]It then follows that $|\rho_\tau(u - v) - \rho_\tau(u)| \le |v|\cdot|\mathbf{1}\{u < 0\} - \tau| + |v|\cdot\mathbf{1}\{|u| < |v|\} \le 3|v|$.

is the minimizer of $\hat{\mathbb{Q}}_T(\tau,\lambda)$ by definition, and $H_0^{-1}\lambda_i(\tau)$ is the unique minimizer of $\mathbb{Q}_{\infty,H_0}(\tau,\lambda)$ as shown above, it then follows from Lemma B.1 of Chernozhukov and Hansen (2006) that $\sup_{\tau\in\mathcal{T}}\|\hat{\lambda}_i(\tau)-H_0^{-1}\lambda_i(\tau)\|=o_P(1)$ for all i. Moreover,

$$\sup_{\tau\in\mathcal{T}}\|\hat{\lambda}_i(\tau)-H_{NT}^{-1}\lambda_i(\tau)\|\leq\sup_{\tau\in\mathcal{T}}\|\hat{\lambda}_i(\tau)-H_0^{-1}\lambda_i(\tau)\|+\|H_{NT}^{-1}-H_0^{-1}\|\cdot\sup_{\tau\in\mathcal{T}}\|\lambda_i(\tau)\|=o_P(1).$$

■

A.2 Proof of Theorem 2

We first prove a key result about the uniform rate of convergence of the estimated factors:

Lemma 2. *Suppose Assumptions 1 and 2 hold, then:*

$$\max_{1\leq t\leq T}\|\hat{F}_t-H_{NT}'F_t\|=O_P(T^{1/8}/\sqrt{N})+O_P(T^{-5/8})=o_P(T^{-1/2}).$$

In our model $\mathbb{E}[e_{it}^2e_{jt}^2]-\mathbb{E}[e_{it}^2]\mathbb{E}[e_{jt}^2]\neq 0$ for $i\neq j$, so the above result the our proof is slightly different from Bai and Ng (2008), who show that $\max_{1\leq t\leq T}\|\hat{F}_t-H_{NT}'F_t\|=O_P(T^{1/8}/\sqrt{N})+O_P(T^{-7/8})$.

Proof. Define the L_P-norm of any random variable Z by $\|Z\|_p=\left(\mathbb{E}|Z|^p\right)^{1/p}$. For any random variables $Z_1,Z_2,\ldots$, we have

$$\Big\|\max_{1\leq t\leq T}Z_t\Big\|_p\leq\Big\|\max_{1\leq t\leq T}|Z_t|\Big\|_p=\Big(\mathbb{E}\Big[\max_{1\leq t\leq T}|Z_t|^p\Big]\Big)^{1/p}\leq\left(\sum_{t=1}^T\mathbb{E}|Z_t|^p\right)^{1/p}\leq T^{1/p}\cdot\max_{1\leq t\leq T}\|Z_t\|_p. \tag{23}$$

A immediate result of this maximal inequality is that $\max_{1\leq t\leq T}\|F_t\|=O_P(T^{1/8})$ if $\mathbb{E}\|F_t\|^8<\infty$.

Following Bai (2003), we have

$$\hat{F}_t-H_{NT}'F_t=V_{NT}^{-1}\Big(\underbrace{\frac{1}{T}\sum_{\substack{s=1\\s\neq t}}^T\hat{F}_s\gamma_{st}}_{a_t}+\underbrace{\frac{1}{T}\hat{F}_t\gamma_{tt}}_{b_t}+\underbrace{\frac{1}{T}\sum_{s=1}^T\hat{F}_s\eta_{st}}_{c_t}+\underbrace{\frac{1}{T}\sum_{s=1}^T\hat{F}_s\xi_{st}}_{d_t}\Big), \tag{24}$$

where

$$\gamma_{st}=\frac{1}{N}\sum_{i=1}^N e_{is}e_{it},\quad \eta_{st}=F_s'\Lambda'e_t/N,\quad \xi_{st}=F_t'\Lambda'e_s/N,$$

$e_t=[e_{1t},\ldots,e_{Nt}]'$, and V_{NT} is defined as in Section 3.1. Define $u_{it}=\lambda_i(U_{it})-\lambda_i$, note that under our assumptions, $e_{it}=u_{it}'F_t$ are uncorrelated across i and t, and $\mathbb{E}|e_{it}|^p\leq\mathbb{E}\|u_{it}\|^p\cdot\mathbb{E}\|F_t\|^p$ for any finite p. Moreover, $\mathbb{E}\|u_{it}\|^p<\infty$ for all i and any finite p because U_{it} are uniformly distributed over $[0,1]$ and $\mathcal{A}$ is compact by our assumptions.

First, by adding and subtracting terms,

$$a_t = \underbrace{\frac{1}{T}\sum_{\substack{s=1\\s\neq t}}^{T}(\hat{F}_s - H'_{NT})\gamma_{st}}_{a_{1t}} + \underbrace{H'_{NT}\frac{1}{T}\sum_{\substack{s=1\\s\neq t}}^{T}F_s\gamma_{st}}_{a_{2t}}.$$

For a_{1t}, we have

$$\|a_{1t}\| \leq \left(\frac{1}{T}\sum_{s=1}^{T}\|\hat{F}_s - H'_{NT}F_s\|^2\right)^{1/2}\left(\frac{1}{T}\sum_{s\neq t}^{T}\gamma_{st}^2\right)^{1/2} = O_P(C_{NT}^{-1/2})\left(\frac{1}{T}\sum_{s\neq t}^{T}\gamma_{st}^2\right)^{1/2},$$

and since $\mathbb{E}[e_{it}e_{it}e_{jt}e_{js}] = 0$ for $t \neq s$ and $j \neq i$,

$$\mathbb{E}\left[\frac{N}{T}\sum_{s\neq t}^{T}\gamma_{st}^2\right]^2 = \mathbb{E}\left\{\frac{1}{T}\sum_{s\neq t}^{T}\left[N^{-1/2}\sum_{i=1}^{N}e_{is}e_{it}\right]^2\right\}^2 \leq \max_{1\leq s\leq T, s\neq t}\mathbb{E}\left[N^{-1/2}\sum_{i=1}^{N}e_{is}e_{it}\right]^4,$$

It is easy to see that for all $s \neq t$,

$$\mathbb{E}\left[N^{-1/2}\sum_{i=1}^{N}e_{is}e_{it}\right]^4 \leq C\cdot\max_{1\leq i\leq N}\mathbb{E}[e_{it}^8] \leq C\cdot\mathbb{E}\|F_t\|^8\cdot\max_{1\leq i\leq N}\mathbb{E}[u_{it}^8] \leq \infty$$

for some finite C by assumptions. Then by the maximal inequality (23) we have $\max_{1\leq t\leq T}\|a_{1t}\| = O_P(C_{NT}^{-1/2}T^{1/4}N^{-1/2})$. For a_{2t}, note that

$$\frac{1}{T}\sum_{\substack{s=1\\s\neq t}}^{T}F_s\gamma_{st} = \frac{1}{\sqrt{NT}}\left(\frac{1}{\sqrt{NT}}\sum_{s=1,s\neq t}^{T}\sum_{i=1}^{N}F_se_{is}e_{it}\right)$$

and

$$\mathbb{E}\left(\frac{1}{\sqrt{NT}}\sum_{s=1,s\neq t}^{T}\sum_{i=1}^{N}F_se_{is}e_{it}\right)^2 \leq \frac{1}{NT}\sum_{s=1,s\neq t}^{T}\sum_{i=1}^{N}\mathbb{E}\left[\|F_s\|^2e_{is}^2e_{it}^2\right] \leq \max_{1\leq i\leq N}\mathbb{E}\|u_{it}\|^4\cdot\mathbb{E}\|F_t\|^6 < \infty$$

for all t by assumption. Then by the maximal inequality (23) we have $\max_{1\leq t\leq T}\|a_{2t}\| = O_P(N^{-1/2})$. In sum, we have $\max_{1\leq t\leq T}\|a_t\| = O_P(N^{-1/2})$ by Assumption 2.

Second, by adding and subtracting terms,

$$b_t = \underbrace{T^{-1}H'_{NT}F_t\left(\frac{1}{N}\sum_{i=1}^{N}e_{it}^2\right)}_{b_{1t}} + \underbrace{T^{-1}(\hat{F}_t - H'_{NT}F_t)\left(\frac{1}{N}\sum_{i=1}^{N}e_{it}^2\right)}_{b_{2t}}.$$

Since

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}e_{it}^2\right]^4 = \frac{1}{N^4}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{p=1}^{N}\sum_{q=1}^{N}\mathbb{E}\left[e_{it}^2e_{jt}^2e_{pt}^2e_{qt}^2\right] \leq \max_{1\leq i\leq N}\mathbb{E}(e_{it}^8) \leq \mathbb{E}\|F_t\|^8\cdot\max_{1\leq i\leq N}\mathbb{E}\|u_{it}\|^8 < \infty$$

for all t, then by the maximal inequality (23)

$$\max_{1\le t\le T}\|b_{1t}\| \le T^{-1}\max_{1\le t\le T}\|F_t\|\cdot O_P(T^{1/4}) = O_P(T^{-5/8})$$

because $\mathbb{E}\|F_t\|^8<\infty$ implies $\max_{1\le t\le T}\|F_t\| = O_P(T^{1/8})$. Moreover, it follows from (21) that $\max_{1\le t\le T}\|b_{2t}\| = T^{-1}O_P(1)O_P(T^{1/4}) = O_P(T^{-3/4})$. In sum, we have $\max_{1\le t\le T}\|b_t\| = O_P(T^{-5/8})$.

Third,

$$c_t = \underbrace{T^{-1}\sum_{s=1}^T(\hat F_s - H_{NT}'F_s)\eta_{st}}_{c_{1t}} + \underbrace{H_{NT}'T^{-1}\sum_{s=1}^T F_s\eta_{st}}_{c_{2t}}.$$

Note that $T^{-1}\sum_{s=1}^T F_s\eta_{st} = N^{-1/2}\big(T^{-1}\sum_{s=1}^T F_sF_s'\big)\big(N^{-1/2}\sum_{i=1}^N\lambda_i e_{it}\big)$, and it is easy to see that $\mathbb{E}\big(N^{-1/2}\sum_{i=1}^N\lambda_i e_{it}\big)^8<\infty$ for all t under our assumptions, then by the maximal inequality (23) we have $\max_{1\le t\le T}\|c_{2t}\| = O_P(T^{1/8}/\sqrt{N})$. Moreover,

$$\|c_{1t}\| \le \left(T^{-1}\sum_{s=1}^T\eta_{st}^2\right)^{1/2}\left(T^{-1}\sum_{s=1}^T\|\hat F_s - H_{NT}'F_s\|^2\right)^{1/2},$$

and

$$T^{-1}\sum_{s=1}^T\eta_{st}^2 = T^{-1}\sum_{s=1}^T(F_s'\Lambda' e_t/N)^2 \le N^{-1}\|\Lambda' e_t/\sqrt{N}\|^2\left(T^{-1}\sum_{s=1}^T\|F_s\|^2\right),$$

since $\mathbb{E}\|\Lambda' e_t/\sqrt{N}\|^8<\infty$ for all t as shown above, by the maximal inequality (23) we have $\max_{1\le t\le T}|T^{-1}\sum_{s=1}^T\eta_{st}^2| = O_P(T^{1/4}/N)$, and thus $\max_{1\le t\le T}\|c_{1t}\| = O_P(T^{1/8}/\sqrt{N})o_P(1)$. In sum, we have $\max_{1\le t\le T}\|c_t\| = O_P(T^{1/8}/\sqrt{N})$.

