Quantile Factor Models*

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

[1] *Peking University HSBC Business School, 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*

October 20, 2020

Abstract

Quantile factor models (QFM) represent a new class of factor models for high-dimensional panel data. Unlike approximate factor models (AFM), which only extract mean factors, QFM also allow unobserved factors to shift other relevant parts of the distributions of observables. We propose a quantile regression approach, labeled Quantile Factor Analysis (QFA), to consistently estimate all the quantile-dependent factors and loadings. Their asymptotic distributions are established using a kernel-smoothed version of the QFA estimators. Two consistent model selection criteria, based on information criteria and rank minimization, are developed to determine the number of factors at each quantile. QFA estimation remains valid even when the idiosyncratic errors exhibit heavy-tailed distributions. An empirical application illustrates the usefulness of QFA by highlighting the role of extra factors in the forecasts of US GDP growth and inflation rates using a large set of predictors.

Keywords: Factor models, quantile regression, incidental parameters.
JEL codes: C31, C33, C38.

*We are indebted to a co-editor and four anonymous referees for many constructive inputs which have improved the paper. We also thank Dante Amengual, Manuel Arellano, Steve Bond, Guillaume Carlier, Valentina Corradi, Jean-Pierre Florens, Alfred Galichon, Peter R. Hansen, Jerry Hausman, Sophocles Mavroeidis, Bent Nielsen, Olivier Scaillet, Enrique Sentana, Liangjun Su, and participants at several seminars and conferences for helpful comments and suggestions. Financial support from the National Natural Science Foundation of China (Grant No.71703089), The Open Society Foundation, The Oxford Martin School, the Spanish Ministerio de Economía y Competitividad (grants ECO2016-78652, PID2019-104960GB-I00, and Maria de Maeztu MDM 2014-0431), and MadEco-CM (grant S205/HUM-3444) is gratefully acknowledged. The usual disclaimer applies.

1 Introduction

The theory and empirical applications of *approximate factor models* (AFM) have developed very rapidly in economics and finance since their introduction by Chamberlain and Rothschild (1983). As is well known, AFM imply that a panel $\{X_{it}\}$ of N variables (units), each with T observations, has the representation $X_{it} = \lambda_i' f_t + \epsilon_{it}$, where $\lambda_i = [\lambda_{1i}, .., \lambda_{ri}]'$ and $f_t = [f_{1t}, .., f_{rt}]'$ are $r \times 1$ vectors of factor loadings and common factors, respectively, with $r \ll N$, and $\{\epsilon_{it}\}$ are zero-mean weakly dependent idiosyncratic disturbances, uncorrelated with the factors. The availability of fairly straightforward estimation procedures for AFM — e.g. via principal component analysis (PCA) — has triggered their widespread use in research.[1]

Inspired by the generalization of linear regression to quantile regression (QR) models, our starting point in the current paper is to recall that the standard regression interpretation of AFM as linear conditional mean models of X_{it} given f_t (i.e. $\mathbb{E}(X_{it}|f_t) = \lambda_i' f_t$), entails two possibly restrictive features. First, PCA does not capture hidden factors that may shift characteristics (moments or quantiles) of the distribution of X_{it} other than its mean. Second, neither the loadings λ_i nor the factors f_t are allowed to vary across the distribution of each unit in the panel. The relevance of these features has been recently highlighted in the empirical finance, macro and micro literatures. For example, in the former, Amengual and Sentana (2020) find nonlinear tail dependence, co-skewness and co-kurtosis in cross-sectional dependence among monthly returns on individual US stocks; likewise Ando and Bai (2020) (AB 2020, hereafter) show that the common factor structures explaining the upper and lower tails of the asset return distributions in global financial markets have become different since the subprime crisis. On the macro side, Adrian, Boyarchenko, and Giannone (2019) document that only the estimated lower conditional quantiles of the distribution of future GDP growth in the US exhibit strong dependence on current financial conditions. Lastly, in micro theory, de Castro and Galvao (2019) have recently extended the traditional expected utility model of rational behavior to quantile utility preferences, where e.g. factor structures determining hedonic pricing of consumption goods or financial stocks may exhibit large differences across quantiles.

In light of this evidence, our goal here is to develop a common factor methodology for a class of models, coined *quantile factor models* (QFM), which is flexible enough to capture the quantile-dependent objects that standard AFM tools are unable to retrieve. In particular, we

[1] Early applications of AFM abound in aggregation theory, consumer theory, business cycle analysis, finance, monetary economics, and monitoring and forecasting; see, *inter alia*, Bai (2003), Bai and Ng (2008), Stock and Watson (2011). More recently, the characterization of cross-sectional dependence among error terms in Panel Data has relied on the use of a finite number of unobserved common factors which originate from economy-wide shocks affecting units with different intensities (loadings). Interactive fixed-effects models can be easily estimated by PCA (see Bai 2009) or by common correlated effects (see Pesaran 2006), and there are even generalizations of these techniques for nonlinear panel single-index models (see Chen, Fernández-Val, and Weidner 2020). Lastly, the surge of big data and machine learning technologies has made factor models a key tool for dimension reduction and predictive analytics when using very large datasets (see Athey and Imbens 2019 for a survey).

focus on the factor structure of a data generating process (DGP) rewritten in QR format as $X_{it} = \lambda_i'(\tau)f_t(\tau) + u_{it}(\tau)$, with $0 < \tau < 1$, where the conditional quantile satisfies $\mathsf{Q}_{u_{it}(\tau)}[\tau|f_t(\tau)] = 0$, and the number of factors, $r(\tau)$ is also allowed to depend on τ.[2] To estimate and draw inference on the quantile-dependent factors and loadings in QFM, we propose an estimation approach labeled *Quantile Factor Analysis* (QFA) while, to estimate the number of factors at each τ, we derive two novel selection criteria (one based on information criteria and another on rank minimization). Put succinctly, QFM and QFA could be thought of as capturing the same type of flexible generalization that QR techniques represent for linear regression models.

Our proposed QFA estimation procedure relies on the minimization of the standard *check* function in QR (instead of the conventional quadratic loss function used in AFM) to estimate jointly the common factors and the loadings at a given quantile τ, once the number of factors has been selected. However, since the objective function for QFA is not convex in the relevant parameters, we introduce an iterative QR algorithm to yield estimators of the quantile-dependent objects. We then derive their average rates of convergence, and prove the consistency of the two selection criteria. Finally, we establish asymptotic normality for QFA estimators based on smoothed QR (see e.g., Horowitz 1998 and Galvao and Kato 2016). These results are obtained under the assumption that $\{u_{it}(\tau)\}$ are independent across i and t conditional on $\{f_t(\tau)\}$, though our simulation results indicate that QFA still performs well under some mild time-series and cross-sectional dependence.

The main results from the previous analysis can be summarized as follows: (i) the average convergence rates of the QFA estimators are the same as the corresponding rates of the PCA estimators of Bai and Ng (2002) (BN 2002, hereafter), which is a crucial result for proving the consistency of the two selection criteria; (ii) the QFA estimators based on smoothed QR are shown to converge at the parametric rates ($\sqrt{N}$ and $\sqrt{T}$) to normal distributions, as in Bai (2003); (iii) as a byproduct of our approach (and in exchange for some restrictions on the dependence of the idiosyncratic errors; see Assumption 1 below), the QFA estimators inherit certain robustness properties of QR to the presence of outliers and heavy-tailed distributions, which would render PCA invalid; and (iv) the extraction of all quantile-shifting factors (including those affecting the means of observed variables) through QFA can improve the information traditionally provided by PCA and related methods in several applied contexts, as illustrated by our empirical application on density forecasting with a large macro dataset in Section 6.

Related literature

There is a recent literature that attempts to make the AFM setup more flexible. For example, Su and Wang (2017) allow the factor loadings to be time-varying whereas Pelger and Xiong (2018) allow them to be state-dependent. Chen, Hansen, and Scheinkman (2009) pro-

[2]Throughout the paper we use $\mathsf{Q}_W[\tau|Z]$ to denote the conditional quantile of W given Z.

vide a theory for nonlinear PCA based on sieve estimation to retrieve nonlinear factors. Finally, Gorodnichenko and Ng (2017) propose an algorithm for the joint estimation of level and volatility factors simultaneously. Different from these studies which focus on specific DGPs, our approach here is to remain agnostic about the nature of the true DGP and use the conditional quantiles of the observed data to capture nonlinearities in factor models.

In addition, there is an emerging literature on panel quantile models where a few unobservable factors explain the co-movements of a wide range of financial asset returns. In parallel and independent research, there have been two recent papers related to ours. First, Ma, Linton, and Gao (2020) propose sieve and GLS estimation and inference procedures in semiparametric quantile factor models where factor loadings/betas are smooth functions of a small number of observables, under the assumption that the included factors have non-zero means. We depart from these authors in not requiring the loadings to depend on observables and, foremost, in considering not only loadings but also factors to be quantile-dependent objects. Second, AB (2020) use a similar setup to ours, where the unobservable factor structure is also allowed to be quantile dependent. These authors use Bayesian MCMC and frequentist estimation approaches, the latter building upon our proposed iterative procedure. However, we differ from their approach in several respects, most notably: (i) our assumptions on the idiosyncratic errors are less restrictive since we rely on properties of the density, as in QR, while these authors need several moments to exist, (ii) our proofs of the main results are different and can be easily extended to deal with some nonlinear models with smooth object functions, like the probit and logit factor models considered by Chen, Fernández-Val, and Weidner (2020), and (iii) our novel rank-minimization estimator for the number of factors behaves better in finite samples and is computationally more efficient than the information criteria-based methods these authors propose.

