# Estimation of Characteristics-based Quantile Factor Models *

Liang Chen[1], Juan J. Dolado[2], Jesús Gonzalo[3], and Haozi Pan[4]

[1] *HSBC Business School, Peking University, chenliang@phbs.pku.edu.cn*
[2] *Department of Economics, Universidad Carlos III de Madrid, dolado@eco.uc3m.es*
[3] *Department of Economics, Universidad Carlos III de Madrid, jgonzalo@est-eco.uc3m.es*
[4] *HSBC Business School, Peking University, 2001212373@stu.pku.edu.cn*

March 6, 2023

**Abstract**

This paper studies the estimation of characteristic-based quantile factor models where the factor loadings are unknown functions of observed individual characteristics while the idiosyncratic error terms are subject to conditional quantile restrictions. We propose a three-stage estimation procedure that is easily implementable in practice and has nice properties. The convergence rates, the limiting distributions of the estimated factors and loading functions, and a consistent selection criterion for the number of factors at each quantile are derived under general conditions. The proposed estimation methodology is shown to work satisfactorily when: (i) the idiosyncratic errors have heavy tails, (ii) the time dimension of the panel dataset is not large, and (iii) the number of factors exceeds the number of characteristics. Finite sample simulations and an empirical application aimed at estimating the loading functions of the daily returns of a large panel of S&P500 index securities help illustrate these properties.

**Keywords:** quantile factor models, nonparametric quantile regression, principal component analysis.

**JEL codes**: C14, C31, C33.

*We are highly indebted to Yuan Liao for extremely useful comments. We also thank participants at the 2nd Workshop on High Dimensional Data Analysis. Financial support from MICIN/ AEI/10.13039/501100011033, grants PID2019-104960GB-I00, PID2020-118659RB-I00, TED2021-129784BI00, CEX2021-001181-M, and Comunidad de Madrid, grants EPUC3M11 (V PRICIT) and H2019/HUM-5891, is gratefully acknowledged. The usual disclaimer applies.

# 1 Introduction

The generalization of the classical factor analysis has taken place through two main approaches. On the one hand, there has been the development of approximate factor models (**AFM**) where factors are assumed to be unobserved and therefore need to be jointly estimated with their loadings.[1] As a result, AFM suffer from a generic identification problem since both sets of objects can only be identified up to a rotation matrix. On the other hand, a growing literature has emerged in finance trying to explain the cross-sectional co-movements of stock returns on the basis of observable factors. In this setup, the factors are usually approximated using the differences between the returns of portfolios sorted by some observed characteristics, e.g. market capitalization and book-to-market ratio. This popular approach, pioneered by Fama and French (1993), has been extended to include some additional factors, such as *momentum, profitability* and *investment*, together with the well-known Fama-French three factors (see Fama and French (2015)).

Both approaches have pros and cons. In effect, while the *latent factors* approach relies on easily implementable estimation methods, such as the *principal component analysis* (**PCA**), it is often criticized for the lack of interpretation of the estimated factors. Conversely, the Fama-French approach turns out to be unambiguous about this interpretation; yet, their method of constructing the factor proxies quickly becomes unreliable for typical sample sizes as the number of factors grows (see Connor and Linton (2007)).

A setup that tries to take advantage of both approaches while avoiding their shortcomings is the so-called *characteristic-based factor models* (**CFM**) introduced by Connor and Linton (2007) and later extended by Connor, Hagmann, and Linton (2012). According to this framework, the factor loadings are assumed to be smooth nonlinear functions of some observed characteristics of the different units, while the factors remain unobserved, as in AFM. Thereby, the latent factors in CFM can be easily estimated even when the number of factors is not small, whereas their interpretation hinges on the choice of the observed characteristics. Subsequently, Fan, Liao, and Wang (2016) have generalized this framework by allowing both the number of factors to differ from the number of observed characteristics and the factor loadings not to be fully explained by those characteristics. For the estimation of this kind of models, these authors introduced a new methodology called *projected principal component analysis* (**PPCA**), showing that such estimators exhibit faster rates of convergence than the conventional PCA estimators for AFM.

The main goal in this paper is to extend the analysis of Connor et al. (2012) and Fan et al. (2016) to a new class of factor models labeled *characteristic-based quantile factor models*

[1] AFM were first proposed by Chamberlain and Rothschild (1983) to characterize the co-movement of a large set of financial asset returns. The estimation and inference theory of these models and their subsequent extensions have been developed, *inter alia*, by Stock and Watson (2002), Bai and Ng (2002), and Bai (2003); see Fan, Li, and Liao (2021) for a recent overview of this line of research.

(**CQFM**). In relation to CFM, the main difference is that the idiosyncratic errors in CQFM are subject to quantile restrictions instead of mean restrictions. Moreover, in CQFM the latent factors, the loading functions, and the number of factors are all allowed to vary across quantiles, providing in this fashion a more complete picture of how the joint distributions of many asset returns are driven by a few common risk factors.

In particular, our main contribution here is to provide a new three-stage estimation method for CQFM, labeled *quantile-projected principal components analysis* (**QPPCA**), which is computationally simpler than the other available estimation procedures for this kind of factor models. This procedure works as follows. First, at each time period, the outcomes (e.g. stock returns) are projected onto the space of the observed characteristics by means of sieve quantile regressions. Second, using the fitted values from the first step, their factors and loadings are estimated by means of PCA. Finally, the whole loading functions are retrieved by projecting the estimated loadings onto the basis of the sieve space. In addition, we propose a novel estimator for the number of factors at each quantile which is shown to perform satisfactorily for reasonable sample sizes.

The rates of convergence and the limiting distributions of the estimated factors and loading functions in QPPCA are derived under very general conditions, whereas the estimator for the number of factors at each quantile is also shown to be consistent. In particular, a relevant contribution of QPPCA to this literature is that all its asymptotic properties are obtained without assuming any moment restrictions on the idiosyncratic errors. Thus, it becomes a nice tool to analyze data from financial markets where the error distributions are known to have heavy tails. Moreover, we only require the number of cross section observations (denoted as $n$) to diverge in our asymptotic analysis, while the number of time series observations (denoted as $T$) can be taken to be either fixed or diverging.

It is noteworthy that the QPPCA estimation method exhibits several similarities with the PPCA approach of Fan et al. (2016). Yet, the main difference is that, given our less restrictive assumption on the error terms, sieve quantile regressions are implemented in the first step to project the observed outcomes, whereas Fan et al. (2016) uses sieve least square regressions. Thereby, QPPCA estimation turns out to be more robust to heavy tails and outliers, while the consistency of the PPCA estimators requires much stronger moment restrictions on the disturbance terms. However, the robustness of QPPCA estimators entails two potential costs: (i) the average convergence rate of the estimated factors is generally slower than that of the PPCA estimator, unless the strong assumption that the idiosyncratic errors are independent of the observed characteristics is made; and (ii) unlike PPCA, it has to be assumed that the factor loadings are fully explained by the observed characteristics (see below).

A closely related paper to ours is Ma, Linton, and Gao (2021), which also addresses the estimation and inference for CQFM by considering a *semiparametric quantile factor analysis*

(**SQFA**) approach where observed characteristics are potentially allowed to affect stock returns in a nonlinear fashion. Thus, as QPPCA, SQFA extends Connor et al. (2012) to the quantile restriction case. Yet, several important differences exist between the two approaches. First, as in Connor and Linton (2007) and Connor et al. (2012), SQFA assumes that the number of factors is known and is equal to the number of observed characteristics. In contrast, QPPCA does not only allow the number of factors (which can vary across quantiles) to be different from the number of characteristics, but also implies that the number of factors at each quantile can be consistently estimated from the data. Second, while SQFA's initial estimator of the quantile loading functions also relies on sieve quantile regressions, its subsequent steps are based on an iterative minimization algorithm to jointly estimate the factors and loadings. This algorithm can be computationally costly. This potential problem is easily solved with the QPPCA methodology since its second and third steps are based on PCA, which are much easier to compute. Third, the asymptotic results of Ma et al. (2021) are obtained as $n, T \to \infty$, while all our results hold either $T$ is fixed or $T \to \infty$ as $n \to \infty$. Additionally, there are some differences in the assumptions imposed in these two approaches, which will be further discussed in the next sections, once the main theoretical results are presented.

Next, it is important to highlight that the CQFM can be viewed as closely related to the quantile factor models (**QFM**) recently proposed by Chen, Dolado, and Gonzalo (2021) to generalize AFM to quantile regressions. In effect, while no restrictions on the factor loadings are imposed in QFM (except for a standard rank condition), the loadings in CQFM are modeled as unknown functions of some observed characteristics to help interpret the latent factors though, like in SQFA, it also entails the risk of misspecification in the choice of the relevant characteristics. Furthermore, in relation to the quantile factor analysis (**QFA**) estimators proposed by Chen et al. (2021), an additional advantage of the CQFM setup is that the model can be consistently estimated even when $T$ is fixed, while the QFA estimators are only consistent when both $n$ and $T$ go to infinity.

Lastly, we provide an empirical application of the proposed estimators to analyze the behavior of the risk factors and their loadings in a panel dataset of excess stock returns that has been used in other studies. Our main finding is that the use of QPPCA allows for uncovering substantial variations of the estimated loading functions across different quantiles which cannot be obtained using PPCA.

The outline of the rest of the paper is as follows. Section 2 introduces the model and the estimators. Section 3 derives the asymptotic properties of the proposed estimators and presents a novel consistent estimator of the number of factors at each quantile. Section 4 provides several Monte Carlo simulation results for finite samples. Section 5 is devoted to an empirical application of the proposed estimators. Finally, Section 6 concludes. An online appendix gathers detailed proofs of the theorems.

**Notations:** For any matrix $\boldsymbol{C}$, $\|\boldsymbol{C}\|$ and $\|\boldsymbol{C}\|_S$ denote the Frobenius norm and the spectral norm of $\boldsymbol{C}$, respectively; $\lambda_{\min}$ and $\lambda_{\max}$ denote the minimum and maximum eigenvalues of $\boldsymbol{C}$, respectively, when the all eigenvalues are real; and $\boldsymbol{C} > 0$ signifies that $\boldsymbol{C}$ is a positive definite matrix. For two sequences of positive constants $\{a_1, \dots, a_n, \dots\}$ and $\{b_1, \dots, b_n, \dots\}$, $a_n \asymp b_n$ means that $a_n/b_n$ is bounded below and above for all large $n$. The symbol $\lesssim$ means that the left side is bounded by a positive constant times the right side. Finally, for a random vector $(Y, X)$, $\mathsf{Q}_\tau[Y|X = x]$ denotes the $\tau$-quantile of $Y$ given $X = x$.

**Acronyms:** Given the large number of acronyms used throughout the paper, we repeat them here (in the same order as they appear in the main text) to facilitate the reading of the paper: **AFM** (approximate factor model), **CFM** (characteristics-based factor model), **PPCA** (projected principal component analysis), **CQFM** (characteristics-based quantile factor model), **QPPCA** (quantile-projected principal component analysis), **SQFA** (semiparametric quantile factor analysis), **QFM** (quantile factor model), and **QFA** (quantile factor analysis).

# 2 Model and Estimators

## 2.1 Model

For a panel of observed data $\{y_{it}\}_{1\le i\le n, 1\le t\le T}$, Chen et al. (2021) consider the following quantile factor model (QFM):

$$y_{it} = \boldsymbol{\lambda}_i'(\tau)\boldsymbol{f}_t(\tau) + u_{it}(\tau), \quad \tau \in (0, 1), \tag{1}$$

where $\boldsymbol{\lambda}_i(\tau), \boldsymbol{f}_t(\tau) \in \mathbb{R}^R$ are quantile-dependent *unobserved* quantile factor loadings and quantile factors, respectively, $R$ is the number of factors at quantile $\tau$, and $u_{it}(\tau)$ is the idiosyncratic error satisfying $\mathsf{Q}_\tau[u_{it}(\tau)|\boldsymbol{\lambda}_i(\tau), \boldsymbol{f}_t(\tau)] = 0$.[2]

Our focus in this paper is on the CQFM model considered by Ma et al. (2021), which can be viewed as nesting the special case of the QFM in (1) where $\boldsymbol{\lambda}_i(\tau)$ are unrestricted. In particular, let us assume the existence of a vector of *observed* characteristics $\boldsymbol{x}_i = (x_{i1}, x_{i2}, .., x_{iD}) \in \mathbb{R}^D$ for unit $i$ such that

$$\boldsymbol{\lambda}_i(\tau) = \boldsymbol{g}_\tau(\boldsymbol{x}_i), \tag{2}$$

where $\boldsymbol{g}_\tau(\cdot) : \mathbb{R}^D \mapsto \mathbb{R}^R$ is a vector of unknown functions for each $\tau$. As in Connor and Linton (2007), Connor et al. (2012) and Fan et al. (2016), we suppose that the $r$th element of $\boldsymbol{g}_\tau(\boldsymbol{x}_i)$ is given by the following additive function

$$g_{\tau,r}(\boldsymbol{x}_i) = \sum_{d=1}^{D} g_{\tau,rd}(x_{id}),$$

[2]Note that the dependence of $R$ on $\tau$ is suppressed to ease the notations.

where $g_{\tau,r1},\ldots,g_{\tau,rD}$ are unknown functions. As in the related literature, it is assumed that $g_{\tau,r}$ is time-invariant so that the loadings capture the cross-sectional heterogeneity only. As Fan et al. (2016) argued, such a specification is not stringent since in many factor-model applications to stationary time series, the analysis is carried out within each fixed time window with either a fixed or slowly-growing $T$. Yet, even if there are individual characteristics that are time-variant, like e.g. firm size or firm age, following these authors, we expect the conclusions in the current paper to remain valid if some smoothness assumptions are added for the time-varying components of those covariates.

Let $\boldsymbol{Y}$ be the $n\times T$ matrix of $y_{it}$, $\boldsymbol{F}_\tau$ be the $T\times R$ matrix of $\boldsymbol{f}_t(\tau)$, $\boldsymbol{X}$ be the $n\times D$ matrix of $\boldsymbol{x}_i$, $\boldsymbol{G}_\tau(\boldsymbol{X})$ be the $n\times R$ matrix of $\boldsymbol{g}_\tau(\boldsymbol{x}_i)$, $\boldsymbol{U}_\tau$ be the $n\times T$ matrix of $u_{it}(\tau)$. Then, models (1) and (2) can be rewritten in compact matrix form as:

$$\boldsymbol{Y}=\boldsymbol{G}_\tau(\boldsymbol{X})\boldsymbol{F}_\tau'+\boldsymbol{U}_\tau. \tag{3}$$

As already mentioned, the above setup is more general than those considered in the models of Connor et al. (2012) and Ma et al. (2021). In the latter, the dimension of the vector of characteristics is required to be equal to the number of factors ($D=R$), while each of the loading functions is assumed to be linked to only one of the observed characteristics, i.e., $g_{\tau,r}(\boldsymbol{x}_i)=g_{\tau,r}(x_{ir})$ for $r=1,\ldots,R$. Note that these assumptions facilitate the interpretation of the estimated factors, e.g. the first estimated factor would be the value factor, the second one would be the momentum factor, and so on. However, both conditions also could be restrictive in other setups. For example, if $y_{it}$ represents the profit flow of firm $i$ at time $t$ and there are two factors capturing, say, a monetary shock and a fiscal shock, then it seems more reasonable to allow for dependence of the response of the firm's profit to the macro shocks on a wide range of firm characteristics — such as size, leverage, growth, etc. — that exceeds the number of factors. Moreover, a drawback of the two above-mentioned approaches is that they are more difficult to estimate and therefore require algorithms involving multiple iterations, particularly when the number of characteristics is large (say there are tens of characteristics, then there will be tens of factors). In contrast, in our setup, it is potentially easier to generalize CQFM to allow for high dimensional characteristics since the number of factors can be much smaller than the number of characteristics. Lastly, it should be noted that in the previous estimation methods, one needs to assume that the number of factors $R$ is a priori known, while in this paper we will propose a method that consistently estimates $R$ from the data at each quantile (see Section 3.3 below).

Relative to the semiparametric factor models considered by Fan et al. (2016), the most salient difference is that the idiosyncratic errors in CQFM are subject to conditional quantile restrictions, rather than to conditional mean restrictions. From this perspective, as pointed out in Chen et al. (2021), the QFM framework allows to recover different factor structures (including the factors, the loadings, and the number of factors) across different quantiles, even when the distribution of the idiosyncratic errors exhibits heavy tails. Hence, these features make CQFM a useful tool to analyze the co-movement of the financial market variables, where the correlation of the tail risks between different assets becomes the main object of interest. Furthermore, a relevant extension of Fan et al. (2016) with respect to Connor et al. (2012) is that the factor loadings are allowed to be functions of other unobserved random variables, besides the set of observed characteristics. However, allowing for this more general case in the context of QFM would be very challenging. To see this, assume that

$$\boldsymbol{\lambda}_i(\tau) = \boldsymbol{g}_\tau(\boldsymbol{x}_i) + \boldsymbol{\gamma}_i,$$

where $\boldsymbol{\gamma}_i$ is unobserved and independent of $\boldsymbol{x}_i$. Then model (1) can be written as

$$y_{it} = \boldsymbol{g}_\tau(\boldsymbol{x}_i)'\boldsymbol{f}_t(\tau) + \tilde{u}_{it}(\tau) \quad \text{where} \quad \tilde{u}_{it}(\tau) = u_{it}(\tau) + \boldsymbol{\gamma}_i'\boldsymbol{f}_t(\tau).$$

The above-mentioned model can be viewed as a CQFM with *measurement errors*, where the new error terms $\tilde{u}_{it}(\tau)$ no longer satisfy the conditional quantile restrictions, even when $\boldsymbol{x}_i$ and $\boldsymbol{\gamma}_i$ are independent. The insight is that imposing quantile conditional restrictions for the two elements of $\tilde{u}_{it}(\tau)$ does not imply that this conditional restriction should hold for their sum. By contrast, if conditional expectation were applied, as in AFM, both error terms would have zero means allowing the application of PPCA. In fact, dealing with measurement errors is far from being a trivial issue even in standard quantile regressions (see e.g. Hausman, Liu, Luo, and Palmer (2021)). Thus, in what follows, the analysis will be restricted to the case where the factor loadings are fully explained by the observed characteristics, as in Ma et al. (2021), but allowing each factor loading to depend on a host of observable characteristics instead of just a single one. Several data generating processes (DGPs) discussed in the Monte Carlo simulations reported in Section 4 provide examples of CQFM models to be estimated by QPPCA.

### 2.2 Estimators

To simplify the notations even further, in the rest of the paper we suppress the $\tau$-subscripts in the model and use $\boldsymbol{g}(), \boldsymbol{G}(\cdot), \boldsymbol{F}, \boldsymbol{U}$ instead of $\boldsymbol{g}_\tau(\cdot), \boldsymbol{G}_\tau(\cdot), \boldsymbol{F}_\tau, \boldsymbol{U}_\tau$.

Write $\theta_{0t}(\boldsymbol{x}_i) = \boldsymbol{g}(\boldsymbol{x}_i)'\boldsymbol{f}_t = \sum_{r=1}^R g_r(\boldsymbol{x}_i) f_{tr} = \sum_{r=1}^R (\sum_{d=1}^D g_{rd}(x_{id})) f_{tr}$. Let $\Theta$ be a space of continuous functions such that $\theta_{0t} \in \Theta$ for all $t = 1, \ldots, T$, while $\{\Theta_n\}$ is a sequence of sieve spaces approximating $\Theta$. In particular, let us consider the following finite dimensional linear spaces:

$$\Theta_n = \left\{ h : \mathcal{X} \mapsto \mathbb{R}, \quad h(\boldsymbol{x}) = \sum_{d=1}^D \sum_{j=1}^{k_n} a_{jd}\phi_j(x_d) : (a_{11}, \ldots, a_{jd}, \ldots, a_{k_n D}) \in \mathbb{R}^{Dk_n} \right\},$$

where $\mathcal{X} \subset \mathbb{R}^D$ is the support of $\boldsymbol{x}_i$, and $\phi_1, \ldots, \phi_{k_n}$ is a set of continuous basis functions. Write

$$\underbrace{\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)}_{Dk_n \times 1} = [\phi_1(x_{i1}), \ldots, \phi_{k_n}(x_{i1}), \ldots \phi_1(x_{id}), \ldots, \phi_{k_n}(x_{id}), \ldots, \phi_1(x_{iD}), \ldots, \phi_{k_n}(x_{iD})]'.$$

Suppose that for $r = 1, \ldots, R$, there exists $\boldsymbol{b}_{01}, \ldots, \boldsymbol{b}_{0R} \in \mathbb{R}^{Dk_n}$ such that for some constant $\alpha > 0$,

$$\max_{1 \le r \le R} \sup_{\boldsymbol{x} \in \mathcal{X}} \left| g_r(\boldsymbol{x}) - \boldsymbol{b}_{0r}' \boldsymbol{\phi}_{k_n}(\boldsymbol{x}) \right| = O(k_n^{-\alpha}). \tag{4}$$

Then, for $\boldsymbol{B}_0 = (\boldsymbol{b}_{01}, \ldots, \boldsymbol{b}_{0R}) \in \mathbb{R}^{Dk_n \times R}$, $\boldsymbol{a}_{0t} = \boldsymbol{B}_0 \boldsymbol{f}_t$ and $\pi_n \theta_{0t}(\cdot) = \boldsymbol{a}_{0t}' \boldsymbol{\phi}_{k_n}(\cdot)$, we have $\pi_n \theta_{0t} \in \Theta_n$ for all $t$ and

$$\max_{1 \le t \le T} \sup_{\boldsymbol{x} \in \mathcal{X}} \left| \pi_n \theta_{0t}(\boldsymbol{x}) - \theta_{0t}(\boldsymbol{x}) \right| = O(k_n^{-\alpha}). \tag{5}$$

Once the definitions above have been established, the next stage is to introduce our QPPCA estimation method which consists of the following three steps.

**Step 1:** Obtain the sieve estimator of $\theta_{0t}$. Let $\rho_\tau(u) = (\tau - \mathbf{1}\{u \le 0\})u$ be the check function, and define $l(\theta, y_{it}, \boldsymbol{x}_i) = \rho_\tau(y_{it} - \theta(\boldsymbol{x}_i)) - \rho_\tau(y_{it} - \theta_{0t}(\boldsymbol{x}_i))$, $L_n(\theta) = n^{-1} \sum_{i=1}^n l(\theta, y_{it}, \boldsymbol{x}_i)$. Then the sieve estimator $\hat{\theta}_{nt}$ is defined by

$$L_n(\hat{\theta}_{nt}) \le \inf_{\theta \in \Theta_n} L_n(\theta).$$

In practice, $\hat{\theta}_{nt}$ can be obtained by a simple parametric quantile regression as follows:

$$\hat{\boldsymbol{a}}_t = \arg\min_{\boldsymbol{a} \in \mathbb{R}^{Dk_n}} \sum_{i=1}^N \rho_\tau\left(y_{it} - \boldsymbol{a}' \boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)\right) \quad \text{and} \quad \hat{\theta}_{nt}(\cdot) = \hat{\boldsymbol{a}}_t' \boldsymbol{\phi}_{k_n}(\cdot).$$

**Step 2:** Write $\hat{y}_{it} = \hat{\theta}_{nt}(\boldsymbol{x}_i) = \hat{\boldsymbol{a}}_t' \boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)$ and let $\hat{\boldsymbol{Y}}$ be the $n \times T$ matrix of $\hat{y}_{it}$. Then, the estimator of $\boldsymbol{F}$, denoted as $\hat{\boldsymbol{F}}$, is the matrix of eigenvectors (multiplied by $\sqrt{T}$) associated with $R$ largest eigenvalues of the $T \times T$ matrix $\hat{\boldsymbol{Y}}' \hat{\boldsymbol{Y}}$. Moreover, the estimator of the characteristics-based loading matrix $\boldsymbol{G}(\boldsymbol{X})$ is given by $\hat{\boldsymbol{G}}(\boldsymbol{X}) = \hat{\boldsymbol{Y}} \hat{\boldsymbol{F}}/T$. It is well known that these estimators are the ones that minimize the objective function: $L_{nT}(\boldsymbol{G}(\boldsymbol{X}), \boldsymbol{F}) = \|\hat{\boldsymbol{Y}} - \boldsymbol{G}(\boldsymbol{X})' \boldsymbol{F}\|^2$, subject to the standard normalizations, namely, $\boldsymbol{F}' \boldsymbol{F}/T = \boldsymbol{I}_R$ and $\boldsymbol{G}(\boldsymbol{X})' \boldsymbol{G}(\boldsymbol{X})/n$ is diagonal (see Stock and Watson (2002)).[3]

**Step 3:** Estimate the factor loading functions: $g_r(\cdot)$ for $r = 1, \ldots, R$. Let $\boldsymbol{A}_0 = (\boldsymbol{a}_{01}, \ldots, \boldsymbol{a}_{0T})$ and $\hat{\boldsymbol{A}} = (\hat{\boldsymbol{a}}_{01}, \ldots, \hat{\boldsymbol{a}}_{0T})$. Intuitively, $\hat{\boldsymbol{A}} \approx \boldsymbol{A}_0 = \boldsymbol{B}_0 \boldsymbol{F}' \approx \boldsymbol{B}_0 \hat{\boldsymbol{F}}'$, as a result of which $\boldsymbol{B}$ can be

[3]Note, however, that the estimator is invariant to the rotation transformations of the sieve bases.

simply estimated as

$$\hat{\boldsymbol{B}} = \hat{\boldsymbol{A}}\hat{\boldsymbol{F}}/T. \tag{6}$$

The estimator of $\boldsymbol{g}(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathcal{X}$ is given by $\hat{\boldsymbol{g}}(\boldsymbol{x})' = \boldsymbol{\phi}_{k_n}(\boldsymbol{x})'\hat{\boldsymbol{B}}$.