Finally, by Bai (2003)

$$\|d_t\| \le \frac{1}{\sqrt{N}}\left(T^{-1}\sum_{s=1}^T\|\hat F_s - H_{NT}'F_s\|^2\right)^{1/2}\left(T^{-1}\sum_{s=1}^T\left\|\Lambda' e_s/\sqrt{N}\right\|^2\right)^{1/2}\|F_t\| + \frac{1}{\sqrt{NT}}\left\|\frac{1}{\sqrt{NT}}\sum_{s=1}^T\sum_{i=1}^N F_s\lambda_i' e_{is}\right\|\|F_t\|.$$

It is easy to see that $T^{-1}\sum_{s=1}^T\left\|\Lambda' e_s/\sqrt{N}\right\|^2$ and $\frac{1}{\sqrt{NT}}\sum_{s=1}^T\sum_{i=1}^N F_s\lambda_i' e_{is}$ are both $O_P(1)$ by the stated assumptions, and since $\max_{1\le t\le T}\|F_t\| = O_P(T^{1/8})$, we have $\max_{1\le t\le T}\|d_t\| = O_P(N^{-1/2}C_{NT}^{-1/2}T^{1/8}) + O_P(N^{-1/2}T^{-1/2}T^{1/8}) = O_P(N^{-1/2}T^{-3/8})$. Combining all above results gives: $\max_{1\le t\le T}\|\hat F_t - H_{NT}'F_t\| = O_P(T^{1/8}/\sqrt{N}) + O_P(T^{-5/8})$. □

To simply the notations, we suppress the subscription i and write $X_t, \lambda(\tau), \hat\lambda(\tau)$ instead of $X_{it}, \lambda_i(\tau), \hat\lambda_i(\tau)$. For any $D\in\mathcal{D}$, define

$$\mathbb{S}_{T,D}(\tau,\lambda) = T^{-1}\sum_{t=1}^T\varphi_\tau(X_t - \lambda' D' F_t)D'F_t, \quad \mathbb{S}_{\infty,D}(\tau,\lambda) = \mathbb{E}[\varphi_\tau(X_t - \lambda' D' F_t)D'F_t],$$

$$\text{and } \mathbb{G}_T(\tau,\lambda,D)=\sqrt{T}[\mathbb{S}_{T,D}(\tau,\lambda)-\mathbb{S}_{\infty,D}(\tau,\lambda)].$$

The following lemmas hold under Assumptions 1 and 2:

Lemma 3. *Define* $\hat{\mathbb{S}}_T(\tau,\lambda)=T^{-1}\sum_{t=1}^T\varphi_\tau(X_t-\lambda'\hat{F}_t)\hat{F}_t$, *then* $\sup_{\tau\in\mathcal{T}}\|\sqrt{T}\hat{\mathbb{S}}_T(\tau,\hat{\lambda}(\tau))\|=o_P(1)$.

Proof. First, we have

$$\sup_{1\le t\le T}\|\hat{F}_t\|\le\sup_{1\le t\le T}\|F_t\|\cdot\|H_{NT}\|+\sup_{1\le t\le T}\|\hat{F}_t-H_{NT}F_t\|,$$

so $\sup_{1\le t\le T}\|\hat{F}_t\|=o_P(T^{1/2})$ because $\sup_{1\le t\le T}\|F_t\|=O_P(T^{1/8})$ and $\sup_{1\le t\le T}\|\hat{F}_t-H_{NT}F_t\|$ is $o_P(T^{-1/2})$ by Lemma 2. Then it follows by Theorem 2.1 of Koenker (2005) that $\|\hat{\mathbb{S}}_T(\tau,\hat{\lambda}_i(\tau))\|\le r\cdot\sup_{1\le t\le T}\|\hat{F}_t\|/T+O_P(T^{-1})=o_P(T^{-1/2})$ for all $\tau\in\mathcal{T}$. □

Lemma 4. $\sup_{\tau\in\mathcal{T}}\|\mathbb{G}_T(\tau,\hat{\lambda}(\tau),H_{NT})-\mathbb{G}_T(\tau,H_0^{-1}\lambda(\tau),H_0)\|=o_P(1)$.

Proof. Define the empirical process

$$\tilde{\mathbb{G}}_T(\tau,\theta,D)=\frac{1}{\sqrt{T}}\sum_{t=1}^T\Big\{\varphi_\tau(X_t-\theta'F_t)D'F_t-\mathbb{E}[\varphi_\tau(X_t-\theta'F_t)D'F_t]\Big\},$$

and the compact set $\Theta=\{\theta\in\mathbb{R}^k:\theta=D\lambda,\lambda\in\mathcal{A},D\in\mathcal{D}\}$. By Theorems 2 and 3 of Andrews (1994), it is easy to see that the class of functions $\{(\mathbf{1}\{X_t<\theta'F_t\}-\tau)D'F_t:\tau\in\mathcal{T},\theta\in\Theta,D\in\mathcal{D}\}$ satisfies the Pollard's entropy condition with an square integrable envelop $\sup_{D\in\mathcal{D}}\|D\|\cdot\|F_t\|$, thus by Theorem 1 of Andrews (1994), $\tilde{\mathbb{G}}_T(\tau,\theta,D)$ is ρ-stochastic equicontinuous with the pseudometric:

$$\rho[(\tau_1,\theta_1,D_1),(\tau_2,\theta_2,D_2)]=\sqrt{\max_{1\le j\le r}\mathbb{E}\Big(\varphi_{\tau_1}(X_t-\theta_1'F_t)D_1^{(j\cdot)}F_t-\varphi_{\tau_2}(X_t-\theta_2'F_t)D_2^{(j\cdot)}F_t\Big)^2}$$

where $D^{(\cdot j)}$ denotes the jth column of D. Note that

$$\begin{aligned}&\rho[(\tau_1,\theta_1,D_1),(\tau_2,\theta_2,D_2)]\\ \le\ &\max_{j=1,\dots,r}\sqrt{\tau^2\mathbb{E}[(D_1^{(j\cdot)}-D_2^{(j\cdot)})F_t]^2}+\max_{j=1,\dots,r}\sqrt{\mathbb{E}[\mathbf{1}\{X_t<\theta_1F_t\}(D_1^{(j\cdot)}-D_2^{(j\cdot)})F_t]^2}\\ &+\max_{j=1,\dots,r}\sqrt{\mathbb{E}[(\mathbf{1}\{X_t<\theta_1F_t\}-\mathbf{1}\{X_t<\theta_2F_t\})D_2^{(j\cdot)}F_t]^2},\end{aligned}$$

and the first two terms on the right-hand side of above inequality are bounded by $C\cdot\|D_1-D_2\|\cdot\sqrt{\mathbb{E}\|F_t\|^2}$ for some $C>0$. For the third term, by Hölder's inequality and Assumption 1(vi) we have

$$\begin{aligned}&\max_{j=1,\dots,r}\sqrt{\mathbb{E}[(\mathbf{1}\{X_t<\theta_1F_t\}-\mathbf{1}\{X_t<\theta_2F_t\})D_2^{(j\cdot)}F_t]^2}\\ \le\ &\max_{j=1,\dots,r}(\mathbb{E}|\mathbf{1}\{X_t<\theta_1F_t\}-\mathbf{1}\{X_t<\theta_2F_t\}|)^{\frac{\epsilon}{2(2+\epsilon)}}(\mathbb{E}[D_2^{(j\cdot)}F_t]^{2+\epsilon})^{\frac{1}{2+\epsilon}}\\ \le\ &\|D_2\|\cdot(\mathbb{E}\|F_t\|^{2+\epsilon})^{\frac{1}{2+\epsilon}}\cdot(\mathbb{E}\|F_t\|)^{\frac{\epsilon}{2(2+\epsilon)}}\cdot(\bar{f}\cdot\|\theta_1-\theta_2\|)^{\frac{\epsilon}{2(2+\epsilon)}}.\end{aligned}$$

Then by uniform consistency of $\hat{\lambda}(\tau)$ and Lemma 1(ii),

$$\delta = \sup_{\tau\in\mathcal{T}} \rho[(\tau, \hat{\lambda}(\tau)' H_{NT}', H_{NT}), (\tau, \lambda(\tau)', H_0)] = o_P(1),$$

and thus

$$\begin{aligned} & \sup_{\tau\in\mathcal{T}} \|\mathbb{G}_T(\tau, \hat{\lambda}(\tau), H_{NT}) - \mathbb{G}_T(\tau, H_0^{-1}\lambda(\tau), H_0)\| \\ = & \sup_{\tau\in\mathcal{T}} \|\tilde{\mathbb{G}}_T(\tau, \hat{\lambda}(\tau)' H_{NT}', H_{NT}) - \tilde{\mathbb{G}}_T(\tau, \lambda(\tau)', H_0)\| \\ \leq & \sup_{\rho[(\tau_1,\theta_1,D_1),(\tau_2,\theta_2,D_2)]\leq\delta} \|\tilde{\mathbb{G}}_T(\tau_1, \theta_1, D_1) - \tilde{\mathbb{G}}_T(\tau_2, \theta_2, D_2)\| \end{aligned}$$

which is $o_P(1)$ by the stochastic continuity of $\tilde{\mathbb{G}}_T(\tau, \theta, D)$. □

Lemma 5. *Let $r = [r_1, \ldots, r_T]$, where r_t is a $r \times 1$ vector of real numbers for each t. For any $D \in \mathcal{D}$ and $\lambda \in \mathcal{A}$, define:*

$$\mathbb{U}_{T,D}(\lambda, r) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \Big\{ \big[\mathbf{1}\{X_t - \lambda' D' F_t < 0\} - \mathbf{1}\{X_t - \lambda' D' F_t < \lambda' r_t\}\big] D' F_t \Big\},$$

then $\sup_{\lambda\in\mathcal{A}} \left\| \mathbb{U}_{T,H_{NT}}(\lambda, r)|_{r_t = \hat{F}_t - H_{NT}' F_t} \right\| = o_P(1)$.

Proof. Let $D^{(\cdot j)}$ be the jth column of D, and let $\mathbb{U}_{T,D,j}$ be the jth element of $\mathbb{U}_{T,D}$, i.e.,

$$\mathbb{U}_{T,D,j}(\lambda, r) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \Big\{ \big[\mathbf{1}\{X_t - \lambda' D' F_t < 0\} - \mathbf{1}\{X_t - \lambda' D' F_t < \lambda' r_t\}\big] D^{(\cdot j)} F_t \Big\},$$

then it suffices to show that $\sup_{\lambda\in\mathcal{A}} \left\| \mathbb{U}_{T,H_{NT},j}(\lambda, r)|_{r_t = \hat{F}_t - H_{NT}' F_t} \right\| = o_P(1)$ for each $j = 1, \ldots, r$. We can write:

$$-\mathbb{U}_{T,D,j}(\lambda, r) = \mathbb{U}^1_{T,D,j}(\lambda, r) + \mathbb{U}^2_{T,D,j}(\lambda, r),$$

where

$$\mathbb{U}^1_{T,D,j}(\lambda, r) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \Big\{ \big[\mathbf{1}\{X_t - \lambda' D' F_t < \lambda' r_t\} - \mathbf{1}\{X_t - \lambda' D' F_t < 0\}\big] D^{(\cdot j)} F_t \mathbf{1}\{D^{(\cdot j)} F_t \leq 0\} \Big\},$$

$$\mathbb{U}^2_{T,D,j}(\lambda, r) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \Big\{ \big[\mathbf{1}\{X_t - \lambda' D' F_t < \lambda' r_t\} - \mathbf{1}\{X_t - \lambda' D' F_t < 0\}\big] D^{(\cdot j)} F_t \mathbf{1}\{D^{(\cdot j)} F_t > 0\} \Big\}.$$