Finally, we refer to another ongoing line of research in asset pricing, coined the "idiosyncratic volatility puzzle" by Ang, Hodrick, Xing, and Zhang (2006). This approach assumes a genuine factor structure in the the idiosyncratic volatility processes of a panel of asset returns, and basically consists of applying PCA (or cross-sectional averages) to the squared residuals, once mean factors have been removed from the original variables (a procedure labeled PCA-SQ hereafter).[3] For example, if the DGP were to be known, PCA-SQ would be optimal for some of our illustrative examples of QFM discussed in subsection 2.2 below. Yet, this procedure will not be able to extract the whole QFM structure in other instances where the QFA approach achieves this goal. Moreover, PCA-SQ will also fail when the distributions of the idiosyncratic errors exhibit heavy tails.

Structure of the Paper

The outline of the paper is as follows: Section 2 defines QFM and provides a list of simple il-

[3]See, e.g., Barigozzi and Hallin (2016), Herskovic, Kelly, Lustig, and Van Nieuwerburgh (2016) and Renault, Van Der Heijden, and Werker (2017).

lustrative examples where this methodology could be applied. In Section 3, we present the QFA estimator and its computational algorithm, establish the average rates of convergence of the quantile-dependent objects, and propose two consistent selection criteria to choose the number of factors at each quantile. Section 4 introduces a kernel-smoothed version of the QFA estimators to derive their asymptotic distributions. Section 5 contains some Monte Carlo simulation results to evaluate the performance of QFA in finite samples with different assumptions about the idiosyncratic error terms. Section 6 provides an empirical application of QFA estimation, regarding a large panel of US macroeconomic variables, where we document the relevance of extra factors to density forecasting. Finally, Section 7 concludes with a summary and some directions for further research. Proofs of the main results are collected in the Appendix, while an Online Appendix contains additional results.

Notation

The Frobenius norm is denoted as $\|\cdot\|$. For a matrix A with real eigenvalues, $\rho_j(A)$ denotes the jth largest eigenvalue. For a real number a, let $\mathsf{sgn}(a) = 1$ if $a \geq 0$ and $\mathsf{sgn}(a) = -1$ if $a < 0$. For a square matrix A whose jth diagonal element is denoted as A_{jj}, define $\mathsf{sgn}(A)$ as a diagonal matrix whose jth diagonal element is equal to $\mathsf{sgn}(A_{jj})$. Following van der Vaart and Wellner (1996), the symbol $\lesssim$ means "left side bounded by a positive constant times the right side" (the symbol $\gtrsim$ is defined similarly), and $D(\cdot, g, \mathcal{G})$ denotes the packing number of space $\mathcal{G}$ endowed with semimetric g.

2 The Model and Some Illustrative Examples

This section starts by introducing the main definitions to be used throughout the paper. Next, we show how to derive the QFM representation of several illustrative DGPs exhibiting different factor structures.

2.1 Quantile Factor Models

Suppose that the observed variable X_{it}, with $i = 1, 2, .., N$ and $t = 1, 2, ..., T$, has the following QFM structure at some $\tau \in (0, 1)$:

$$\mathsf{Q}_{X_{it}}[\tau | f_t(\tau)] = \lambda_i'(\tau) f_t(\tau) \quad \text{almost surely,}$$

where the common factors $f_t(\tau)$ is a $r(\tau) \times 1$ vector of unobservable random variables, and $\lambda_i(\tau)$ is a $r(\tau) \times 1$ vector of non-random factor loadings with $r(\tau) \ll N$. Note that in the QFM defined above, the factors, the loadings, and the number of factors are all allowed to be quantile-dependent.

Alternatively, the above equation implies that

$$X_{it} = \lambda_i'(\tau) f_t(\tau) + u_{it}(\tau), \tag{1}$$

where the quantile-dependent idiosyncratic error $u_{it}(\tau)$ satisfies the quantile restrictions:

$$P[u_{it}(\tau) \leq 0 | f_t(\tau)] = \tau \quad \text{almost surely.}$$

2.2 Examples

We next provide a few illustrative examples of how QFM can be derived from different specifications of location-scale shift models and related ones. The idea behind these simple illustrations is to show instances where, if the true DGPs were to be unknown rather than known, the standard AFM methodology might fail to capture the full factor structure, whereas the alternative QFM approach would succeed in doing so.

Example 1. ***Location-shift model.*** *$X_{it} = \alpha_i f_{1t} + \epsilon_{it}$, where $\{\epsilon_{it}\}$ are zero-mean i.i.d errors independent of $\{f_{1t}\}$ with cumulative distribution function (CDF) F_ϵ. Let $\mathsf{Q}_\epsilon(\tau) = \mathsf{F}_\epsilon^{-1}(\tau) = \inf\{c : \mathsf{F}_\epsilon(c) \leq \tau\}$ be the quantile function of ϵ_{it}, and assume that the median of ϵ_{it} is 0, i.e., $\mathsf{Q}_\epsilon(0.5) = 0$. Then, this simple model has a QFM representation (1) by defining $\lambda_i(\tau) = [\mathsf{Q}_\epsilon(\tau), \alpha_i]'$, $f_t(\tau) = [1, f_{1t}]'$ for $\tau \neq 0.5$, and $\lambda_i(\tau) = \alpha_i$, $f_t(\tau) = f_{1t}$ for $\tau = 0.5$. However, note that the standard estimation method (PCA) for this AFM may not be consistent if the distribution of ϵ_{it} has heavy tails. For example, Assumption C of BN (2002) requires $\mathbb{E}[\epsilon_{it}^8] < \infty$, which is not satisfied if, e.g. ϵ_{it} follows the standard Cauchy or some Pareto distributions.*

Example 2. ***Location-scale-shift model (same sign-restricted factor).*** *$X_{it} = \alpha_i f_{1t} + \eta_i f_{1t}\epsilon_{it}$, where $\eta_i f_{1t} > 0$ for all i, t and $\{\epsilon_{it}\}$ are defined as in Example 1. This model has a QFM representation (1) by defining $\lambda_i(\tau) = \eta_i \mathsf{Q}_\epsilon(\tau) + \alpha_i$ and $f_t(\tau) = f_{1t}$ for all τ, such that the loadings of the factor f_{1t} are the only quantile-dependent objects.*

Example 3. ***Location-scale-shift model (different factors).*** *$X_{it} = \alpha_i' f_{1t} + (\eta_i' f_{2t})\epsilon_{it}$, where $\{\epsilon_{it}\}$ are defined as in Example 1, $\alpha_i, f_{1t} \in \mathbb{R}^{r_1}$, η_i, $f_{2t} \in \mathbb{R}^{r_2}$, and $\eta_i' f_{2t} > 0$. When f_{1t} and f_{2t} do not share common elements, this model has a QFM representation (1) with $\lambda_i(\tau) = [\alpha_i', \eta_i' \mathsf{Q}_\epsilon(\tau)]'$, $f_t(\tau) = [f_{1t}', f_{2t}']$ for $\tau \neq 0.5$, and $\lambda_i(\tau) = \alpha_i$, $f_t(\tau) = f_{1t}$ for $\tau = 0.5$.*

Example 4. ***Location-scale-shift model with an idiosyncratic error and its cube.*** *$X_{it} = \alpha_i f_{1t} + f_{2t}\epsilon_{it} + c_i f_{3t}\epsilon_{it}^3$, where ϵ_{it} is a standard normal random variable whose CDF is denoted as $\Phi(\cdot)$. Let f_{2t}, f_{3t}, c_i be positive, then X_{it} has an equivalent representation in form of (1) with $\lambda_i(\tau) = [\alpha_i, \Phi^{-1}(\tau), c_i\Phi^{-1}(\tau)^3]'$, $f_t(\tau) = (f_{1t}, f_{2t}, f_{3t})'$ for $\tau \neq 0.5$, and $\lambda_i(\tau) = \alpha_i$, $f_t(\tau) = f_{1t}$ for $\tau = 0.5$. In particular, if $c_i = 1$ for all i and noticing that the mapping $\tau \mapsto \Phi^{-1}(\tau)^3$ is strictly increasing, then we have $\mathsf{Q}_{X_{it}}[\tau | f_t(\tau)] = \alpha_i f_{1t} + \Phi^{-1}(\tau) \cdot [f_{2t} + f_{3t}\Phi^{-1}(\tau)^2]$. Hence, there* *exists a QFM representation* (1) *with* $\lambda_i(\tau) = [\alpha_i, \Phi^{-1}(\tau)]'$ *and* $f_t(\tau) = [f_{1t}, f_{2t} + f_{3t}\Phi^{-1}(\tau)^2]'$, *where note that the second factor in* $f_t(\tau)$, $f_{2t}+f_{3t}\Phi^{-1}(\tau)^2$ *is quantile dependent even for* $\tau \neq 0.5$.

Not surprisingly, if the researcher was unaware that the data had been generated according to the above DGPs, the standard PCA methodology would only work in Example 1, insofar as the idiosyncratic errors satisfy certain moment conditions. In the remaining examples, PCA can only consistently estimate the factors shifting the locations, failing to capture those extra factors which shift quantiles other than the means, or their corresponding quantile-varying loadings. In the sequel, QFA is hence proposed as a novel estimation procedure capable of estimating both sets of quantile-dependent objects in QFM.