**Remark 1.** *The basic idea of the projections is to smooth the observations $\{y_{it}\}_{1\leq i\leq n}$ for each given $t$ against its associated covariates, as in step 1, to then compute the factors of the var-cov matrix of the projections in step 2; finally using the estimated factors and the estimates of the sieve functions in step 1, the loading functions are easily obtained in step 3. The main difference between this three-stage estimation method and the PPCA estimating approach of Fan et al. (2016) is how we project $\boldsymbol{y}_t$ onto the space of $\boldsymbol{X}$ in the first step, namely, how $\boldsymbol{a}_{01}, \ldots, \boldsymbol{a}_{0T}$ are estimated. The use of sieve quantile regressions instead of the least squares projections is a natural choice given that the idiosyncratic errors in CQFM are subject to conditional quantile restrictions. When the distributions of the errors are symmetric around 0, the QPPCA estimators at $\tau = 0.5$ can be viewed as a robust version of the PPCA estimators since the consistency of the QPPCA estimators does not rely on moment restrictions of the errors (see Theorem 1 below).*

**Remark 2.** *The SQFA estimation method advocated by Ma et al. (2021) chooses $\boldsymbol{B}$ and $\boldsymbol{F}$ in an iterative fashion to minimize the following objective function:*

$$L_{nT}(\boldsymbol{B}, \boldsymbol{F}) = \sum_{i=1}^{n}\sum_{t=1}^{T}\rho_\tau(y_{it} - \boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)'\boldsymbol{B}\boldsymbol{f}_t),$$

*while the quantile factor analysis (**QFA**) proposed by Chen et al. (2021) for QFM relies on a similar approach which estimates the factor loadings $\boldsymbol{G}(\boldsymbol{X})$ and $\boldsymbol{F}$ jointly. Accordingly, both SQFA and QFA require $n$ and $T$ go to infinity in order to establish the consistency of the estimators. By contrast, as will be shown in the next section, the consistency of the QPPCA estimators can be established either when $T$ is fixed or $T$ goes to infinity along with $n$.*

# 3 Asymptotic Properties of the Estimators

In this section, we derive the rates of convergence and the asymptotic distributions of the QPPCA estimators proposed in the previous section. To simplify the discussion, the number of factors is taken to be known in the first two subsections while this assumption is relaxed in the last subsection, where a consistent estimator of $R$ is introduced.

As in Chen et al. (2021) and Ma et al. (2021), the quantile factors are treated as nonrandom constants in the asymptotic analysis. Hence, the conditional quantile restrictions on the idiosyncratic errors imply that

$$P[u_{it} \leq 0 | \boldsymbol{x}_i = \boldsymbol{x}] = \tau \text{ for any } \boldsymbol{x} \in \mathcal{X}. \tag{7}$$

Alternatively, all the assumptions and results to be presented below could be understood as being conditional on the realizations of the factors.

Lastly, it should be noted that all the results to be presented below hold either when: (i) $T$ is fixed and $n \to \infty$, or (ii) $n, T \to \infty$. The first case is also called the *high-dimension-low-sample-size* setup in the statistics literature (see Shen, Shen, and Marron (2013) and Jung and Marron (2009)). One of the main insights of Fan et al. (2016) is that dimensionality is a blessing rather than a curse in the context of CFM, implying that their PPCA estimators are consistent even when $T$ is fixed. Our results below extend the finite-$T$-consistency results of Fan et al. (2016) to CQFM.

### 3.1 Rates of convergence

Suppose that the observed data $\{y_{it}\}$ are generated by (3) and that $\{u_{it}\}$ satisfy (7). Let

$$\varepsilon_n = \sqrt{k_n/n} \vee k_n^{-\alpha} \quad \text{and} \quad \varepsilon_{nT} = \sqrt{\ln T} \vee 1 \cdot \varepsilon_n.$$

For any $\theta_1, \theta_2 \in \Theta$, define the pseudo-metric $d(\theta_1, \theta_2) \equiv \sqrt{\mathbb{E}\left(\theta_1(\boldsymbol{x}_i) - \theta_2(\boldsymbol{x}_i)\right)^2}$. The following set of conditions are required to establish the uniform rate of convergence of $\hat{\theta}_{n1}, \ldots, \hat{\theta}_{nT}$, which is a crucial result to prove the other theorems.

**Assumption 1.** *Let $M$ be a generic bounded constant.*
*(i) Define $\boldsymbol{z}_i = (u_{i1}, \ldots, u_{iT}, \boldsymbol{x}_i)$. Then, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ are i.i.d. Moreover, the distributions of $(u_{i1}, \boldsymbol{x}_i), \ldots, (u_{iT}, \boldsymbol{x}_i)$ are identical for each $i$.*
*(ii) Equation (4) holds for some $\alpha \geq 1$.*
*(iii) $\mathcal{X} \subset \mathbb{R}^D$ is bounded, and $\sup_{\theta \in \Theta} \sup_{\boldsymbol{x} \in \mathcal{X}} |\theta(\boldsymbol{x})| < M$. $\|\boldsymbol{f}_t\| < M$ for all $t = 1, \ldots, T$.*
*(iv) The conditional density of $u_{it}$ given $\boldsymbol{x}_i = \boldsymbol{x}$, denoted as $\mathsf{f}(\cdot|\boldsymbol{x})$, satisfies: $0 < \inf_{\mathcal{X}} \mathsf{f}(0|\boldsymbol{x}) \leq \sup_{\mathcal{X}} \mathsf{f}(0|\boldsymbol{x}) < \infty$ and $\sup_{\mathcal{X}} |\mathsf{f}(z|\boldsymbol{x}) - \mathsf{f}(0|\boldsymbol{x})| \to 0$ as $|z| \to 0$.*
*(v) As $n \to \infty$, $k_n \to \infty$ and $\varepsilon_{nT} \to 0$.*

Although, in principle, Assumption 1(i) is stronger than those in Fan et al. (2016) and Ma et al. (2021), it can be relaxed to allow for weak cross-sectional dependence — see Remark 3 below for the details. Assumption 1(ii) is a general condition on the sieve approximations that can be easily verified using more primitive conditions. For instance, it holds if $\Theta$ is an $\alpha-$smooth Hölder space (see Chen (2007) for further examples). Assumption 1(iii) and Assumption 1(iv) are also standard in sieve quantile regressions, noting that the last assumption imposes very mild restrictions on the size of $T$ when it goes to infinity jointly with $n$.

**Proposition 1.** *Under Assumption 1, when either $T$ is fixed or $T \to \infty$ as $n \to \infty$, it holds that* $\max_{1 \le t \le T} d(\hat{\theta}_{nt}, \theta_{0t}) = O_P(\varepsilon_{nT})$.

**Remark 3.** *The proof of Proposition 1 is based on Corollary 1 of Chen and Shen (1998). In particular, we show that*

$$P\left[\max_t d(\hat{\theta}_{nt}, \theta_{0t}) \ge C\varepsilon_{nT}\right] \le \sum_{t=1}^{T} P\left[d(\hat{\theta}_{nt}, \theta_{0t}) \ge C\varepsilon_{nT}\right] \le c_1 \exp\left\{C^2 \ln T(1 - c_2 n \varepsilon_n^2)\right\}$$

*for any $C \ge 1$ and some constants $c_1, c_2$. Moreover, as shown in Chen and Shen (1998), the above inequality holds when the observations are generated from a stationary uniform ($\phi$-) mixing sequence with $\phi(j) \lesssim j^{-\zeta}$ for some $\zeta > 1$. Thus, in line with Connor and Korajczyk (1993), Lee and Robinson (2016) and Ma et al. (2021), one can assume the existence of a reordering of the cross-sectional units such that their dependence can be characterized by the uniform mixing condition mentioned above, and the conclusion of Proposition 1 will still hold.*

To establish the convergence rates of the estimated factors and loading functions, some further assumptions are required.

**Assumption 2.** *Let $M$ be a generic bounded constant.*
*(i) Let $\mathbf{\Sigma}_\phi = \mathbb{E}[\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)']$. Then, there exist constants $c_1, c_2$ such that $0 < c_1 \le \lambda_{\min}(\mathbf{\Sigma}_\phi) \le \lambda_{\max}(\mathbf{\Sigma}_\phi) \le c_2 < \infty$ for all $n$.*
*(ii) $k_n^2/n \to 0$ as $n \to \infty$.*
*(iii) There exist a constant $c > 0$ such that $\lambda_{\min}(\boldsymbol{F}'\boldsymbol{F}/T) > c$ for all $T$.*
*(iv) $\hat{\mathbf{\Sigma}}_g \equiv n^{-1}\sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)' \xrightarrow{P} \mathbf{\Sigma}_g > 0$ as $n \to \infty$.*
*(v) The eigenvalues of $\mathbf{\Sigma}_g \cdot \boldsymbol{F}'\boldsymbol{F}/T$ are distinct.*

The conditions in Assumption 2 are all standard in the literature on factor models and sieve estimation. In particular, Assumption 2(ii) strengthens Assumption 1(v), and Assumption 2(iii) implicitly requires that $T \ge R$. In comparison, Assumption A0 of Ma et al. (2021) imposes that $\liminf_{T\to\infty} |T^{-1}\sum_{t=1}^{T} f_{tr}| > 0$ for all $r = 1, \ldots, R$, which excludes the possibility that the underlying time series generating $\boldsymbol{F}$ has zero mean. The following theorem gives the rates of convergence of the estimated factors and loading functions.

**Theorem 1.** *Let $\hat{\mathbf{\Omega}}$ be the diagonal matrix whose elements are the eigenvalues of $\hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}}/(nT)$, and define $\hat{\boldsymbol{H}} = \hat{\mathbf{\Sigma}}_g(\boldsymbol{F}'\hat{\boldsymbol{F}}/T)\hat{\mathbf{\Omega}}^{-1}$. Then, under Assumptions 1 and 2, the following results hold either when $T$ is fixed or $T \to \infty$ as $n \to \infty$,:*
*(i) $\|\hat{\boldsymbol{F}} - \boldsymbol{F}\hat{\boldsymbol{H}}\|/\sqrt{T} = O_P(\varepsilon_{nT})$.*
*(ii) $\|\hat{\boldsymbol{G}}(\boldsymbol{X}) - \boldsymbol{G}(\boldsymbol{X})(\hat{\boldsymbol{H}}')^{-1}\|/\sqrt{n} = O_P(\varepsilon_{nT})$.*
*(iii) $\sup_{\boldsymbol{x}\in\mathcal{X}} \|\hat{\boldsymbol{g}}(\boldsymbol{x}) - \hat{\boldsymbol{H}}^{-1}\boldsymbol{g}(\boldsymbol{x})\| = O_P(\sqrt{k_n}\varepsilon_{nT})$.*

A few remarks on this result are relevant. First, it is worth highlighting that Theorem 1 (and Theorem 2 below) is obtained without imposing any restrictions on the time-series dependence of the idiosyncratic errors, while both Fan et al. (2016) and Ma et al. (2021) impose some kind of weak-time-series-dependence conditions. Second, while our setup does not require any moment restrictions on $u_{it}$, Assumption 3.4 of Fan et al. (2016) needs the error terms to have exponential tails. Third, the price to pay for these nice properties is that the convergence rates given in Theorem 1 are generally slower than those of Fan et al. (2016), mainly due to our use of sieve quantile estimators rather than sieve least square estimators. In fact, in the proof of Theorem 1, we only use the uniform convergence rate of $\hat{\boldsymbol{a}}_t$ (see Lemma 1 in the Appendix). Yet, by exploring the Bahadur representation of $\hat{\boldsymbol{a}}_t$, the convergence rate of the estimated loading functions can be improved when $T$ is large (see Theorem 3 below), whilst the convergence rate of the estimated factors can be greatly improved even when $T$ is fixed if the following extra assumptions are imposed.

**Assumption 3.** *Let $L$ be a generic bounded constant and let $\mathsf{f}(\cdot)$ denote the p.d.f. of $u_{it}$.*
*(i) For each $i$, $\boldsymbol{x}_i$ is independent of $(u_{i1},\ldots,u_{iT})$.*
*(ii) $|\mathsf{f}(c)-\mathsf{f}(0)|\leq L|c|$ for any $c$ in a neighborhood of $0$.*
*(iii) Equation (4) holds for some $\alpha\geq 3$.*

Assumption 3(i) essentially requires that the observed characteristics only affect the location but not the scale of the distributions of $y_{it}$. In such a case, the leading term in the Bahadur representation of $\hat{\boldsymbol{a}}_t$ has a similar structure to the least square estimators (see Lemma 2 in the Appendix). Thus, this assumption implies an improved convergence rate of $\hat{\boldsymbol{F}}$ which happens to be as fast as that of the PPCA estimators (see Theorem 4.1 of Fan et al. (2016)).

**Theorem 2.** *Let $\eta_{nT}=\sqrt{\ln(k_n^{-1/4}\varepsilon_{nT}^{-1/2})}\cdot k_n^{5/4}\varepsilon_{nT}^{1/2}n^{-1/2}$. Under Assumptions 1 to 3, we have*

$$\|\hat{\boldsymbol{F}}-\boldsymbol{F}\hat{\boldsymbol{H}}\|/\sqrt{T}=O_P\left(n^{-1/2}\vee k_n^{-\alpha}\vee\eta_{nT}\vee\varepsilon_{nT}^2\right).$$

*Moreover, if $T\asymp n^{\gamma_1}$ and $k_n\asymp n^{1/(6+\gamma_2)}$ for some $\gamma_1\geq 0$ and $\gamma_2>0$, then*

$$\|\hat{\boldsymbol{F}}-\boldsymbol{F}\hat{\boldsymbol{H}}\|/\sqrt{T}=O_P\left(n^{-1/2}\vee k_n^{-\alpha}\right).$$

**Remark 4.** *The term $\eta_{nT}$ in Theorem 2 represents the higher-order terms in the Bahadur representation of $\hat{\boldsymbol{a}}_t$. When $\alpha$ is large, $\eta_{nT}$ is approximately equal to $k_n^{3/2}n^{-3/4}$. Note that this slightly unusual expression of $\eta_{nT}$ is mainly due to the non-smoothness of the check function. Indeed, similar terms can be found in Theorem 2 of Horowitz and Lee (2005), Theorem 3.2 of Kato et al. (2012) and Theorem 2 of Ma et al. (2021).*

### 3.2 Asymptotic distribution

Define $\boldsymbol{\Sigma}_{\mathsf{f}\phi} = \mathbb{E}[\mathsf{f}(0|\boldsymbol{x}_i)\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)']$ and $\sigma_{k_n}^2 = \boldsymbol{\phi}_{k_n}'(\boldsymbol{x})\boldsymbol{\Sigma}_{\mathsf{f}\phi}^{-1}\boldsymbol{\Sigma}_{\phi}\boldsymbol{\Sigma}_{\mathsf{f}\phi}^{-1}\boldsymbol{\phi}_{k_n}(\boldsymbol{x})$.

**Assumption 4.** *Let $L$ be a generic bounded constant.*
*(i) $u_{i1}, \ldots, u_{iT}$ are independent conditional on $\boldsymbol{x}_i$.*
*(ii) $|\mathsf{f}(c|\boldsymbol{x}) - \mathsf{f}(0|\boldsymbol{x})| \leq L|c|$ for any $c$ in a neighborhood of $0$ and any $\boldsymbol{x} \in \mathcal{X}$.*
*(iii) There exist constants $c_1, c_2$ such that $0 < c_1 \leq \lambda_{\min}(\boldsymbol{\Sigma}_{\mathsf{f}\phi}) \leq \lambda_{\max}(\boldsymbol{\Sigma}_{\mathsf{f}\phi}) \leq c_2 < \infty$ for all $k_n$.*
*(iv) $(nT)^{1/2}k_n^{1/2-\alpha}\sigma_{k_n}^{-1} = o(1)$ and $(nT)^{1/2}k_n^{1/2}\eta_{nT}\sigma_{k_n}^{-1} = o(1)$.*

Assumption 4(i) is adopted for simplicity, though it could be replaced by $\beta-$mixing conditions at the cost of getting more complex asymptotic covariance matrices. When $\sigma_{k_n} \asymp k_n^{1/2}$ and $T$ is fixed, Assumption 4(iv) essentially requires that $n^{1/2}k_n^{-\alpha} = o(1)$ and $n^{1/2}\eta_{nT} = o(1)$, or $k_n^6 \ll n \ll k_n^{2\alpha}$. As a result, we need (4) to hold with $\alpha > 3$. The remaining conditions in Assumption 4 are standard — see, e.g. Assumptions 3 and 5 of Horowitz and Lee (2005). We are now in the position of establishing the asymptotic distribution of the estimated loading functions.

**Theorem 3.** *Under Assumptions 1, 2, and 4, it holds that for any $\boldsymbol{x} \in \mathcal{X}$*

$$\boldsymbol{\Sigma}_{T,\tau}^{-1/2}(\hat{\boldsymbol{H}}')^{-1} \cdot \frac{\sqrt{nT}}{\sigma_{k_n}}\Big(\hat{\boldsymbol{g}}(\boldsymbol{x}) - (\boldsymbol{F}'\hat{\boldsymbol{F}}/T)'\boldsymbol{g}(\boldsymbol{x})\Big) \xrightarrow{d} N(0, \boldsymbol{I}_R),$$

*where $\boldsymbol{\Sigma}_{T,\tau} = \tau(1-\tau)(\boldsymbol{F}'\boldsymbol{F}/T)$.*

The asymptotic distribution of $\hat{\boldsymbol{F}}$ is more difficult to derive, especially when $n$ and $T$ go to infinity simultaneously. For this reason, instead of focusing on $\hat{\boldsymbol{F}}$, let us consider the following updated estimator for the factors:

$$\tilde{\boldsymbol{F}} = \hat{\boldsymbol{Y}}'\hat{\boldsymbol{G}}(\boldsymbol{X}) \cdot (\hat{\boldsymbol{G}}(\boldsymbol{X})'\hat{\boldsymbol{G}}(\boldsymbol{X}))^{-1}.$$

In addition, let $\tilde{\boldsymbol{H}} = (\boldsymbol{G}(\boldsymbol{X})'\hat{\boldsymbol{G}}(\boldsymbol{X})/n)(\hat{\boldsymbol{G}}(\boldsymbol{X})'\hat{\boldsymbol{G}}(\boldsymbol{X})/n)^{-1}$ and

$$\boldsymbol{\Xi}_\tau = \tau(1-\tau) \cdot \boldsymbol{\Sigma}_g^{-1}\mathbb{E}[\boldsymbol{g}(\boldsymbol{x}_i)\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)']\boldsymbol{\Sigma}_{\mathsf{f}\phi}^{-1}\boldsymbol{\Sigma}_{\phi}\boldsymbol{\Sigma}_{\mathsf{f}\phi}^{-1}\mathbb{E}[\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)']\boldsymbol{\Sigma}_g^{-1}.$$

Then, the following additional assumption is needed to derive the asymptotic distribution of $\tilde{\boldsymbol{f}}_t$.

**Assumption 5.** *Conditions (i) to (iii) of Assumption 4 hold and $n^{1/2}k_n^{-\alpha}/\|\boldsymbol{\Xi}_\tau\|^{1/2} = o(1)$, $n^{1/2}\eta_{nT}/\|\boldsymbol{\Xi}_\tau\|^{1/2} = o(1)$, $\varepsilon_{nT}\sqrt{k_n} = o(1)$.*

**Theorem 4.** *Under Assumptions 1, 2, and 5, it holds for all $t = 1, \ldots, T$,*

$$\boldsymbol{\Xi}_\tau^{-1/2}(\hat{\boldsymbol{H}}')^{-1}\sqrt{n}(\tilde{\boldsymbol{f}}_t - \tilde{\boldsymbol{H}}'\boldsymbol{f}_t) \xrightarrow{d} N(0, \boldsymbol{I}_R).$$

When $\|\boldsymbol{\Xi}_\tau\| \asymp k_n$, the convergence rate of $\tilde{\boldsymbol{f}}_t$ is $O_P(\sqrt{n/k_n})$, and Assumption 5 requires that $n^{1/2}k_n^{-\alpha-1/2} = o(1)$ and $n^{1/2}\eta_{nT}k_n^{-1/2} = o(1)$, or $k_n^4 \ll n \ll k_n^{2\alpha+1}$. As a result, we need (4) to hold with $\alpha \geq 2$. Alternatively, if Assumption 3(i) holds, it can be shown that

$$\|\boldsymbol{\Xi}_\tau - \tau(1-\tau)\cdot\boldsymbol{\Sigma}_g^{-1}\cdot \mathsf{f}^{-2}(0)\| = O(k_n^{-\alpha}).$$

In this case, the convergence rate of $\tilde{\boldsymbol{f}}_t$ is $\sqrt{n}$ for each $t$.

**Remark 5.** *As in Proposition 1 of Bai (2003), it can be shown that $\hat{\boldsymbol{H}}$, $\tilde{\boldsymbol{H}}$ and $\boldsymbol{F}'\hat{\boldsymbol{F}}/T$ all converge in probability to some positive definite matrices as $n, T \to \infty$. In particular, if $\boldsymbol{F}'\boldsymbol{F}/T = \boldsymbol{I}_R$ and $\hat{\boldsymbol{\Sigma}}_g$ is diagonal, the probability limits of $\hat{\boldsymbol{H}}$, $\tilde{\boldsymbol{H}}$ and $\boldsymbol{F}'\hat{\boldsymbol{F}}/T$ are all equal to $\boldsymbol{I}_R$.*

**Remark 6.** *Note that both Theorems 3 and 4 are fulfilled when $T$ is fixed as well as when $T \to \infty$ as $n \to \infty$. In the latter case, if $n \asymp T$, Chen et al. (2021) show that the estimators of the quantile factors are $\sqrt{n}$-consistent and asymptotically normal under more general conditions. Thus, whenever $T$ is as large as $n$ and the quantile factors are the main objects of interest, the estimators of Chen et al. (2021) should be preferable. However, if $T$ is small and $n$ is large, Theorem 4 above shows that the QPPCA estimators proposed here remain consistent and asymptotically normal.*

### 3.3 Estimating the number of factors

Given that $\hat{\boldsymbol{Y}} = \boldsymbol{\Phi}(\boldsymbol{X})\hat{\boldsymbol{A}} \approx \boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{A}_0 = \boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{B}_0\boldsymbol{F}' \approx \boldsymbol{G}(\boldsymbol{X})\boldsymbol{F}'$, the rank of $\hat{\boldsymbol{Y}}$ is asymptotically equal to $R$. Let $\hat{\rho}_1, \ldots, \hat{\rho}_{\bar{R}}$ be the $\bar{R}$ largest eigenvalues of $\hat{\boldsymbol{Y}}\hat{\boldsymbol{Y}}'/(nT)$ in descending order. Then, the estimator of $R$ is given by the number of non-vanishing eigenvalues of $\hat{\boldsymbol{Y}}\hat{\boldsymbol{Y}}'/(nT)$, i.e.

$$\hat{R} = \sum_{j=1}^{\bar{R}} \mathbf{1}\{\hat{\rho}_j > p_n\}, \tag{8}$$

where $\{p_n\}$ is a sequence of non-increasing positive constants. The following theorem provides conditions on the threshold $p_n$ to establish the consistency of $\hat{R}$ which, following Chen et al. (2021), is denoted as the rank minimization estimator of the number of factors.

**Theorem 5.** *Suppose that $\bar{R} \geq R$ and $p_n \to 0$, $p_n\varepsilon_{nT}^{-1} \to \infty$ as $n \to \infty$, then under Assumptions 1 and 2, we have $P[\hat{R} = R] \to 1$ as $n \to \infty$.*

To prove Theorem 5, we show that the largest $R$ eigenvalues of $\hat{\boldsymbol{Y}}\hat{\boldsymbol{Y}}'/(nT)$ converge in probability to some positive constants, while the remaining eigenvalues are all $O_P(\varepsilon_{nT})$. Then, the decreasing sequence $\{p_n\}$ is chosen to dominate the vanishing eigenvalues in the limit. Again, this result also holds even when $T$ is fixed.

In theory, the choice of $p_n$ is determined by $\alpha$, which depends on the smoothness of the unknown quantile loading functions. Thus, a conservative choice of $p_n$ can rely on assuming that $\alpha = 1$. In this case, $\varepsilon_{nT} = (k_n^{1/2} n^{-1/2} \vee k_n^{-1}) \ln T$, and the optimal choice of $k_n$ is $k_n^* \asymp n^{1/3}$. Hence, to satisfy the condition of Theorem 5, we need $p_n \gg n^{-1/3} \ln T$. The following choice is recommended in practice:

$$p_n = d \cdot \hat{\rho}_1^{1/2} \cdot n^{-1/4} \ln T, \tag{9}$$

where $d$ is a positive constant and $\hat{\rho}_1^{1/2}$ plays the role of a normalization factor.

**Remark 7.** *Alternatively, to avoid the choice of the threshold sequence $\{p_n\}$, use could be made of the approach proposed by Ahn and Horenstein (2013) to estimate the number of factor by maximizing the ratios of consecutive eigenvalues, i.e.*

$$\tilde{R} = \underset{j=1,\ldots,\bar{R}}{\arg\max} \frac{\hat{\rho}_j}{\hat{\rho}_{j+1}}.$$

*This is the estimator considered by Fan et al. (2016) in the context of AFM where the error terms are required to be sub-Gaussian. For this reason, a formal proof of the consistency of this estimator in the context of QFM is technically challenging, being left for further research.*

# 4 Simulations

In this section, we run a few Monte Carlo simulations to study the behavior in finite samples of the QPPCA estimators regarding the estimation of the number of factors, the factors themselves and their loading functions. In most cases, unless otherwise explicitly said, we suppose that the number of characteristics is $D = 5$ and that all of them, $\{x_{id}, d = 1, \ldots, 5\}$, are drawn independently from the uniform distribution: $U[-1, 1]$.