Define

$$l_{r,\lambda,T} = \min_{1\leq t\leq T} \lambda' r_t \quad \text{and} \quad u_{r,\lambda,T} = \max_{1\leq t\leq T} \lambda' r_t,$$

and

$$\mathbb{R}^1_{T,D,j}(\lambda, \gamma) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \Big\{ \big[\mathbf{1}\{X_t - \lambda' D' F_t < \gamma\} - \mathbf{1}\{X_t - \lambda' D' F_t < 0\}\big] D^{(\cdot j)} F_t \mathbf{1}\{D^{(\cdot j)} F_t \leq 0\} \Big\},$$

$$\mathbb{R}^2_{T,D,j}(\lambda,\gamma) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\Big\{\big[\mathbf{1}\{X_t-\lambda' D' F_t<\gamma\}-\mathbf{1}\{X_t-\lambda' D' F_t<0\}\big]D^{(\cdot j)}F_t\mathbf{1}\{D^{(\cdot j)}F_t>0\}\Big\},$$

then we have

$$\big\|\mathbb{U}^1_{T,D,j}(\lambda,r)\big\| \le \max\Big[\big\|\mathbb{R}^1_{T,D,j}(\lambda,u_{r,\lambda,T})\big\|,\big\|\mathbb{R}^1_{T,D,j}(\lambda,l_{r,\lambda,T})\big\|\Big],$$

$$\big\|\mathbb{U}^2_{T,D,j}(\lambda,r)\big\| \le \max\Big[\big\|\mathbb{R}^2_{T,D,j}(\lambda,l_{r,\lambda,T})\big\|,\big\|\mathbb{R}^2_{T,D,j}(\lambda,u_{r,\lambda,T})\big\|\Big].$$

Adding and subtracting terms, we have

$$\begin{aligned}
&\mathbb{R}^1_{T,D,j}(\lambda,\gamma)\\
=\;& \underbrace{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\Big\{\mathbf{1}\{X_t-\lambda' D' F_t<\gamma\}D^{(\cdot j)}F_t\mathbf{1}\{D^{(\cdot j)}F_t\le 0\}-\mathbb{E}\Big[\mathbf{1}\{X_t-\lambda' D' F_t<\gamma\}D^{(\cdot j)}F_t\mathbf{1}\{D^{(\cdot j)}F_t\le 0\}\Big]\Big\}}_{\mathbb{W}_{T,D}(\lambda' D',\gamma)}\\
-\;& \underbrace{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\Big\{\mathbf{1}\{X_t-\lambda' D' F_t<0\}D^{(\cdot j)}F_t\mathbf{1}\{D^{(\cdot j)}F_t\le 0\}-\mathbb{E}\Big[\mathbf{1}\{X_t-\lambda' D' F_t<0\}D^{(\cdot j)}F_t\mathbf{1}\{D^{(\cdot j)}F_t\le 0\}\Big]\Big\}}_{\mathbb{W}_{T,D}(\lambda' D',0)}\\
+\;& \underbrace{\sqrt{T}\cdot\mathbb{E}\Big[\mathbf{1}\{X_t-\lambda' D' F_t<\gamma\}D^{(\cdot j)}F_t\mathbf{1}\{D^{(\cdot j)}F_t\le 0\}\Big]-\sqrt{T}\cdot\mathbb{E}\Big[\mathbf{1}\{X_t-\lambda' D' F_t<0\}D^{(\cdot j)}F_t\mathbf{1}\{D^{(\cdot j)}F_t\le 0\}\Big]}_{R_D(\lambda,\gamma)}.
\end{aligned}$$

For simplicity, write $g_t = D^{(\cdot j)}F_t$. First, note that the class of functions

$$\{(\theta,\gamma)\mapsto \mathbf{1}\{X_t-\theta' F_t<\gamma\}\cdot g_t\mathbf{1}\{g_t\le 0\}\}$$

satisfies Pollard's entropy condition with envelop function $\|g_t\|$ and $\mathbb{E}[|g_t|^{2+\epsilon}]<\infty$ according to Theorem 3 of Andrews (1994), then by theorem 1 of Andrews (1994) the empirical process $\mathbb{W}_{T,D}$ defined as above is d-stochastic equicontinuous, where

$$\begin{aligned}
d[(\theta_1,\gamma_1),(\theta_2,\gamma_2)] &= \sqrt{\mathbb{E}\big|[\mathbf{1}\{X_t-\theta_1' F_t<\gamma_1\}-\mathbf{1}\{X_t-\theta_2' F_t<\gamma_2\}]g_t\mathbf{1}\{g_t\le 0\}\big|^2}\\
&\le \|D\|\cdot(\mathbb{E}\|F_t\|^{2+\epsilon})^{\frac{1}{2+\epsilon}}\cdot\big[\bar f(|\gamma_1-\gamma_2|+\|\theta_1-\theta_2\|\cdot\mathbb{E}\|F_t\|)\big]^{\frac{\epsilon}{2(2+\epsilon)}}
\end{aligned}$$

by Hölder's inequality. Second, we have

$$\sup_{\lambda\in\mathcal{A}}|u_{r,\lambda,T}| \le \sup_{\lambda\in\mathcal{A}}\Big|\min_{1\le t\le T}\lambda' r_t\Big| \le \sup_{\lambda\in\mathcal{A}}\max_{1\le t\le T}\Big|\lambda' r_t\Big| \le \sup_{\lambda\in\mathcal{A}}\|\lambda\|\cdot\max_{1\le t\le T}\|r_t\|,$$

and thus

$$\sup_{\lambda\in\mathcal{A}}|\hat u_{\lambda,T}| \le \sup_{\lambda\in\mathcal{A}}\|\lambda\|\cdot\max_{1\le t\le T}\|\hat F_t - H'_{NT}F_t\| = o_P(1)$$

by Lemma 2, where $\hat u_{\lambda,T} = u_{r,\lambda,T}|_{r=\hat F - H'_{NT}F}$. Therefore, $\sup_{\lambda\in\mathcal{A}} d[(\lambda' D',\hat u_{\lambda,T}),(\lambda' D',0)] = o_P(1)$ by the above inequality about d, and $\sup_{\lambda\in\mathcal{A}}\big|\mathbb{W}_{T,D}(\lambda' D',\hat u_{\lambda,T})-\mathbb{W}_{T,D}(\lambda' D',0)\big| = o_P(1)$ by d-stochastic equicontinuity of $\mathbb{W}_{T,D}$.

Next, it is easy to see that $|R_D(\lambda,\gamma)| \leq \bar{f}\cdot\|D\|\cdot\mathbb{E}\|F_t\|\cdot\sqrt{T}|\gamma|$, and so

$$\sup_{\lambda\in\mathcal{A}}|R_D(\lambda,\hat{u}_{\lambda,T}))| \leq \bar{f}\cdot\|D\|\cdot\sup_{\lambda\in\mathcal{A}}\|\lambda\|\cdot\mathbb{E}\|F_t\|\cdot\sqrt{T}\max_{1\leq t\leq T}\|\hat{F}_t - H'_{NT}F_t\| = o_P(1)$$

by Lemma 2. As a result,

$$\sup_{\lambda\in\mathcal{A}}\left|\mathbb{R}^1_{T,D,j}(\lambda,\hat{u}_{\lambda,T})\right| \leq \sup_{\lambda\in\mathcal{A}}\left|\mathbb{W}_{T,D}(\lambda'D',\hat{u}_{\lambda,T}) - \mathbb{W}_{T,D}(\lambda'D',0)\right| + \sup_{\lambda\in\mathcal{A}}|R_D(\lambda,\hat{u}_{\lambda,T}))| = o_P(1),$$

and we can show that $\sup_{\lambda\in\mathcal{A}}\left|\mathbb{R}^1_{T,D,j}(\lambda,\hat{l}_{\lambda,T})\right| = o_P(1)$ in a similar way, which implies that

$$\sup_{\lambda\in\mathcal{A}}\left|\mathbb{U}^1_{T,D,j}(\lambda,r)\right|_{r_t=\hat{F}_t - H'_{NT}F_t} = o_P(1).$$

Similarly, it can be shown that $\sup_{\lambda\in\mathcal{A}}\left|\mathbb{U}^2_{T,D,j}(\lambda,r)\right|_{r_t=\hat{F}_t - H'_{NT}F_t} = o_P(1)$, and finally we can conclude that

$$\sup_{\lambda\in\mathcal{A}}\left|\mathbb{U}_{T,D,j}(\lambda,r)\right|_{r_t=\hat{F}_t - H'_{NT}F_t} \leq \sup_{\lambda\in\mathcal{A}}\left|\mathbb{U}^1_{T,D,j}(\lambda,r)\right|_{r_t=\hat{F}_t - H'_{NT}F_t} + \sup_{\lambda\in\mathcal{A}}\left|\mathbb{U}^2_{T,D,j}(\lambda,r)\right|_{r_t=\hat{F}_t - H'_{NT}F_t} = o_P(1),$$

and the desired result follows by letting $D = H_{NT}$. □

Lemma 6. *For any $D \in \mathcal{D}$, define*

$$\hat{\mathbb{H}}_{T,D}(\tau,\lambda) = \sqrt{T}[\mathbb{S}_{T,D}(\tau,\lambda) - \hat{\mathbb{S}}_T(\tau,\lambda)] = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\{\varphi_\tau(X_t - \lambda'D'F_t)D'F_t - \varphi_\tau(X_t - \lambda'\hat{F}_t)\hat{F}_t\right\},$$

then $\sup_{\tau\in\mathcal{T}}\left\|\hat{\mathbb{H}}_{T,H_{NT}}(\tau,\hat{\lambda}(\tau))\right\| = o_P(1)$.

Proof. Adding and subtracting terms, for any $D \in \mathcal{D}$, we have:

$$\begin{aligned}
\hat{\mathbb{H}}_{T,H_{NT}}(\tau,\lambda) &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\{\varphi_\tau(X_t - \lambda'H'_{NT}F_t)D'F_t - \varphi_\tau(X_t - \lambda'\hat{F}_t)\hat{F}_t\right\} \\
&= \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\{\left[\mathbf{1}\{X_t < \lambda'H'_{NT}F_t\} - \mathbf{1}\{X_t < \lambda'\hat{F}_t\}\right]H'_{NT}F_t\right\} - \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\varphi_\tau(X_t - \lambda'\hat{F}_t)(\hat{F}_t - H'_{NT}F_t) \\
&= \mathbb{U}_{T,H_{NT}}(\lambda,r)|_{r_t=\hat{F}_t - H'_{NT}F_t} - \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\varphi_\tau(X_t - \lambda'\hat{F}_t)(\hat{F}_t - H'_{NT}F_t),
\end{aligned}$$

it then follows from Lemma 2 and 5 that

$$\sup_{\tau\in\mathcal{T}}\left\|\hat{\mathbb{H}}_{T,H_{NT}}(\tau,\hat{\lambda}(\tau))\right\| \leq \sup_{\lambda\in\mathcal{A}}\left\|\mathbb{U}_{T,H_{NT}}(\lambda,r)_{r_t=\hat{F}_t - H'_{NT}F_t}\right\| + 2T^{-1/2}\sum_{t=1}^{T}\|\hat{F}_t - H'_{NT}F_t\| = o_P(1)$$

because $T^{-1/2}\sum_{t=1}^{T}\|\hat{F}_t - H'_{NT}F_t\| \leq \sqrt{T}\max_{1\leq t\leq T}\|\hat{F}_t - H'_{NT}F_t\| = o_P(1)$. □

Proof of Theorem 2: first note that For any $D \in \mathcal{D}$, we have the following expansion for each

$\tau \in \mathcal{T}$:

$$\mathbb{S}_{\infty,D}(\tau,\hat{\lambda}(\tau)) = \mathbb{S}_{\infty,D}(\tau, D^{-1}\lambda(\tau)) + D'\mathbb{E}[\mathsf{f}_X(\lambda^*(\tau)'D'F_t|F_t)F_tF_t']D\cdot[\hat{\lambda}(\tau) - D^{-1}\lambda(\tau)],$$

where $\lambda^*(\tau)$ is on the line connecting $D^{-1}\lambda(\tau)$ and $\hat{\lambda}(\tau)$ for each τ. Then, by uniform continuity of $\mathsf{f}_X(x|f)$ and uniform convergence of $\hat{\lambda}(\tau)$ for $H_{NT}^{-1}\lambda(\tau)$, we have

$$\mathbb{S}_{\infty,H_{NT}}(\tau,\hat{\lambda}(\tau)) = H_{NT}'[J(\lambda(\tau)) + o_P(1)]H_{NT}\cdot[\hat{\lambda}(\tau) - H_{NT}^{-1}\lambda(\tau)] \tag{25}$$

uniformly over $\mathcal{T}$ since $\mathbb{S}_{\infty,D}(\tau, D^{-1}\lambda(\tau)) = 0$.