3 Estimators and their Asymptotic Properties

Consider a sample of observations $\{X_{it}\}$ generated by (1) for $i = 1, \ldots, N$, and $t = 1, \ldots, T$, where the realized values of $\{f_t(\tau)\}$ are $\{f_{0t}(\tau)\}$ and the true values of $\{\lambda_i(\tau)\}$ are $\{\lambda_{0i}(\tau)\}$. We take a fixed-effects approach by treating $\{\lambda_{0i}(\tau)\}$ and $\{f_{0t}(\tau)\}$ as parameters to be estimated, and our asymptotic analysis is conditional on $\{f_{0t}(\tau)\}$. In Section 3.1, we start by analyzing the estimation of $\{\lambda_{0i}(\tau)\}$ and $\{f_{0t}(\tau)\}$ where $r(\tau)$ is assumed to be known, while Section 3.2 deals with the estimation of $r(\tau)$ for each quantile. To simplify the notation, we suppress the dependence of $f_{0t}(\tau), \lambda_{0i}(\tau), r(\tau)$ and $u_{it}(\tau)$ on τ in the rest of the paper.

3.1 Estimating Factors and Loadings

A well-known result in the literature on factor models is that $\{\lambda_{0i}\}$ and $\{f_{0t}\}$ cannot be separately identified without imposing normalizations (see BN 2002). Without loss of generality, we choose the following normalizations in QFM:

$$\frac{1}{T}\sum_{t=1}^{T} f_t f_t' = \mathbb{I}_r, \quad \frac{1}{N}\sum_{i=1}^{N} \lambda_i \lambda_i' \text{ is diagonal with non-increasing diagonal elements.} \tag{2}$$

Let $M = (N+T)r$, $\theta = (\lambda_1', \ldots, \lambda_N', f_1', \ldots, f_T')'$, and $\theta_0 = (\lambda_{01}', \ldots, \lambda_{0N}', f_{01}', \ldots, f_{0T}')'$ denotes the vector of true parameters, where the dependence of θ and θ_0 on M is also suppressed to save notation. Let $\mathcal{A}, \mathcal{F} \subset \mathbb{R}^r$ and define:

$$\Theta^r = \left\{\theta \in \mathbb{R}^M : \lambda_i \in \mathcal{A}, f_t \in \mathcal{F} \text{ for all } i, t, \{\lambda_i\} \text{ and } \{f_t\} \text{ satisfy the normalizations in (2)}\right\}.$$

Further, define:

$$\mathbb{M}_{NT}(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \rho_\tau(X_{it} - \lambda_i' f_t),$$

where $\rho_\tau(u) = (\tau - \mathbf{1}\{u \leq 0\})u$ is the check function. The QFA estimator of θ_0 is defined as:

$$\hat{\theta} = (\hat{\lambda}'_1, \ldots, \hat{\lambda}'_N, \hat{f}'_1, \ldots, \hat{f}'_T)' = \arg\min_{\theta \in \Theta^r} \mathbb{M}_{NT}(\theta).$$

It can be easily seen that the way in which our estimator is related to the PCA estimator studied by BN (2002) and Bai (2003) is analogous to how QR is related to standard least-squares regressions. However, unlike Bai (2003)'s PCA estimator, our estimator $\hat{\theta}$ does not yield an analytical closed form. This makes it difficult not only to find a computational algorithm that would yield the estimator, but also to derive its asymptotic properties. In the sequel, we introduce a computational algorithm called *iterative quantile regression* (IQR) that can effectively locate the stationary points of the object function. In parallel, Theorem 1 shows that $\hat{\theta}$ achieves the same convergence rate as the PCA estimators for AFM.

To describe this algorithm, let $\Lambda = (\lambda_1, \ldots, \lambda_N)'$, $F = (f_1, \ldots, f_T)'$, and define the object function in terms of the following averages:

$$\mathbb{M}_{i,T}(\lambda, F) = \frac{1}{T}\sum_{t=1}^{T} \rho_\tau(X_{it} - \lambda' f_t) \quad \text{and} \quad \mathbb{M}_{t,N}(\Lambda, f) = \frac{1}{N}\sum_{i=1}^{N} \rho_\tau(X_{it} - \lambda'_i f).$$

Note that $\mathbb{M}_{NT}(\theta) = N^{-1}\sum_{i=1}^{N} \mathbb{M}_{i,T}(\lambda_i, F) = T^{-1}\sum_{t=1}^{T} \mathbb{M}_{t,N}(\Lambda, f_t)$, and the main difficulty in finding the global minimum of $\mathbb{M}_{NT}$ is that this object function is not convex in θ. However, for given F, $\mathbb{M}_{i,T}(\lambda, F)$ happens to be convex in λ for each i and likewise, for given Λ, $\mathbb{M}_{t,N}(\Lambda, f)$ is convex in f for each t. Thus, both optimization problems can be efficiently solved by several linear programming methods (see Chapter 6 of Koenker 2005). Based on this observation, we propose the following iterative procedure:

Iterative quantile regression (IQR):

Step 1: Choose random starting parameters: $F^{(0)}$.

Step 2: Given $F^{(l-1)}$, solve $\lambda_i^{(l-1)} = \arg\min_\lambda \mathbb{M}_{i,T}(\lambda, F^{(l-1)})$ for $i = 1, \ldots, N$; given $\Lambda^{(l-1)}$, solve $f_t^{(l)} = \arg\min_f \mathbb{M}_{t,N}(\Lambda^{(l-1)}, f)$ for $t = 1, \ldots, T$.

Step 3: For $l = 1, \ldots, L$, iterate the second step until $\mathbb{M}_{NT}(\theta^{(L)})$ is close to $\mathbb{M}_{NT}(\theta^{(L-1)})$, where $\theta^{(l)} = (\text{vech}(\Lambda^{(l)})', \text{vech}(F^{(l)})')'$.

Step 4: Normalize $\Lambda^{(L)}$ and $F^{(L)}$ so that they satisfy the normalizations in (2).

To examine the connection between the IQR algorithm and the PCA estimator of Bai (2003), suppose that $r = 1$, and replace the check function in the IQR algorithm by the quadratic loss function. Then, it is easy to show that the second step of the algorithm above yields $\Lambda^{(l-1)} = (X'F^{(l-1)})/\|F^{(l-1)}\|^2$ and $F^{(l)} = (X\Lambda^{(l-1)})/\|\Lambda^{(l-1)}\|^2 = XX'F^{(l-1)}/C_{l-1}$, where X is the $T \times N$ matrix with elements $\{X_{it}\}$, and $C_l = \|F^{(l)}\|^2 \cdot \|\Lambda^{(l)}\|^2$. Thus, with proper normalizations at each step, the iterative procedure is equivalent to the well-known *power method* of Hotelling (1933), and the sequence $F^{(0)}, F^{(1)}, \ldots$ will converge to the eigenvector associated with the largest eigenvalue of XX'. In the more general case where $r > 1$, if we replace the check function in the IQR algorithm by the quadratic loss function and normalize $F^{(l-1)}, \Lambda^{(l-1)}$ to satisfy (2) at step 2, it can be shown that the above iterative procedure is similar to the method of *orthogonal iteration* (see Section 7.3.2 of Golub and Van Loan 2013) for calculating the eigenvectors associated with the r largest eigenvalues of XX', which is the PCA estimator of Bai (2003). Therefore, the IQR algorithm and its corresponding QFA estimator can be viewed as extensions of PCA to QFM.

Similar algorithms have been proposed in the machine learning literature to reduce the dimensions of binary data, where the check function is replaced by some smooth nonlinear link functions (see, e.g. Collins, Dasgupta, and Schapire 2001). However, unlike PCA, whether such methods guarantee finding the global minimum remains an important open question which is hard to address. Nonetheless, in all of our Monte Carlo simulations we found that the QFA estimators of the factors using the IQR algorithm always converge to the space of the true factors, which is somewhat reassuring in this respect.

To prove the consistency of the QFA estimator $\hat{\theta}$, the following assumptions are made:

Assumption 1. *(i) $\mathcal{A}$ and $\mathcal{F}$ are compact sets and $\theta_0 \in \Theta^r$. In particular, $N^{-1}\sum_{i=1}^{N}\lambda_{0i}\lambda_{0i}' = diag(\sigma_{N1}, \ldots, \sigma_{Nr})$ with $\sigma_{N1} \geq \sigma_{N2} \cdots \geq \sigma_{Nr}$, and $\sigma_{Nj} \to \sigma_j$ as $N \to \infty$ for $j = 1, \ldots, r$ with $\infty > \sigma_1 > \sigma_2 \cdots > \sigma_r > 0$.*
(ii) The conditional density function of u_{it} given $\{f_{0t}\}$, denoted as f_{it}, is continuous, and satisfies that: for any compact set $C \subset \mathbb{R}$ and any $u \in C$, there exists a positive constant $\underline{\mathsf{f}} > 0$ (depending on C) such that $\mathsf{f}_{it}(u) \geq \underline{\mathsf{f}}$ for all i, t.
(iii) Given $\{f_{0t}, 1 \leq t \leq T\}$, $\{u_{it}, 1 \leq i \leq N, 1 \leq t \leq T\}$ are independent across i and t.