## 4.1 Estimating the number of factors

Consider the following DGP:

$$y_{it} = \sum_{r=1}^{3} \lambda_{ir} f_{tr} + \left(x_{i1}^2 + x_{i2}^2 + x_{i3}^2\right) u_{it},$$

where $f_{t1} = 1$, $f_{t2}, f_{t3} \sim i.i.d\ N(0, 1)$. Note that the chosen DGP is a location-scale shift model where the scale is driven by a subset of the five characteristics. This type of heteroskedasticity implies that the quantile loading functions exhibit variations across quantiles, unlike a pure location-shift model where the loading functions would be the same (up to a constant) for different quantiles.

Let $g_1(x) = sin(2\pi x)$, $g_2(x) = sin(\pi x)$ and $g_3(x) = cos(\pi x)$, such that

$$\lambda_{i1} = \sum_{d=1,3,5} g_1(x_{id}), \quad \lambda_{i2} = \sum_{d=1,2} g_2(x_{id}), \quad \lambda_{i3} = \sum_{d=3,4} g_3(x_{id}).$$

As for the idiosyncratic component, $u_{it}$ are i.i.d. draws from three alternative distributions: (i) the standard normal distribution, $N(0,1)$, (ii) the Student's t distribution with 3 degrees of freedom, $t(3)$, and (iii) the standard Cauchy distribution, Cauchy(0,1). In the first-step, we set $k_n = n^{1/3}$ in the quantile sieve estimation, and make use of the *Chebyshev polynomials of the second kind* as the basis functions. Moreover, in order to implement the rank minimization estimator for the number of factors in (8), the threshold $p_n$ is chosen as in (9), with $d = 1/4$.

First, Table 1 displays the results of the number of factors estimated with the rank minimization criterion for $\tau \in \{0.25, 0.5, 0.75\}$, $T \in \{5, 10\}$ and $n \in \{50, 100, 200, 1000\}$ from 1000 simulation replications. For each combination of $\tau$, $n$ and $T$, the reported results represent: [frequency of $\hat{R} < R$; frequency of $\hat{R} = R$; frequency of $\hat{R} > R$]. Next, for comparison, Table 2 reports the corresponding results when the number of factors is estimated using the Ahn and Horenstein (2013)'s eigen-ratio estimator discussed in Remark 7.

There are three main takeaways from these simulation results. First, both selection criteria accurately estimate the number of factors when $T$ is small and $n$ is large, supporting our previous claim about their consistency even when $T$ is fixed. Second, when $n$ is large (=1000), both estimators perform well, even when the errors follow the standard Cauchy distribution. Hence, this result also provides support for the claim that our estimator is consistent in the absence of moment restrictions on the error terms. Third, although both estimators yield similar results when $n = 1000$, the rank minimization estimator outperforms the eigen-ratio estimator when $n$ is not sufficiently large.

## 4.2 Estimating the factors

### 4.2.1 Comparison of QPPCA with PCA, PPCA and QFA

Following Chen et al. (2021), we consider the following DGP:

$$y_{it} = \lambda_{i1} f_{t1} + \lambda_{i2} f_{t2} + (\lambda_{i3} f_{t3}) u_{it},$$

where $f_{t3} = |h_t|, f_{t1}, f_{t2}, h_t \sim i.i.d\ N(0,1)$. As before, let $g_1(x) = sin(2\pi x)$, $g_2(x) = sin(\pi x)$ but now $g_3(x) = |cos(\pi x)|$. The factor loading functions and the error terms are also generated as in Subsection 4.1. Note that, in this DGP, there are two location shift factors, $f_{t1}$ and $f_{t2}$, that affect the mean of $y_{it}$ and only one scale shift factor $f_{t3}$ that affects the variance of $y_{it}$.

First, we focus on the estimation of the two location factors: $f_{t1}$ and $f_{t2}$. Four competing estimation methods are considered: (i) our the proposed method with $\tau = 0.5$ (QPPCA); (ii) the quantile factor analysis estimator (QFA) of Chen et al. (2021) with $\tau = 0.5$; (iii) the projection estimator proposed by Fan et al. (2016) (PPCA); and (iv) the standard estimator of Bai and Ng (2002) for AFM (PCA). For the first two methods, the choices of $k_n$ and the basis functions are again the same as in Subsections 4.1 and 4.2.

Regarding the choices of $n$ and $T$, two different scenarios are considered:

(i) Fix $T = 10, 50$ and let $n$ increase from 50 to 500.

(ii) Fix $n = 100, 200$ and let $T$ increase from 5 to 200.

For each estimation method, the number of factors ($R = 2$ at $\tau = 0.5$) is assumed to be known, and we report the average Frobenius error as a measure of fit: $\|\hat{\boldsymbol{F}} - \boldsymbol{F}\hat{\boldsymbol{H}}\|/\sqrt{T}$ from 1000 replications, where $\hat{\boldsymbol{H}}$ represents the associated rotation matrix for each estimator.

The results for the first scenario (fixed $T$ and increasing $n$) are plotted in Figure 1. As can be inspected, for small $T$ ($T = 10$), the PCA and QFA estimators perform worse than the PPCA and QPPCA estimators when $u_{it}$ is either drawn from the $N(0, 1)$ or $t(3)$ distributions. Moreover, when the distribution is a standard Cauchy, the QPPCA estimator performs much better than its competitors. These findings agree again with our previous theoretical results showing that this estimator is consistent even when $T$ is fixed or the moments of $u_{it}$ do not exist.

When $T$ is relatively large ($T = 50$) and the distribution of $u_{it}$ has a thin tail, like a $N(0, 1)$ random variable, all the estimators behave similarly, as long as $n \geq 100$. However, if $u_{it}$ follows the $t(3)$ distribution, the PCA estimator is subject to a much larger estimation error than the alternative procedures. In the extreme case of the standard Cauchy distribution, the two methods based on quantile regressions are the obvious winners, with the performances of the QFA and QPPCA estimators being very similar insofar $n \geq 200$.

The results for the second scenario (fixed $n$ and increasing $T$) are displayed in Figure 2. The main takeaway from this simulation exercise is that the QPPCA estimator provides the most robust approach against heavy-tailed distributions when $T$ is small, while only the QFA estimator performs slightly better as $T$ increases.

Next, we proceed to estimate all the three factors jointly, paying particular attention to the results for the scale factor $f_{t3}$. Since this last factor is absent when $\tau = 0.5$, for brevity we only provide simulations for $\tau = 0.25, 0.75$, and sample sizes where $T \in \{10, 50\}$ and $n \in \{50, 100, 200\}$. In each of these setups, the three estimated factors by the four different approaches are denoted as $\hat{F}^{\tau}_{QPPCA}, \hat{F}^{\tau}_{QFA}, \hat{F}_{PPCA}, \hat{F}_{PCA}$. Subsequently, each of the true factors is regressed on these estimated factors and the adjusted $R^2$s are computed as a measure of goodness of fit. The whole procedure is repeated 1000 times and the averages of the adjusted $R^2$s are reported in Tables 3 to 5.

Table 3 displays the results for the QPPCA estimator. As can be observed, it performs well in estimating all the three factors. It should be noted, however, that the estimates of the scale factor $f_{t3}$ are not as good as the estimates of the two location factors, $f_{t1}, f_{t2}$, when $n$ is small, though the fit improves substantially as $n$ increases. Table 4, in turn, presents the results for the QFA estimator, whereas Table 5 presents the corresponding results for the PCA and PPCA estimators. The main finding from Table 4 is that the QFA estimator performs poorly in estimating the scale factor $f_{t3}$ when $T$ is small ($T = 10$), while it performs similarly to the QPPCA estimator when $T$ is relatively large ($T = 50$). Finally and not surprisingly, the main conclusion from Table 5 is that both the PCA and PPCA estimators fail to capture the scale factor $f_{t3}$ in all instances since they are designed for AFM but not for QFM.

#### 4.2.2 Comparison of QPPCA with SQFA

In the previous subsection the SQFA estimator proposed by Ma et al. (2021) was not included in the set of comparisons since its performance is close to that of the QFA estimator whenever the number of characteristics is larger or equal to the number of factors ($D \geq R$). Yet, in this subsection, we study how they differ when the number of characteristics is smaller than the number of factors ($D < R$).

To do so, we consider the following location-scale model as the DGP:

$$y_{it} = \lambda_{i1} f_{t1} + \lambda_{i2} f_{t2} + (\lambda_{i3} f_{t3}) u_{it},$$

where $f_{t3} = |h_t|, f_{t1}, f_{t2}, h_t \sim i.i.d\ N(0, 1)$. Now, the number of characteristics is 2 and, as in the previous simulations, all characteristics $x_{id}$ ($i = 1, ..., n$ and $d = 1, 2$) are independently drawn from the uniform distribution: $U[-1, 1]$. Let $g_1(x) = sin(2\pi x),\ g_2(x) = sin(\pi x)\ and\ g_3(x) = |cos(\pi x)|$. Moreover, let $\lambda_{i1} = \sum_{d=1,2} g_1(x_{id}),\ \lambda_{i2} = \sum_{d=1,2} g_2(x_{id})\ and\ \lambda_{i3} = \sum_{d=1,2} g_3(x_{id})$. Again, $u_{it}$ are generated from three different distributions discussed in Subsections 4.1 and 4.2.

For each estimator, we consider $\tau \in \{0.25, 0.5, 075\}$, $T \in \{10, 50\}$, $n \in \{50, 100, 200, 500\}$. Note that, when $\tau = 0.5$, there are only two location factors because $f_{t3}$ does not affect the median of $y_{it}$. By contrast, when $\tau = 0.25, 0.75$ there will be two location factors and one scale factor. Moreover, to simplify the analysis, $R$ is assumed to be known. For each $\tau$, $R$ factors are estimated using QPPCA. Note that the SQFA method chooses the number of factors as the number of characteristics by default, implying that only two factors will be estimated. Moreover, the choices of the basis functions and $k_n$ are the same as in Subsection 4.1.

As before, we proceed to regress each of the true factors on the estimated factors and compute the adjusted $R^2$s. The whole procedure is repeated 1000 times and the averages of the adjusted $R^2$s are reported in Tables 6 and 7 for the QPPCA and SQFA estimators, respectively. When it comes to the estimation of the volatility factor, $f_{t3}$, it is not surprising to check that the

QPPCA estimator outperforms the SQFA estimator since the latter is restricted to estimating only $D = 2$ factors. Thus, the main finding here is that the QPPCA estimator performs better than the SQFA estimator in estimating the location and scale shift factors whenever the number of factors exceeds the number of characteristics.

### 4.3 Estimating the loading functions

Finally, we close this section by analyzing the estimation of the quantile loading functions. To do so, consider the following DGP:

$$y_{it} = \lambda_{i1} f_{t1} + (\lambda_{i2} f_{t2}) u_{it},$$

where $f_{t2} = |g_t|$ and $f_{t1}, g_t \sim i.i.d\ N(0, 1)$. The number of characteristics is 2 and all characteristics $x_{id}$ ($i = 1, ..., n$ and $d = 1, 2$) are independently drawn from uniform distribution: $U[-1, 1]$. Let $g_{11}(x) = sin(2\pi x), g_{21}(x) = 0, g_{12}(x) = sin(\pi x)$, $g_{22}(x) = cos^2(\pi x)$ and $\lambda_{i1} = g_{11}(x_{i1}) + g_{12}(x_{i2})$, $\lambda_{i2} = g_{21}(x_{i1}) + g_{22}(x_{i2})$. This time, $u_{it}$ are independently drawn from the $N(0, 1)$ and the $t(3)$ distributions, excluding Cauchy. Note that this DGP involves one location factor, $f_{1t}$, as well as a single scale factor, $f_{2t}$. Moreover, $g_{11}$ and $g_{12}$ are related to the location factor and they are quantile-invariant, while $g_{21}$ and $g_{22}$ are associated with the scale factor and they are quantile-dependent. In particular, suppose the $\tau$th quantile of $u_{it}$ is $\mathsf{Q}_\tau$, then the true loading function is given by $g_{\tau,2d}(x) = \mathsf{Q}_\tau \cdot g_{2d}(x)$.

Given its better performance in the simulations above, we only consider our QPPCA estimator in this simulation exercise and set $n = 500, T = 10$. Once again, the choices of basis functions and $k_n$ are as in Subsection 4.1. For each distribution of the error terms, we estimate the loading functions at $\tau = 0.25, 0.5, 0.75$ by taking 201 equidistant points within the interval [-1,1] and compute the estimated function values at these points. By repeating this procedure 1000 times, the lower 5% and 95% quantiles of these replications are reported at each point. When the error term follows the $N(0, 1)$ distribution, the estimated loading functions of the first and second characteristics are plotted in Figures 3 and 4, respectively, while the corresponding loading functions when the errors follow the $t(3)$ distribution are displayed in Tables 5 an 6. In line with our theoretical results, the main lesson to be drawn from these simulations is the good performance of the QPPCA estimator in retrieving the true loading functions even when $T$ is not large.

## 5 Empirical Application

In this section, the QPPCA estimation method is applied to investigate the factor structure of security returns. Following Fan et al. (2016), we use a dataset that includes information on the daily returns of S&P500 index securities with complete daily closing price records from 2005 to 2013.[4] The sample consists of 355 stocks, whose book value and market capitalization are drawn from Compustat. Moreover, as is conventional in this literature, the 1-month US treasury bond rate is used as the risk-free rate to compute the daily excess return of each stock.

Following Connor et al. (2012), Fan et al. (2016), and Ma et al. (2021), four characteristics are considered: *size, value, momentum* and *volatility*, which are standardized to have zero means and unit standard deviations. Similar to Fan et al. (2016), we analyze the data corresponding to the first quarter of 2006, which includes $T = 62$ observations. Once again, the second Chebyshev polynomials are used as the basis functions in the sieve regressions and $k_n = 4$.

First, Table 8 shows the estimated number of mean factors using the eigen-ratio estimator proposed by Fan et al. (2016) and the estimated numbers of quantile factors using the QPPCA rank minimization estimator for $\tau \in \{0.05, 0.25, 0.5, 0.75, 0.95\}$. The five largest eigenvalues of $\hat{Y}\hat{Y}'$ and the threshold $p_n$ are also displayed in this table. In addition, the estimated numbers of quantile factors using the QFA rank minimization estimator are reported in the last column. Overall, the results provide strong evidence in favor of the existence of a single location factor and one factor at each quantile.

Second, Table 9 shows the correlation coefficients between the estimated location factors by PPCA and the estimated quantile factors by QPPCA for the above-mentioned values of $\tau$. The sample means of each estimated factor are also reported in the last column. Figure 7 in turn provides plots of these factors which exhibit different means and high correlations. As can be inspected, the PPCA factor is highly correlated with the QPPCA factor corresponding to $\tau = 0.5$, but misses the factors at the more extreme quantiles.

Third, Figure 8 shows the estimated loading functions of the four characteristics using PPCA and QPPCA at $\tau = 0.5$ where values of each standardized covariate appears in the horizontal axis. The fact that both methods yield similar estimated loading functions indicates that the idiosyncratic errors of the stock returns have symmetric distributions so that the mean and the median coincide.

Fourth, given its better performance, Figure 9 plots the estimated loading functions of the four characteristics under consideration using QPPCA at different quantiles. In general, these functions feature considerable variation both across the values of the characteristics and the quantiles. A few salient findings emerge. First, the loading functions of size and volatility seem to behave monotonically at all quantiles while, for value and momentum, they exhibit a clear non-linear pattern, mostly looking U-shaped. By the way, the shapes of the loading function resemble those reported by Ma et al. (2021) except for value (i.e. a value stock refers to shares of a company that appears to trade at a lower price relative to its fundamentals, such as dividends, earnings, or sales) which these authors find to have an inverted U-shape. Next, for all

[4]This dataset is downloaded from CRSP (Center for Research in Security Prices).

characteristics, their quantile loading functions at $\tau = 0.25$ and $\tau = 0.5$ are very close. Lastly, there is strong evidence that the loading functions at the tails ($\tau = 0.05, 0.95$) have greater curvatures than at the remaining quantiles. In sum, this empirical evidence points out that the estimated loading functions vary substantially across different quantiles, a fact that cannot be uncovered using the PPCA method. Yet, this is a useful finding since deviating from the efficient market hypothesis, factors contributing to alpha generation can have different relevance depending on the distribution of excess returns. Thus, the standard asset valuation techniques based on CAPM and the Fama-French factors should take these features into consideration to create a portfolio delivering excess returns over time beating the market.

Finally, it is worth highlighting that QPPCA allows estimating the conditional quantile of excess returns $\mathsf{Q}_\tau[y_{it}|\boldsymbol{x}_i] = \boldsymbol{g}_\tau(\boldsymbol{x}_i)'\boldsymbol{f}_t$, yielding $\hat{y}_{it}(\tau) = \hat{\boldsymbol{a}}_t'\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)$ as its estimator, where $\hat{\boldsymbol{a}}_t$ is obtained from the cross-sectional quantile regressions in step 1 of the three-step procedure introduced in Section 2. One could interpret $\hat{y}_{it}(\tau)$ as the "quantile return" which is interesting from the perspective of empirical applications since it is idiosyncratic free, that is, much less noisy than the realized return $y_{it}$. Just as the literature on asset pricing has increasingly appreciated the concept of "expected returns" because it is noiseless (see e.g. Elton (1999)), "quantile returns" are also interesting on their own and could perhaps help provide a better explanation of the distribution of returns, an issue which remains high in our research agenda.

# 6 Conclusions

This paper proposes a three-stage estimation method for characteristic-based quantile factor models (CQFM). The convergence rates of the proposed estimators, labeled QPPCA, are established, and the asymptotic distributions of the estimated factors and loading functions are derived under very general conditions. Compared with the existing estimation methods of CQFM, not only QPPCA estimators are easier to implement in practice, but also they are consistent for fixed $T$ as long as $n$ goes to infinity, as well as being robust to heavy tails and outliers in the distribution of the idiosyncratic errors. Moreover, the number of quantile factors are allowed to be different from the number of the characteristics, and this number can be consistently estimated using a new rank-minimization estimator proposed in this paper.

Simulation results show that the proposed estimators perform satisfactorily in finite samples, especially when the number of cross-section observations is large. An application of the estimators to a dataset consisting of individual stock returns reveals that the quantile factor loadings are nonlinear functions of some observed characteristics and that these functions exhibit considerable variations across quantiles. We conjecture that this leads to the concept of *quantile returns* which generalizes the standard concept of *expected returns*, typically proxied by averages of realized returns.

The methodology associated with QPPCA is useful to derive the convergence rates and asymptotic properties of the (average) quantile returns which remains high in our ongoing research agenda. Moreover, for the tractability of the problem, it has been assumed that the quantile factor loadings can be fully explained by the observed characteristics. Admittedly, this is a restrictive assumption. Relaxing this assumption and allowing the factor loadings to be functions of other unobserved characteristics is a challenging task in the context of quantile regressions. This interesting question is also left for future research.

# A Proofs of the Main Results

### Proof of Proposition 1:

*Proof.* For any $\theta \in \Theta$, define $K(\theta, \theta_{0t}) = \mathbb{E}(L_n(\theta)) = \mathbb{E}[l(\theta, y_{it}, \boldsymbol{x}_i)]$. Under Assumption 1(iv), it can be shown that $K(\theta, \theta_{0t}) \asymp d(\theta, \theta_{0t})^2$. For the finite-dimensional linear sieve spaces $\Theta_n$, it can be shown that Condition A.3 of Chen and Shen (1998) is satisfied with $\delta_n = \sqrt{k_n/n}$ (see Section 3.3 of Chen (2007)). By the definition of $d$ and the properties of the check function, it is easy to see that,[5]

$$\begin{aligned}\sup_{\theta\in\Theta_n, d(\theta,\theta_{0t})\le\varepsilon} \operatorname{Var}\left[l(\theta, y_{it}, \boldsymbol{x}_i)\right] &\le \sup_{\theta\in\Theta_n, d(\theta,\theta_{0t})\le\varepsilon} \mathbb{E}\left[l(\theta, y_{it}, \boldsymbol{x}_i)\right]^2 \\ &\lesssim \sup_{\theta\in\Theta_n, d(\theta,\theta_{0t})\le\varepsilon} \mathbb{E}\left(\theta(\boldsymbol{x}_i) - \theta_{0t}(\boldsymbol{x}_i)\right)^2 \le \varepsilon^2.\end{aligned}$$

Thus, Condition A.2 of Chen and Shen (1998) is also satisfied. By Assumption 1(iii) we have $\sup_{\theta\in\Theta} |l(\theta, y_{it}, \boldsymbol{x}_i)| \lesssim \sup_{\theta\in\Theta}\sup_{\mathcal{X}} |\theta(\boldsymbol{x}) - \theta_{0t}(\boldsymbol{x})| < \infty$. Assumption 1(ii) implies that $d(\pi_n\theta_{0t}, \theta_{0t}) = \sqrt{\mathbb{E}\left(\pi_n\theta_{0t}(\boldsymbol{x}_i) - \theta_{0t}(\boldsymbol{x}_i)\right)^2} = O(k_n^{-\alpha})$. Therefore, it follows from Corollary 1 of Chen and Shen (1998) that

$$P\left[\max_t d(\hat{\theta}_{nt}, \theta_{0t}) \ge C\varepsilon_{nT}\right] \le \sum_{t=1}^{T} P\left[d(\hat{\theta}_{nt}, \theta_{0t}) \ge C\varepsilon_{nT}\right] \le c_1 \exp\left\{C^2 \ln T(1 - c_2 n\varepsilon_n^2)\right\}$$

for any $C \ge 1$. Therefore, the desired result follows from the above inequality since $n\varepsilon_n^2 \ge k_n$. □

**Lemma 1.** *If Assumption 1 and Assumption 2(i) hold, and $\varepsilon_n$ is defined as in Assumption 1, then:*
*(i)* $\max_{1\le t\le T} \|\hat{\boldsymbol{a}}_t - \boldsymbol{a}_{0t}\| = O_P(\varepsilon_{nT})$;
*(ii) Let* $\hat{\boldsymbol{V}} \equiv \hat{\boldsymbol{Y}} - \boldsymbol{G}(\boldsymbol{X})\boldsymbol{F}'$, *then* $(nT)^{-1/2}\|\hat{\boldsymbol{V}}\| = O_P(\varepsilon_{nT})$.

*Proof.* By Assumption 1 and Assumption 2(i),

$$\begin{aligned}d(\hat{\theta}_{nt}, \theta_{0t})^2 &= \int_{\mathcal{X}} \left(\hat{\theta}_{nt}(\boldsymbol{x}) - \theta_{0t}(\boldsymbol{x})\right)^2 d\mathsf{F}_x(\boldsymbol{x}) = \int_{\mathcal{X}} \left(\hat{\theta}_{nt}(\boldsymbol{x}) - \pi_n\theta_{0t}(\boldsymbol{x})\right)^2 d\mathsf{F}_x(\boldsymbol{x}) + O_P(\varepsilon_{nT}k_n^{-\alpha}) \\ &= (\hat{\boldsymbol{a}}_t - \boldsymbol{a}_{0t})'\boldsymbol{\Sigma}_\phi(\hat{\boldsymbol{a}}_t - \boldsymbol{a}_{0t}) + O_P(\varepsilon_{nT}k_n^{-\alpha}) \ge c_1\|\hat{\boldsymbol{a}}_t - \boldsymbol{a}_{0t}\|^2 + O_P(\varepsilon_{nT}k_n^{-\alpha})\end{aligned}$$

where $c_1 > 0$, and the $O_P(\varepsilon_{nT}k_n^{-\alpha})$ in the above equation is uniform in $t$. It then follows from Proposition 1 that $\max_{1\le t\le T} \|\hat{\boldsymbol{a}}_t - \boldsymbol{a}_{0t}\|^2 = O_P(\varepsilon_{nT}^2)$.

[5] Note that $|\rho_\tau(u_1) - \rho_\tau(u_2)| \le 2|u_1 - u_2|$.