Second, by definition we have:

$$\sqrt{T}\mathbb{S}_{\infty,H_{NT}}(\tau,\hat{\lambda}(\tau)) = \sqrt{T}\hat{\mathbb{S}}_T(\tau,\hat{\lambda}(\tau)) - \mathbb{G}_T(\tau,\hat{\lambda}(\tau),H_{NT}) + \hat{\mathbb{H}}_{T,H_{NT}}(\tau,\hat{\lambda}(\tau)), \tag{26}$$

and combining Lemmas 1, 4, 6, (25), and (26) gives:

$$[H_0'J(\lambda(\tau))H_0 + o_P(1)]\cdot\sqrt{T}[\hat{\lambda}(\tau) - H_{NT}^{-1}\lambda(\tau)] = -\mathbb{G}_T(\tau, H_0^{-1}\lambda(\tau), H_0) + o_P(1) \tag{27}$$

uniformly in $\tau \in \mathcal{T}$. It then follows from (27) and Assumption 2(iii) that[7]

$$\sup_{\tau\in\mathcal{T}}\| - \mathbb{G}_T(\tau, H_0^{-1}\lambda(\tau), H_0) + o_P(1)\| \geq (\rho^* + o_P(1))\cdot\sup_{\tau\in\mathcal{T}}\sqrt{T}\|\hat{\lambda}(\tau) - H_{NT}^{-1}\lambda(\tau)\|. \tag{28}$$

Since the mapping $\tau \mapsto \lambda(\tau)$ is continuous due to implicit function theorem and Assumption 1(v)[8], the process $\mathbb{V}_T(\cdot) = \mathbb{G}_T(\cdot, H_0^{-1}\lambda(\cdot), H_0)$ is $\tilde{\rho}$-stochastic equicontinuous with

$$\tilde{\rho}[\tau_1,\tau_2] = \rho[(\tau_1, \lambda(\tau_1)'H_0', H_0), (\tau_2, \lambda(\tau_2)'H_0', H_0)]$$

where ρ is defined in Lemma 3. Then by stochastic equicontinuity and a standard multivariate central limit theorem, we have

$$\mathbb{V}_T(\tau) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\varphi_\tau(X_t - \lambda(\tau)'F_t)H_0'F_t$$

converges weakly to a zero mean Gaussian process $\mathbb{V}_\infty(\tau)$ defined by its covariance matrix

$$\Sigma(\tau_1,\tau_2) = \mathbb{E}[\mathbb{V}_\infty(\tau_1)\mathbb{V}_\infty(\tau_2)] = [\min(\tau_1,\tau_2) - \tau_1\tau_2]H_0'\Sigma_F H_0.$$

It then follows from (28) that $\sup_{\tau\in\mathcal{T}}\sqrt{T}\|\hat{\lambda}(\tau) - H_{NT}^{-1}\lambda(\tau)\|$ is $O_P(1)$, and thus from (27) we can conclude that $[H_0'J(\lambda(\cdot)H_0]\cdot\sqrt{T}[\hat{\lambda}(\cdot) - H_{NT}^{-1}\lambda(\cdot)]$ converges weakly to $\mathbb{V}_\infty(\cdot)$ in $\ell^\infty(\mathcal{T})$. The desired result follows by noting that $H_0'\Sigma_F H_0 = I_r$. ■

[7] For a symmetric positive definite matrix A and a non-zero vector a, $\|Aa\| = \sqrt{a'^2a} = \sqrt{(a/\|a\|)'^2(a/\|a\|)}\cdot\|a\| \geq \sqrt{\rho(A^2)}\|a\| = \rho(A)\|a\|$, where $\rho(\cdot)$ is the minimum eigenvalue.

[8] See Angrist et al (2006).

A.3 Proof of Theorem 3

Proof of Theorem 3: Again, for simplicity, we suppress the subscript i. Recall that:

$$J_{H_0}(\lambda(\tau)) = \mathbb{E}\big[f_{X|F}(\lambda(\tau)'F_t|F_t)H_0'F_tF_t'H_0\big]$$

and

$$\hat{J}(\hat{\lambda}(\tau)) = \frac{1}{2h_T \cdot T}\sum_{t=1}^{T}\Big\{\mathbf{1}\{|X_t - \hat{\lambda}(\tau)'\hat{F}_t| \leq h_T\}\hat{F}_t\hat{F}_t'\Big\}.$$

Define

$$J(\hat{\lambda}(\tau)) = \frac{1}{2h_T \cdot T}\sum_{t=1}^{T}\Big\{\mathbf{1}\{|X_t - \hat{\lambda}(\tau)'H_{NT}'F_t| \leq h_T\}H_0'F_tF_t'H_0\Big\}.$$

It can be shown that $\sup_{\tau\in\mathcal{T}}\|J_{H_0}(\lambda(\tau)) - J(\hat{\lambda}(\tau))\| = o_P(1)$ by Assumptions 1(v) and $\mathbb{E}\|F_t\|^4 < \infty$, uniform consistency of $\hat{\lambda}(\tau)$ for $H_0^{-1}\lambda(\tau)$, Lemma 1(ii), and the definition of density functions[9]. Thus, the uniform consistency of $\hat{J}(\hat{\lambda}(\tau))$ follows from

$$\sup_{\tau\in\mathcal{T}}\|\hat{J}(\hat{\lambda}(\tau)) - J(\hat{\lambda}(\tau))\| = o_P(1). \tag{29}$$

To prove (29), note that

$$\begin{aligned}
& 2h_T\big(\hat{J}(\hat{\lambda}(\tau)) - J(\hat{\lambda}(\tau))\big) \\
= & \underbrace{\frac{1}{T}\sum_{t=1}^{T}\Big\{\mathbf{1}\{|X_t - \hat{\lambda}(\tau)'\hat{F}_t| \leq h_T\}(\hat{F}_t\hat{F}_t' - H_0'F_tF_t'H_0)\Big\}}_{I} \\
& \underbrace{+\frac{1}{T}\sum_{t=1}^{T}\big[\mathbf{1}\{|X_t - \hat{\lambda}(\tau)'\hat{F}_t| \leq h_T\} - \mathbf{1}\{|X_t - \hat{\lambda}(\tau)'H_{NT}'F_t| \leq h_T\}\big]H_0'F_tF_t'H_0}_{II}.
\end{aligned}$$

[9]The details of the proof is similar to that of equation (A.8) in Angrist et al (2006) and is therefore omitted.

First, we have

$$
\begin{aligned}
\|I\| &\leq \frac{1}{T}\sum_{t=1}^{T}\|\hat{F}_t\hat{F}_t' - H_0'F_tF_t'H_0\| \leq \frac{1}{T}\sum_{t=1}^{T}\|\hat{F}_t - H_0'F_t\|\|H_0'F_t\| + \frac{1}{T}\sum_{t=1}^{T}\|\hat{F}_t - H_0'F_t\|\|\hat{F}_t\| \\
&\leq 2\|H_0\|\cdot\frac{1}{T}\sum_{t=1}^{T}\|\hat{F}_t - H_0'F_t\|\|F_t\| + \frac{1}{T}\sum_{t=1}^{T}\|\hat{F}_t - H_0'F_t\|^2 \\
&\leq 2\|H_0\|\cdot\frac{1}{T}\sum_{t=1}^{T}\|\hat{F}_t - H_{NT}'F_t\|\|F_t\| + 2\|H_0\|\cdot\|H_{NT}-H_0\|\cdot\frac{1}{T}\sum_{t=1}^{T}\|F_t\|^2 + 2\frac{1}{T}\sum_{t=1}^{T}\|\hat{F}_t - H_{NT}'F_t\|^2 \\
&\quad +2\|H_{NT}-H_0\|^2\cdot\frac{1}{T}\sum_{t=1}^{T}\|F_t\|^2 \\
&\leq 2\|H_0\|\sqrt{\frac{1}{T}\sum_{t=1}^{T}\|\hat{F}_t - H_{NT}'F_t\|^2}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\|F_t\|^2} + O_P(\|H_{NT}-H_0\|) + O_P(C_{NT}^{-1}) \\
&= \sqrt{\frac{1}{T}\sum_{t=1}^{T}\|\hat{F}_t - H_{NT}'F_t\|^2}\cdot O_P(1) + \|H_{NT}-H_0\|\cdot O_P(1),
\end{aligned}
$$

then by Assumptions 2(ii), 3 and Lemma 1(i) we have $\|I\|/h_T = o_P(1)$ uniformly in τ.

Second, define $G_{ij,t}$ as the ith row and jth column of $H_0'F_tF_t'H_0$, and we first consider the case $i=j$ such that $G_{i,t} = G_{ii,t} \geq 0$. It is easy to see that:

$$
\begin{aligned}
II_{i,i} &= \frac{1}{T}\sum_{t=1}^{T}\left[\mathbf{1}\{|X_t - \hat{\lambda}(\tau)'\hat{F}_t| \leq h_T\} - \mathbf{1}\{|X_t - \hat{\lambda}(\tau)'H_{NT}'F_t| \leq h_T\}\right]G_{i,t} \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\Big[\mathbf{1}\big\{|X_t - \hat{\lambda}(\tau)'H_{NT}'F_t| \leq h_T + \hat{k}_T(\tau)\big\} - \mathbf{1}\{|X_t - \hat{\lambda}(\tau)'H_{NT}'F_t| \leq h_T\}\Big]G_{i,t}.
\end{aligned}
$$

Where $\hat{k}_T(\tau) = \max_{1\leq t\leq T}|\hat{\lambda}(\tau)'(\hat{F}_t - H_{NT}'F_t)|$. Now consider the following empirical process

$$
\mathbb{C}_T(\theta,h) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\Big\{\mathbf{1}\{|X_t - \theta'F_t| \leq h\}\cdot G_{i,t} - \mathbb{E}[\mathbf{1}\{|X_t - \theta'F_t| \leq h\}\cdot G_{i,t}]\Big\}.
$$

The right hand side of the last inequality is equal to

$$
\begin{aligned}
&\underbrace{T^{-1/2}\Big[\mathbb{C}_T(\theta,h)|_{h=h_T+\hat{k}_T(\tau),\theta=H_{NT}\hat{\lambda}(\tau)} - \mathbb{C}_T(\lambda,h)|_{h=h_T,\theta=H_{NT}\hat{\lambda}(\tau)}\Big]}_{III} \\
+\ &\underbrace{\mathbb{E}\big[\mathbf{1}\{|X_t - \theta'F_t| \leq h\}\cdot G_{i,t}\big]_{h=h_T+\hat{k}_T(\tau),\theta=H_{NT}\hat{\lambda}(\tau)} - \mathbb{E}\big[\mathbf{1}\{|X_t - \theta'F_t| \leq h\}\cdot G_{i,t}\big]_{h=h_T,\theta=H_{NT}\hat{\lambda}(\tau)}}_{IV}.
\end{aligned}
$$