Assumptions 1(i) is essentially the *strong factors* assumption that is standard in the literature (see assumption B of Bai 2003). The requirement that $\sigma_1, \ldots, \sigma_r$ are distinct is similar to Assumption G of Bai (2003), which is a convenient assumption to order the factors. Assumptions 1(ii) and (iii) resemble assumptions (C1) and (C2) in AB (2020), except that we do not require moments of u_{it} to exist. Also notice that assumption (iii), which allows for both cross-sectional and time series heteroskedasticity, requires the idiosyncratic errors to be mutually independent. This strong assumption stems from the use of Hoeffding's inequality in the proofs of some results, which provides a sub-Gaussian tail bound for the sum of bounded independent random variables. There have been attempts to relax this assumption (see Remark 1.4 below) but it is difficult to characterize the minimal set of conditions that the error terms should satisfy to achieve the sub-Gaussian inequality required in our proofs. Notice, however, that in exchange for the independence assumption, we can dispense with the bounded moment conditions in the idiosyncratic terms, whose violation renders PCA invalid. At any rate, in subsection 5.2 we run some Monte Carlo simulations on the performance of our QFA estimation when error terms are allowed to exhibit mild cross-sectional and serial dependence to check how robust are our results to these features.

Write $\hat{\Lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_N)'$, $\Lambda_0 = (\lambda_{01}, \ldots, \lambda_{0N})'$, $\hat{F} = (\hat{f}_1, \ldots, \hat{f}_T)'$, $F_0 = (f_{01}, \ldots, f_{0T})'$, and let $L_{NT} = \min\{\sqrt{N}, \sqrt{T}\}$. The following theorem provides the average rate of convergence of $\hat{\Lambda}$ and $\hat{F}$.

Theorem 1. *Let* $\hat{\mathsf{S}} = \mathsf{sgn}(\hat{F}'F_0/T)$. *Then under Assumption 1,*

$$\|\hat{\Lambda} - \Lambda_0\hat{\mathsf{S}}\|/\sqrt{N} = O_P(L_{NT}^{-1}) \quad \textit{and} \quad \|\hat{F} - F_0\hat{\mathsf{S}}\|/\sqrt{T} = O_P(L_{NT}^{-1}).$$

Proof. See Appendix A.1. □

Note that the sign matrix $\hat{\mathsf{S}}$ appears above due to the intrinsic sign indeterminacy of factors and loadings — that is, the factor structure remains unchanged if a factor and its loading are both multiplied by -1 (see e.g. Theorem 1.b of Stock and Watson 2002 for a similar result). In particular, it can be shown that $\|\hat{F}'F_0/T - \hat{\mathsf{S}}\| = O_P(L_{NT}^{-1})$.

Remark 1.1: Since our proof strategy is substantially different from that of BN (2002), we briefly sketch here its main underlying ideas. To facilitate the discussion, for any $\theta_a, \theta_b \in \Theta^r$ define the semimetric d by:

$$d(\theta_a, \theta_b) = \sqrt{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\lambda'_{ai}f_{at} - \lambda'_{bi}f_{bt})^2} = \frac{1}{\sqrt{NT}}\left\|\Lambda_a F'_a - \Lambda_b F'_b\right\|,$$

and let $\omega_{it}(\lambda_i, f_t) = \rho_\tau(X_{it} - \lambda'_i f_t) - \rho_\tau(X_{it} - \lambda'_{0i}f_{0t})$,

$$\mathbb{M}^*_{NT}(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\omega_{it}(\lambda_i, f_t), \quad \bar{\mathbb{M}}^*_{NT}(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}[\omega_{it}(\lambda_i, f_t)].$$

The semimetric d plays an important role in our asymptotic analysis. We first show that $d(\hat{\theta}, \theta_0) = o_P(1)$. Next, it can be shown that:

$$\bar{\mathbb{M}}^*_{NT}(\hat{\theta}) \gtrsim d^2(\hat{\theta}, \theta_0), \tag{3}$$

and that, for sufficiently small $\delta > 0$,

$$\mathbb{E}\left[\sup_{\theta\in\Theta^r(\delta)}\left|\mathbb{M}^*_{NT}(\theta) - \bar{\mathbb{M}}^*_{NT}(\theta)\right|\right] \lesssim \frac{\delta}{L_{NT}}, \tag{4}$$

where $\Theta^r(\delta) = \{\theta \in \Theta^r : d(\theta, \theta_0) \leq \delta\}$. By the definition of $\hat{\theta}$, we have $\mathbb{M}^*_{NT}(\hat{\theta}) \leq 0$, which implies $\bar{\mathbb{M}}^*_{NT}(\hat{\theta}) \leq -[\mathbb{M}^*_{NT}(\hat{\theta}) - \bar{\mathbb{M}}^*_{NT}(\hat{\theta})]$. It then follows from (3) and (4) that $d(\hat{\theta}, \theta_0) = O_P(L_{NT}^{-1})$ (see Theorem 3.2.5 of van der Vaart and Wellner 1996). Finally, the desired results follow from the fact that $\|\hat{\Lambda} - \Lambda_0\hat{\mathsf{S}}\|/\sqrt{N} + \|\hat{F} - F_0\hat{\mathsf{S}}\|/\sqrt{T} \lesssim d(\hat{\theta}, \theta_0)$.

Inequality (3) follows easily from a Taylor expansion of $\bar{\mathbb{M}}^*_{NT}(\hat{\theta})$ around θ_0, together with Assumption 1(ii). It is worth stressing that the proof of (4) requires the chaining argument which is commonly used in the theory of empirical processes. In particular, using Hoeffding's inequality and the fact that $|\rho_\tau(u) - \rho_\tau(v)| \leq 2|u - v|$, it can be shown that, for any given $\theta_a, \theta_b \in \Theta^r$,

$$P\left[\sqrt{NT}\left|\mathbb{M}^*_{NT}(\theta_a) - \bar{\mathbb{M}}^*_{NT}(\theta_a) - \mathbb{M}^*_{NT}(\theta_b) + \bar{\mathbb{M}}^*_{NT}(\theta_b)\right| \geq c\right] \leq e^{-\frac{c^2}{Kd^2(\theta_a,\theta_b)}} \tag{5}$$

for some constant K. Then, along the lines of Theorem 2.2.4 of van der Vaart and Wellner (1996), it holds that the left-hand side of (4) is bounded (up to a positive constant) by $\int_0^\delta \sqrt{\log D(\epsilon, d, \Theta^r(\delta))}d\epsilon/\sqrt{NT}$. Finally, we can prove that $\int_0^\delta \sqrt{\log D(\epsilon, d, \Theta^r(\delta))}d\epsilon \lesssim \delta\sqrt{M}$, from which inequality (4) follows.

Remark 1.2: Compared to BN (2002), recall that, in exchange for Assumption 1(iii) we do not require any moment of u_{it} to be finite. Thus, for the canonical AFM (e.g., Example 1) when the idiosyncratic errors have median equal to zero and satisfy Assumption 1(iii), the QFA estimator for $\tau = 0.5$ can be interpreted as a least absolute deviation (LAD) estimator which is robust to heavy tails and outliers. In relation to this issue, it is important to point out that the LAD estimator is related to robust PCA in the machine learning literature that aims to recover a low rank matrix from a large panel of observables. For example, the *Principal Components Pursuit* method proposed by Candès, Li, Ma, and Wright (2011) features a combination of the L_1 norm (as in LAD) and a nuclear norm on the low rank matrix (see Chapter 3 of Vidal, Ma, and Sastry 2016 and Bai and Ng 2019 for other robust PCA methods). In Section 5 below, we will illustrate the robustness properties of our QFA estimator at $\tau = 0.5$ by Monte Carlo simulations.

Remark 1.3: As in Theorem 1 of BN (2002), if the true factors and factor loadings are not assumed to satisfy the normalization (2), it can still be proven that the QFA estimator is consistent for the true factors and factor loadings up to a rotation matrix. In particular, define $\Sigma_{T,F} = T^{-1}\sum_{t=1}^T f_{0t}f'_{0t}$, $\Sigma_{N,\Lambda} = N^{-1}\sum_{i=1}^N \lambda_{0i}\lambda'_{0i}$, and $H_{NT} = \Sigma_{T,F}^{-1/2}\Gamma_{NT}$, where Γ_{NT} is the matrix of eigenvectors of $\Sigma_{T,F}^{1/2}\Sigma_{N,\Lambda}\Sigma_{T,F}^{1/2}$. Then it can be shown that

$$\|\hat{\Lambda} - \Lambda_0(H'_{NT})^{-1}\hat{\mathsf{S}}\|/\sqrt{N} = O_P(L_{NT}^{-1}) \quad \text{and} \quad \|\hat{F} - F_0H_{NT}\hat{\mathsf{S}}\|/\sqrt{T} = O_P(L_{NT}^{-1}).$$

The proof of this result is identical to that of Theorem 1, since it is easy to see that $\Lambda_0(H'_{NT})^{-1}$ and F_0H_{NT} satisfy the normalization (2). Notice, however, that the rotation matrix H_{NT} is slightly different from the rotation matrix of BN (2002) and Bai (2003) because the QFA estimator is implicitly defined as the minimizer of the check function, while the PCA estimator is explicitly defined through the eigenequation. Moreover, since both λ_{0i} and f_{0t} are τ-dependent, H_{NT} also varies across quantiles, though we did not explicitly make this matrix quantile dependent in the previous discussion to simplify notation.