Next, note that

$$\begin{aligned}(nT)^{-1}\|\hat{\boldsymbol{V}}\|^2 &\leq \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(\hat{\theta}_{nt}(\boldsymbol{x}_i)-\pi_n\theta_{0t}(\boldsymbol{x}_i)\right)^2+O_P(k_n^{-2\alpha}) \\ &= \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left((\hat{\boldsymbol{a}}_t-\boldsymbol{a}_{0t})'\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)\right)^2+O_P(k_n^{-2\alpha}) \\ &\leq T^{-1}\sum_{t=1}^{T}\|\hat{\boldsymbol{a}}_t-\boldsymbol{a}_{0t}\|^2\cdot\lambda_{\max}\left(\hat{\boldsymbol{\Sigma}}_\phi\right)+O_P(k_n^{-2\alpha}) \\ &\leq \max_{1\leq t\leq T}\|\hat{\boldsymbol{a}}_t-\boldsymbol{a}_{0t}\|^2\cdot\lambda_{\max}\left(\hat{\boldsymbol{\Sigma}}_\phi\right)+O_P(k_n^{-2\alpha})\end{aligned}$$

where $\hat{\boldsymbol{\Sigma}}_\phi \equiv n^{-1}\sum_{i=1}^n \boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)'$. Since Assumption 1(iii) implies that $\sup_{\mathcal{X}}\|\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)\| = \sqrt{k_n}$, similar to the proof of Theorem 1 in Newey (1997), one can show that $\|\hat{\boldsymbol{\Sigma}}_\phi - \boldsymbol{\Sigma}_\phi\| = o_P(1)$ under Assumption 2, and therefore we have $\lambda_{\max}(\hat{\boldsymbol{\Sigma}}_\phi) = O_P(1)$. This completes the proof. □

### Proof of Theorem 1:

*Proof.* Write $\hat{\boldsymbol{Y}} = \boldsymbol{G}(\boldsymbol{X})\boldsymbol{F}' + \hat{\boldsymbol{V}}$ where $\hat{\boldsymbol{V}}$ is as defined in Lemma 1. Let $\boldsymbol{\Omega}_R$ be the diagonal matrix whose elements are the eigenvalues of $\boldsymbol{\Sigma}_g \cdot \boldsymbol{F}'\boldsymbol{F}/T$. Note that

$$\begin{aligned}\hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}}/(nT) = \boldsymbol{F}\boldsymbol{G}(\boldsymbol{X})'\boldsymbol{G}(\boldsymbol{X})\boldsymbol{F}'/(nT) + \hat{\boldsymbol{V}}'\boldsymbol{G}(\boldsymbol{X})\boldsymbol{F}'/(nT) \\ + \boldsymbol{F}\boldsymbol{G}(\boldsymbol{X})'\hat{\boldsymbol{V}}/(nT) + \hat{\boldsymbol{V}}'\hat{\boldsymbol{V}}/(nT).\end{aligned} \tag{A.1}$$

It then follows from Assumption 2(iv), Assumption 1(i) and Lemma 1 that:

$$\begin{aligned}&\|\hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}}/(nT) - \boldsymbol{F}\boldsymbol{\Sigma}_g\boldsymbol{F}'/T\| \\ \leq\ & o_P(1) + 2\|\hat{\boldsymbol{V}}\|/\sqrt{nT}\cdot\|\boldsymbol{G}(\boldsymbol{X})\|/\sqrt{n}\cdot\|\boldsymbol{F}\|/\sqrt{T} + \|\hat{\boldsymbol{V}}\|^2/(nT) \\ =\ & o_P(1) + O_P(\varepsilon_{nT}).\end{aligned}$$

By the Wielandt-Hoffman inequality, we have $\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\| = o_P(1)$. It then follows from Assumption 2(iii) and 2(iv) that $\lambda_{\min}(\hat{\boldsymbol{\Omega}}) > 0$ with probability approaching 1.

By the definition of $\hat{\boldsymbol{F}}$, $\hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}}/(nT)\hat{\boldsymbol{F}} = \hat{\boldsymbol{F}}\hat{\boldsymbol{\Omega}}$, it then follows from (A.1) that

$$\hat{\boldsymbol{F}} = \boldsymbol{F}\hat{\boldsymbol{H}} + \hat{\boldsymbol{V}}'\boldsymbol{G}(\boldsymbol{X})\boldsymbol{F}'\hat{\boldsymbol{F}}/(nT)\hat{\boldsymbol{\Omega}}^{-1} + \boldsymbol{F}\boldsymbol{G}(\boldsymbol{X})'\hat{\boldsymbol{V}}\hat{\boldsymbol{F}}/(nT)\hat{\boldsymbol{\Omega}}^{-1} + \hat{\boldsymbol{V}}'\hat{\boldsymbol{V}}/(nT)\hat{\boldsymbol{F}}\hat{\boldsymbol{\Omega}}^{-1}. \tag{A.2}$$

Thus, it follows from (A.2) and Lemma 1 that

$$\|\hat{\boldsymbol{F}} - \boldsymbol{F}\hat{\boldsymbol{H}}\|/\sqrt{T} \leq 2O_P(1)\cdot\frac{\|\hat{\boldsymbol{V}}\|}{\sqrt{nT}}\cdot\frac{\|\boldsymbol{F}\|}{\sqrt{T}}\cdot\frac{\|\hat{\boldsymbol{F}}\|}{\sqrt{T}}\cdot\frac{\|\boldsymbol{G}(\boldsymbol{X})\|}{\sqrt{n}} + O_P(1)\cdot\frac{\|\hat{\boldsymbol{F}}\|}{\sqrt{T}}\cdot\frac{\|\hat{\boldsymbol{V}}\|^2}{nT} = O_P(\varepsilon_{nT}).$$

Then the first part of Theorem 1 follows.

Next, similar to the proof of Proposition 1 in Bai (2003) it can be shown that $\hat{\boldsymbol{H}} \to \boldsymbol{H} > 0$. Thus, $\hat{\boldsymbol{H}}$ is invertible with probability approaching 1. Note that $\hat{\boldsymbol{G}}(\boldsymbol{X}) = \hat{\boldsymbol{Y}}\hat{\boldsymbol{F}}/T = \boldsymbol{G}(\boldsymbol{X})\boldsymbol{F}'\hat{\boldsymbol{F}}/T + \hat{\boldsymbol{V}}\hat{\boldsymbol{F}}/T$. Write $\boldsymbol{F} = \hat{\boldsymbol{F}}\hat{\boldsymbol{H}}^{-1} + \boldsymbol{F} - \hat{\boldsymbol{F}}\hat{\boldsymbol{H}}^{-1}$, then

$$\hat{\boldsymbol{G}}(\boldsymbol{X}) = \boldsymbol{G}(\boldsymbol{X})(\hat{\boldsymbol{H}}')^{-1} + \boldsymbol{G}(\boldsymbol{X})(\boldsymbol{F} - \hat{\boldsymbol{F}}\hat{\boldsymbol{H}}^{-1})'\hat{\boldsymbol{F}}/T + \hat{\boldsymbol{V}}\hat{\boldsymbol{F}}/T,$$

and thus

$$\|\hat{\boldsymbol{G}}(\boldsymbol{X}) - \boldsymbol{G}(\boldsymbol{X})(\hat{\boldsymbol{H}}')^{-1}\|\sqrt{n} \le \frac{\|\boldsymbol{G}(\boldsymbol{X})\|}{\sqrt{n}} \cdot \frac{\|\boldsymbol{F} - \hat{\boldsymbol{F}}\hat{\boldsymbol{H}}^{-1}\|}{\sqrt{T}} \cdot \frac{\|\hat{\boldsymbol{F}}\|}{\sqrt{T}} + \frac{\|\hat{\boldsymbol{V}}\|}{\sqrt{nT}} \cdot \frac{\|\hat{\boldsymbol{F}}\|}{\sqrt{T}} = O_P(\varepsilon_{nT}).$$

Then the second part of Theorem 1 follows.

Finally, note that $\hat{\boldsymbol{B}} = \hat{\boldsymbol{A}}\hat{\boldsymbol{F}}/T = \boldsymbol{B}_0(\boldsymbol{F}'\hat{\boldsymbol{F}}/T) + (\hat{\boldsymbol{A}} - \boldsymbol{A}_0)\hat{\boldsymbol{F}}/T$. It follows from Proposition 1 that

$$\|\hat{\boldsymbol{B}} - \boldsymbol{B}_0(\boldsymbol{F}'\hat{\boldsymbol{F}}/T)\| \le \frac{\|\hat{\boldsymbol{A}} - \boldsymbol{A}_0\|}{\sqrt{T}} \cdot \frac{\|\hat{\boldsymbol{F}}\|}{\sqrt{T}} = O_P(\varepsilon_{nT}). \tag{A.3}$$

Thus, for any $\boldsymbol{x} \in \mathcal{X}$,

$$\begin{aligned} \hat{\boldsymbol{g}}(\boldsymbol{x})' &= \boldsymbol{\phi}_{k_n}(\boldsymbol{x})'\hat{\boldsymbol{B}} = \boldsymbol{\phi}_{k_n}(\boldsymbol{x})'\boldsymbol{B}_0(\boldsymbol{F}'\hat{\boldsymbol{F}}/T) + \boldsymbol{\phi}_{k_n}(\boldsymbol{x})'\left(\hat{\boldsymbol{B}} - \boldsymbol{B}_0(\boldsymbol{F}'\hat{\boldsymbol{F}}/T)\right) \\ &= \boldsymbol{g}(\boldsymbol{x})'(\hat{\boldsymbol{H}}^{-1})' + (\boldsymbol{\phi}_{k_n}(\boldsymbol{x})'\boldsymbol{B}_0 - \boldsymbol{g}(\boldsymbol{x})')(\boldsymbol{F}'\hat{\boldsymbol{F}}/T) + \boldsymbol{\phi}_{k_n}(\boldsymbol{x})'\left(\hat{\boldsymbol{B}} - \boldsymbol{B}_0(\boldsymbol{F}'\hat{\boldsymbol{F}}/T)\right) + O_P(\varepsilon_{nT}). \end{aligned}$$

Thus, it follows from (A.3) and Assumption 1 that

$$\sup_{\mathcal{X}} \left\|\hat{\boldsymbol{g}}(\boldsymbol{x}) - \hat{\boldsymbol{H}}^{-1}\boldsymbol{g}(\boldsymbol{x})\right\| \le O_P(k_n^{-\alpha}) + \sup_{\mathcal{X}} \|\boldsymbol{\phi}_{k_n}(\boldsymbol{x})\| \cdot O_P(\varepsilon_{nT}) = O_P(\sqrt{k_n}\varepsilon_{nT}).$$

This completes the proof. □

**Lemma 2.** *Let* $\xi_{it} = \theta_{0t}(\boldsymbol{x}_i) - \pi_n\theta_{0t}(\boldsymbol{x}_i) = \boldsymbol{g}(\boldsymbol{x}_i)'\boldsymbol{f}_t - \boldsymbol{a}_{0t}'\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)$ *and* $\psi_{it} = \mathsf{F}(-\xi_{it}) - \mathbf{1}\{u_{it} \le -\xi_{it}\}$*. If Assumptions 1 to 3 hold, then*

$$\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{\boldsymbol{a}}_t - \boldsymbol{a}_{0t} - \mathsf{f}^{-1}(0) \cdot \hat{\boldsymbol{\Sigma}}_{\phi}^{-1} \cdot \frac{1}{n}\sum_{i=1}^{n}\psi_{it}\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)\right\|^2} = O_P\left(k_n^{-\alpha}\right) + O_P\left(\eta_{nT}\right).$$

*Proof.* **Step 1:** For any $\boldsymbol{a} \in \mathbb{R}^{Dk_n}$ define:

$$\boldsymbol{m}_t(\boldsymbol{a}) = \frac{1}{n}\sum_{i=1}^{n}\left[\tau - \mathbf{1}\{u_{it} \le (\boldsymbol{a} - \boldsymbol{a}_{0t})'\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i) - \xi_{it}\}\right]\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i),$$

$$\boldsymbol{m}_t^*(\boldsymbol{a}) = \frac{1}{n}\sum_{i=1}^n \left[\tau - \mathsf{F}\left((\boldsymbol{a}-\boldsymbol{a}_{0t})'\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i) - \xi_{it}\right)\right]\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i).$$

Since $\mathsf{F}\left(-\xi_{it}\right) = \tau - \mathsf{f}\left(-\xi_{it}^*\right)\cdot\xi_{it}$ where $\xi_{it}^*$ is between 0 and $\xi_{it}$, it follows that

$$\boldsymbol{m}_t^*(\boldsymbol{a}_{0t}) = \frac{1}{n}\sum_{i=1}^n \mathsf{f}\left(-\xi_{it}^*\right)\cdot\xi_{it}\cdot\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i). \tag{A.4}$$

Taylor Expansion of $\boldsymbol{m}_t^*(\hat{\boldsymbol{a}}_t)$ around $\boldsymbol{a}_{0t}$ gives

$$\boldsymbol{m}_t^*(\hat{\boldsymbol{a}}_t) = \boldsymbol{m}_t^*(\boldsymbol{a}_{0t}) - \boldsymbol{M}_t^*(\tilde{\boldsymbol{a}}_t)\cdot(\hat{\boldsymbol{a}}_t - \boldsymbol{a}_{0t}) \tag{A.5}$$

where $\tilde{\boldsymbol{a}}_t$ is between $\boldsymbol{a}_{0t}$ and $\hat{\boldsymbol{a}}_t$ and

$$\boldsymbol{M}_t^*(\tilde{\boldsymbol{a}}_t) = -\frac{\partial\boldsymbol{m}_t^*(\boldsymbol{a})}{\partial\boldsymbol{a}'}|_{\boldsymbol{a}=\tilde{\boldsymbol{a}}_t} = \frac{1}{n}\sum_{i=1}^n \mathsf{f}\left((\tilde{\boldsymbol{a}}_t-\boldsymbol{a}_{0t})'\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)-\xi_{it}\right)\cdot\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)'. \tag{A.6}$$

By Assumption 3(ii) one can write

$$\boldsymbol{M}_t^*(\tilde{\boldsymbol{a}}_t) = \mathsf{f}(0)\cdot\hat{\boldsymbol{\Sigma}}_\phi + n^{-1}\boldsymbol{\Phi}(\boldsymbol{X})'\boldsymbol{D}_t^*\boldsymbol{\Phi}(\boldsymbol{X}), \tag{A.7}$$

where $\hat{\boldsymbol{\Sigma}}_\phi = n^{-1}\boldsymbol{\Phi}(\boldsymbol{X})'\boldsymbol{\Phi}(\boldsymbol{X})$ and $\boldsymbol{D}_t^*$ is a $n\times n$ diagonal matrix whose diagonal elements are bounded by in absolute values by $L\left|(\tilde{\boldsymbol{a}}_t-\boldsymbol{a}_{0t})'\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)-\xi_{it}\right|$. Note that by Lemma 1,

$$\begin{aligned}\max_{1\le t\le T}\|\boldsymbol{D}_t^*\|_S &\lesssim \max_{i,t}\left|(\tilde{\boldsymbol{a}}_t-\boldsymbol{a}_{0t})'\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)-\xi_{it}\right| \\ &\le \max_{1\le t\le T}\|\hat{\boldsymbol{a}}_t-\boldsymbol{a}_{0t}\|\cdot O_P(\sqrt{k_n}) + O_P(k_n^{-\alpha}) = O_P(\sqrt{k_n}\varepsilon_{nT}).\end{aligned} \tag{A.8}$$

Moreover, one can write

$$\boldsymbol{m}_t^*(\hat{\boldsymbol{a}}_t) = \boldsymbol{m}_t(\hat{\boldsymbol{a}}_t) - \tilde{\boldsymbol{m}}_t(\boldsymbol{a}_{0t}) + \left[\tilde{\boldsymbol{m}}_t(\boldsymbol{a}_{0t}) - \tilde{\boldsymbol{m}}_t(\hat{\boldsymbol{a}}_t)\right] \tag{A.9}$$

where $\tilde{\boldsymbol{m}}_t(\boldsymbol{a}) = \boldsymbol{m}_t(\boldsymbol{a}) - \boldsymbol{m}_t^*(\boldsymbol{a})$. It then follows from (A.5) (A.7) and (A.9) that

$$\begin{aligned}&\hat{\boldsymbol{a}}_t - \boldsymbol{a}_{0t} - \mathsf{f}^{-1}(0)\cdot\hat{\boldsymbol{\Sigma}}_\phi^{-1}\cdot\tilde{\boldsymbol{m}}_t(\boldsymbol{a}_{0t}) = \mathsf{f}^{-1}(0)\cdot\hat{\boldsymbol{\Sigma}}_\phi^{-1} \\ &\qquad\left\{\boldsymbol{m}_t^*(\boldsymbol{a}_{0t}) - \boldsymbol{m}_t(\hat{\boldsymbol{a}}_t) - \left[\tilde{\boldsymbol{m}}_t(\boldsymbol{a}_{0t}) - \tilde{\boldsymbol{m}}_t(\hat{\boldsymbol{a}}_t)\right] - n^{-1}\boldsymbol{\Phi}(\boldsymbol{X})'\boldsymbol{D}_t^*\boldsymbol{\Phi}(\boldsymbol{X})(\hat{\boldsymbol{a}}_t-\boldsymbol{a}_{0t})\right\},\end{aligned}$$

where

$$\tilde{\boldsymbol{m}}_t(\boldsymbol{a}_{0t}) = \frac{1}{n}\sum_{i=1}^n\left[\mathsf{F}\left(-\xi_{it}\right) - \mathbf{1}\{u_{it}\le-\xi_{it}\}\right]\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i) = \frac{1}{n}\sum_{i=1}^n\psi_{it}\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i).$$

Since $\mathsf{f}(0)$ is bounded below, and $\lambda_{\min}(\hat{\boldsymbol{\Sigma}}_\phi)$ is bounded below with probability approaching 1, it suffices to show that

$$\max_{1\le t\le T}\|\boldsymbol{m}_t^*(\boldsymbol{a}_{0t})\| = O_P(k_n^{-\alpha}), \tag{A.10}$$

$$\max_{1\le t\le T}\|\boldsymbol{m}_t(\hat{\boldsymbol{a}}_t)\| = O_P(k_n^{3/2}/n), \tag{A.11}$$

$$\frac{1}{T}\sum_{t=1}^{T}\|\tilde{\boldsymbol{m}}_t(\boldsymbol{a}_{0t}) - \tilde{\boldsymbol{m}}_t(\hat{\boldsymbol{a}}_t)\|^2 = O_P\left(\eta_{nT}^2\right), \tag{A.12}$$

$$\max_{1\le t\le T}\left\|n^{-1}\boldsymbol{\Phi}(\boldsymbol{X})'\boldsymbol{D}_t^*\boldsymbol{\Phi}(\boldsymbol{X})(\hat{\boldsymbol{a}}_t - \boldsymbol{a}_{0t})\right\| = O_P(\sqrt{k_n}\varepsilon_{nT}^2). \tag{A.13}$$

**Step 2:** By (A.4) and Assumption 1,

$$\begin{aligned}
&\max_{1\le t\le T}\|\boldsymbol{m}_t^*(\boldsymbol{a}_{0t})\| \\
= \quad &\max_{1\le t\le T}\left\|\frac{1}{n}\sum_{i=1}^{N}\mathsf{f}\left(-\xi_{it}^*\right)\cdot\xi_{it}\cdot\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)\right\| \\
\le \quad &\max_{1\le t\le T}\left\|\frac{1}{n}\sum_{i=1}^{N}\mathsf{f}\left(0\right)\cdot\xi_{it}\cdot\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)\right\| + O_P\left(k_n^{1/2-2\alpha}\right).
\end{aligned}$$

Define $z_{it} = \mathsf{f}(0)\cdot\xi_{it}$ and $\boldsymbol{z}_t = (z_{1t},\ldots,z_{Nt})'$, then

$$\frac{1}{n}\sum_{i=1}^{N}\mathsf{f}\left(0\right)\cdot\xi_{it}\cdot\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i) = N^{-1}\boldsymbol{\Phi}(\boldsymbol{X})'\boldsymbol{z}_t$$

and

$$\begin{aligned}
&\max_{1\le t\le T}\left\|\frac{1}{n}\sum_{i=1}^{N}\mathsf{f}\left(0\right)\cdot\xi_{it}\cdot\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)\right\| \\
= \quad &\max_{1\le t\le T}\left\|N^{-1}\boldsymbol{\Phi}(\boldsymbol{X})'\boldsymbol{z}_t\right\| \le \left\|N^{-1/2}\boldsymbol{\Phi}(\boldsymbol{X})\right\|_S\cdot\max_{1\le t\le T}\left\|N^{-1/2}\boldsymbol{z}_t\right\| = O_P(k_n^{-\alpha}).
\end{aligned}$$

In sum, we have

$$\max_{1\le t\le T}\|\boldsymbol{m}_t^*(\boldsymbol{a}_{0t})\| = O_P(k_n^{1/2-2\alpha}) + O_P(k_n^{-\alpha}) = O_P(k_n^{-\alpha}),$$

which gives (A.10).

**Step 3:** Similar to the proof of Lemma A4 of Horowitz and Lee (2005) it can be shown that

$$\max_{1\le t\le T}\|\boldsymbol{m}_t(\hat{\boldsymbol{a}}_t)\| = O_P(k_n^{3/2}/n),$$

which gives (A.11).

**Step 4:** By (A.8) and Lemma 1

$$\max_{1\le t\le T}\left\|n^{-1}\boldsymbol{\Phi}(\boldsymbol{X})'\boldsymbol{D}_t^*\boldsymbol{\Phi}(\boldsymbol{X})(\hat{\boldsymbol{a}}_t-\boldsymbol{a}_{0t})\right\| \le \|\boldsymbol{\Phi}(\boldsymbol{X})/\sqrt{n}\|_S^2\cdot\max_{1\le t\le T}\|\boldsymbol{D}_t^*\|_S\cdot\max_{1\le t\le T}\|\hat{\boldsymbol{a}}_t-\boldsymbol{a}_{0t}\| = O_P(\sqrt{k_n}\varepsilon_{nT}^2),$$

which gives (A.13).

**Step 5:** Define:

$$\delta_{1t}(\boldsymbol{\alpha}) = \frac{1}{n}\sum_{i=1}^n\left[\mathbf{1}\{u_{it}\le(\boldsymbol{a}-\boldsymbol{a}_{0t})'\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)-\xi_{it}\}-\mathbf{1}\{u_{it}\le-\xi_{it}\}\right]\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i),$$

$$\delta_{2t}(\boldsymbol{\alpha}) = \frac{1}{n}\sum_{i=1}^n\left[\mathsf{F}\left((\boldsymbol{a}-\boldsymbol{a}_{0t})'\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)-\xi_{it}\right)-\mathsf{F}\left(-\xi_{it}\right)\right]\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i),$$

$$\tilde{\delta}_{1t}(\boldsymbol{\alpha}) = \delta_{1t}(\boldsymbol{\alpha})-\mathbb{E}[\delta_{1t}(\boldsymbol{\alpha})],\qquad \tilde{\delta}_{2t}(\boldsymbol{\alpha}) = \delta_{2t}(\boldsymbol{\alpha})-\mathbb{E}[\delta_{2t}(\boldsymbol{\alpha})].$$

Note that $\mathbb{E}[\delta_{1t}(\boldsymbol{\alpha})]=\mathbb{E}[\delta_{2t}(\boldsymbol{\alpha})]$ because $\delta_{2t}(\boldsymbol{\alpha})=\mathbb{E}[\delta_{1t}(\boldsymbol{\alpha})|\boldsymbol{x}_i]$. Then $\tilde{\boldsymbol{m}}_t(\hat{\boldsymbol{a}}_t)-\tilde{\boldsymbol{m}}_t(\boldsymbol{a}_{0t}) = \tilde{\delta}_{2t}(\hat{\boldsymbol{a}}_t)-\tilde{\delta}_{1t}(\hat{\boldsymbol{a}}_t)$, and

$$\frac{1}{T}\sum_{t=1}^T\|\tilde{\boldsymbol{m}}_t(\hat{\boldsymbol{a}}_t)-\tilde{\boldsymbol{m}}_t(\boldsymbol{a}_{0t})\|^2 \le \frac{1}{T}\sum_{t=1}^T\left\|\tilde{\delta}_{1t}(\hat{\boldsymbol{a}}_t)\right\|^2+\frac{1}{T}\sum_{t=1}^T\left\|\tilde{\delta}_{2t}(\hat{\boldsymbol{a}}_t)\right\|^2. \tag{A.14}$$

In what follows, we will show that

$$\frac{1}{T}\sum_{t=1}^T\left\|\tilde{\delta}_{1t}(\hat{\boldsymbol{a}}_t)\right\|^2 = O_P\left(\ln(k_n^{-1/4}\varepsilon_{nT}^{-1/2})\cdot k_n^{5/2}\varepsilon_{nT}n^{-1}\right), \tag{A.15}$$

$$\frac{1}{T}\sum_{t=1}^T\left\|\tilde{\delta}_{2t}(\hat{\boldsymbol{a}}_t)\right\|^2 = O_P\left(\ln(k_n^{-1/2}\varepsilon_{nT}^{-1})\cdot k_n^{3}\varepsilon_{nT}^2n^{-1}\right), \tag{A.16}$$

which imply (A.12) and therefore complete the proof. We will focus on the proof of (A.15) since the proof of (A.16) is similar.

Let $\phi_{jd}(\boldsymbol{x}_i)$ be the $jd$th element of $\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)$ for $j=1,\dots,k_n; d=1,\dots,D$, and define

$$\Delta_{it}(\boldsymbol{\alpha},\boldsymbol{x}_i) = \mathbf{1}\{u_{it}\le(\boldsymbol{a}-\boldsymbol{a}_{0t})'\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)-\xi_{it}\}-\mathbf{1}\{u_{it}\le-\xi_{it}\}.$$

Then for some $C>0$, with probability approach 1,

$$\frac{1}{T}\sum_{t=1}^T\left\|\tilde{\delta}_{1t}(\hat{\boldsymbol{a}}_t)\right\|^2 \le \frac{1}{n}\cdot\frac{1}{T}\sum_{t=1}^T\sum_{j=1}^{k_n}\sum_{d=1}^D\sup_{\|\boldsymbol{a}-\boldsymbol{a}_{0t}\|\le C\varepsilon_{nT}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\{\Delta_{it}(\boldsymbol{\alpha},\boldsymbol{x}_i)\phi_{jd}(\boldsymbol{x}_i)-\mathbb{E}[\Delta_{it}(\boldsymbol{\alpha},\boldsymbol{x}_i)\phi_{jd}(\boldsymbol{x}_i)]\}\right|^2$$

We will show that

$$\mathbb{E}\left[\sup_{\|\boldsymbol{a}-\boldsymbol{a}_{0t}\|\leq C\varepsilon_{nT}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\Delta_{it}(\boldsymbol{\alpha},\boldsymbol{x}_i)\phi_{jd}(\boldsymbol{x}_i)-\mathbb{E}[\Delta_{it}(\boldsymbol{\alpha},\boldsymbol{x}_i)\phi_{jd}(\boldsymbol{x}_i)]\}\right|^2\right]$$

$$= O\left(\ln(k_n^{-1/4}\varepsilon_{nT}^{-1/2})\cdot k_n^{3/2}\varepsilon_{nT}\right) \quad \text{(A.17)}$$

uniformly in $t$ and $j$, from which (A.15) follows.