Since $\mathbb{C}_T(\theta,h)$ is stochastic equicontinous when $\mathbb{E}\|F_t\|^4 < \infty$ by Theorem 1 of Andrews (1994), it then follows that $\|III\|$ is $o_P(T^{-1/2})$ uniformly in τ given that

$$
\sup_{\tau\in\mathcal{T}}|\hat{k}_T(\tau)| \leq \sup_{\lambda\in\mathcal{A}}\|\lambda\|\cdot\max_{1\leq t\leq T}\|\hat{F}_t - H_{NT}'F_t\| = o_P(1). \tag{30}
$$

Next, note that

$$
\begin{aligned}
& \mathbb{E}\big[\mathbf{1}\{|X_t - \theta' F_t| \le h_1\} \cdot G_{i,t}\big] - \mathbb{E}\big[\mathbf{1}\{|X_t - \theta' F_t| \le h_2\} \cdot G_{i,t}\big] \\
= \; & \mathbb{E}\Big[\Big(f_{X|F}(\theta' F_t + h^*)(h_1 - h_2) - f_{X|F}(\theta' F_t + h^{**})(h_1 - h_2)\Big) G_{i,t}\Big],
\end{aligned}
$$

where h^* and h^{**} are points on the lines connecting h_1 and h_2. We then have $\|IV\| \le 2\bar{f} \cdot \mathbb{E}\|G_{i,t}\| \cdot |\hat{k}_T(\tau)| = o_P(T^{-1/2})$ uniformly in τ by (30). Combining the above results and Assumption 3, (29) follows directly and thus the first statement in Theorem 3 is proved. The second statement follows trivially by Slutsky's Theorem and the fact that $\frac{1}{T}\sum_{t=1}^{T} \|\hat{F}_t \hat{F}_t' - H_0' F_t F_t' H_0\| = o_P(1)$. ■

A.4 Proof of Theorem 4

To simplify the notations, we suppress the dependence of various object on k, τ, i.e., $\hat{F} = \hat{F}(k, \tau)$, $F_t^0 = F_t^0(\tau)$, $l(z) = l_\tau(z)$, etc. Write $Z_{it} = X_{it} - \lambda_i' F_t$, $U_{it} = Z_{it}^0 = X_{it} - \lambda_i^{0\prime} F_t^0$, and $\hat{Z}_{it} = X_{it} - \hat{\lambda}_i' \hat{F}_t$. Finally, for a matrix C, $\|C\|_S$ denotes the spectral norm of C: $\|C\|_S = \sqrt{\rho_1(C'C)}$, where ρ_1 denote the largest eigenvalue. Note that we have $\|C\|_S \le \|C\| \le \sqrt{\text{rank}(C)} \|C\|_S$.

Using Taylor expansion we have

$$
l(\hat{Z}_{it}) - l(U_{it}) = \partial_z l(U_{it})(\hat{Z}_{it} - U_{it}) + 0.5 \partial_{z^2} l(\tilde{Z}_{it})(\hat{Z}_{it} - U_{it})^2 \ge \partial_z l(U_{it})(\hat{Z}_{it} - U_{it}) + 0.5 b_{NT}^* (\hat{Z}_{it} - U_{it})^2,
$$

where $\tilde{Z}_{it}$ is between $\hat{Z}_{it}$ and U_{it} and the inequality follows from Assumption 4. Then for $k \ge r$

$$
0 \ge \sum_{i=1}^{N}\sum_{t=1}^{T} l(\hat{Z}_{it}) - \sum_{i=1}^{N}\sum_{t=1}^{T} l(U_{it}) \ge 0.5 b_{NT}^* \sum_{i=1}^{N}\sum_{t=1}^{T} \Big[(\hat{Z}_{it} - U_{it})^2 + 2\partial_z l(U_{it})/b_{NT}^* (\hat{Z}_{it} - U_{it})\Big].
$$

Note that $\sum_{i=1}^{N}\sum_{t=1}^{T}(\hat{Z}_{it} - U_{it})^2 = \sum_{i=1}^{N}\sum_{t=1}^{T}(\hat{\lambda}_i' \hat{F}_t - \lambda_i^{0\prime} F_t^0)^2 = \|\hat{F}\hat{\Lambda}' - F^0 \Lambda^{0\prime}\|^2$, and $\sum_{i=1}^{N}\sum_{t=1}^{T} \partial_z l(U_{it})(\hat{Z}_{it} - U_{it}) = \text{Tr}[\partial_z l * (F^0 \Lambda^{0\prime} - \hat{F}\hat{\Lambda}')']$, where $\partial_z l$ is a $T \times N$ matrix with elements $\partial_z l(U_{it})$. So the above inequality can be written as

$$
\|\hat{F}\hat{\Lambda}' - F^0 \Lambda^{0\prime}\|^2 + 2/b_{NT}^* \cdot \text{Tr}[\partial_z l * (F^0 \Lambda^{0\prime} - \hat{F}\hat{\Lambda}')'] \le 0. \tag{31}
$$

Since for any $T \times N$ matrices A, B

$$
|\,\text{Tr}[AB']| \le \text{rank}(AB') \|AB'\|_S \le (\min\{\text{rank}(A), \text{rank}(B)\}) \|A\|_S \|B\|_S,
$$

we have

$$
\big|\,\text{Tr}[\partial_z l * (F^0 \Lambda^{0\prime} - \hat{F}\hat{\Lambda}')']\big| \le (r + k) \cdot \|\partial_z l\|_S \cdot \|F^0 \Lambda^{0\prime} - \hat{F}\hat{\Lambda}'\|_S
$$

because $\text{rank}(F^0 \Lambda^{0\prime} - \hat{F}\hat{\Lambda}') \le \text{rank}(F^0 \Lambda^{0\prime}) + \text{rank}(\hat{F}\hat{\Lambda}') \le r + k$. First, it is easy to see that $\|F^0 \Lambda^{0\prime} - \hat{F}\hat{\Lambda}'\|_S \le \sqrt{r + k} \|F^0 \Lambda^{0\prime} - \hat{F}\hat{\Lambda}'\|$. Second, $\|\partial_z l\|_S \le \|\partial_z l - \mathbb{E}[\partial_z l]\|_S + \|\mathbb{E}[\partial_z l]\|_S$. Similar to Lemma D.6 of Fernandez-Val and Weidner (2015), we can show that

$$
\|\partial_z l - \mathbb{E}[\partial_z l]\|_S = O_P(\sqrt{NT} \cdot T^{-5/8}) + O_P(\sqrt{NT} \cdot T^{-1/2} N^{-1/8}) + O_P(\sqrt{NT} \cdot T^{-1/8} N^{-1/4}).
$$

Moreover, $\|\mathbb{E}[\partial_z l]\|_S \leq \sqrt{NT} \cdot \max_{1\leq t\leq T} |\mathbb{E}[\partial_z l(U_{it})]|$, and it follows from standard proof for kernel density estimators that $\max_{1\leq t\leq T} |\mathbb{E}[\partial_z l(U_{it})]| = O(c_{NT}^d)$. Thus, from Assumption 4(v) we have $1/b_{NT}^* \cdot \|\partial_z l\|_S = o_P(\sqrt{NT})$. Plugging all the above results into (29) gives

$$(NT)^{-1}\|\hat{F}\hat{\Lambda}' - F^0\Lambda^{0'}\|^2 + o_P(1) \cdot (NT)^{-1/2}\|\hat{F}\hat{\Lambda}' - F^0\Lambda^{0'}\| \leq 0,$$

which implies

$$(NT)^{-1/2}\|\hat{F}\hat{\Lambda}' - F^0\Lambda^{0'}\| = o_P(1). \tag{32}$$

Finally, define $M_{\hat{F}} = I - \hat{F}(\hat{F}'\hat{F})^{-1}\hat{F}'$, we have

$$\|M_{\hat{F}}(\hat{F}\hat{\Lambda}' - F^0\Lambda^{0'})\| \leq \sqrt{\operatorname{rank}\left(M_{\hat{F}}(\hat{F}\hat{\Lambda}' - F^0\Lambda^{0'})\right)} \cdot \|M_{\hat{F}}\|_S \cdot \|(\hat{F}\hat{\Lambda}' - F^0\Lambda^{0'})\|_S.$$

Since $\operatorname{rank}(M_{\hat{F}}) = T - k$, $\operatorname{rank}(\hat{F}\hat{\Lambda}' - F^0\Lambda^{0'}) \leq r + k$, $\|M_{\hat{F}}\|_S = 1$ and $\|(\hat{F}\hat{\Lambda}' - F^0\Lambda^{0'})\|_S \leq \|(\hat{F}\hat{\Lambda}' - F^0\Lambda^{0'})\|$, it follows that

$$(NT)^{-1/2}\|M_{\hat{F}}F^0\Lambda^{0'}\| = \sqrt{\operatorname{Tr}\left[\frac{F^{0'}M_{\hat{F}}F^0}{T} \cdot \frac{\Lambda^{0'}\Lambda^0}{N}\right]} = o_P(1).$$

Because $N^{-1}\Lambda^{0'}\Lambda^0$ converges to a full rank matrix by Assumption 4, then

$$\left\|\frac{F^{0'}M_{\hat{F}}F^0}{T}\right\| = o_P(1),$$

which implies

$$F^{0'}F^0/T - (F^{0'}\hat{F}/T)(\hat{F}'F^0/T) = o_P(1).$$

Consequently,

$$\|P_{\hat{F}} - P_{F^0}\|^2 = \operatorname{Tr}[P_{\hat{F}}] + \operatorname{Tr}[P_{F^0}] - 2\operatorname{Tr}[P_{\hat{F}} \cdot P_{F^0}] = (k-r) + \operatorname{Tr}\left[(F^{0'}F^0/T)^{-1}\left(F^{0'}F^0/T - (F^{0'}\hat{F}/T)(\hat{F}'F^0/T)\right)\right],$$

which is equal to $k - r + o_P(1)$ since $F^{0'}F^0/T$ converges to a positive definite matrix by Assumption 4. The proof is then complete by setting $k = r$. ■