Remark 1.4: Compared to BN (2002), our Assumption 1(iii) is admittedly strong. However, note that this assumption is made conditional on $\{f_{0t}\}$, so cross-sectional and temporal dependence of u_{it} due to the common factors are still allowed for. Moreover, the independence assumption is only used to establish the sub-Gaussian inequality (5). Thus, Assumption 1(iii) could be relaxed as long as the sub-Gaussian inequality holds.[4]

3.2 Selecting the Number of Factors

While in the previous subsection the number of quantile-dependent factors $r(\tau)$ was assumed to be known at each τ, we now propose two different methods to select the correct number of factors at each quantile with probability approaching one. The first procedure selects the number of factors by rank minimization while the second one uses information criteria (IC). As before, the dependence of the quantile-dependent objects on τ, including $r(\tau)$, is suppressed for simplicity.

3.2.1 Model Selection by Rank Minimization

Let k be a positive integer larger than r, and $\mathcal{A}^k$ and $\mathcal{F}^k$ be compact subsets of $\mathbb{R}^k$. In particular, let us assume that $[\lambda_{0i}' \quad \mathbf{0}_{1\times(k-r)}]' \in \mathcal{A}^k$ for all i.

Let $\lambda_i^k, f_t^k \in \mathbb{R}^k$ for all i, t and write $\theta^k = (\lambda_1^{k\prime}, \ldots, \lambda_N^{k\prime}, f_1^{k\prime}, \ldots, f_T^{k\prime})'$, $\Lambda^k = (\lambda_1^k, \ldots, \lambda_N^k)'$, $F^k = (f_1^k, \ldots, f_T^k)'$. Consider the following normalizations:

$$\frac{1}{T}\sum_{t=1}^{T} f_t^k f_t^{k\prime} = \mathbb{I}_k, \quad \frac{1}{N}\sum_{i=1}^{N} \lambda_i^k \lambda_i^{k\prime} \text{ is diagonal with non-increasing diagonal elements.} \tag{6}$$

Define $\Theta^k = \{\theta^k : \lambda_i^k \in \mathcal{A}^k, f_t^k \in \mathcal{F}^k, \text{ and } \lambda_i^k, f_t^k \text{ satisfy (6)}\}$, and

$$\hat{\theta}^k = (\hat{\lambda}_1^{k\prime}, \ldots, \hat{\lambda}_N^{k\prime}, \hat{f}_1^{k\prime}, \ldots, \hat{f}_T^{k\prime})' = \underset{\theta^k \in \Theta^k}{\arg\min} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \rho_\tau(X_{it} - \lambda_i^{k\prime} f_t^k).$$

Moreover, define $\hat{\Lambda}^k = (\hat{\lambda}_1^k, \ldots, \hat{\lambda}_N^k)'$ and write

$$(\hat{\Lambda}^k)'\hat{\Lambda}^k/N = \operatorname{diag}\left(\hat{\sigma}_{N,1}^k, \ldots, \hat{\sigma}_{N,k}^k\right).$$

[4] See van de Geer (2002) for the properties of Hoeffding inequalities for martingales.

The first estimator of the number of factors r is defined as:

$$\hat{r}_{\text{rank}} = \sum_{j=1}^{k} \mathbf{1}\{\hat{\sigma}_{N,j}^{k} > P_{NT}\},$$

where P_{NT} is a sequence that goes to 0 as $N, T \to \infty$. In other words, $\hat{r}_{\text{rank}}$ is equal to the number of diagonal elements of $(\hat{\Lambda}^k)'\hat{\Lambda}^k/N$ that are larger than the threshold P_{NT}. We call $\hat{r}_{\text{rank}}$ the *rank-minimization estimator* because, as discussed in Remark 2.1 below, it can be interpreted as a rank estimator of $(\hat{\Lambda}^k)'\hat{\Lambda}^k/N$.

It can then be shown that:

Theorem 2. *Under Assumption 1, $P[\hat{r}_{rank} = r] \to 1$ as $N, T \to \infty$ if $k > r$, $P_{NT} \to 0$ and $P_{NT}L_{NT}^2 \to \infty$.*

Proof. See Appendix A.2. □

Remark 2.1: In the proof of Theorem 2, we show that, for $k > r$, it holds that (up to sign)

$$\left\|\hat{F}^{k,r} - F_0\right\| / \sqrt{T} = O_P(L_{NT}^{-1}) \quad \text{and} \quad \left\|\hat{\Lambda}^k - \Lambda_0^*\right\| / \sqrt{N} = O_P(L_{NT}^{-1}),$$

where $\hat{F}^{k,r}$ is the first r columns of $\hat{F}^k$ and $\Lambda_0^* = [\Lambda_0, \mathbf{0}_{N\times(k-r)}]$. It then follows from Assumption 1 that $\hat{\sigma}_{N,j}^k \xrightarrow{p} \sigma_j > 0$ for $j = 1, \ldots, r$ and $\hat{\sigma}_{N,j}^k = N^{-1}\sum_{i=1}^{N}(\hat{\lambda}_{i,j}^k)^2 = O_P(1/L_{NT}^2)$ for $j = r+1, \ldots, k$. Thus, the first r diagonal components of $(\hat{\Lambda}^k)'\hat{\Lambda}^k/N$ converge in probability to positive constants while the remaining diagonal components are all $O_P(1/L_{NT}^2)$. In other words, $(\hat{\Lambda}^k)'\hat{\Lambda}^k/N$ converges in probability to a matrix with rank r, and P_{NT} can be viewed as a cutoff value to choose the asymptotic rank of $(\hat{\Lambda}^k)'\hat{\Lambda}^k/N$.

3.2.2 Model Selection by Information Criteria

The second estimator of r is similar to the IC-based estimator of BN (2002). Let l denote a positive integer smaller or equal to k, and $\mathcal{A}^l$ and $\mathcal{F}^l$ be compact subsets of $\mathbb{R}^l$. In particular, for $l > r$, assume that $[\lambda_{0i}' \quad \mathbf{0}_{1\times(l-r)}]' \in \mathcal{A}^l$ for all i. Moreover, we can define $\Theta^l, \hat{\theta}^l, \hat{f}_t^l, \hat{\lambda}_i^l, \hat{F}^l$ and $\hat{\Lambda}^l$ in a similar fashion.

Define the IC-based estimator of r as follows:

$$\hat{r}_{\text{IC}} = \underset{1\le l\le k}{\arg\min} \left[\mathbb{M}_{NT}(\hat{\theta}^l) + l \cdot P_{NT}\right].$$

We can show that:

Theorem 3. *Suppose Assumption 1 holds, and assume that for any compact set $C \subset \mathbb{R}$ and any $u \in C$, there exists $\bar{\mathsf{f}} > 0$ (depending on C) such that $\mathsf{f}_{it}(u) \leq \bar{\mathsf{f}}$ for all i, t. Then $P[\hat{r}_{IC} = r] \to 1$ as $N, T \to \infty$ if $k > r$, $P_{NT} \to 0$ and $P_{NT} L_{NT}^2 \to \infty$.*

Proof. See Appendix A.3. □

Remark 3.1: AB (2020) obtain a similar result, but the difference with ours is that we only need the density function of the idiosyncratic errors to be uniformly bounded above and below, while AB (2020) requires all the moments of the errors to be bounded. The reason why we can obtain the same result here with less restrictions is that our proof is based on the novel argument discussed in Remark 1.1 and on the average convergence rate of the estimators, while the proof of AB (2020) depends on the uniform convergence rate of the estimators.

Remark 3.2: Let X denote the $T \times N$ matrix of observed variables, and let $\check{F}^l, \check{\Lambda}^l$ denote the matrices of PCA estimators of BN (2002) when the number of factors is specified as l. Then BN (2002)'s estimator of r can be written as:

$$\hat{r} = \arg\min_{1 \leq l \leq k} \left[(NT)^{-1} \| X - \check{F}^l \check{\Lambda}^{l\prime} \|^2 + l \cdot P_{NT} \right],$$

where $k > r$ and P_{NT} is defined as in Theorem 2 above. It can be shown that IC-based estimator $\hat{r}$ is equivalent to the number of diagonal elements in $\check{\Lambda}^{k\prime} \check{\Lambda}^k / N$ that are larger than P_{NT}. Thus, the two seemly different estimators of the number of factors are equivalent in AFM. However, due to the differences of the object functions, such equivalence does not hold any longer in QFM.

Remark 3.3: The choice of P_{NT} for $\hat{r}_{\text{rank}}$ and $\hat{r}_{\text{IC}}$ can be different in practice. In particular, it can differ from the penalties used by BN (2002). For example, AB (2020) choose

$$P_{NT} = \log\left(\frac{NT}{N+T}\right) \cdot \frac{N+T}{NT}$$

for $\hat{r}_{\text{IC}}$, similar to IC_{p1} of BN (2002). However, as shown in AB's (2020) simulation results, this choice does not perform too well, even for N, T as large as 300.

Remark 3.4: Even though both $\hat{r}_{\text{rank}}$ and $\hat{r}_{\text{IC}}$ yield consistent estimators of r, the computational burden of $\hat{r}_{\text{rank}}$ is much lower than that of $\hat{r}_{\text{IC}}$, because for $\hat{r}_{\text{rank}}$ we only estimate the model once, while for $\hat{r}_{\text{IC}}$ the model needs to be estimated k times. Thus, in the simulations and empirical applications below we will focus on $\hat{r}_{\text{rank}}$, while we refer to AB (2020) for the corresponding simulation results of $\hat{r}_{\text{IC}}$. In particular, we find that the choice

$$P_{NT} = \hat{\sigma}_{N,1}^k \cdot \left(L_{NT}^2\right)^{-1/3} \tag{7}$$

for $\hat{r}_{\text{rank}}$ works fairly well as long as $\min\{N, T\}$ is 100. As a result, we use this value in all of our simulations and applications below.