Define $\mathcal{H}_{\varepsilon_{nT}} = \{h(\boldsymbol{a},\boldsymbol{x}_i) \equiv \Delta_{it}(\boldsymbol{\alpha},\boldsymbol{x}_i)\phi_{jd}(\boldsymbol{x}_i) - \mathbb{E}[\Delta_{it}(\boldsymbol{\alpha},\boldsymbol{x}_i)\phi_{jd}(\boldsymbol{x}_i)] : \|\boldsymbol{a}-\boldsymbol{a}_{0t}\| \leq C\varepsilon_{nT}\}$, and for any $h \in \mathcal{H}_{\varepsilon_{nT}}$ define $\mathbb{G}_n h = n^{-1/2}\sum_{i=1}^{n} h(\boldsymbol{a},\boldsymbol{x}_i)$. Write

$$\sup_{\|\boldsymbol{a}-\boldsymbol{a}_{0t}\|\leq C\varepsilon_n}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\Delta_{it}(\boldsymbol{\alpha},\boldsymbol{x}_i)\phi_{jd}(\boldsymbol{x}_i)-\mathbb{E}[\Delta_{it}(\boldsymbol{\alpha},\boldsymbol{x}_i)\phi_{jd}(\boldsymbol{x}_i)]\}\right| = \|\mathbb{G}_n h\|_{\mathcal{H}_{\varepsilon_{nT}}},$$

then the left-hand side of (A.17) can be written as $\mathbb{E}\,\|\mathbb{G}_n h\|^2_{\mathcal{H}_{\varepsilon_{nT}}}$. Let $N(\mathcal{H}_{\varepsilon_{nT}}, L_2(Q), \epsilon)$ be the covering number of $\mathcal{H}_{\varepsilon_{nT}}$, where $L_2(Q)$ is the $L_2$ norm for functions and $Q$ is any probability measure on $\mathcal{X}$. Similar to the proof of (A.12) in Kato et al. (2012), it can be shown that $N(\mathcal{H}_{\varepsilon_{nT}}, L_2(Q), 2\epsilon) \leq (A/\epsilon)^{c_1 k_n}$ for some bounded constant $c_1$ and $A \geq 3\sqrt{e}$ that do not depend on $t$ and $j$. Moreover, it is easy to show that $\sup_{h\in\mathcal{H}_{\varepsilon_{nT}}}\mathbb{E}[h^2(\boldsymbol{a},\boldsymbol{x}_i)] \leq c_2^2\sqrt{k_n}\varepsilon_n$ for some bounded constant $c_2$. Then, applying Proposition B.1 of Kato et al. (2012), we have

$$\mathbb{E}\,\|\mathbb{G}_n h\|_{\mathcal{H}_{\varepsilon_{nT}}} \leq c_3\left[\cdot\ln(c_4 k_n^{-1/4}\varepsilon_{nT}^{-1/2})\cdot k_n/\sqrt{n} + \sqrt{\ln(c_4 k_n^{-1/4}\varepsilon_{nT}^{-1/2})}\cdot k_n^{3/4}\varepsilon_{nT}^{1/2}\right]$$

$$\leq c_5\sqrt{\ln(k_n^{-1/4}\varepsilon_{nT}^{-1/2})}\cdot k_n^{3/4}\varepsilon_{nT}^{1/2}, \quad \text{(A.18)}$$

where $c_3, c_4, c_5$ are bounded constants that do not depend on $t$ and $j$. Finally, (A.17) follows by noting that (see Chapter 6 of Ledoux and Talagrand 1991)

$$\mathbb{E}\,\|\mathbb{G}_n h\|^2_{\mathcal{H}_{\varepsilon_{nT}}} \leq \left(\mathbb{E}\,\|\mathbb{G}_n h\|_{\mathcal{H}_{\varepsilon_{nT}}}\right)^2 + O(n^{-1}).$$

This completes the proof. □

## Proof of Theorem 2:

*Proof.* Let $\boldsymbol{\Psi}$ be the $n \times T$ matrix of $\psi_{it}$, then the result of Lemma 2 can be written as

$$\left\|\hat{\boldsymbol{A}} - \boldsymbol{A}_0 - \mathsf{f}(0)^{-1}\cdot\hat{\boldsymbol{\Sigma}}_\phi^{-1}\boldsymbol{\Phi}'(\boldsymbol{X})\boldsymbol{\Psi}/n\right\|/\sqrt{T} = O_P\left(k_n^{-\alpha}\right) + O_P\left(\eta_{nT}\right). \quad \text{(A.19)}$$

From (A.2) and Lemma 1 we have

$$\|\hat{\boldsymbol{F}} - \boldsymbol{F}\hat{\boldsymbol{H}}\|/\sqrt{T} \leq O_P(1) \cdot \|\boldsymbol{F}\boldsymbol{G}(\boldsymbol{X})'\hat{\boldsymbol{V}}/(nT)\|_S + O_P(\varepsilon_{nT}^2). \tag{A.20}$$

Define $\boldsymbol{R}(\boldsymbol{X}) = \boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{B}_0 - \boldsymbol{G}(\boldsymbol{X})$, then by Assumption 1(ii) $\|\boldsymbol{R}(\boldsymbol{X})\|/\sqrt{n} = O_P(k_n^{-\alpha})$. Moreover, we can write

$$\begin{aligned}
\hat{\boldsymbol{V}} &= \hat{\boldsymbol{Y}} - \boldsymbol{G}(\boldsymbol{X})\boldsymbol{F}' \\
&= \boldsymbol{\Phi}(\boldsymbol{X})\hat{\boldsymbol{A}} - \boldsymbol{G}(\boldsymbol{X})\boldsymbol{F}' \\
&= \boldsymbol{\Phi}(\boldsymbol{X})\hat{\boldsymbol{A}} - \boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{A}_0 + \boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{A}_0 - \boldsymbol{G}(\boldsymbol{X})\boldsymbol{F}' \\
&= \boldsymbol{\Phi}(\boldsymbol{X})(\hat{\boldsymbol{A}} - \boldsymbol{A}_0) + \boldsymbol{R}(\boldsymbol{X})\boldsymbol{F}'.
\end{aligned}$$

Thus,

$$\begin{aligned}
& \boldsymbol{F}\boldsymbol{G}(\boldsymbol{X})'\hat{\boldsymbol{V}}/(nT) \\
= \ & \boldsymbol{F}(\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{B}_0 - \boldsymbol{R}(\boldsymbol{X}))'[\boldsymbol{\Phi}(\boldsymbol{X})(\hat{\boldsymbol{A}} - \boldsymbol{A}_0) + \boldsymbol{R}(\boldsymbol{X})\boldsymbol{F}']/(nT) \\
= \ & \boldsymbol{F}\boldsymbol{B}_0'\boldsymbol{\Phi}(\boldsymbol{X})'\boldsymbol{\Phi}(\boldsymbol{X})(\hat{\boldsymbol{A}} - \boldsymbol{A}_0)/(nT) - \boldsymbol{F}\boldsymbol{R}(\boldsymbol{X})'\boldsymbol{\Phi}(\boldsymbol{X})(\hat{\boldsymbol{A}} - \boldsymbol{A}_0)/(nT) \\
& + \boldsymbol{F}\boldsymbol{G}(\boldsymbol{X})'\boldsymbol{R}(\boldsymbol{X})\boldsymbol{F}'/(nT).
\end{aligned}$$

It then follows from Theorem 1 and Lemma 1 that

$$\|\boldsymbol{F}\boldsymbol{G}(\boldsymbol{X})'\hat{\boldsymbol{V}}/(nT)\|_S \leq \|\boldsymbol{F}\boldsymbol{B}_0'\boldsymbol{\Phi}(\boldsymbol{X})'\boldsymbol{\Phi}(\boldsymbol{X})(\hat{\boldsymbol{A}} - \boldsymbol{A}_0)/(nT)\|_S + O_P(k_n^{-\alpha}).$$

The above inequality and (A.20) imply that

$$\|\hat{\boldsymbol{F}} - \boldsymbol{F}\hat{\boldsymbol{H}}\|/\sqrt{T} \leq \|\boldsymbol{F}\boldsymbol{B}_0'\boldsymbol{\Phi}(\boldsymbol{X})'\boldsymbol{\Phi}(\boldsymbol{X})(\hat{\boldsymbol{A}} - \boldsymbol{A}_0)/(nT)\|_S + O_P(k_n^{-\alpha}) + O_P(\varepsilon_{nT}^2). \tag{A.21}$$

By (A.19) and Assumption 1(ii), we have

$$\begin{aligned}
& \|\boldsymbol{F}\boldsymbol{B}_0'\boldsymbol{\Phi}(\boldsymbol{X})'\boldsymbol{\Phi}(\boldsymbol{X})(\hat{\boldsymbol{A}} - \boldsymbol{A}_0)/(nT)\|_S \\
\leq \ & \mathsf{f}(0)^{-1}\|\boldsymbol{B}_0'\boldsymbol{\Phi}(\boldsymbol{X})'\boldsymbol{\Phi}(\boldsymbol{X})\hat{\boldsymbol{\Sigma}}_\phi^{-1}\boldsymbol{\Phi}'(\boldsymbol{X})\boldsymbol{\Psi}/(n^2T^{1/2})\|_S + O_P\left(k_n^{-\alpha} + \eta_{nT}\right) \\
= \ & \mathsf{f}(0)^{-1}\|\boldsymbol{B}_0'\boldsymbol{\Phi}'(\boldsymbol{X})\boldsymbol{\Psi}/(nT^{1/2})\|_S + O_P\left(k_n^{-\alpha} + \eta_{nT}\right) \\
\leq \ & \mathsf{f}(0)^{-1}\|\boldsymbol{G}'(\boldsymbol{X})\boldsymbol{\Psi}/(nT^{1/2})\| + \|\boldsymbol{G}(\boldsymbol{X}) - \boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{B}_0\|/\sqrt{n} \cdot \|\boldsymbol{\Psi}\|/\sqrt{nT} + O_P\left(k_n^{-\alpha} + \eta_{nT}\right) \\
= \ & \mathsf{f}(0)^{-1}\|\boldsymbol{G}'(\boldsymbol{X})\boldsymbol{\Psi}/(nT^{1/2})\| + O_P\left(k_n^{-\alpha} + \eta_{nT}\right).
\end{aligned}$$

Note that

$$\|\boldsymbol{G}'(\boldsymbol{X})\boldsymbol{\Psi}/(nT^{1/2})\| = \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{x}_i)\psi_{it}\right\|^2} = O_P(n^{-1/2})$$

because it is easy to see that $\mathbb{E}\left\|n^{-1/2}\sum_{i=1}^n \boldsymbol{g}(\boldsymbol{x}_i)\psi_{it}\right\|^2 < \infty$ for all $t$. It then follows from (A.21) that

$$\|\hat{\boldsymbol{F}} - \boldsymbol{F}\hat{\boldsymbol{H}}\|/\sqrt{T} = O_P(n^{-1/2}) + O_P(k_n^{-\alpha}) + O_P(\eta_{nT}) + O_P(\varepsilon_{nT}^2).$$

This completes the proof. $\square$

**Lemma 3.** *Under Assumptions 1, 2 and 4, we have*

$$\left\|\hat{\boldsymbol{A}} - \boldsymbol{A}_0 - \boldsymbol{\Sigma}_{\mathsf{f}\phi}^{-1}\boldsymbol{\Phi}'(\boldsymbol{X})\boldsymbol{\Psi}(\boldsymbol{X})/n\right\|/\sqrt{T} = O_P\left(k_n^{-\alpha}\right) + O_P\left(\eta_{nT}\right).$$

*where* $\psi_{it}(\boldsymbol{x}_i) = \mathsf{F}\left(-\xi_{it}|\boldsymbol{x}_i\right) - \mathbf{1}\{u_{it} \leq -\xi_{it}\}$ *and* $\boldsymbol{\Psi}(\boldsymbol{X})$ *is the* $n \times T$ *matrix of* $\psi_{it}(\boldsymbol{x}_i)$.

*Proof.* The proof is similar to the proof of Lemma 2. Therefore, it is omitted to save space. $\square$

## Proof of Theorem 3:

*Proof.* By the proof of Theorem 1, for any $\boldsymbol{x} \in \mathcal{X}$,

$$\hat{\boldsymbol{g}}(\boldsymbol{x}) = (\boldsymbol{F}'\hat{\boldsymbol{F}}/T)'\boldsymbol{g}(\boldsymbol{x}) + (\boldsymbol{F}'\hat{\boldsymbol{F}}/T)'(\boldsymbol{B}_0'\boldsymbol{\phi}_{k_n}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{x})) + (\hat{\boldsymbol{B}} - \boldsymbol{B}_0(\boldsymbol{F}'\hat{\boldsymbol{F}}/T))'\boldsymbol{\phi}_{k_n}(\boldsymbol{x}).$$

Moreover,

$$\hat{\boldsymbol{B}} - \boldsymbol{B}_0(\boldsymbol{F}'\hat{\boldsymbol{F}}/T) = (\hat{\boldsymbol{A}} - \boldsymbol{A}_0)\boldsymbol{F}\hat{\boldsymbol{H}}/T + (\hat{\boldsymbol{A}} - \boldsymbol{A}_0)(\hat{\boldsymbol{F}} - \boldsymbol{F}\hat{\boldsymbol{H}})/T.$$

Thus, by Lemma 1 and Theorem 1,

$$\hat{\boldsymbol{g}}(\boldsymbol{x}) - (\boldsymbol{F}'\hat{\boldsymbol{F}}/T)'\boldsymbol{g}(\boldsymbol{x}) = \hat{\boldsymbol{H}}'\boldsymbol{F}'(\hat{\boldsymbol{A}} - \boldsymbol{A}_0)'\boldsymbol{\phi}_{k_n}(\boldsymbol{x})/T + O_P(k_n^{-\alpha}) + O_P(\varepsilon_{nT}^2\sqrt{k_n}).$$

It then follows from Lemma 3 that

$$\hat{\boldsymbol{g}}(\boldsymbol{x}) - (\boldsymbol{F}'\hat{\boldsymbol{F}}/T)'\boldsymbol{g}(\boldsymbol{x}) = \hat{\boldsymbol{H}}'\boldsymbol{F}'\boldsymbol{\Psi}'(\boldsymbol{X})\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{\Sigma}_{\mathsf{f}\phi}^{-1}\boldsymbol{\phi}_{k_n}(\boldsymbol{x})/(nT) + O_P(k_n^{1/2-\alpha}) + O_P(\sqrt{k_n}\eta_{nT}).$$

Define $\boldsymbol{d}_T(\boldsymbol{x}_i) = T^{-1}\sum_{t=1}^T \boldsymbol{f}_t\psi_{it}(\boldsymbol{x}_i)$, $q(\boldsymbol{x}_i) = \boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)'\boldsymbol{\Sigma}_{\mathsf{f}\phi}^{-1}\boldsymbol{\phi}_{k_n}(\boldsymbol{x})$, then we can write

$$\boldsymbol{F}'\boldsymbol{\Psi}'(\boldsymbol{X})\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{\Sigma}_{\mathsf{f}\phi}^{-1}\boldsymbol{\phi}_{k_n}(\boldsymbol{x})/(nT) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{d}_T(\boldsymbol{x}_i)q(\boldsymbol{x}_i).$$

Note that $\mathbb{E}[\boldsymbol{d}_T(\boldsymbol{x}_i)q(\boldsymbol{x}_i)] = 0$ because $\mathbb{E}[\boldsymbol{d}_T(\boldsymbol{x}_i)|\boldsymbol{x}_i] = 0$, and it is easy to show that

$$\begin{aligned}\mathbb{E}[\boldsymbol{d}_T(\boldsymbol{x}_i)\boldsymbol{d}_T(\boldsymbol{x}_i)'q^2(\boldsymbol{x}_i)] &= \tau(1-\tau)(\boldsymbol{F}'\boldsymbol{F}/T^2)\boldsymbol{\phi}_{k_n}'(\boldsymbol{x})\boldsymbol{\Sigma}_{\mathsf{f}\phi}^{-1}\boldsymbol{\Sigma}_{\phi}\boldsymbol{\Sigma}_{\mathsf{f}\phi}^{-1}\boldsymbol{\phi}_{k_n}(\boldsymbol{x}) + o(1)\\ &= \tau(1-\tau)(\boldsymbol{F}'\boldsymbol{F}/T^2)\sigma_{k_n}^2 + o(1).\end{aligned}$$

Thus, we have

$$\boldsymbol{\Sigma}_{T,\tau}^{-1/2}(\hat{\boldsymbol{H}}')^{-1}\cdot\frac{\sqrt{nT}}{\sigma_{k_n}}\Big(\hat{\boldsymbol{g}}(\boldsymbol{x})-(\boldsymbol{F}'\hat{\boldsymbol{F}}/T)'\boldsymbol{g}(\boldsymbol{x})\Big)=\boldsymbol{\Sigma}_{T,\tau}^{-1/2}\cdot\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sqrt{T}\boldsymbol{d}_T(\boldsymbol{x}_i)q(\boldsymbol{x}_i)/\sigma_{k_n}$$
$$+O_P(k_n^{1/2-\alpha}+\sqrt{k_n}\eta_{nT})\sqrt{nT}\sigma_{k_n}^{-1}. \quad \text{(A.22)}$$

Finally, it follows from the Lyapunov's CLT and Assumption 4(iv) that

$$\boldsymbol{\Sigma}_{T,\tau}^{-1/2}(\hat{\boldsymbol{H}}')^{-1}\cdot\frac{\sqrt{nT}}{\sigma_{k_n}}\Big(\hat{\boldsymbol{g}}(\boldsymbol{x})-(\boldsymbol{F}'\hat{\boldsymbol{F}}/T)'\boldsymbol{g}(\boldsymbol{x})\Big)\xrightarrow{d}N(0,\boldsymbol{I}_R).$$

This completes the proof. □

## Proof of Theorem 4:

*Proof.* Define $\boldsymbol{R}(\boldsymbol{X})=\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{B}_0-\boldsymbol{G}(\boldsymbol{X})$, we can write

$$\hat{\boldsymbol{Y}}=\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{A}_0+\boldsymbol{\Phi}(\boldsymbol{X})(\hat{\boldsymbol{A}}-\boldsymbol{A}_0)=\boldsymbol{G}(\boldsymbol{X})\boldsymbol{F}'+\boldsymbol{R}(\boldsymbol{X})\boldsymbol{F}'+\boldsymbol{\Phi}(\boldsymbol{X})(\hat{\boldsymbol{A}}-\boldsymbol{A}_0).$$

Thus,

$$\tilde{\boldsymbol{F}}=\hat{\boldsymbol{Y}}'\hat{\boldsymbol{G}}(\boldsymbol{X})\cdot(\hat{\boldsymbol{G}}(\boldsymbol{X})'\hat{\boldsymbol{G}}(\boldsymbol{X}))^{-1}=\boldsymbol{F}(\boldsymbol{G}(\boldsymbol{X})'\hat{\boldsymbol{G}}(\boldsymbol{X})/n)(\hat{\boldsymbol{G}}(\boldsymbol{X})'\hat{\boldsymbol{G}}(\boldsymbol{X})/n)^{-1}$$
$$+\boldsymbol{F}(\boldsymbol{R}(\boldsymbol{X})'\hat{\boldsymbol{G}}(\boldsymbol{X})/n)(\hat{\boldsymbol{G}}(\boldsymbol{X})'\hat{\boldsymbol{G}}(\boldsymbol{X})/n)^{-1}+(\hat{\boldsymbol{A}}-\boldsymbol{A}_0)'(\boldsymbol{\Phi}(\boldsymbol{X})'\hat{\boldsymbol{G}}(\boldsymbol{X})/n)(\hat{\boldsymbol{G}}(\boldsymbol{X})'\hat{\boldsymbol{G}}(\boldsymbol{X})/n)^{-1},$$

and

$$\tilde{\boldsymbol{f}}_t-\tilde{\boldsymbol{H}}'\boldsymbol{f}_t=(\hat{\boldsymbol{G}}(\boldsymbol{X})'\hat{\boldsymbol{G}}(\boldsymbol{X})/n)^{-1}(\hat{\boldsymbol{G}}(\boldsymbol{X})'\boldsymbol{R}(\boldsymbol{X})/n)\boldsymbol{f}_t$$
$$+(\hat{\boldsymbol{G}}(\boldsymbol{X})'\hat{\boldsymbol{G}}(\boldsymbol{X})/n)^{-1}(\hat{\boldsymbol{G}}(\boldsymbol{X})'\boldsymbol{\Phi}(\boldsymbol{X})/n)(\hat{\boldsymbol{a}}_t-\boldsymbol{a}_{0t}).$$

It is easy to see from Theorem 1 and Assumption 1(ii) that the first term on the right-hand side of the above equation is $O_P(k_n^{-\alpha})$. Moreover, by Lemma 3, the second term can be written as

$$(\hat{\boldsymbol{G}}(\boldsymbol{X})'\hat{\boldsymbol{G}}(\boldsymbol{X})/n)^{-1}\cdot(\hat{\boldsymbol{G}}(\boldsymbol{X})'\boldsymbol{\Phi}(\boldsymbol{X})/n)\cdot\boldsymbol{\Sigma}_{\mathsf{f}\phi}^{-1}\cdot\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)\psi_{it}(\boldsymbol{x}_i)+O_P(k_n^{-\alpha})+O_P(\eta_{nT}).$$

By Theorem 1 we can show that

$$\|(\hat{\boldsymbol{G}}(\boldsymbol{X})'\hat{\boldsymbol{G}}(\boldsymbol{X})/n)^{-1}-\hat{\boldsymbol{H}}'\boldsymbol{\Sigma}_g^{-1}\hat{\boldsymbol{H}}\|=O_P(\varepsilon_{nT}),$$

$$\|(\hat{\boldsymbol{G}}(\boldsymbol{X})'\boldsymbol{\Phi}(\boldsymbol{X})/n)-\hat{\boldsymbol{H}}^{-1}\mathbb{E}[\boldsymbol{g}(\boldsymbol{x}_i)\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)']\|_S=O_P(\varepsilon_{nT}),$$

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)\psi_{it}(\boldsymbol{x}_i) \right\| = O_P(\sqrt{k_n/n}),$$

it then follows from Assumption 4(iii) that

$$(\hat{\boldsymbol{H}}')^{-1}\sqrt{n}(\tilde{\boldsymbol{f}}_t - \tilde{\boldsymbol{H}}'\boldsymbol{f}_t) = \boldsymbol{\Sigma}_g^{-1}\mathbb{E}[\boldsymbol{g}(\boldsymbol{x}_i)\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)']\boldsymbol{\Sigma}_{\mathsf{f}\phi}^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)\psi_{it}(\boldsymbol{x}_i)\right) + O_P(\varepsilon_{nT}k_n^{1/2}) + O_P(n^{1/2}k_n^{-\alpha}) + O_P(n^{1/2}\eta_{nT}).$$

By the Lyapunov's CLT we can show that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\phi}_{k_n}(\boldsymbol{x}_i)\psi_{it}(\boldsymbol{x}_i) \xrightarrow{d} N(0, \tau(1-\tau)\boldsymbol{\Sigma}_{\phi}),$$

then the desired result follows from Assumption 5. □

## Proof of Theorem 5:

*Proof.* First, note that

$$\begin{aligned}
& \|\boldsymbol{\Phi}(\boldsymbol{X})\hat{\boldsymbol{A}}\hat{\boldsymbol{A}}'\boldsymbol{\Phi}(\boldsymbol{X})' - \boldsymbol{G}(\boldsymbol{X})\boldsymbol{F}'\boldsymbol{F}\boldsymbol{G}(\boldsymbol{X})'\|/(nT) \\
\leq\ & 2\|\boldsymbol{G}(\boldsymbol{X})\boldsymbol{F}'\|/\sqrt{nT} \cdot \|\boldsymbol{\Phi}(\boldsymbol{X})\hat{\boldsymbol{A}} - \boldsymbol{G}(\boldsymbol{X})\boldsymbol{F}'\|/\sqrt{nT} + \|\boldsymbol{\Phi}(\boldsymbol{X})\hat{\boldsymbol{A}} - \boldsymbol{G}(\boldsymbol{X})\boldsymbol{F}'\|^2/(nT) \\
=\ & O_P(1) \cdot \|\hat{\boldsymbol{V}}\|/\sqrt{nT} + \|\hat{\boldsymbol{V}}\|^2/(nT).
\end{aligned}$$

It then follows from Lemma 1(ii) that

$$\|\boldsymbol{\Phi}(\boldsymbol{X})\hat{\boldsymbol{A}}\hat{\boldsymbol{A}}'\boldsymbol{\Phi}(\boldsymbol{X})' - \boldsymbol{G}(\boldsymbol{X})\boldsymbol{F}'\boldsymbol{F}\boldsymbol{G}(\boldsymbol{X})'\|/(nT) = O_P(\varepsilon_{nT}). \tag{A.23}$$

Second, Assumption 2(iii) and (iv) imply that the largest $R$ eigenvalues of $\boldsymbol{G}(\boldsymbol{X})\boldsymbol{F}'\boldsymbol{F}\boldsymbol{G}(\boldsymbol{X})'/(nT)$, which are also the $R$ eigenvalues of $(\boldsymbol{F}'\boldsymbol{F}/T)\cdot\boldsymbol{G}(\boldsymbol{X})'\boldsymbol{G}(\boldsymbol{X})/n$, converge in probability to the $R$ eigenvalues of $(\boldsymbol{F}'\boldsymbol{F}/T)\cdot\boldsymbol{\Sigma}_g$. Also, note that the remaining eigenvalues of $\boldsymbol{G}(\boldsymbol{X})\boldsymbol{F}'\boldsymbol{F}\boldsymbol{G}(\boldsymbol{X})'/(nT)$ are all 0, it then follows from (A.23) and the Wielandt-Hoffman inequality that $\hat{\rho}_j = O_P(\varepsilon_{nT})$ for $j = R+1, \ldots, \bar{R}$, and $\hat{\rho}_j$ converges in probability in some positive constant for $j = 1, \ldots, R$. The desired result then follows because $P[\hat{\rho}_j > p_n] \to 1$ for $j = 1, \ldots, R$ and $P[\hat{\rho}_j > p_n] \to 0$ for $j = R+1, \ldots, \bar{R}$. □