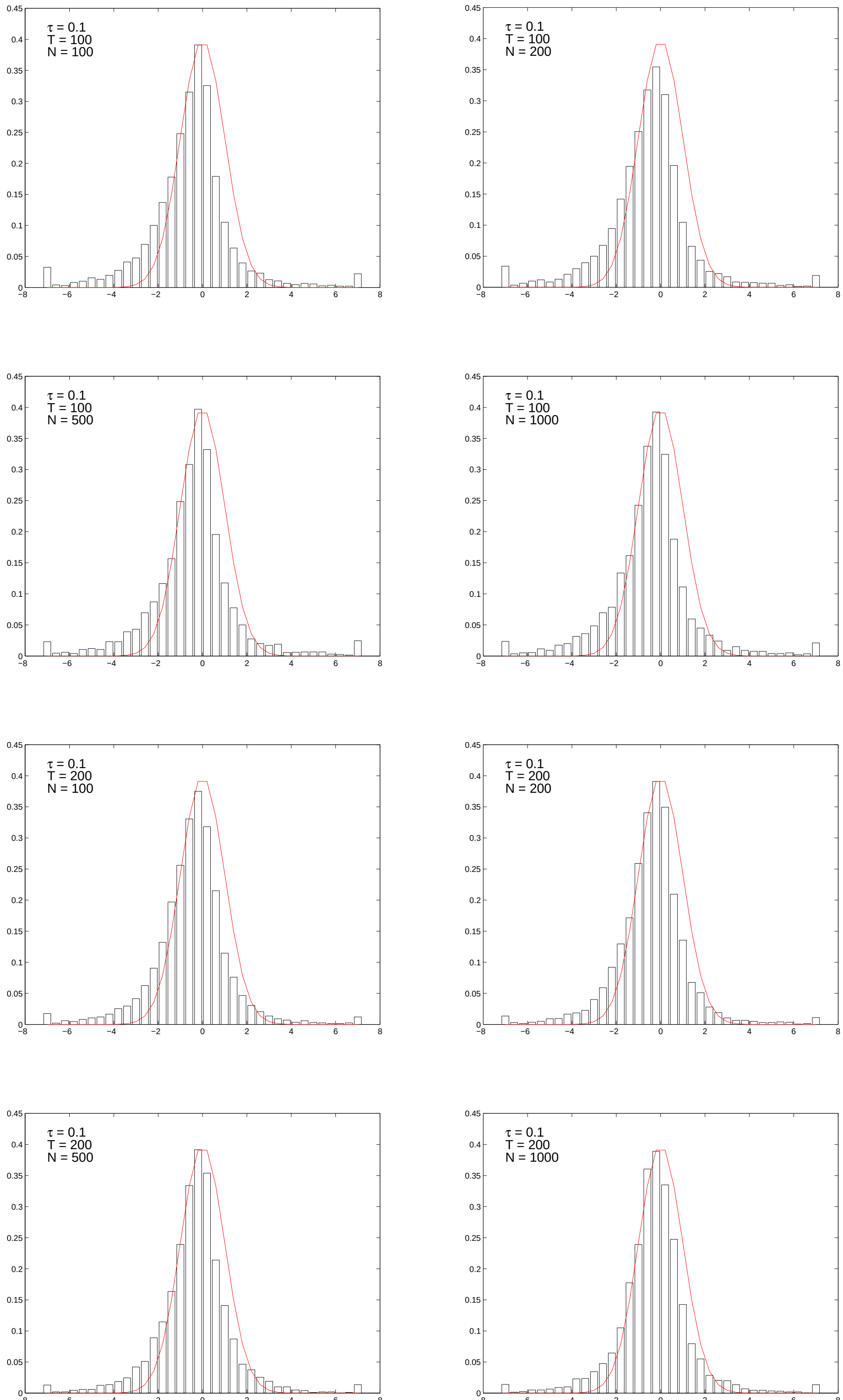

Figure 1: Histograms of $[\tau(1-\tau)]^{-1/2}\cdot\hat{J}(\hat{\lambda}_1(\tau))\cdot\sqrt{T}[\hat{\lambda}_1(\tau)-H_{NT}^{-1}\lambda_1(\tau)]$ and the density function of $\mathcal{N}(0,1)$ for $\tau = 0.1$.

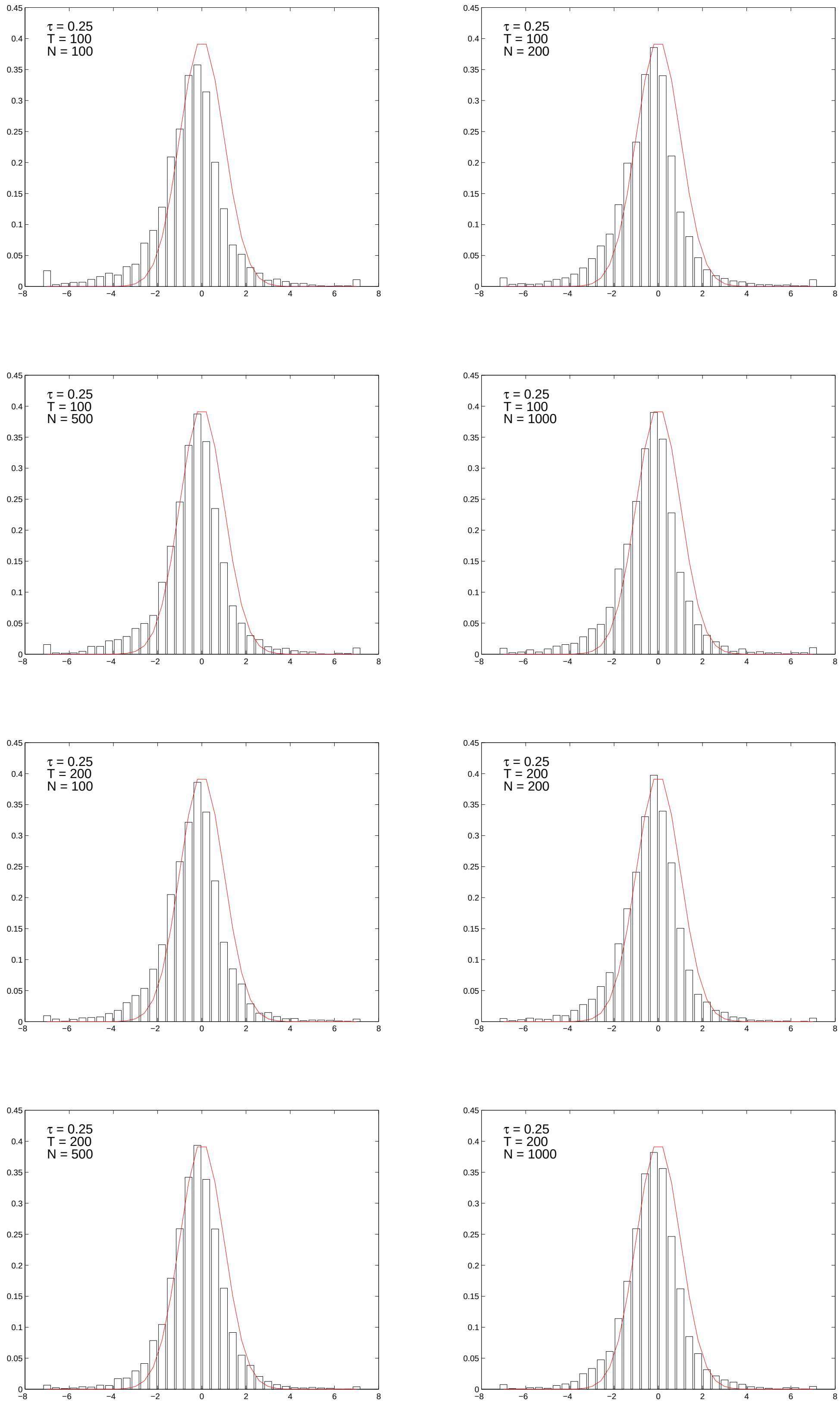

Figure 2: Histograms of $[\tau(1-\tau)]^{-1/2}\cdot\hat{J}(\hat{\lambda}_1(\tau))\cdot\sqrt{T}[\hat{\lambda}_1(\tau)-H_{NT}^{-1}\lambda_1(\tau)]$ and the density function of $\mathcal{N}(0,1)$ for $\tau = 0.25$.

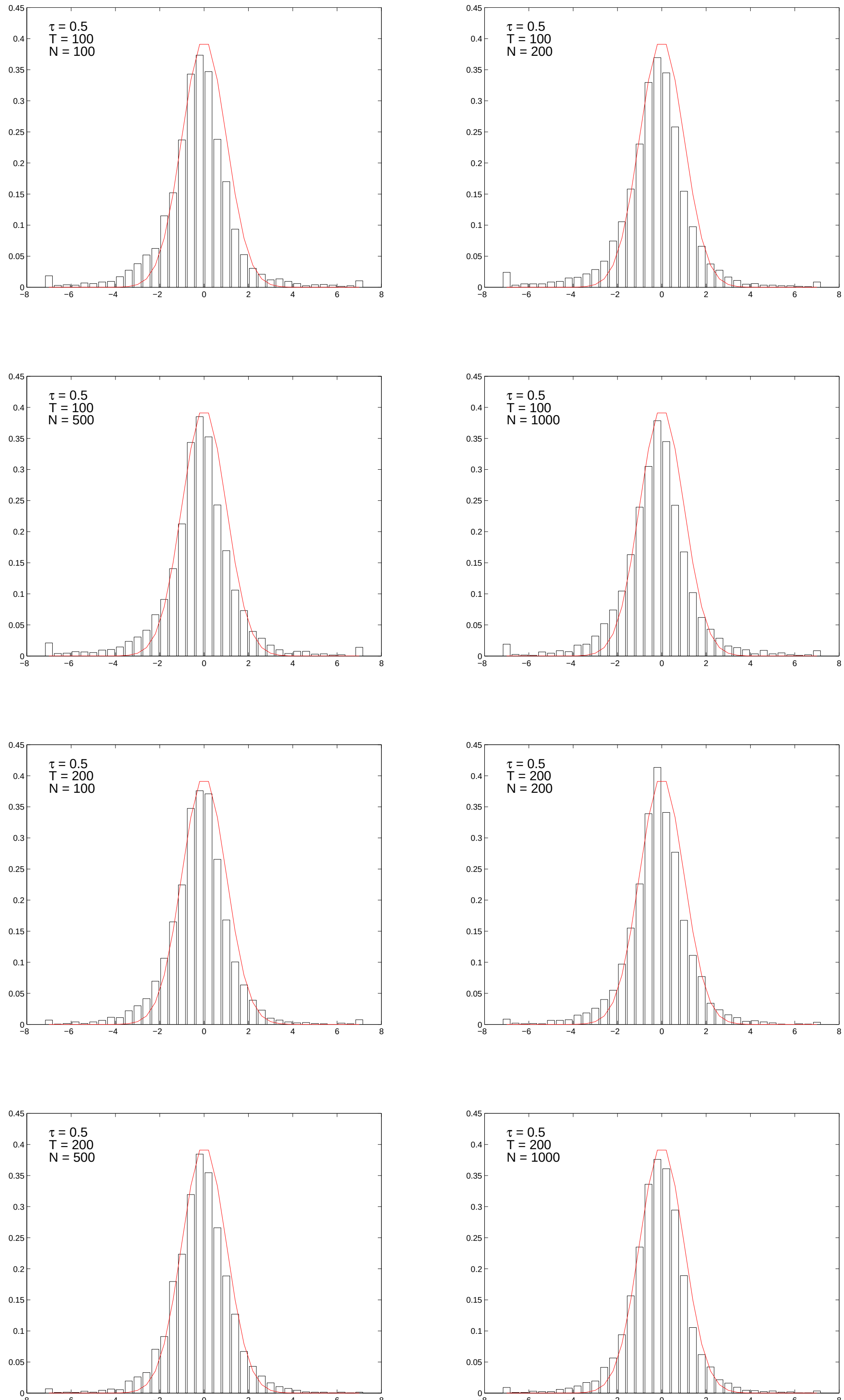

Figure 3: Histograms of $[\tau(1-\tau)]^{-1/2}\cdot\hat{J}(\hat{\lambda}_1(\tau))\cdot\sqrt{T}[\hat{\lambda}_1(\tau)-H_{NT}^{-1}\lambda_1(\tau)]$ and the density function of $\mathcal{N}(0,1)$ for $\tau=0.5$.