4 Estimators Based on Smoothed Quantile Regressions

The non-smoothness of the check function and the incidental-parameters problem make it difficult to derive the asymptotic distribution of the QFA estimator $\hat{\theta}$. As in the asymptotic analysis of conventional QR, one way to overcome these difficulties is to expand the expected score function (which is smooth and continuously differentiable) and obtain a stochastic expansion for $\hat{\lambda}_i - \mathsf{S}\lambda_{0i}$; yet, the following term appears in the expansion which may be non-negligible:

$$\frac{1}{T}\sum_{t=1}^{T}\left\{\left(\mathbf{1}\{X_{it} \le \hat{\lambda}_i'\hat{f}_t\} - \mathbb{E}[\mathbf{1}\{X_{it} \le \hat{\lambda}_i'\hat{f}_t\}]\right)\hat{f}_t - \left(\mathbf{1}\{X_{it} \le \lambda_{0i}'f_{0t}\} - \tau\right)f_{0t}\right\}. \tag{8}$$

Consequently, the next step is to show that (8) is a higher-order term (i.e. $o_P(T^{-1/2})$) which does not affect the asymptotic distribution of $\hat{\lambda}_i$. However, due to the presence of the indicator functions in (8), this is not straightforward. To see this, let us consider a similar problem for the PCA estimators of AFM. Let $\check{\lambda}_i$ and $\check{f}_t$ be the PCA estimators. In the stochastic expansion of $\check{\lambda}_i - \lambda_{0i}$, the analogous term to (8) happens to be: $T^{-1}\sum_{t=1}^{T}\epsilon_{it}(\check{f}_t - f_{0t})$, where ϵ_{it} is the idiosyncratic error in the AFM. Note that, based on the result $T^{-1}\sum_{t=1}^{T}\|\check{f}_t - f_{0t}\|^2 = O_P(L_{NT}^{-2})$, one can only show that:

$$\left\|\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}(\check{f}_t - f_{0t})\right\| \le \sqrt{\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}^2}\cdot\sqrt{\frac{1}{T}\sum_{t=1}^{T}\|\check{f}_t - f_{0t}\|^2} = O_P(L_{NT}^{-1}).$$

Hence, one has to use instead the stochastic expansion of $\check{f}_t - f_{0t}$ to show that $T^{-1}\sum_{t=1}^{T}\epsilon_{it}(\check{f}_t - f_{0t}) = O_P(L_{NT}^{-2})$ (see the proof of Lemma B.1 of Bai 2003). Likewise, to prove that (8) is $o_P(T^{-1/2})$, establishing the convergence rate of $\hat{f}_t - \hat{\mathsf{S}}f_{0t}$ is not enough, and the stochastic expansion of $\hat{f}_t - \hat{\mathsf{S}}f_{0t}$ is required. However, due the non-smoothness of the indicator functions, it is not immediate how to explore this stochastic expansion in (8).

To overcome this difficulty, we proceed by defining a new estimator of θ_0, denoted as $\tilde{\theta}$, which relies on the following smoothed quantile regressions (SQR):

$$\tilde{\theta} = (\tilde{\lambda}_1', \ldots, \tilde{\lambda}_N', \tilde{f}_1', \ldots, \tilde{f}_T')' = \underset{\theta\in\Theta^r}{\arg\min}\, \mathbb{S}_{NT}(\theta),$$

where

$$\mathbb{S}_{NT}(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[\tau - K\left(\frac{X_{it} - \lambda_i'f_t}{h}\right)\right](X_{it} - \lambda_i'f_t),$$

such that $K(z) = 1 - \int_{-1}^{z} k(z)dz$, $k(z)$ is a continuous kernel function with support $[-1,1]$, and h is a bandwidth parameter that goes to 0 as N, T grow.

Define

$$\Phi_i = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathsf{f}_{it}(0) f_{0t} f_{0t}' \quad \text{and} \quad \Psi_t = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathsf{f}_{it}(0)\lambda_{0i}\lambda_{0i}'$$

for all i, t. We impose the following assumptions:

Assumption 2. *Let $m \geq 8$ be a positive integer,*
(i) $\Phi_i > 0$ and $\Psi_t > 0$ for all i, t.
(ii) λ_{0i} is an interior point of $\mathcal{A}$ and f_{0t} is an interior point of $\mathcal{F}$ for all i, t.
(iii) $k(z)$ is symmetric around 0 and twice continuously differentiable. $\int_{-1}^{1} k(z)dz = 1$, $\int_{-1}^{1} z^j k(z)dz = 0$ for $j = 1, \ldots, m-1$ and $\int_{-1}^{1} z^m k(z)dz \neq 0$.
(iv) f_{it} is $m+2$ times continuously differentiable. Let $\mathsf{f}_{it}^{(j)}(u) = (\partial/\partial u)^j \mathsf{f}_{it}(u)$ for $j = 1, \ldots, m+2$. For any compact set $C \subset \mathbb{R}$ and any $u \in C$, there exists $-\infty < \underline{l} < \bar{l} < \infty$ (depending on C) such that $\underline{l} \leq \mathsf{f}_{it}^{(j)}(u) \leq \bar{l}$ and $\underline{\mathsf{f}} \leq \mathsf{f}_{it}(u) \leq \bar{l}$ for $j = 1, \ldots, m+2$ and for all i, t.
(v) As $N, T \to \infty$, $N \propto T$, $h \propto T^{-c}$ and $m^{-1} < c < 1/6$.

The above conditions are standard in SQR, with the exception of (v). Note that, as in Galvao and Kato (2016), we require $k(z)$ to be a higher-order kernel function to control the higher-order terms in the stochastic expansions of the estimators. However, Galvao and Kato (2016) assume that $m^{-1} < c < 1/3$ (or $m \geq 4$), while we need $m^{-1} < c < 1/6$ (or $m \geq 8$). This difference arises from the fact that the incidental parameters (λ_{0i} and f_{0t}) in QFM enter the model interactively, while no interactive fixed-effects appear in the panel quantile models considered by these authors.

Then, it is shown that the following result holds:

Theorem 4. *Let $\tilde{\mathsf{S}} = \mathsf{sgn}(\tilde{F}'F_0/T)$. Then under Assumptions 1 and 2,*

$$\sqrt{T}(\tilde{\lambda}_i - \tilde{\mathsf{S}}\lambda_{0i}) \xrightarrow{d} \mathcal{N}(0, \tau(1-\tau)\Phi_i^{-2}) \quad \textit{and} \quad \sqrt{N}(\tilde{f}_t - \tilde{\mathsf{S}} f_{0t}) \xrightarrow{d} \mathcal{N}(0, \tau(1-\tau)\Psi_t^{-1}\Sigma_\Lambda\Psi_t^{-1})$$

for each i and t, where $\Sigma_\Lambda = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$.

Proof. See the Online Appendix. □

Remark 4.1: Similar to the proof of Theorem 1, it holds that

$$\|\tilde{\Lambda} - \Lambda_0\tilde{\mathsf{S}}\|/\sqrt{N} = O_P(L_{NT}^{-1}) + O_P(h^{m/2}) \quad \text{and} \quad \|\tilde{F} - F_0\tilde{\mathsf{S}}\|/\sqrt{T} = O_P(L_{NT}^{-1}) + O_P(h^{m/2}),$$

where the extra $O_P(h^{m/2})$ term is due to the approximation bias of the smoothed check function. However, Assumption 2(v) implies that $1/L_{NT} \gg h^{m/2}$, and then it follows that average convergence rates of $\tilde{\Lambda}$ and $\tilde{F}$ are both L_{NT}.

Remark 4.2: Similar to Theorems 1 and 2 of Bai (2003), we show that the new estimator is free of incidental-parameter biases. This implies that the asymptotic distribution of $\tilde{\lambda}_i$ is the same as if $\{f_{0t}\}$ were observed, and likewise the asymptotic distribution of $\tilde{f}_t$ is the same as if $\{\lambda_{0i}\}$ were observed. The proof of this result is not trivial. To see why, first define $\varrho(u) = [\tau - K(u/h)]u$ and $\mathbb{S}_{i,T}(\lambda, F) = T^{-1}\sum_{t=1}^{T}\varrho(X_{it} - \lambda' f_t)$, then we can write $\tilde{\lambda}_i = \arg\min_{\lambda\in\mathcal{A}}\mathbb{S}_{i,T}(\lambda, \tilde{F})$. Expanding $\partial\mathbb{S}_{i,T}(\tilde{\lambda}_i, \tilde{F})/\partial\lambda$ around $(\tilde{\mathsf{S}}\lambda_{0i}, F_0\tilde{\mathsf{S}})$ yields

$$\left(\frac{1}{T}\sum_{t=1}^{T}\varrho^{(2)}(u_{it})f_{0t}f_{0t}'\right)(\tilde{\lambda}_i - \tilde{\mathsf{S}}\lambda_{0i}) \approx \frac{1}{T}\sum_{t=1}^{T}\varrho^{(1)}(u_{it})\tilde{\mathsf{S}}f_{0t} + \frac{1}{T}\sum_{t=1}^{T}\varrho^{(1)}(u_{it})(\tilde{f}_t - \tilde{\mathsf{S}}f_{0t}) - \frac{1}{T}\sum_{t=1}^{T}\varrho^{(2)}(u_{it})f_{0t}\lambda_{0i}'(\tilde{f}_t - \tilde{\mathsf{S}}f_{0t}), \quad (9)$$

where $\varrho^{(j)}(u) = (\partial/\partial u)^j\varrho(u)$. The key step is to show that the last two terms on the right-hand side of the above equation are both $o_P(1/\sqrt{T})$. This is relatively easier for the PCA estimator of Bai (2003), since $(\tilde{f}_t - \tilde{\mathsf{S}}f_{0t})$ has an analytical form (like e.g. in equation A.1 of Bai 2003). In our case, we would also need a stochastic expansion for $(\tilde{f}_t - \tilde{\mathsf{S}}f_{0t})$, which in turn depends on the stochastic expansion of $(\tilde{\lambda}_i - \tilde{\mathsf{S}}\lambda_{0i})$ due to the nature of factor models. As in Chen et al. (2020), this problem can be partly solved by showing that the expected Hessian matrix is asymptotically block-diagonal (see Lemma S.6 in the Online Appendix). However, the proof of Chen et al. (2020) is only applicable to a special infeasible normalization, namely $\sum_{i=1}^{N}\lambda_{0i}\lambda_i = \sum_{t=1}^{T}f_{0t}f_t'$, while our proof of Lemma S.6 allows for normalization (2) and can be generalized to any of the other normalizations considered by Bai and Ng (2013) that uniquely pin down the rotation matrix.