## B Figures and Tables

Table 1: Estimating the number of factors: rank minimization estimator

| | $T$ | $n$ | $N(0,1)$ | | | $t(3)$ | | | Cauchy(0,1) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\tau = 0.25$ | 5 | 50 | [0.13 | 0.65 | 0.23] | [0.03 | 0.41 | 0.56] | [0.01 | 0.10 | 0.89] |
| | 5 | 100 | [0.10 | 0.72 | 0.19] | [0.02 | 0.44 | 0.54] | [0.00 | 0.03 | 0.97] |
| | 5 | 200 | [0.23 | 0.77 | 0.00] | [0.12 | 0.82 | 0.06] | [0.00 | 0.17 | 0.83] |
| | 5 | 1000 | [0.17 | 0.83 | 0.00] | [0.16 | 0.84 | 0.00] | [0.06 | 0.81 | 0.13] |
| | 10 | 50 | [0.17 | 0.76 | 0.07] | [0.03 | 0.50 | 0.47] | [0.02 | 0.06 | 0.92] |
| | 10 | 100 | [0.08 | 0.89 | 0.03] | [0.03 | 0.65 | 0.46] | [0.00 | 0.03 | 0.97] |
| | 10 | 200 | [0.07 | 0.93 | 0.00] | [0.05 | 0.95 | 0.00] | [0.00 | 0.24 | 0.76] |
| | 10 | 1000 | [0.03 | 0.97 | 0.00] | [0.02 | 0.98 | 0.00] | [0.01 | 0.98 | 0.01] |
| $\tau = 0.5$ | 5 | 50 | [0.19 | 0.71 | 0.10] | [0.09 | 0.56 | 0.35] | [0.00 | 0.15 | 0.85] |
| | 5 | 100 | [0.17 | 0.76 | 0.08] | [0.07 | 0.59 | 0.34] | [0.00 | 0.20 | 0.80] |
| | 5 | 200 | [0.23 | 0.77 | 0.00] | [0.19 | 0.80 | 0.01] | [0.06 | 0.75 | 0.19] |
| | 5 | 1000 | [0.18 | 0.82 | 0.00] | [0.15 | 0.85 | 0.00] | [0.13 | 0.87 | 0.00] |
| | 10 | 50 | [0.20 | 0.78 | 0.03] | [0.08 | 0.76 | 0.15] | [0.00 | 0.13 | 0.87] |
| | 10 | 100 | [0.12 | 0.87 | 0.01] | [0.05 | 0.87 | 0.08] | [0.00 | 0.24 | 0.76] |
| | 10 | 200 | [0.05 | 0.95 | 0.00] | [0.05 | 0.95 | 0.00] | [0.03 | 0.94 | 0.03] |
| | 10 | 1000 | [0.01 | 0.99 | 0.00] | [0.02 | 0.98 | 0.00] | [0.02 | 0.99 | 0.00] |
| $\tau = 0.75$ | 5 | 50 | [0.11 | 0.68 | 0.21] | [0.04 | 0.41 | 0.56] | [0.01 | 0.09 | 0.90] |
| | 5 | 100 | [0.10 | 0.71 | 0.19] | [0.02 | 0.42 | 0.56] | [0.00 | 0.04 | 0.96] |
| | 5 | 200 | [0.22 | 0.78 | 0.00] | [0.14 | 0.81 | 0.05] | [0.00 | 0.15 | 0.85] |
| | 5 | 1000 | [0.18 | 0.82 | 0.00] | [0.17 | 0.83 | 0.00] | [0.04 | 0.82 | 0.15] |
| | 10 | 50 | [0.15 | 0.78 | 0.08] | [0.04 | 0.50 | 0.46] | [0.01 | 0.05 | 0.94] |
| | 10 | 100 | [0.11 | 0.86 | 0.04] | [0.03 | 0.65 | 0.32] | [0.00 | 0.03 | 0.97] |
| | 10 | 200 | [0.06 | 0.94 | 0.00] | [0.05 | 0.94 | 0.01] | [0.01 | 0.27 | 0.73] |
| | 10 | 1000 | [0.02 | 0.98 | 0.00] | [0.02 | 0.98 | 0.00] | [0.02 | 0.97 | 0.01] |

Note: the DGP is $y_{it} = \sum_{r=1}^{3} \lambda_{ir} f_{tr} + \left(x_{i1}^2 + x_{i2}^2 + x_{i3}^2\right) u_{it}$, where $f_{t1} = 1$, $f_{t2}, f_{t3} \sim i.i.d\ N(0,1)$. The number of characteristics is 5 and all characteristics $x_{id}$ are drawn independently from the uniform distribution: $U[-1,1]$. $g_1(x) = sin(2\pi x)$, $g_2(x) = sin(\pi x)$ and $g_3(x) = cos(\pi x)$, and $\lambda_{i1} = \sum_{d=1,3,5} g_1(x_{id})$, $\lambda_{i2} = \sum_{d=1,2} g_2(x_{id})$, $\lambda_{i3} = \sum_{d=3,4} g_3(x_{id})$. $u_{it}$ are i.i.d variables drawn from three different distributions. In the first step quantile sieve estimation, $k_n = n^{1/3}$ and we use the *Chebyshev polynomials of the second kind* as the basis functions. For the estimator of the number of factors, the threshold $p_n$ is chosen as in (9) with $d = 1/4$. The reported results are [frequency of $\hat{R} < R$; frequency of $\hat{R} = R$; frequency of $\hat{R} > R$] from 1000 replications.

Table 2: Estimating the number of factors: eigen-ratio estimator

| | $T$ | $n$ | $N(0,1)$ | | | $t(3)$ | | | Cauchy(0,1) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\tau = 0.25$ | 5 | 50 | [0.57 | 0.25 | 0.19] | [0.54 | 0.22 | 0.25] | [0.54 | 0.17 | 0.29] |
| | 5 | 100 | [0.58 | 0.33 | 0.09] | [0.58 | 0.27 | 0.15] | [0.59 | 0.15 | 0.26] |
| | 5 | 200 | [0.44 | 0.54 | 0.01] | [0.54 | 0.43 | 0.04] | [0.62 | 0.24 | 0.14] |
| | 5 | 1000 | [0.23 | 0.77 | 0.00] | [0.31 | 0.69 | 0.00] | [0.56 | 0.42 | 0.02] |
| | 10 | 50 | [0.46 | 0.37 | 0.17] | [0.45 | 0.18 | 0.37] | [0.47 | 0.07 | 0.46] |
| | 10 | 100 | [0.37 | 0.59 | 0.04] | [0.46 | 0.42 | 0.11] | [0.60 | 0.09 | 0.31] |
| | 10 | 200 | [0.09 | 0.91 | 0.00] | [0.19 | 0.80 | 0.01] | [0.59 | 0.31 | 0.11] |
| | 10 | 1000 | [0.01 | 0.99 | 0.00] | [0.03 | 0.97 | 0.00] | [0.17 | 0.83 | 0.00] |
| $\tau = 0.5$ | 5 | 50 | [0.58 | 0.28 | 0.14] | [0.57 | 0.22 | 0.20] | [0.50 | 0.20 | 0.30] |
| | 5 | 100 | [0.58 | 0.33 | 0.09] | [0.57 | 0.28 | 0.15] | [0.56 | 0.21 | 0.22] |
| | 5 | 200 | [0.42 | 0.57 | 0.01] | [0.46 | 0.51 | 0.03] | [0.54 | 0.41 | 0.06] |
| | 5 | 1000 | [0.21 | 0.79 | 0.00] | [0.23 | 0.77 | 0.00] | [0.28 | 0.72 | 0.00] |
| | 10 | 50 | [0.41 | 0.46 | 0.13] | [0.46 | 0.33 | 0.21] | [0.42 | 0.10 | 0.48] |
| | 10 | 100 | [0.30 | 0.66 | 0.04] | [0.36 | 0.57 | 0.07] | [0.51 | 0.24 | 0.26] |
| | 10 | 200 | [0.06 | 0.94 | 0.00] | [0.11 | 0.89 | 0.00] | [0.22 | 0.76 | 0.02] |
| | 10 | 1000 | [0.01 | 0.99 | 0.00] | [0.02 | 0.98 | 0.00] | [0.03 | 0.97 | 0.00] |
| $\tau = 0.75$ | 5 | 50 | [0.58 | 0.25 | 0.17] | [0.54 | 0.22 | 0.24] | [0.55 | 0.17 | 0.28] |
| | 5 | 100 | [0.57 | 0.32 | 0.10] | [0.59 | 0.24 | 0.17] | [0.56 | 0.20 | 0.24] |
| | 5 | 200 | [0.43 | 0.55 | 0.02] | [0.52 | 0.43 | 0.04] | [0.65 | 0.21 | 0.14] |
| | 5 | 1000 | [0.24 | 0.76 | 0.00] | [0.33 | 0.67 | 0.00] | [0.55 | 0.44 | 0.01] |
| | 10 | 50 | [0.46 | 0.36 | 0.18] | [0.44 | 0.20 | 0.37] | [0.47 | 0.05 | 0.48] |
| | 10 | 100 | [0.36 | 0.59 | 0.06] | [0.46 | 0.40 | 0.14] | [0.63 | 0.09 | 0.28] |
| | 10 | 200 | [0.11 | 0.89 | 0.00] | [0.19 | 0.80 | 0.01] | [0.58 | 0.31 | 0.11] |
| | 10 | 1000 | [0.01 | 0.99 | 0.00] | [0.03 | 0.97 | 0.00] | [0.16 | 0.83 | 0.01] |

Note: the DGP is $y_{it} = \sum_{r=1}^{3} \lambda_{ir} f_{tr} + \left(x_{i1}^2 + x_{i2}^2 + x_{i3}^2\right) u_{it}$, where $f_{t1} = 1$, $f_{t2}, f_{t3} \sim i.i.d\ N(0,1)$. The number of characteristics is 5 and all characteristics $x_{id}$ are drawn independently from the uniform distribution: $U[-1,1]$. $g_1(x) = sin(2\pi x)$, $g_2(x) = sin(\pi x)$ and $g_3(x) = cos(\pi x)$, and $\lambda_{i1} = \sum_{d=1,3,5} g_1(x_{id})$, $\lambda_{i2} = \sum_{d=1,2} g_2(x_{id})$, $\lambda_{i3} = \sum_{d=3,4} g_3(x_{id})$. $u_{it}$ are i.i.d variables drawn from three different distributions. In the first step quantile sieve estimation, $k_n = n^{1/3}$ and we use the *Chebyshev polynomials of the second kind* as the basis functions. The estimator for the number of factors is the integer that maximizes the eigen-ratios. The reported results are [frequency of $\hat{R} < R$; frequency of $\hat{R} = R$; frequency of $\hat{R} > R$] from 1000 replications.

Table 3: Factor estimation using QPPCA

| | $T$ | $n$ | $N(0,1)$ $f_{1t}$ | $f_{2t}$ | $f_{3t}$ | $t(3)$ $f_{1t}$ | $f_{2t}$ | $f_{3t}$ | Cauchy(0,1) $f_{1t}$ | $f_{2t}$ | $f_{3t}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\tau = 0.25$ | 10 | 50 | 0.859 | 0.879 | 0.574 | 0.738 | 0.745 | 0.630 | 0.386 | 0.370 | 0.609 |
| | 10 | 100 | 0.971 | 0.956 | 0.857 | 0.938 | 0.890 | 0.835 | 0.670 | 0.566 | 0.767 |
| | 10 | 200 | 0.989 | 0.983 | 0.924 | 0.978 | 0.959 | 0.911 | 0.862 | 0.767 | 0.867 |
| | 10 | 500 | 0.997 | 0.995 | 0.979 | 0.994 | 0.990 | 0.972 | 0.968 | 0.940 | 0.950 |
| | 50 | 50 | 0.893 | 0.909 | 0.417 | 0.751 | 0.796 | 0.499 | 0.086 | 0.069 | 0.375 |
| | 50 | 100 | 0.976 | 0.968 | 0.824 | 0.957 | 0.940 | 0.797 | 0.623 | 0.407 | 0.654 |
| | 50 | 200 | 0.990 | 0.986 | 0.901 | 0.982 | 0.977 | 0.892 | 0.919 | 0.838 | 0.821 |
| | 50 | 500 | 0.997 | 0.995 | 0.973 | 0.995 | 0.992 | 0.967 | 0.984 | 0.975 | 0.941 |
| $\tau = 0.75$ | 10 | 50 | 0.861 | 0.876 | 0.581 | 0.749 | 0.749 | 0.623 | 0.383 | 0.362 | 0.605 |
| | 10 | 100 | 0.971 | 0.955 | 0.858 | 0.933 | 0.894 | 0.834 | 0.682 | 0.573 | 0.768 |
| | 10 | 200 | 0.989 | 0.983 | 0.921 | 0.979 | 0.960 | 0.905 | 0.867 | 0.777 | 0.867 |
| | 10 | 500 | 0.997 | 0.995 | 0.979 | 0.994 | 0.990 | 0.974 | 0.973 | 0.937 | 0.950 |
| | 50 | 50 | 0.893 | 0.911 | 0.420 | 0.749 | 0.794 | 0.493 | 0.081 | 0.066 | 0.380 |
| | 50 | 100 | 0.977 | 0.967 | 0.824 | 0.958 | 0.938 | 0.794 | 0.617 | 0.400 | 0.656 |
| | 50 | 200 | 0.990 | 0.986 | 0.901 | 0.982 | 0.976 | 0.894 | 0.915 | 0.832 | 0.818 |
| | 50 | 500 | 0.997 | 0.995 | 0.972 | 0.995 | 0.992 | 0.967 | 0.984 | 0.974 | 0.938 |

Note: the DGP is $Y_{it} = \lambda_{i1} f_{t1} + \lambda_{i2} f_{t2} + (\lambda_{i3} f_{t3}) u_{it}$, where $f_{t3} = |h_t|, f_{t1}, f_{t2}, h_t \sim i.i.d\ N(0,1)$. The number of characteristics is 5 and all characteristics $x_{id}$ are independently drawn from the uniform distribution: $U[-1,1]$. $g_1(x) = sin(2\pi x),\ g_2(x) = sin(\pi x)\ and\ g_3(x) = |cos(\pi x)|$. The factor loading functions are generated as $\lambda_{i1} = \sum_{d=1,3,5} g_1(x_{id}),\ \lambda_{i2} = \sum_{d=1,2} g_2(x_{id})$ and $\lambda_{i3} = \sum_{d=3,4} g_3(x_{id})$. $\{u_{it}\}$ are i.i.d draws from three different distributions. 3 factors are estimated at each $\tau$ using the proposed method in this paper, and the reported results are the averages of the adjusted $R^2$ of regressing the true factors on the estimated factors from 1000 replications.

Table 4: Factor estimation using QFA

| | $T$ | $n$ | $N(0,1)$ $f_{1t}$ | $f_{2t}$ | $f_{3t}$ | $t(3)$ $f_{1t}$ | $f_{2t}$ | $f_{3t}$ | Cauchy(0,1) $f_{1t}$ | $f_{2t}$ | $f_{3t}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\tau = 0.25$ | 10 | 50 | 0.887 | 0.821 | 0.561 | 0.808 | 0.706 | 0.528 | 0.516 | 0.418 | 0.449 |
| | 10 | 100 | 0.898 | 0.833 | 0.586 | 0.822 | 0.727 | 0.574 | 0.525 | 0.427 | 0.501 |
| | 10 | 200 | 0.904 | 0.841 | 0.624 | 0.834 | 0.735 | 0.584 | 0.525 | 0.443 | 0.504 |
| | 10 | 500 | 0.908 | 0.840 | 0.643 | 0.841 | 0.740 | 0.608 | 0.513 | 0.420 | 0.512 |
| | 50 | 50 | 0.964 | 0.948 | 0.786 | 0.935 | 0.902 | 0.725 | 0.724 | 0.537 | 0.473 |
| | 50 | 100 | 0.983 | 0.976 | 0.884 | 0.972 | 0.956 | 0.848 | 0.871 | 0.767 | 0.669 |
| | 50 | 200 | 0.992 | 0.988 | 0.936 | 0.986 | 0.977 | 0.911 | 0.935 | 0.853 | 0.802 |
| | 50 | 500 | 0.996 | 0.994 | 0.965 | 0.994 | 0.989 | 0.951 | 0.963 | 0.906 | 0.880 |
| $\tau = 0.75$ | 10 | 50 | 0.875 | 0.835 | 0.551 | 0.808 | 0.719 | 0.523 | 0.510 | 0.414 | 0.447 |
| | 10 | 100 | 0.898 | 0.938 | 0.595 | 0.820 | 0.730 | 0.583 | 0.523 | 0.420 | 0.506 |
| | 10 | 200 | 0.904 | 0.846 | 0.616 | 0.828 | 0.736 | 0.600 | 0.520 | 0.429 | 0.497 |
| | 10 | 500 | 0.899 | 0.838 | 0.625 | 0.843 | 0.742 | 0.616 | 0.528 | 0.433 | 0.489 |
| | 50 | 50 | 0.964 | 0.947 | 0.785 | 0.935 | 0.901 | 0.722 | 0.722 | 0.551 | 0.486 |
| | 50 | 100 | 0.983 | 0.975 | 0.884 | 0.972 | 0.956 | 0.846 | 0.874 | 0.760 | 0.672 |
| | 50 | 200 | 0.992 | 0.988 | 0.935 | 0.986 | 0.978 | 0.911 | 0.931 | 0.852 | 0.799 |
| | 50 | 500 | 0.996 | 0.994 | 0.964 | 0.994 | 0.989 | 0.949 | 0.964 | 0.903 | 0.878 |

Note: the DGP is $Y_{it} = \lambda_{i1} f_{t1} + \lambda_{i2} f_{t2} + (\lambda_{i3} f_{t3}) u_{it}$, where $f_{t3} = |h_t|, f_{t1}, f_{t2}, h_t \sim i.i.d\ N(0,1)$. The number of characteristics is 5 and all characteristics $x_{id}$ are independently drawn from the uniform distribution: $U[-1,1]$. $g_1(x) = sin(2\pi x)$, $g_2(x) = sin(\pi x)$ *and* $g_3(x) = |cos(\pi x)|$. The factor loading functions are generated as $\lambda_{i1} = \sum_{d=1,3,5} g_1(x_{id})$, $\lambda_{i2} = \sum_{d=1,2} g_2(x_{id})$ and $\lambda_{i3} = \sum_{d=3,4} g_3(x_{id})$. $\{u_{it}\}$ are i.i.d draws from three different distributions. 3 factors are estimated at each $\tau$ using the QFA proposed by Chen et al. (2021), and the reported results are the averages of the adjusted $R^2$ of regressing the true factors on the estimated factors from 1000 replications.

Table 5: Factor estimation using PCA and PPCA

| | | | $N(0,1)$ | | | $t(3)$ | | | Cauchy(0,1) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $T$ | $n$ | $f_{1t}$ | $f_{2t}$ | $f_{3t}$ | $f_{1t}$ | $f_{2t}$ | $f_{3t}$ | $f_{1t}$ | $f_{2t}$ | $f_{3t}$ |
| PCA | 10 | 50 | 0.955 | 0.921 | 0.420 | 0.847 | 0.723 | 0.455 | 0.271 | 0.250 | 0.392 |
| | 10 | 100 | 0.964 | 0.929 | 0.450 | 0.858 | 0.757 | 0.514 | 0.286 | 0.289 | 0.410 |
| | 10 | 200 | 0.970 | 0.944 | 0.478 | 0.871 | 0.751 | 0.530 | 0.289 | 0.285 | 0.422 |
| | 10 | 500 | 0.975 | 0.944 | 0.493 | 0.879 | 0.767 | 0.568 | 0.292 | 0.303 | 0.433 |
| | 50 | 50 | 0.973 | 0.957 | 0.084 | 0.894 | 0.781 | 0.079 | 0.003 | -0.001 | 0.032 |
| | 50 | 100 | 0.986 | 0.977 | 0.131 | 0.937 | 0.862 | 0.116 | 0.032 | 0.031 | 0.066 |
| | 50 | 200 | 0.993 | 0.988 | 0.149 | 0.961 | 0.901 | 0.141 | 0.044 | 0.048 | 0.075 |
| | 50 | 500 | 0.997 | 0.994 | 0.166 | 0.977 | 0.933 | 0.161 | 0.055 | 0.054 | 0.091 |
| PPCA | 10 | 50 | 0.949 | 0.962 | 0.382 | 0.843 | 0.866 | 0.379 | 0.277 | 0.282 | 0.387 |
| | 10 | 100 | 0.989 | 0.984 | 0.374 | 0.960 | 0.930 | 0.379 | 0.321 | 0.314 | 0.406 |
| | 10 | 200 | 0.995 | 0.993 | 0.382 | 0.983 | 0.969 | 0.383 | 0.318 | 0.309 | 0.409 |
| | 10 | 500 | 0.998 | 0.997 | 0.400 | 0.994 | 0.989 | 0.402 | 0.321 | 0.317 | 0.417 |
| | 50 | 50 | 0.953 | 0.963 | 0.060 | 0.858 | 0.882 | 0.054 | 0.003 | 0.001 | 0.029 |
| | 50 | 100 | 0.987 | 0.982 | 0.095 | 0.962 | 0.947 | 0.085 | 0.036 | 0.031 | 0.062 |
| | 50 | 200 | 0.994 | 0.992 | 0.110 | 0.982 | 0.974 | 0.100 | 0.048 | 0.049 | 0.072 |
| | 50 | 500 | 0.998 | 0.997 | 0.130 | 0.994 | 0.990 | 0.114 | 0.058 | 0.056 | 0.090 |

Note: the DGP is $Y_{it} = \lambda_{i1} f_{t1} + \lambda_{i2} f_{t2} + (\lambda_{i3} f_{t3}) u_{it}$, where $f_{t3} = |h_t|, f_{t1}, f_{t2}, h_t \sim i.i.d\ N(0,1)$. The number of characteristics is 5 and all characteristics $x_{id}$ are independently drawn from the uniform distribution: $U[-1,1]$. $g_1(x) = sin(2\pi x),\ g_2(x) = sin(\pi x)\ and\ g_3(x) = |cos(\pi x)|$. The factor loading functions are generated as $\lambda_{i1} = \sum_{d=1,3,5} g_1(x_{id}),\ \lambda_{i2} = \sum_{d=1,2} g_2(x_{id})$ and $\lambda_{i3} = \sum_{d=3,4} g_3(x_{id})$. $\{u_{it}\}$ are i.i.d draws from three different distributions. 3 factors are estimated using PCA and PPCA respectively, and the reported results are the averages of the adjusted $R^2$ of regressing the true factors on the estimated factors from 1000 replications.