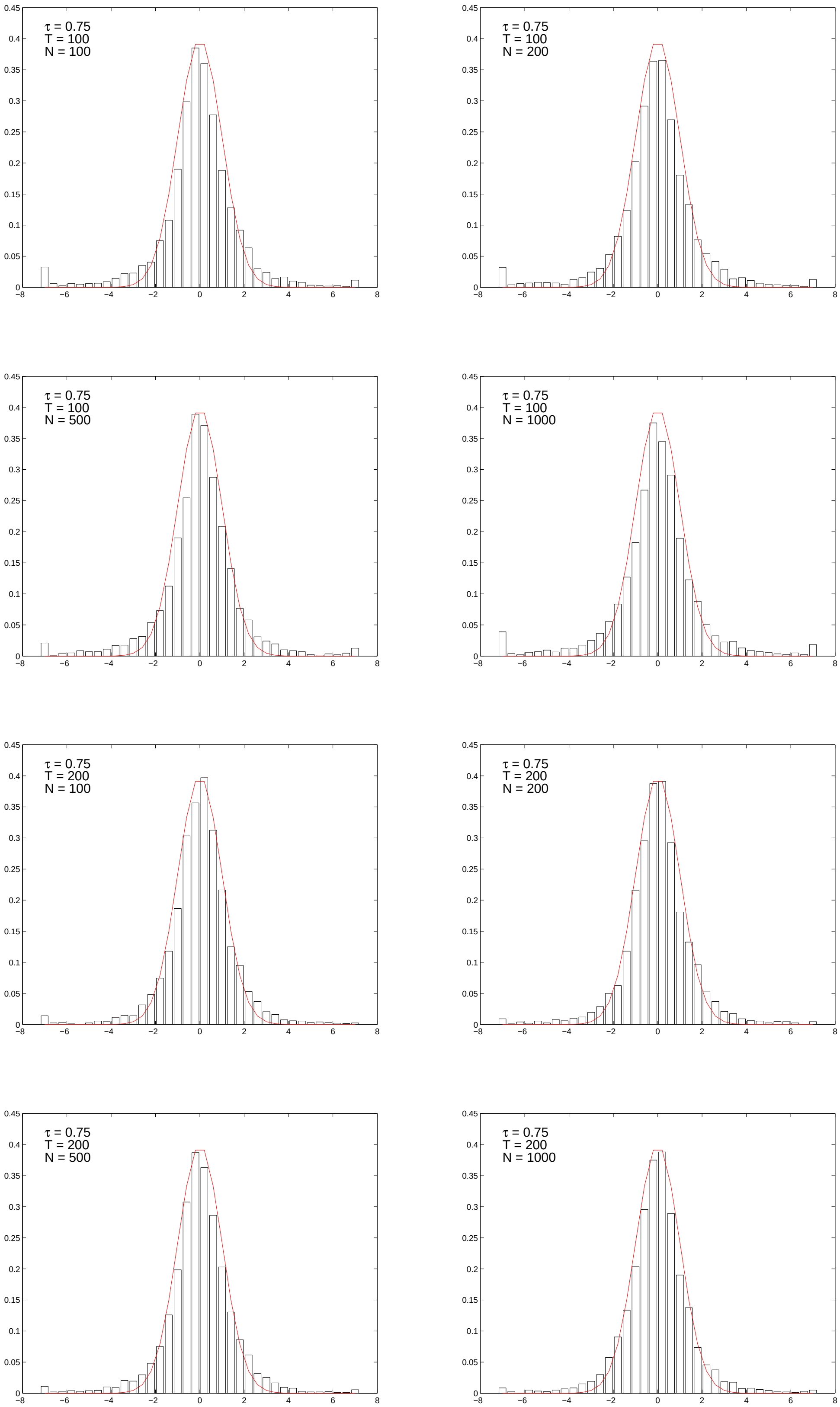

Figure 4: Histograms of $[\tau(1-\tau)]^{-1/2}\cdot\hat{J}(\hat{\lambda}_1(\tau))\cdot\sqrt{T}[\hat{\lambda}_1(\tau)-H_{NT}^{-1}\lambda_1(\tau)]$ and the density function of $\mathcal{N}(0,1)$ for $\tau = 0.75$.

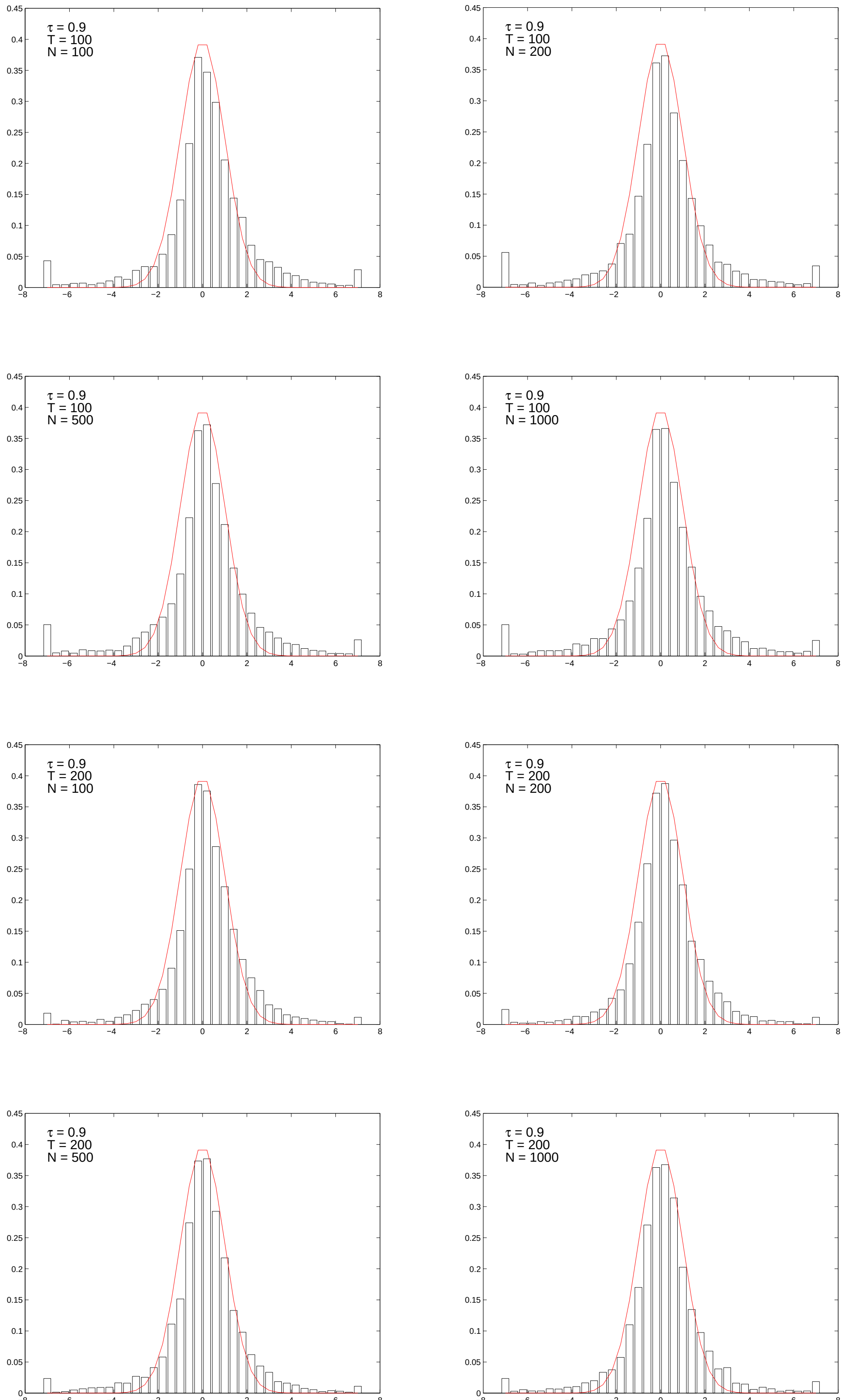

Figure 5: Histograms of $[\tau(1-\tau)]^{-1/2}\cdot\hat{J}(\hat{\lambda}_1(\tau))\cdot\sqrt{T}[\hat{\lambda}_1(\tau)-H_{NT}^{-1}\lambda_1(\tau)]$ and the density function of $\mathcal{N}(0,1)$ for $\tau=0.9$.

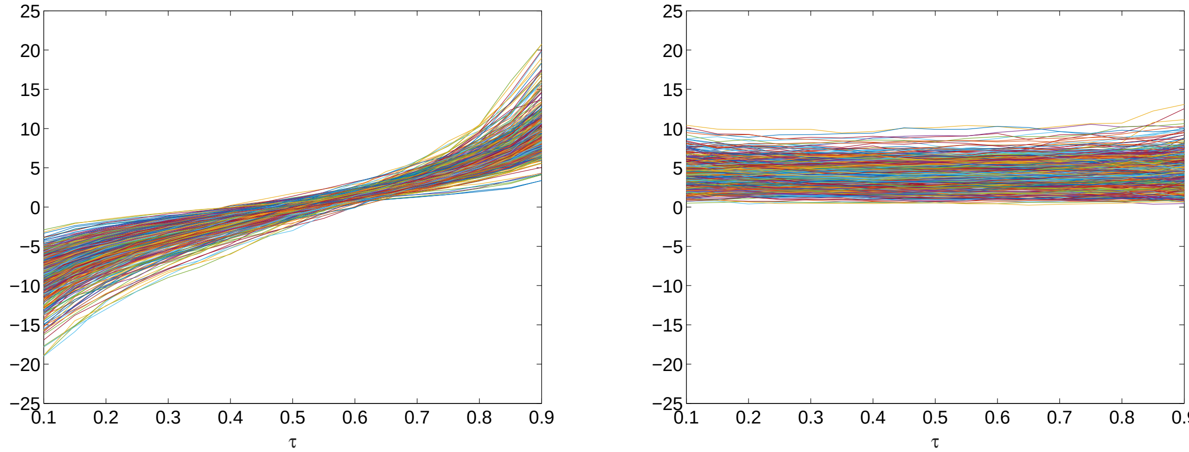

Figure 6: Common stock returns: estimated quantile factor loading processes for the constant (left) and $\hat{F}_t$ (right).

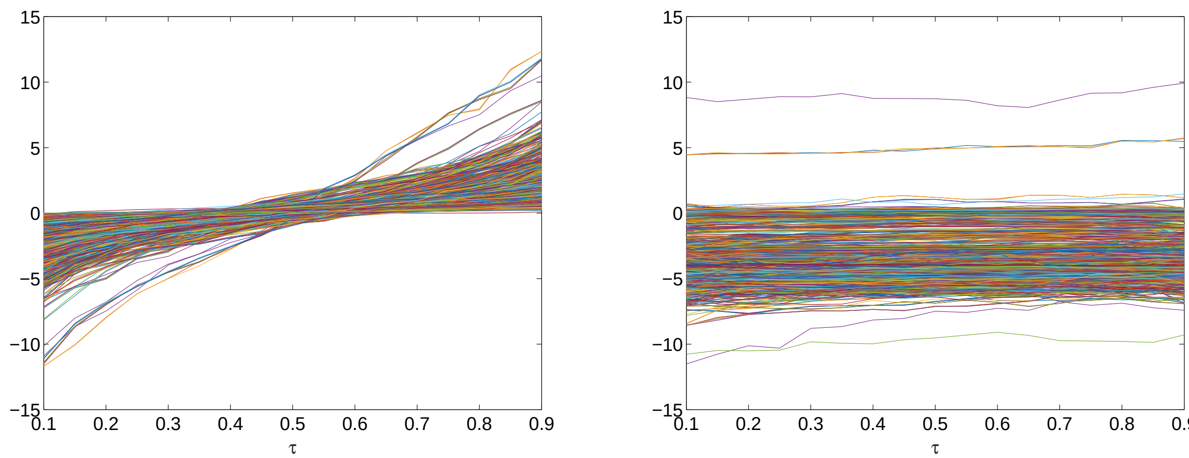

Figure 7: Mutual fund returns: estimated quantile factor loading processes for the constant (left) and $\hat{F}_t$ (right).