Remark 4.3: In line with Remark 1.3, if the true parameters did not satisfy the normalizations (2), the results of Theorem 3 should be stated as

$$\sqrt{T}\left(\tilde{\lambda}_i - \tilde{\mathsf{S}}H_{NT}^{-1}\lambda_{0i}\right) \xrightarrow{d} \mathcal{N}\left(0, \tau(1-\tau)H^{-1}\Phi_i^{-1}\Sigma_F\Phi_i^{-1}(H^{-1})'\right),$$

$$\sqrt{N}\left(\tilde{f}_t - \tilde{\mathsf{S}}H_{NT}'f_{0t}\right) \xrightarrow{d} \mathcal{N}\left(0, \tau(1-\tau)H'\Psi_t^{-1}\Sigma_\Lambda\Psi_t^{-1}H\right),$$

where H_{NT} is defined in Remark 1.3, $\Sigma_F = \lim_{T\to\infty}\Sigma_{T,F}$, $\Sigma_\Lambda = \lim_{N\to\infty}\Sigma_{N,\Lambda}$, $H = \Sigma_F^{-1/2}\Gamma$, and Γ is the matrix of eigenvectors of $\Sigma_F^{1/2}\Sigma_\Lambda\Sigma_F^{1/2}$.

Remark 4.4: Let $l(z)$ be a continuous kernel function with support $[-1, 1]$ where $l^{(j)}(z) = \partial^j l(z)/\partial z^j$ exists and $\sup_{z\in(-1,1)}|l^{(j)}(z)|$ is bounded for $j = 1, 2$. Let b a bandwidth. Estimators for the asymptotic variance matrices of $\tilde{\lambda}_i$ and $\tilde{f}_t$ can be simply constructed as

$$\tilde{V}_{\lambda_i} = \tau(1-\tau)\tilde{\Phi}_i^{-2} \text{ where } \tilde{\Phi}_i = \frac{1}{Tb}\sum_{t=1}^{T} l(\tilde{u}_{it}/b)\cdot \tilde{f}_t\tilde{f}_t',$$

and

$$\tilde{V}_{f_t} = \tau(1-\tau)\tilde{\Psi}_t^{-1}\tilde{\Sigma}_\Lambda\tilde{\Psi}_t^{-1} \text{ where } \tilde{\Psi}_t = \frac{1}{Nb}\sum_{i=1}^{N} l(\tilde{u}_{it}/b)\cdot \tilde{\lambda}_i\tilde{\lambda}_i', \quad \tilde{\Sigma}_\Lambda = \tilde{\Lambda}'\tilde{\Lambda}/N,$$

with $\tilde{u}_{it} = X_{it} - \tilde{\lambda}_i'\tilde{f}_t$. In Section S.2 of the Online Appendix we show that under Assumptions 1 and 2, the above estimators of the asymptotic covariance matrices are consistent if $b \to 0$ and $Nb^3 \to \infty$. Note that this is different from the usual condition $Nb^2 \to \infty$ in standard quantile regressions (see e.g. Powell 1984 and Angrist, Chernozhukov, and Fernández-Val 2006). Moreover, the above estimators are also consistent for the asymptotic covariance matrices discussed in Remark 4.3.

5 Finite Sample Simulations

We next report the results from several Monte Carlo simulations regarding the performance of the QFM methodology in finite samples. In particular, we focus on four relevant issues: (i) how well our preferred estimator of the number of factors and the QFA estimator perform in relation to other methods when the distribution of the idiosyncratic errors exhibits heavy tails, (ii) how well the QFA estimator performs in estimating the extra quantile factors that can not be captured by the PCA estimator, (iii) how robust the QFA estimation procedure is when the errors terms are serially and cross-sectionally correlated, instead of being independent, and (iv) how well the asymptotic normal approximations (derived in Theorem 4 for the QFA estimators based on SQR) behave in finite samples.

5.1 QFA Estimation with Heavy-tailed Idiosyncratic Errors

As pointed out in Remark 1.2, consistency of the QFA estimator does not require the moments of the idiosyncratic errors to exist. Hence, at $\tau = 0.5$, QFA can be viewed as a robust QR alternative to the PCA estimators commonly used in practice. By the same token, our proposed estimators of the number of factors should also be robust to outliers and heavy tails. In this subsection we check these results in finite samples by means of a few simulations.

To do so, we consider a three-factor model with the following DGP:

$$X_{it} = \sum_{j=1}^{3} \lambda_{ji} f_{jt} + u_{it},$$

where $f_{1t} = 0.8 f_{1,t-1} + \epsilon_{1t}$, $f_{2t} = 0.5 f_{2,t-1} + \epsilon_{2t}$, $f_{3t} = 0.2 f_{3,t-1} + \epsilon_{3t}$, $\lambda_{ji}, \epsilon_{jt}$ are all independent draws from $\mathcal{N}(0,1)$, and $u_{it} \sim$ i.i.d t_ν for $\nu = 1,2,3$, where t_ν denotes the student's t distribution with ν degrees of freedom.

To select the number of factors, we focus on our preferred rank-minimization estimator defined in subsection 3.2, having chosen P_{NT} as in (7) and $k = 8$. To estimate the factors, we consider the QFA estimator at $\tau = 0.5$ (denoted as $\hat{F}^{0.5}_{QFA}$) using the IQR algorithm. Since PCA is not well suited for this type of error distributions, we choose an alternative benchmark for the QFA performance. This is the estimator proposed by He, Kong, Yu, and Zhang (2020) (denoted as $\hat{F}_{KEN}$ since it uses Kendall's tau matrix) which, to the best of our knowledge, is the only method in the robust PCA literature that can consistently estimate the factors without imposing moment constraints on the idiosyncratic errors.

With $X_t = (X_{1t}, \ldots, X_{Nt})'$, the empirical *spatial Kendall's tau matrix* (see Section 4.4 of Fan, Liu, and Wang 2018) is defined as

$$\hat{K} = \frac{2}{N(N-1)} \sum_{t<t'} \frac{(X_t - X_{t'})(X_t - X_{t'})'}{\|X_t - X_{t'}\|^2}.$$

Then, He et al. (2020) show that the estimated factor loading matrix $\hat{\Lambda}_{KEN}$ is given by the r leading eigenvectors of $\hat{K}$ times $\sqrt{N}$, while the estimator of the factor matrix corresponds to $\hat{F}_{KEN} = X \cdot \hat{\Lambda}_{KEN}/N$.

Table 1 presents the results for $N, T \in \{50, 100, 200\}$ obtained from 1000 simulations. Columns 2 and 3 report the average estimated number of factors and the frequency of choosing the right number of factors using our rank-minimization estimator; columns 4 to 6 display the average adjusted R^2 of regressing each of the true factors on $\hat{F}_{KEN}$ to compute the distance between the space of the estimated factors and the space of the true factors; finally the last three columns show the average adjusted R^2 of regressing each of the true factors on $\hat{F}^{0.5}_{QFA}$.

There are three main takeaways from these simulation results. First, the rank-minimization estimator selects the right number of factors with high accuracy as long as $\min\{N, T\} \geq 100$, and its performance is not affected as the distribution of the errors changes from t_3 to t_1. Second, the QFA estimator performs very well in capturing the space of the true factors, as measured by the adjusted R^2s, which are always higher than 0.9. Finally, for t_3 and t_2 errors, the estimator $\hat{F}_{KEN}$ performs similarly to $\hat{F}^{0.5}_{QFA}$; however, $\hat{F}_{KEN}$ completely breaks down in the case of i.i.d t_1 errors whereas the QFA estimator continues to work well.[5]

[5] The reason why the estimator of He et al. (2020) performs poorly in the case of i.i.d Cauchy errors is that their estimator is designed for models where $(u_{1t}, u_{2t}, \ldots, u_{Nt})$ have a multivariate t distribution, with the tail behavior being controlled by a single random variable ζ_t.