Table 6: Factor estimation using QPPCA: $R = 3, D = 2$

| | $T$ | $n$ | $N(0,1)$ $f_{1t}$ | $f_{2t}$ | $f_{3t}$ | $t(3)$ $f_{1t}$ | $f_{2t}$ | $f_{3t}$ | Cauchy(0,1) $f_{1t}$ | $f_{2t}$ | $f_{3t}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\tau = 0.25$ | 10 | 50 | 0.731 | 0.880 | 0.606 | 0.588 | 0.806 | 0.617 | 0.348 | 0.481 | 0.602 |
| | 10 | 100 | 0.932 | 0.948 | 0.830 | 0.877 | 0.902 | 0.799 | 0.655 | 0.668 | 0.731 |
| | 10 | 200 | 0.969 | 0.978 | 0.916 | 0.947 | 0.960 | 0.901 | 0.778 | 0.837 | 0.847 |
| | 10 | 500 | 0.990 | 0.993 | 0.974 | 0.983 | 0.989 | 0.968 | 0.929 | 0.941 | 0.942 |
| | 50 | 50 | 0.665 | 0.875 | 0.485 | 0.488 | 0.782 | 0.473 | 0.100 | 0.202 | 0.390 |
| | 50 | 100 | 0.934 | 0.957 | 0.811 | 0.892 | 0.921 | 0.766 | 0.473 | 0.531 | 0.602 |
| | 50 | 200 | 0.969 | 0.982 | 0.906 | 0.949 | 0.970 | 0.889 | 0.785 | 0.839 | 0.796 |
| | 50 | 500 | 0.990 | 0.993 | 0.968 | 0.984 | 0.989 | 0.961 | 0.952 | 0.963 | 0.931 |
| $\tau = 0.5$ | 10 | 50 | 0.643 | 0.889 | 0.152 | 0.524 | 0.837 | 0.165 | 0.364 | 0.664 | 0.201 |
| | 10 | 100 | 0.927 | 0.949 | 0.127 | 0.907 | 0.935 | 0.136 | 0.807 | 0.845 | 0.171 |
| | 10 | 200 | 0.968 | 0.981 | 0.128 | 0.955 | 0.974 | 0.136 | 0.917 | 0.940 | 0.158 |
| | 10 | 500 | 0.990 | 0.994 | 0.135 | 0.987 | 0.991 | 0.126 | 0.979 | 0.986 | 0.142 |
| | 50 | 50 | 0.697 | 0.913 | -0.013 | 0.581 | 0.870 | -0.011 | 0.279 | 0.682 | 0.005 |
| | 50 | 100 | 0.945 | 0.968 | 0.004 | 0.929 | 0.956 | 0.003 | 0.857 | 0.899 | 0.004 |
| | 50 | 200 | 0.973 | 0.984 | 0.011 | 0.967 | 0.980 | 0.012 | 0.945 | 0.968 | 0.014 |
| | 50 | 500 | 0.991 | 0.994 | 0.018 | 0.989 | 0.993 | 0.017 | 0.984 | 0.989 | 0.018 |
| $\tau = 0.75$ | 10 | 50 | 0.718 | 0.878 | 0.603 | 0.609 | 0.804 | 0.629 | 0.356 | 0.473 | 0.596 |
| | 10 | 100 | 0.932 | 0.948 | 0.834 | 0.874 | 0.900 | 0.791 | 0.636 | 0.664 | 0.737 |
| | 10 | 200 | 0.970 | 0.980 | 0.922 | 0.943 | 0.962 | 0.907 | 0.796 | 0.833 | 0.848 |
| | 10 | 500 | 0.991 | 0.993 | 0.975 | 0.984 | 0.987 | 0.968 | 0.933 | 0.941 | 0.943 |
| | 50 | 50 | 0.663 | 0.872 | 0.498 | 0.485 | 0.779 | 0.476 | 0.102 | 0.203 | 0.392 |
| | 50 | 100 | 0.935 | 0.956 | 0.813 | 0.889 | 0.920 | 0.762 | 0.450 | 0.510 | 0.608 |
| | 50 | 200 | 0.969 | 0.981 | 0.906 | 0.951 | 0.970 | 0.890 | 0.792 | 0.845 | 0.800 |
| | 50 | 500 | 0.990 | 0.993 | 0.969 | 0.984 | 0.989 | 0.962 | 0.951 | 0.964 | 0.931 |

Note: the DGP is $Y_{it} = \lambda_{i1} f_{t1} + \lambda_{i2} f_{t2} + (\lambda_{i3} f_{t3}) u_{it}$, where $f_{t3} = |h_t|, f_{t1}, f_{t2}, h_t \sim i.i.d\ N(0,1)$. The number of characteristics is 2 and all characteristics $x_{id}$ are independently drawn from the uniform distribution: $U[-1,1]$. $g_1(x) = sin(2\pi x)$, $g_2(x) = sin(\pi x)$ *and* $g_3(x) = |cos(\pi x)|$. $\lambda_{i1} = \sum_{d=1,2} g_1(x_{id})$, $\lambda_{i2} = \sum_{d=1,2} g_2(x_{id})$ *and* $\lambda_{i3} = \sum_{d=1,2} g_3(x_{id})$. $\{u_{it}\}$ are i.i.d draws from three different distributions. 3 factors are estimated at each $\tau$ using the method proposed in this paper, and the reported results are the averages of the adjusted $R^2$ of regressing the true factors on the estimated factors from 1000 replications.

Table 7: Factor estimation using SQFA: $R = 3, D = 2$

| | | | $N(0,1)$ | | | $t(3)$ | | | Cauchy(0,1) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $T$ | $n$ | $f_{1t}$ | $f_{2t}$ | $f_{3t}$ | $f_{1t}$ | $f_{2t}$ | $f_{3t}$ | $f_{1t}$ | $f_{2t}$ | $f_{3t}$ |
| $\tau = 0.25$ | 10 | 50 | 0.406 | 0.826 | 0.205 | 0.364 | 0.793 | 0.218 | 0.295 | 0.649 | 0.277 |
| | 10 | 100 | 0.595 | 0.698 | 0.195 | 0.578 | 0.690 | 0.208 | 0.556 | 0.656 | 0.231 |
| | 10 | 200 | 0.623 | 0.706 | 0.223 | 0.604 | 0.680 | 0.238 | 0.573 | 0.676 | 0.277 |
| | 10 | 500 | 0.630 | 0.682 | 0.233 | 0.614 | 0.687 | 0.239 | 0.596 | 0.686 | 0.301 |
| | 50 | 50 | 0.311 | 0.845 | 0.040 | 0.267 | 0.804 | 0.058 | 0.191 | 0.662 | 0.101 |
| | 50 | 100 | 0.516 | 0.680 | 0.036 | 0.503 | 0.660 | 0.043 | 0.471 | 0.627 | 0.063 |
| | 50 | 200 | 0.523 | 0.691 | 0.058 | 0.518 | 0.676 | 0.067 | 0.487 | 0.649 | 0.095 |
| | 50 | 500 | 0.578 | 0.656 | 0.061 | 0.553 | 0.651 | 0.068 | 0.543 | 0.626 | 0.102 |
| $\tau = 0.5$ | 10 | 50 | 0.383 | 0.849 | 0.133 | 0.351 | 0.819 | 0.132 | 0.314 | 0.749 | 0.142 |
| | 10 | 100 | 0.584 | 0.695 | 0.150 | 0.573 | 0.688 | 0.143 | 0.531 | 0.674 | 0.154 |
| | 10 | 200 | 0.584 | 0.713 | 0.156 | 0.559 | 0.716 | 0.157 | 0.539 | 0.678 | 0.158 |
| | 10 | 500 | 0.615 | 0.689 | 0.157 | 0.600 | 0.663 | 0.157 | 0.598 | 0.649 | 0.152 |
| | 50 | 50 | 0.277 | 0.865 | -0.014 | 0.236 | 0.843 | -0.013 | 0.185 | 0.773 | -0.014 |
| | 50 | 100 | 0.509 | 0.679 | 0.007 | 0.471 | 0.669 | 0.007 | 0.439 | 0.625 | 0.005 |
| | 50 | 200 | 0.514 | 0.688 | 0.015 | 0.493 | 0.680 | 0.016 | 0.452 | 0.669 | 0.016 |
| | 50 | 500 | 0.557 | 0.639 | 0.022 | 0.544 | 0.636 | 0.023 | 0.503 | 0.623 | 0.022 |
| $\tau = 0.75$ | 10 | 50 | 0.402 | 0.816 | 0.200 | 0.374 | 0.804 | 0.213 | 0.316 | 0.630 | 0.268 |
| | 10 | 100 | 0.606 | 0.688 | 0.191 | 0.564 | 0.900 | 0.195 | 0.556 | 0.666 | 0.226 |
| | 10 | 200 | 0.590 | 0.691 | 0.221 | 0.584 | 0.962 | 0.215 | 0.582 | 0.672 | 0.281 |
| | 10 | 500 | 0.638 | 0.691 | 0.231 | 0.622 | 0.987 | 0.259 | 0.620 | 0.675 | 0.285 |
| | 50 | 50 | 0.318 | 0.837 | 0.039 | 0.268 | 0.779 | 0.049 | 0.191 | 0.658 | 0.099 |
| | 50 | 100 | 0.525 | 0.671 | 0.039 | 0.499 | 0.920 | 0.044 | 0.465 | 0.624 | 0.067 |
| | 50 | 200 | 0.528 | 0.688 | 0.057 | 0.510 | 0.970 | 0.064 | 0.481 | 0.654 | 0.101 |
| | 50 | 500 | 0.574 | 0.652 | 0.063 | 0.564 | 0.989 | 0.067 | 0.543 | 0.627 | 0.096 |

Note: the DGP is $Y_{it} = \lambda_{i1} f_{t1} + \lambda_{i2} f_{t2} + (\lambda_{i3} f_{t3}) u_{it}$, where $f_{t3} = |h_t|, f_{t1}, f_{t2}, h_t \sim i.i.d\ N(0,1)$. The number of characteristics is 2 and all characteristics $x_{id}$ are independently drawn from the uniform distribution: $U[-1,1]$. $g_1(x) = sin(2\pi x)$, $g_2(x) = sin(\pi x)$ $and$ $g_3(x) = |cos(\pi x)|$. $\lambda_{i1} = \sum_{d=1,2} g_1(x_{id})$, $\lambda_{i2} = \sum_{d=1,2} g_2(x_{id})$ $and$ $\lambda_{i3} = \sum_{d=1,2} g_3(x_{id})$. $\{u_{it}\}$ are i.i.d draws from three different distributions. 2 factors are estimated at each $\tau$ using the method proposed by Ma et al. (2021), and the reported results are the averages of the adjusted $R^2$ of regressing the true factors on the estimated factors from 1000 replications.

Table 8: Estimated numbers of factors

| | | Five largest eigenvalues of $\hat{Y}\hat{Y}'$ | | | | | $p_n$ | $\hat{r}$ | $\hat{r}_{QFA}$ |
|---|---|---|---|---|---|---|---|---|---|
| mean(PPCA) | | 0.929 | 0.090 | 0.081 | 0.066 | 0.043 | | 1 | |
| quantile | $\tau$=0.5 | 0.887 | 0.094 | 0.084 | 0.053 | 0.043 | 0.224 | 1 | 1 |
| | $\tau$=0.25 | 1.713 | 0.110 | 0.098 | 0.059 | 0.047 | 0.311 | 1 | 1 |
| | $\tau$=0.75 | 2.706 | 0.115 | 0.087 | 0.074 | 0.067 | 0.391 | 1 | 1 |
| | $\tau$=0.05 | 8.415 | 0.311 | 0.173 | 0.161 | 0.138 | 0.690 | 1 | 1 |
| | $\tau$=0.95 | 13.715 | 0.567 | 0.428 | 0.291 | 0.246 | 0.880 | 1 | 1 |

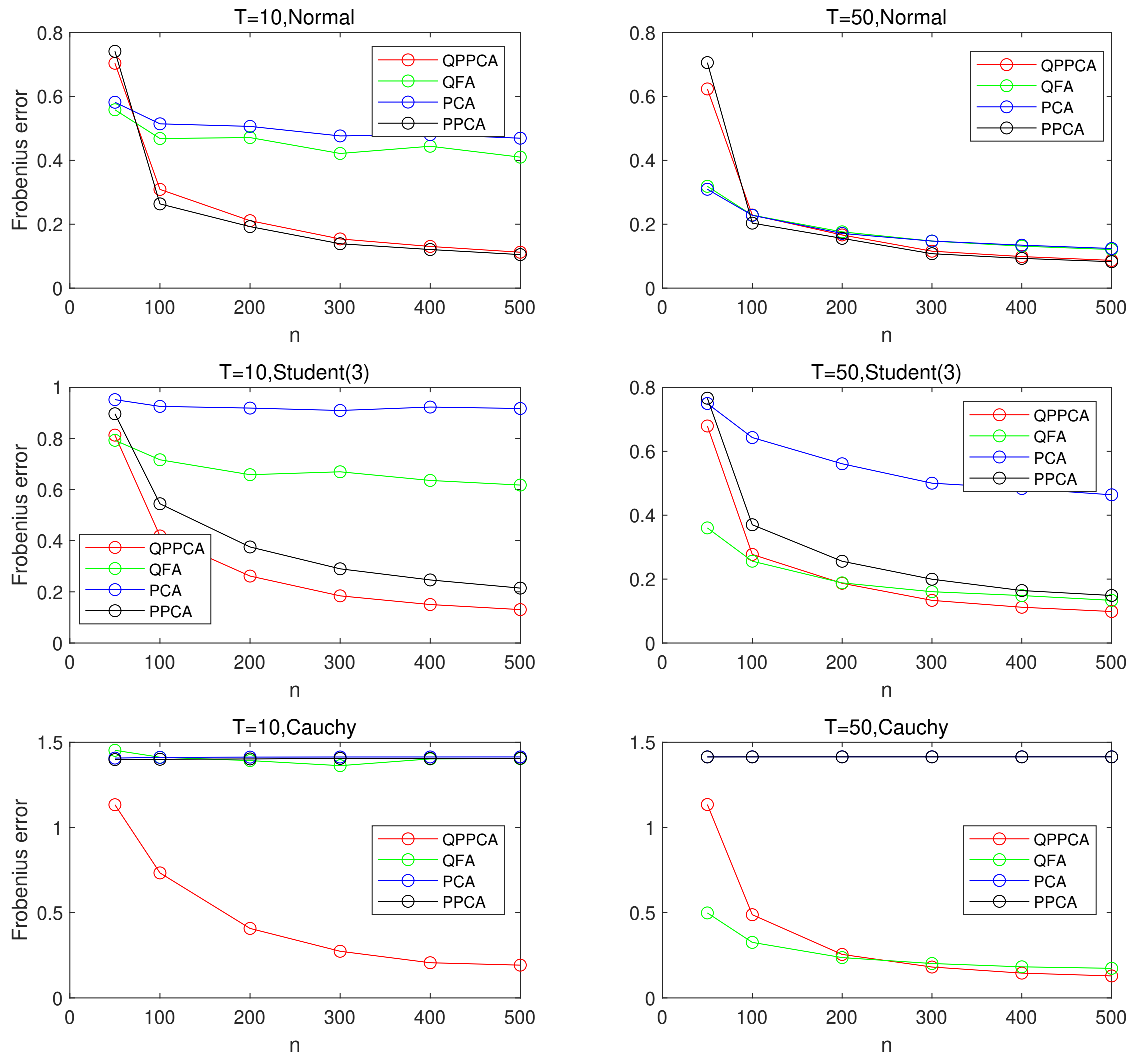
Note: this table shows the estimated numbers of factors using the eigen-ratio estimator proposed by Fan et al. (2016), the proposed estimator in this paper, and the rank-minimization estimator proposed by Chen et al. (2021) for different $\tau$s. Column 3 to Column 7 give the 5 largest eigenvalues of $\hat{Y}\hat{Y}'$, where $\hat{Y}$ is the matrix of fitted values in the first-step sieve regressions, and $p_n$ is the threshold value defined in (9).

Table 9: Correlations and means of estimated factors

| | $\tau$=0.05 | $\tau$=0.25 | $\tau$=0.5 | $\tau$=0.75 | $\tau$=0.95 | PPCA | Mean |
|---|---|---|---|---|---|---|---|
| $\tau$=0.05 | 1 | 0.922 | 0.852 | 0.767 | 0.611 | 0.863 | 0.943 |
| $\tau$=0.25 | | 1 | 0.975 | 0.924 | 0.753 | 0.973 | 0.738 |
| $\tau$=0.5 | | | 1 | 0.971 | 0.814 | 0.990 | -0.121 |
| $\tau$=0.75 | | | | 1 | 0.877 | 0.979 | -0.784 |
| $\tau$=0.95 | | | | | 1 | 0.862 | -0.943 |
| PPCA | | | | | | 1 | -0.231 |

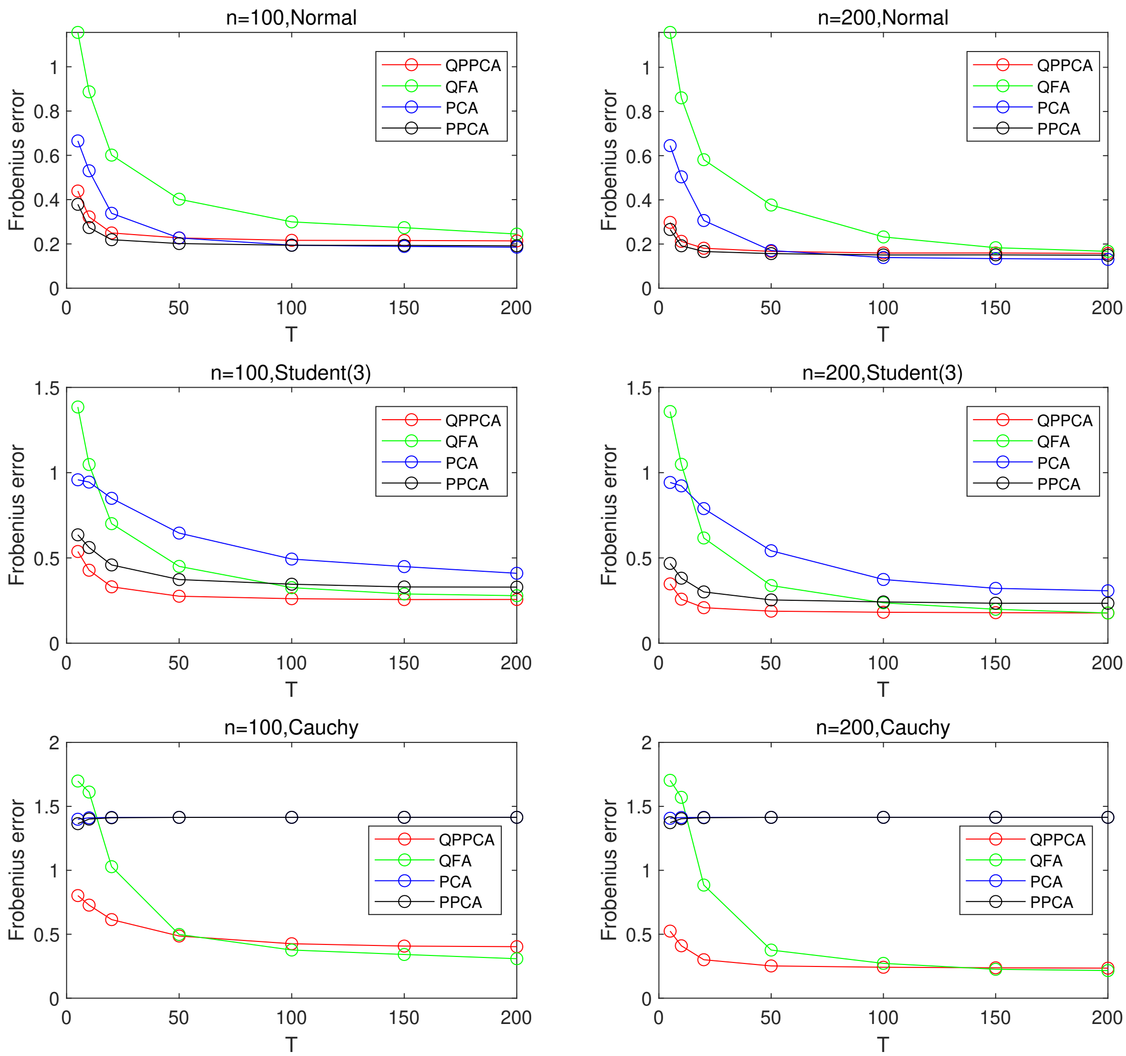
Note: this table shows the correlations and sample means of the estimated mean factor using PPCA and the estimated quantile factors at different $\tau$s using QP-PCA.

Figure 1: Estimation of factors: fixed $T$ and increasing $n$.



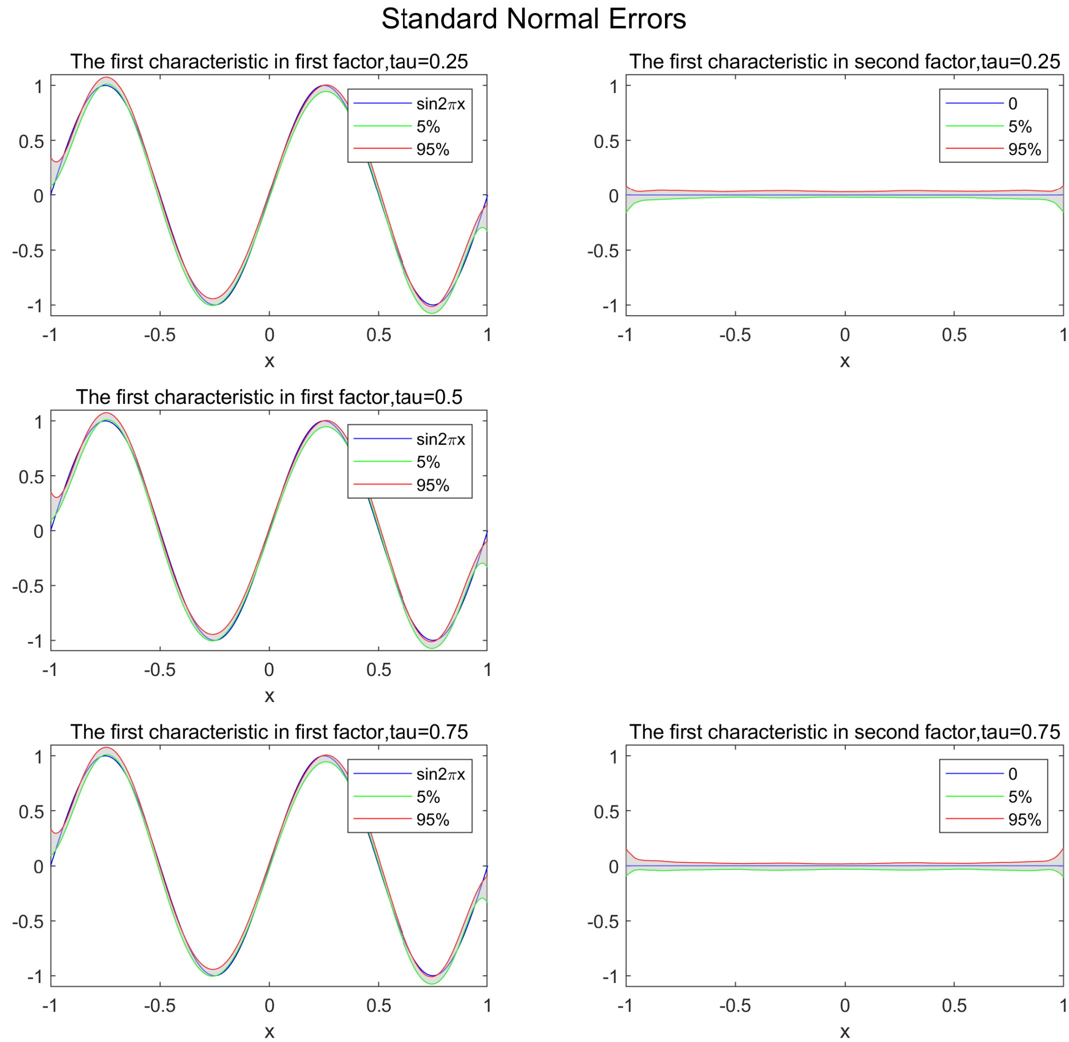

Note: the DGP is $Y_{it} = \lambda_{i1} f_{t1} + \lambda_{i2} f_{t2} + (\lambda_{i3} f_{t3}) u_{it}$, where $f_{t3} = |h_t|, f_{t1}, f_{t2}, h_t \sim i.i.d\ N(0,1)$. The number of characteristics is 5 and all characteristics $x_{id}$ ($i = 1, ...N$ and $d = 1, 2, 3, 4, 5$) are independently drawn from the uniform distribution: $U[-1,1]$. $g_1(x) = sin(2\pi x)$, $g_2(x) = sin(\pi x)$ $and$ $g_3(x) = |cos(\pi x)|$. The factor loading functions are generated as $\lambda_{i1} = \sum_{d=1,3,5} g_1(x_{id})$, $\lambda_{i2} = \sum_{d=1,2} g_2(x_{id})$ and $\lambda_{i3} = \sum_{d=3,4} g_3(x_{id})$. $\{u_{it}\}$ are i.i.d draws from three different distributions. The mean factors ($f_{t1}$ and $f_{t2}$) are estimated by four methods: PCA, PPCA, QFA and QPPCA at $\tau = 0.5$. The reported results are the average Frobenius errors: $\|\hat{\boldsymbol{F}} - \boldsymbol{F}\hat{\boldsymbol{H}}\|/\sqrt{T}$ from 1000 repetitions, where $\hat{\boldsymbol{H}}$ is the associated rotation matrix for each estimator.

Figure 2: Estimation of factors: fixed $n$ and increasing $T$.