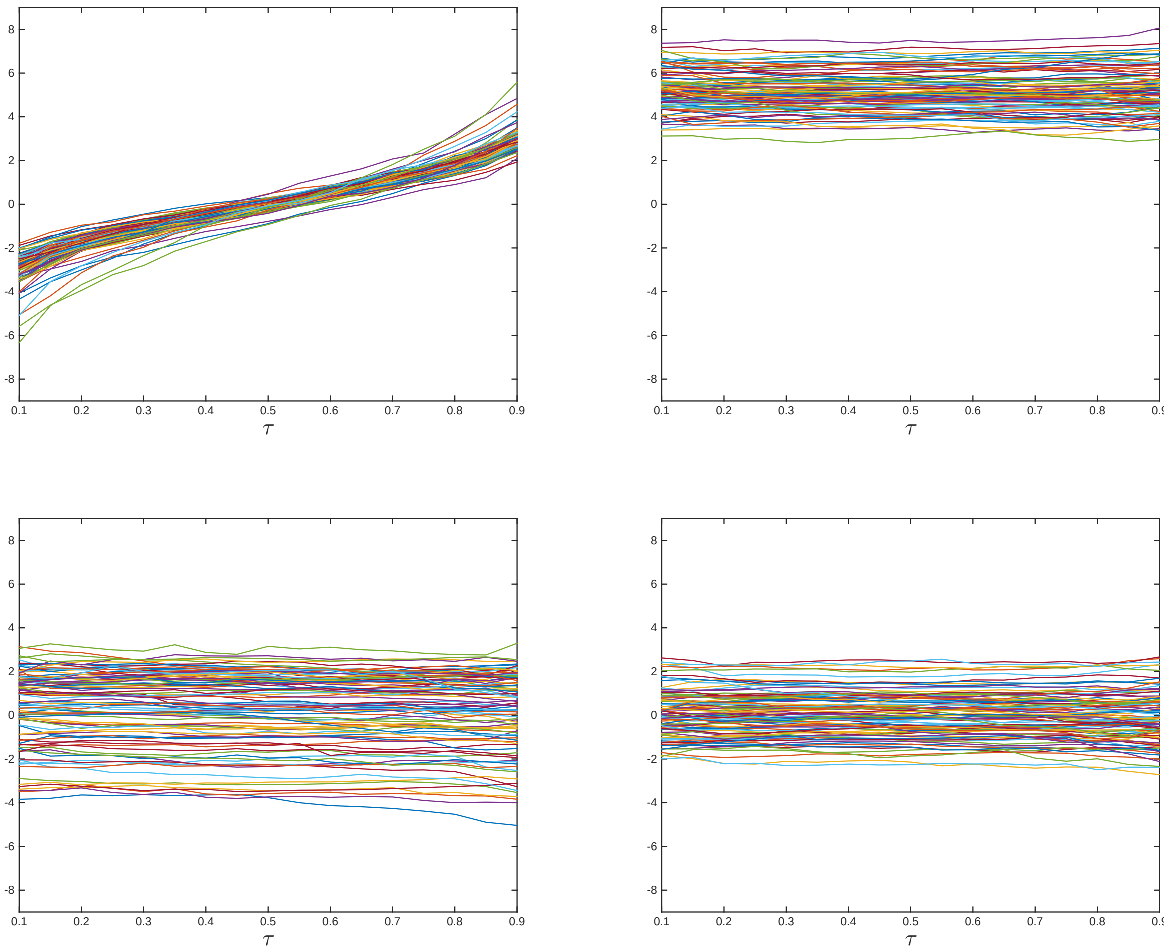

Figure 8: FF portfolios: estimated quantile factor loading processes for the constant (upper left), $\hat{F}_{1t}$ (upper right), $\hat{F}_{2t}$ (lower right), and $\hat{F}_{3t}$ (lower right),.

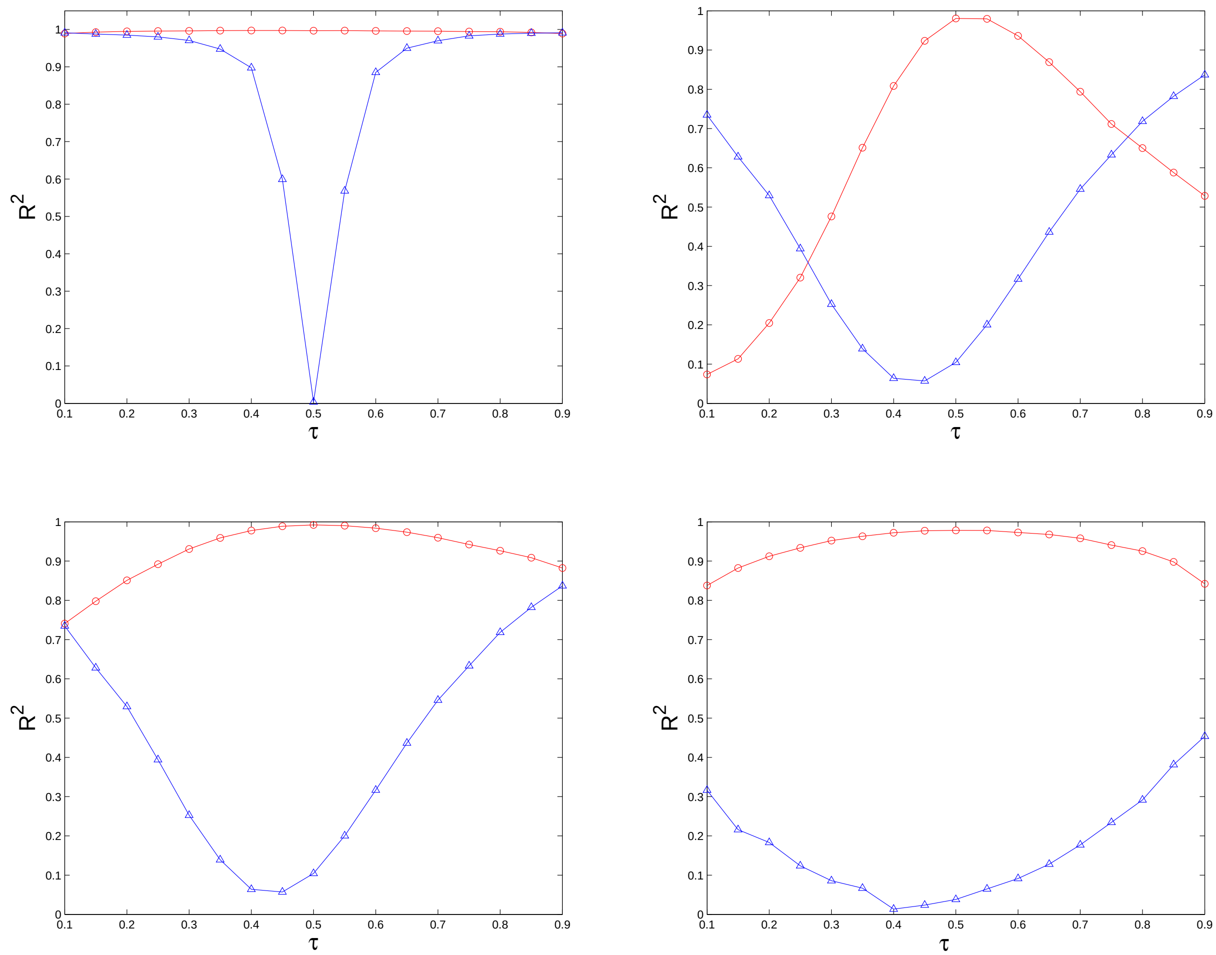

Figure 9: R^2 of regressing $\hat{F}_{PC}$ on $\hat{F}_{QR}$ (red) and regressing a constant factor on $\hat{F}_{QR}$ (blue) for $\tau = 0.1, 0.15, \ldots, 0.9$. **Upper left**: simulated dataset from a location-shift model; **upper right**: stock returns; **lower left**: mutual fund returns; **lower right**: FF portfolios.

References

Abrevaya, J. and C. M. Dahl (2008). The effects of birth inputs on birthweight: evidence from quantile estimation on panel data. *Journal of Business & Economic Statistics 26*(4), 379–397.

Ahn, S. C. and A. R. Horenstein (2013). Eigenvalue ratio test for the number of factors. *Econometrica 81*(3), 1203–1227.

Ando, T. and R. S. Tsay (2011). Quantile regression models with factor-augmented predictors and information criterion. *The Econometrics Journal 14*(1), 1–24.

Andrews, D. W. (1994). Empirical process methods in econometrics. *Handbook of Econometrics 4*, 2247–2294.

Angrist, J., V. Chernozhukov, and I. Fernández-Val (2006a). Quantile regression under misspecification, with an application to the u.s. wage structure. *Econometrica 74*(2), 539–563.

Angrist, J., V. Chernozhukov, and I. Fernández-Val (2006b). Quantile regression under misspecification, with an application to the us wage structure. *Econometrica 74*(2), 539–563.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica 71*(1), 135–171.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica 77*(4), 1229–1279.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

Bai, J. and S. Ng (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica 74*(4), 1133–1150.

Bai, J. and S. Ng (2008a). Extremum estimation when the predictors are estimated from large panels. *Annals of Economics and Finance 9*(2), 201–222.

Bai, J. and S. Ng (2008b). *Large dimensional factor analysis.* Now Publishers Inc.

Bai, J. and S. Ng (2013). Principal components estimation and identification of static factors. *Journal of Econometrics 176*(1), 18–29.

Canay, I. A. (2011). A simple approach to quantile regression for panel data. *The Econometrics Journal 14*(3), 368–386.

Chamberlain, G. and M. Rothschild (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica 51*(5), 1281–304.

Chen, M., I. Fernández-Val, and M. Weidner (2014). Nonlinear panel models with interactive effects. *Working paper*.

Chernozhukov, V. (2005). Extremal quantile regression. *Annals of Statistics*, 806–839.

Chernozhukov, V. and C. Hansen (2006). Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics 132*(2), 491–525.

Chernozhukov, V. and C. Hansen (2008). Instrumental variable quantile regression: A robust inference approach. *Journal of Econometrics 142*(1), 379–398.

Durbin, J. (1973). *Distribution theory for tests based on the sample distribution function*, Volume 9. Siam.

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics 33*(1), 3–56.

Fernández-Val, I. and M. Weidner (2015). Individual and time effects in nonlinear panel models with large N, T. *Working paper*.

Gouriéroux, C. and J. Jasiak (2008). Dynamic quantile models. *Journal of Econometrics 147*(1), 198–205.

Graham, B. S., J. Hahn, and J. L. Powell (2009). The incidental parameter problem in a non-differentiable panel data model. *Economics Letters 105*(2), 181–182.

Harding, M. and C. Lamarche (2014). Estimating and testing a quantile regression model with interactive effects. *Journal of Econometrics 178*, 101–113.

Horowitz, J. L. (1998). Bootstrap methods for median regression models. *Econometrica*, 1327–1351.

Kato, K. and A. F. Galvao (2011). Smoothed quantile regression for panel data. *Working paper*.

Kato, K., A. F. Galvao, and G. V. Montes-Rojas (2012). Asymptotics for panel quantile regression models with individual effects. *Journal of Econometrics 170*(1), 76–91.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis 91*(1), 74–89.

Koenker, R. (2005). *Quantile Regression.* Number 38. Cambridge university press.

Koenker, R. and Z. Xiao (2002). Inference on the quantile regression process. *Econometrica 70*(4), 1583–1612.

Lamarche, C. (2010). Robust penalized quantile regression estimation for panel data. *Journal of Econometrics 157*(2), 396–408.

Powell, J. L. (1986). Censored regression quantiles. *Journal of Econometrics 32*(1), 143–155.

Rosen, A. M. (2012). Set identification via quantile restrictions in short panels. *Journal of Econometrics 166*(1), 127–137.

Stock, J. H. and M. W. Watson (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*(460), 1167–1179.

Stock, J. H. and M. W. Watson (2011). Dynamic factor models. *Oxford Handbook of Economic Forecasting 1*, 35–59.

Xiao, Z. and R. Koenker (2009). Conditional quantile estimation for generalized autoregressive conditional heteroscedasticity models. *Journal of the American Statistical Association 104*(488).