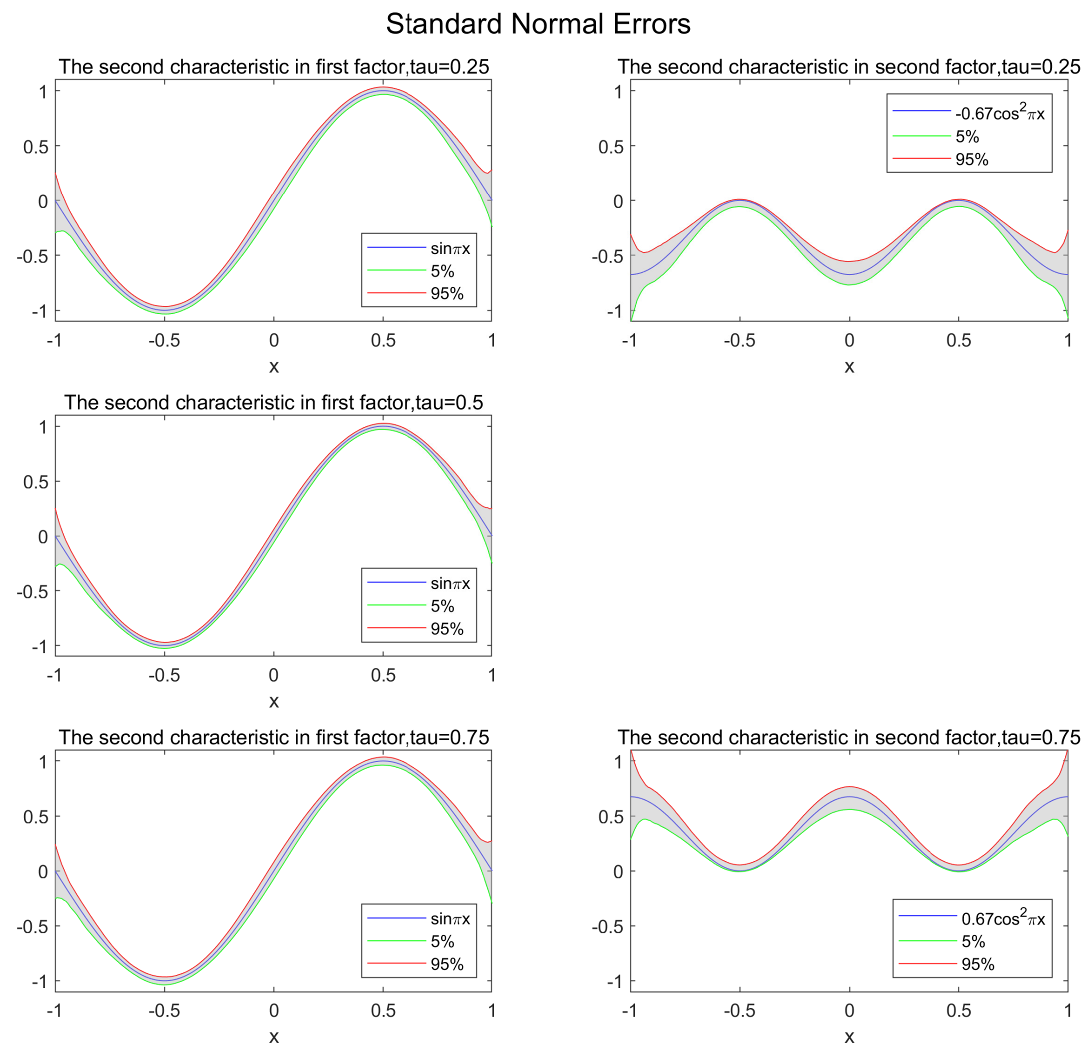

Note: the DGP is $Y_{it} = \lambda_{i1} f_{t1} + \lambda_{i2} f_{t2} + (\lambda_{i3} f_{t3}) u_{it}$, where $f_{t3} = |h_t|, f_{t1}, f_{t2}, h_t \sim i.i.d\ N(0,1)$. The number of characteristics is 5 and all characteristics $x_{id}$ ($i = 1, ...N$ and $d = 1, 2, 3, 4, 5$) are independently drawn from the uniform distribution: $U[-1,1]$. $g_1(x) = sin(2\pi x)$, $g_2(x) = sin(\pi x)$ $and$ $g_3(x) = |cos(\pi x)|$. The factor loading functions are generated as $\lambda_{i1} = \sum_{d=1,3,5} g_1(x_{id})$, $\lambda_{i2} = \sum_{d=1,2} g_2(x_{id})$ and $\lambda_{i3} = \sum_{d=3,4} g_3(x_{id})$. $\{u_{it}\}$ are i.i.d draws from three different distributions. The mean factors ($f_{t1}$ and $f_{t2}$) are estimated by four methods: PCA, PPCA, QFA and QPPCA at $\tau = 0.5$. The reported results are the average Frobenius errors: $\|\hat{\boldsymbol{F}} - \boldsymbol{F}\hat{\boldsymbol{H}}\|/\sqrt{T}$ from 1000 repetitions, where $\hat{\boldsymbol{H}}$ is the associated rotation matrix for each estimator.

Figure 3: Loading function of the first characteristic when error term is $N(0,1)$



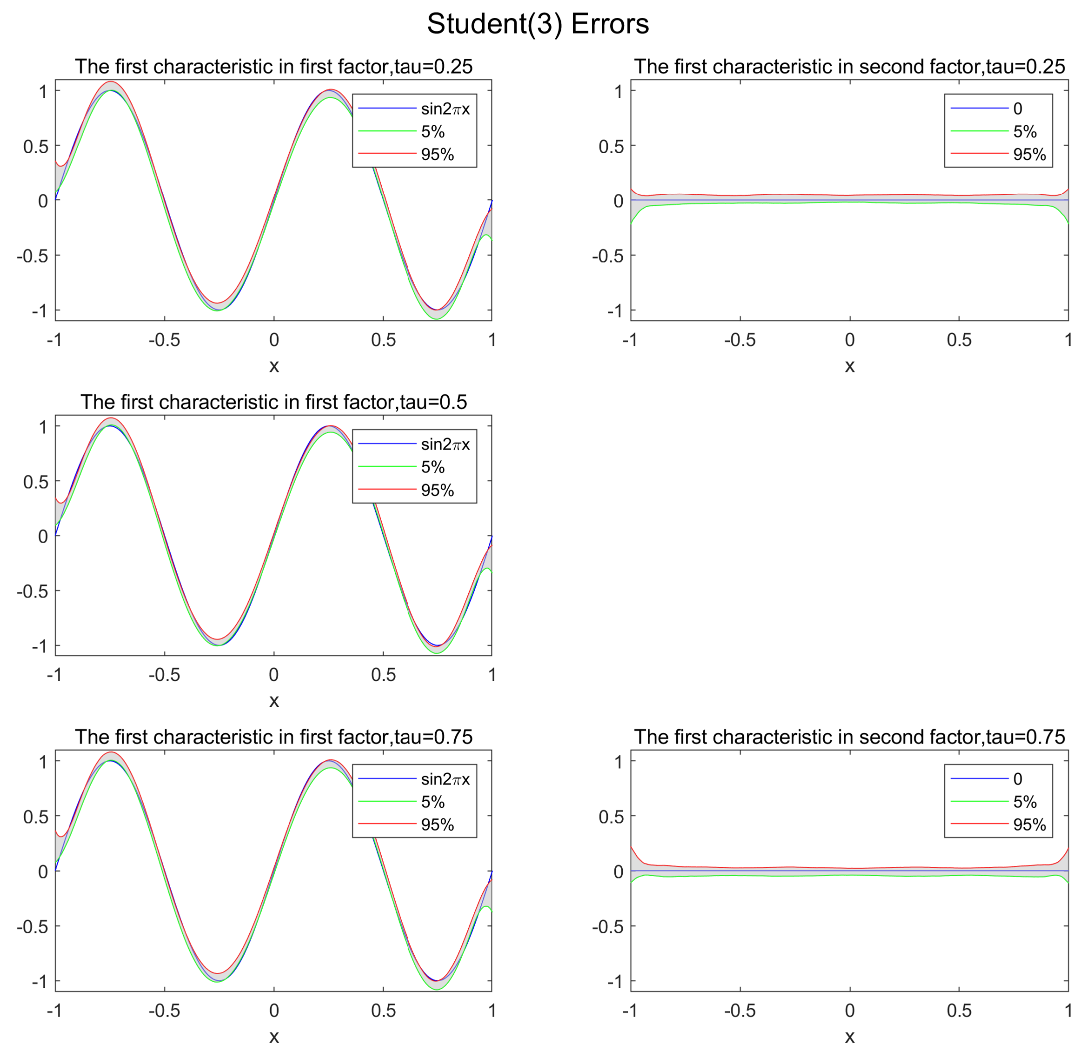

Note: the DGP is: $y_{it} = \lambda_{i1} f_{t1} + (\lambda_{i2} f_{t2}) u_{it}$, where $f_{t2} = |g_t|$ and $f_{t1}, g_t \sim i.i.d\ N(0,1)$. $n = 500, T = 10$. The number of characteristics is 2 and all characteristics $x_{id}$ ($i = 1, ...N$ and $d = 1, 2$) are independently drawn from uniform distribution: $U[-1,1]$. $g_{11}(x) = sin(2\pi x), g_{21}(x) = 0, g_{12}(x) = sin(\pi x)$, $g_{22}(x) = cos^2(\pi x)$, and $\lambda_{i1} = g_{11}(x_{i1}) + g_{12}(x_{i2}), \lambda_{i2} = g_{21}(x_{i1}) + g_{22}(x_{i2})$. $u_{it}$ are drawn independently from the standard normal distribution. The left panel are the estimation results for $g_{11,\tau}(x) = \sin(2\pi x)$ and the right panel are the estimation results for $g_{21,\tau}(x) = 0$ with $\tau \in \{0.25, 0.75\}$. For each graph, the blue line is the true function, the red line and the green line are the 95% and 5% empirical quantiles from 1000 replications.

Figure 4: Loading function of the second characteristic when error term is N(0,1)



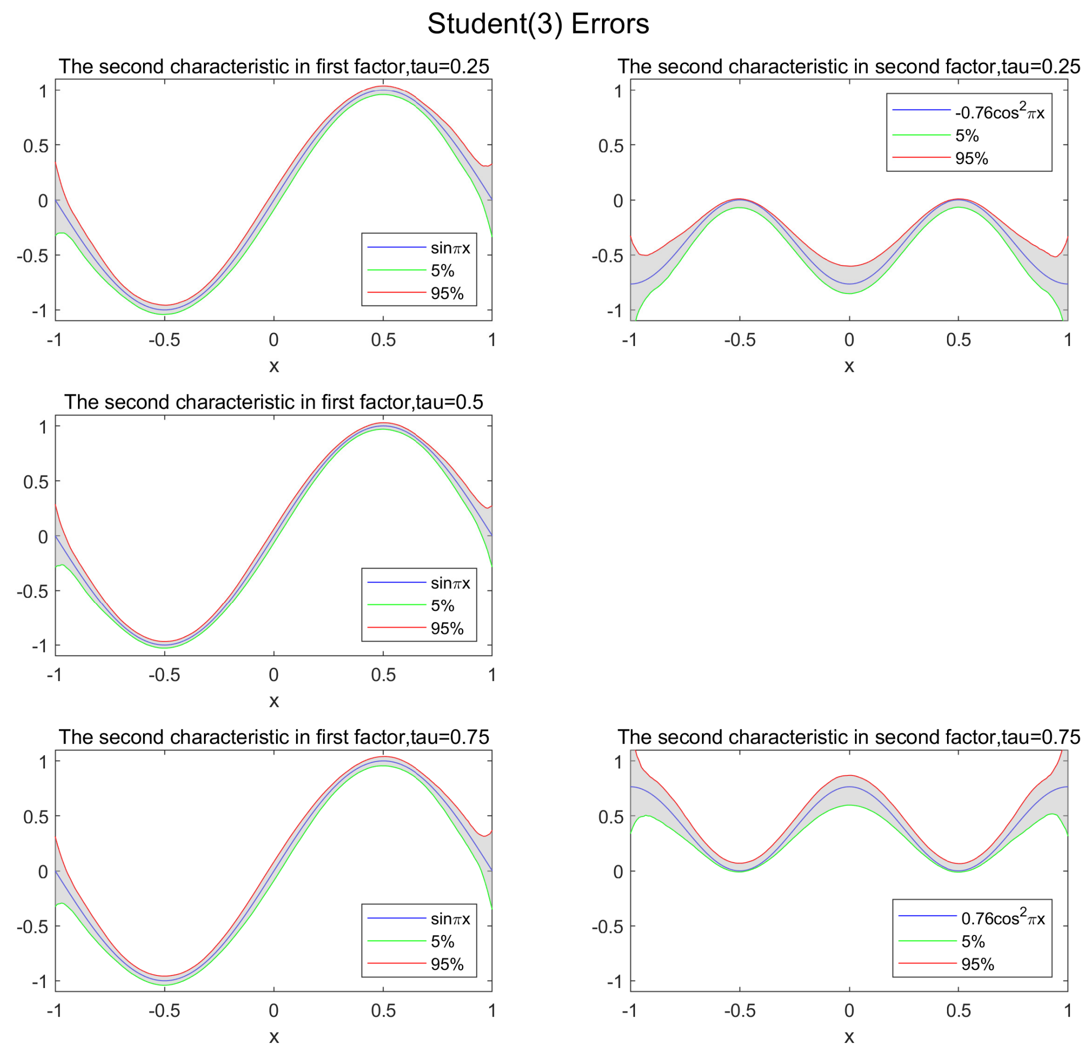

Note: the DGP is: $y_{it} = \lambda_{1i} f_{1t} + (\lambda_{2i} f_{2t}) u_{it}$, where $f_{2t} = |g_t|$ and $f_{1t}, g_t \sim i.i.d\ N(0,1)$. $n = 500, T = 10$. The number of characteristics is 2 and all characteristics $x_{id}$ ($i = 1, ...N$ and $d = 1, 2$) are independently drawn from uniform distribution: $U[-1, 1]$. $g_{11}(x) = sin(2\pi x), g_{21}(x) = 0, g_{12}(x) = sin(\pi x)$, $g_{22}(x) = cos^2(\pi x)$, and $\lambda_{1i} = g_{11}(x_{1i}) + g_{12}(x_{2i}), \lambda_{2i} = g_{21}(x_{1i}) + g_{22}(x_{2i})$. $u_{it}$ are drawn independently from the standard normal distribution. The left panel are the estimation results for $g_{12,\tau}(x) = \sin(\pi x)$ and the right panel are the estimation results for $g_{22,\tau}(x)$ with $\tau \in \{0.25, 0.75\}$. For each graph, the blue line is the true function, the red line and the green line are the 95% and 5% empirical quantiles from 1000 replications.

Figure 5: Loading function of first characteristic when error term is $t(3)$



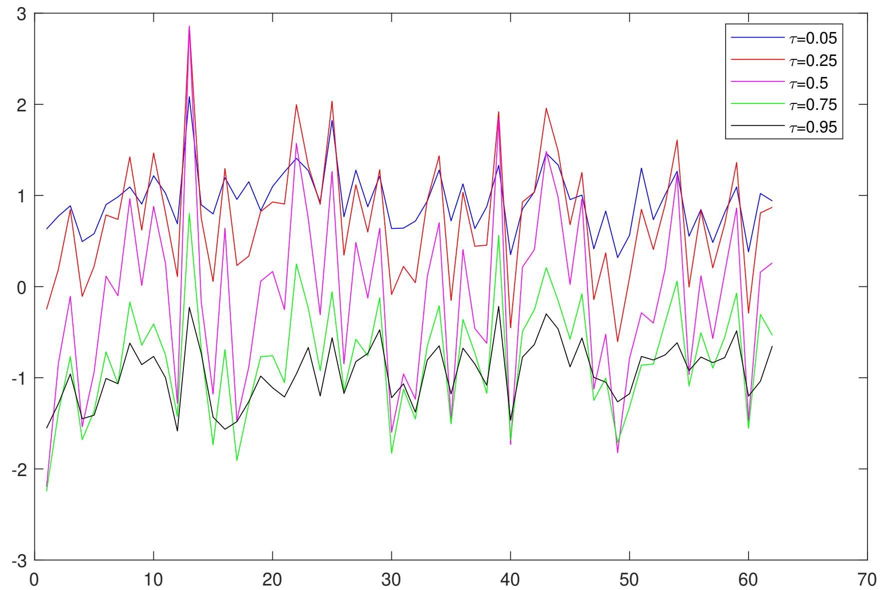

Note: the DGP is: $y_{it} = \lambda_{i1} f_{t1} + (\lambda_{i2} f_{t2}) u_{it}$, where $f_{t2} = |g_t|$ and $f_{t1}, g_t \sim i.i.d\ N(0, 1)$. $n = 500, T = 10$. The number of characteristics is 2 and all characteristics $x_{id}$ $(i = 1, ...N$ and $d = 1, 2)$ are independently drawn from uniform distribution: $U[-1, 1]$. $g_{11}(x) = sin(2\pi x), g_{21}(x) = 0, g_{12}(x) = sin(\pi x)$, $g_{22}(x) = cos^2(\pi x)$, and $\lambda_{i1} = g_{11}(x_{i1}) + g_{12}(x_{i2}), \lambda_{i2} = g_{21}(x_{i1}) + g_{22}(x_{i2})$. $u_{it}$ are drawn independently from the student's t distribution with 3 degrees of freedom. The left panel are the estimation results for $g_{11,\tau}(x) = \sin(2\pi x)$ and the right panel are the estimation results for $g_{21,\tau}(x) = 0$ with $\tau \in \{0.25, 0.75\}$. For each graph, the blue line is the true function, the red line and the green line are the 95% and 5% empirical quantiles from 1000 replications.

Figure 6: Loading function of second characteristic when error term is $t(3)$



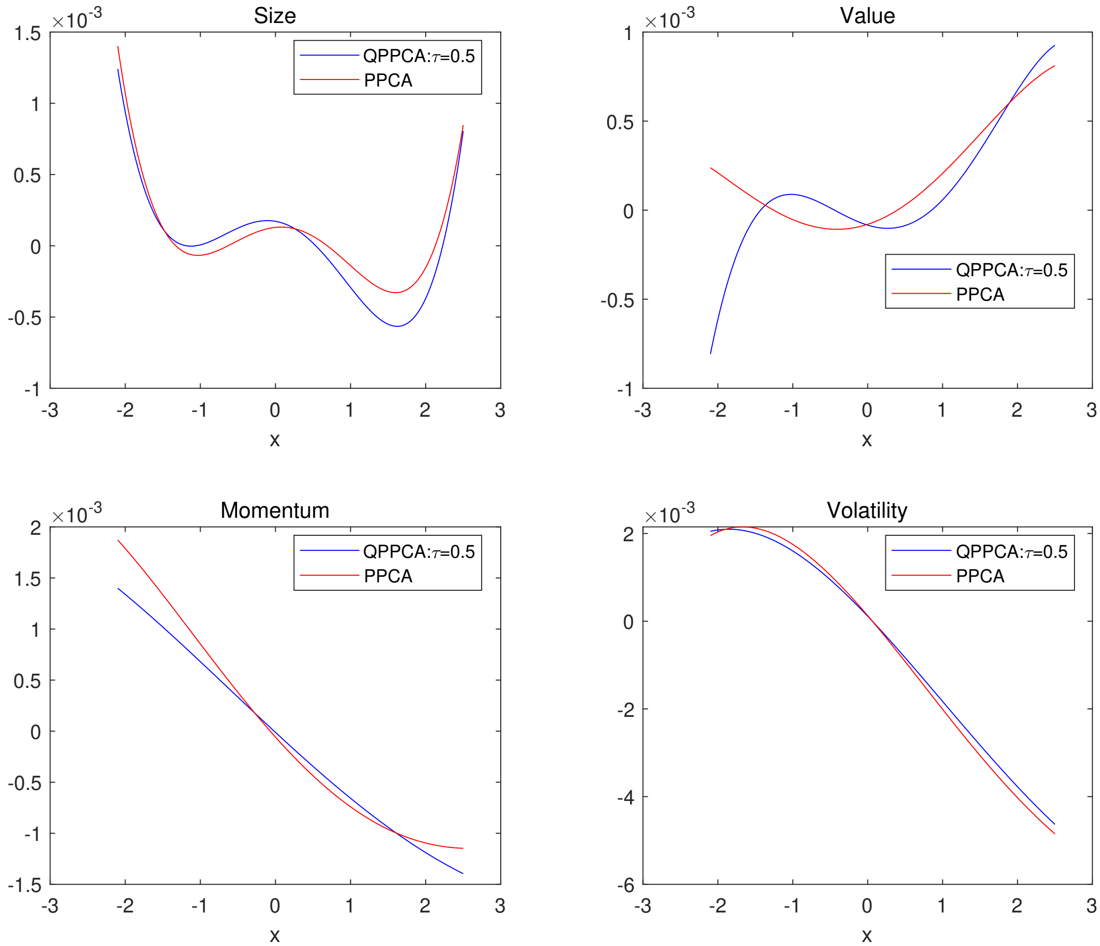

Note: the DGP is: $y_{it} = \lambda_{i1} f_{t1} + (\lambda_{i2} f_{t2}) u_{it}$, where $f_{t2} = |g_t|$ and $f_{t1}, g_t \sim i.i.d\ N(0,1)$. $n = 500, T = 10$. The number of characteristics is 2 and all characteristics $x_{id}$ ($i = 1, ...N$ and $d = 1, 2$) are independently drawn from uniform distribution: $U[-1, 1]$. $g_{11}(x) = sin(2\pi x), g_{21}(x) = 0, g_{12}(x) = sin(\pi x)$, $g_{22}(x) = cos^2(\pi x)$, and $\lambda_{i1} = g_{11}(x_{i1}) + g_{12}(x_{i2}), \lambda_{i2} = g_{21}(x_{i1}) + g_{22}(x_{i2})$. $u_{it}$ are drawn independently from the student's t distribution with 3 degrees of freedom. The left panel are the estimation results for $g_{12,\tau}(x) = \sin(\pi x)$ and the right panel are the estimation results for $g_{22,\tau}(x)$ with $\tau \in \{0.25, 0.75\}$. For each graph, the blue line is the true function, the red line and the green line are the 95% and 5% empirical quantiles from 1000 replications.

Figure 7: Estimated loading functions using QPPCA for different quantiles



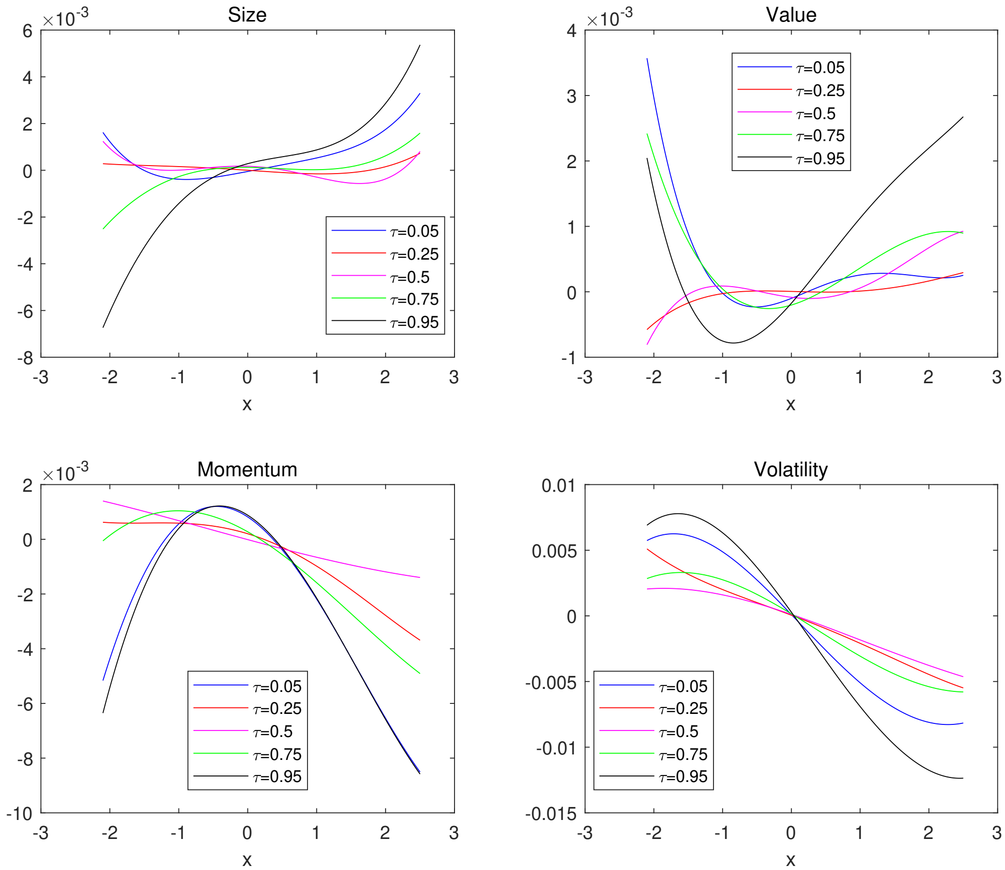

Note: This figure plots the estimated quantile factors at different quantiles using QPPCA.

Figure 8: Estimated loading functions using PPCA and QPPCA for $\tau = 0.5$




Note: this figure plots the estimated quantile factor loading functions of the four characteristics using PPCA and QPPCA at $\tau = 0.5$

Figure 9: Estimated loading functions using QPPCA for different quantiles




Note: This figure plots the estimated quantile factor loading functions of the four characteristics using QPPCA at different $\tau$s.

# References


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

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

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

Chamberlain, G. and M. Rothschild (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica 51*(5), 1281–1304.

Chen, L., J. J. Dolado, and J. Gonzalo (2021). Quantile factor models. *Econometrica 89*(2), 875–910.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics 6*, 5549–5632.

Chen, X. and X. Shen (1998). Sieve extremum estimates for weakly dependent data. *Econometrica*, 289–314.

Connor, G., M. Hagmann, and O. Linton (2012). Efficient semiparametric estimation of the Fama–French model and extensions. *Econometrica 80*(2), 713–754.

Connor, G. and R. A. Korajczyk (1993). A test for the number of factors in an approximate factor model. *the Journal of Finance 48*(4), 1263–1291.

Connor, G. and O. Linton (2007). Semiparametric estimation of a characteristic-based factor model of common stock returns. *Journal of Empirical Finance 14*(5), 694–717.

Elton, E. J. (1999). Expected return, realized return, and asset pricing tests. *The Journal of Finance 54*(4), 1199–1220.

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.

Fama, E. F. and K. R. French (2015). A five-factor asset pricing model. *Journal of Financial Economics 116*(1), 1–22.

Fan, J., K. Li, and Y. Liao (2021). Recent developments in factor models and applications in econometric learning. *Annual Review of Financial Economics 13*, 401–430.

Fan, J., Y. Liao, and W. Wang (2016). Projected principal component analysis in factor models. *Annals of Statistics 44*(1), 219.

Hausman, J., H. Liu, Y. Luo, and C. Palmer (2021). Errors in the dependent variable of quantile regression models. *Econometrica 89*(2), 849–873.

Horowitz, J. L. and S. Lee (2005). Nonparametric estimation of an additive quantile regression model. *Journal of the American Statistical Association 100*(472), 1238–1249.

Jung, S. and J. S. Marron (2009). Pca consistency in high dimension, low sample size context. *The Annals of Statistics 37*(6B), 4104–4130.

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.

Ledoux, M. and M. Talagrand (1991). *Probability in Banach Spaces: isoperimetry and processes*, Volume 23. Springer Science & Business Media.

Lee, J. and P. M. Robinson (2016). Series estimation under cross-sectional dependence. *Journal of Econometrics 190*(1), 1–17.

Ma, S., O. Linton, and J. Gao (2021). Estimation and inference in semiparametric quantile factor models. *Journal of Econometrics 222*(1), 295–323.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics 79*(1), 147–168.

Shen, D., H. Shen, and J. S. Marron (2013). Consistency of sparse pca in high dimension, low sample size contexts. *Journal of Multivariate Analysis 115*, 317–333.

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.