5.2 QFA Estimation with Dependent Idiosyncratic Errors

Next, to check how restrictive is the independence of error terms, adopted in Assumption 1(iii) for analytical tractability in the proofs (see Remark 1.4 above), we now consider the following DGP, which provides a slight variation of Example 3 in subsection 2.2:

$$X_{it} = \lambda_{1i} f_{1t} + \lambda_{2i} f_{2t} + (\lambda_{3i} f_{3t}) \cdot e_{it}, \tag{10}$$

where $f_{1t} = 0.8 f_{1,t-1} + \epsilon_{1t}$, $f_{2t} = 0.5 f_{2,t-1} + \epsilon_{2t}$, $f_{3t} = |g_t|$, $\lambda_{1i}, \lambda_{2i}, \epsilon_{1t}, \epsilon_{2t}, g_t$ are all independent draws from $\mathcal{N}(0,1)$, and λ_{3i} are independent draws from $U[1,2]$. Following BN (2002), the following specification for e_{it} is used:

$$e_{it} = \beta\ e_{i,t-1} + v_{it} + \rho \cdot \sum_{j=i-J, j\neq i}^{i+J} v_{jt},$$

where v_{it} are independent draws from $\mathcal{N}(0,1)$, except in the second case discussed below where we also allow for heavy tails. The autoregressive coefficient β captures the serial correlation of e_{it}, while the parameters ρ and J capture the cross-sectional correlations of e_{it}. We consider three models embedded in the previous specification:

M1: Independent errors: $\beta = 0$ and $\rho = 0$.
M2: Independent errors with heavy tails: $\beta = \rho = 0$, and $v_{it} \sim$ i.i.d t_3.
M3: Serially and cross-sectionally correlated errors: $\beta = 0.2$ and $\rho = 0.2$, and $J = 3$.

For each model and each $\tau \in \{0.25, 0.5, 0.75\}$, we first estimate $\hat{r}$ using the rank-minimization estimator, having set k and P_{NT} as in the previous subsection. Second, we estimate $\hat{r}$ factors by means of QFA, denoted $\hat{F}^{\tau}_{QFA}$. Finally, we regress each of the true factors on $\hat{F}^{\tau}_{QFA}$ and compute the R^2s. This procedure is repeated 1000 times where, for each τ, the averages of $\hat{r}$ and the R^2s in these repetitions are reported.

Table 2 presents the results of these simulations for $N, T \in \{50, 100, 200\}$. Notice that for $\tau = 0.25, 0.75$, we have $r(\tau) = 3$ while, for $\tau = 0.5$, we get $r(\tau) = 2$, since the factor f_{3t} does not affect the median of X_{it}. For **M1** and **M2**, it can be observed that both the rank-minimization and the QFA estimators perform very well in choosing the number of quantile-dependent factors and in estimating them. It should be noticed that at $\tau = 0.25, 0.75$ the estimation of the scale factor f_{3t} is not as good as the mean factors f_{1t}, f_{2t} for small N and T. However, such differences vanish as N and T increase. For **M3**, it can also be inspected that the QFA estimators still fare satisfactorily, even though the independence assumption is violated in these DGPs. Thus, we conclude that, despite Assumption 1 (iii), these simulations show that QFA estimation still works properly when the error terms are allowed to exhibit mild serial and cross-sectional correlations.

Finally, DGP (10) is useful to illustrate the limitations of applying PCA when there are extra factors (like f_{3t} above) and the true model is considered to be unknown. Table S.1 in the Online Appendix reports the estimated number of factors according to IC_{p1} of BN (2002) and the R^2s of regressing the true factors on the estimated PCA factors. As expected, the R^2s are very high for the mean factors f_{1t} and f_{2t}, but instead are much lower for the extra factor f_{3t}, which can be captured by QFA.

5.3 Normal Approximations of the Estimators Based on SQR

Our last set of simulations is devoted to evaluating the finite sample behavior of the asymptotically normal distributions derived in Theorem 4 for the QFA estimators based on SQR. To do so, we consider the following illustrative DGP:

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

where $f_t \sim i.i.d\, \mathcal{U}(1,2)$ normalized by $F'F/T = 1$, $\lambda_i \sim i.i.d\, \mathcal{N}(0,1)$, $\epsilon_{it} \sim i.i.d\, \mathcal{N}(0,1)$, and $N, T \in \{50, 200\}$. Note that, since the results in Theorem 4 are conditional on the factors and the loadings, both f_t and λ_i are taken as fixed in the simulations. To smooth the indicator function, we use the following eighth-order kernel function (see Muller, 1984):

$$k(z) = \mathbf{1}\{|z| \leq 1\} \cdot \frac{3465}{8192}\left(7 - 105z^2 + 462z^4 - 858z^6 + 715z^8 - 221z^{10}\right),$$

while the Epanechnikov kernel $l(z) = 0.75(1 - z^2) \cdot \mathbf{1}\{|z| \leq 1\}$ is applied to estimate the variance.

Figures S.1 and S.2 in the Online Appendix plot together the density function of the standard normal distribution and the histograms of the standardized estimators of the factors: $\hat{V}_{f_t}^{-1}\sqrt{N}(\tilde{f}_t - f_{0t})$ at $\tau = 0.25$, $t = T/2$ from 1000 repetitions, where $\hat{V}_{f_t}$ is estimated variance using the formula in Remark 4.4.[6] To check how the bandwidths affect the finite distributions of the estimators, we display results for different choices of the parameters h and b defined in Remark 4.4. The results show that the asymptotic distributions in Theorem 4 provide reasonably good approximations for the finite sample distributions of the estimators based on SQR, even for $N = T = 50$, and that they are hardly sensitive to the choice of bandwidths.

6 Empirical Application

In this section we provide empirical evidence showing that QFA could become a useful tool for predictive exercises regarding macro aggregates[7]. In particular, we extend the diffusion-index

[6] We choose the signs of $\tilde{f}_t$ such that $\tilde{\mathsf{S}} = 1$.

[7] Further empirical applications of QFA related to causal analysis with climate data and the economic interpretation of quantile-dependent factors in models of stock returns can be found in Chen, Dolado, and Gonzalo (2020).

forecasting techniques popularized by Stock and Watson (2002) to explore the predictive power of the QFA factors. The main goal of this application is to extract a few common factors (by PCA and QFA) from a large panel of macroeconomic variables, and then analyze the role of both sets of factors in forecasting real GDP growth and the inflation rate.

We use the FRED-QD dataset which is a quarterly panel of 211 US macroeconomic variables from 1960Q1 to 2019Q2 ($N = 211, T = 238$), that emulates the popular dataset used by Stock and Watson (2002), but also contains several additional time series. All variables in this dataset are updated in a timely manner and can be downloaded for free.[8] Before estimation, each series is transformed to be stationary using MATLAB codes that are also available on the FRED-QD data website.[9]

We initially estimate the number of mean factors by PCA (where the maximum is set equal to 8 for all estimators) using the standard PC_{p1} criterion of BN (2002), which selects 8 factors. Next, we apply our rank-minimization estimator using a grid of quantiles ranging from 0.01 to 0.99. The results of this estimator are presented in Table 3 (column 2). Notice that the numbers of QFA factors estimated by the rank-minimization criterion varies significantly across quantiles, pointing to the existence of a nonstandard factor structure in this dataset. Moreover, the remaining columns in Table 3 report the R^2s of regressing each of the QFA factors found at the different quantiles on the 8 PCA factors. Given their high R^2s, it becomes clear that the QFA factors at τ close to 0.5 and 0.75 are all well explained by the PCA mean factors. However, this is not the case for other quantiles, where the R^2s are much lower. For example, the first QFA factor at $\tau = 0.9$ (denoted $\hat{F}^{0.9}_{QFA}$) and those at $\tau = 0.95, 0.99$ (denoted as $\hat{F}^{0.95}_{QFA}$ and $\hat{F}^{0.99}_{QFA}$) contain some extra information that could be potentially helpful for forecasting the above-mentioned variables. Since $\hat{F}^{0.95}_{QFA}$ exhibits a very high correlation with $\hat{F}^{0.9}_{QFA}$ and $\hat{F}^{0.99}_{QFA}$, we exclusively focus on the predictive power of $\hat{F}^{0.9}_{QFA}$ and $\hat{F}^{0.99}_{QFA}$ in the subsequent analysis.

Let y_{t+1} denote the realized value of real GDP growth/inflation at period $t+1$. The forecasting model we consider is as follows:

$$y_{t+1} = \alpha + \sum_{j=0}^{p_{max}} \beta_j y_{t-j} + \gamma' F_t + \epsilon_{t+1},$$

where F_t is vector containing several unobserved common factors extracted from this large dataset. The predicted value of y_{t+1}, based on a vector of estimated factors $\hat{F}_t$, is simply constructed as $\hat{y}_{t+1} = \hat{\alpha} + \sum_{j=0}^{\hat{p}} \hat{\beta}_j y_{t-j} + \hat{\gamma}' \hat{F}_t$, where $\hat{\alpha}$, $\hat{\beta}_j$, $\hat{\gamma}$ are OLS estimates of the coefficients, and $\hat{p}$ is the optimal lag length according to BIC. We compare five different specifications for F_t: (i) $F_t = 0$, which is the benchmark AR model, (ii) AR plus $\hat{F}_t$ only including $\hat{F}_{PCA}$, (iii)

[8] Link to the dataset: http://research.stlouisfed.org/econ/mccracken/. We refer to McCracken and Ng (2016) for further details.

[9] Given that the variables in this dataset are measured in different units, they are standardized to have zero mean and variance equal 1 before estimating the number of factors and the factors themselves.

AR plus $\hat{F}_t$ including $\hat{F}_{PCA}$ and $\hat{F}^{0.9}_{QFA}$, (iv) AR plus $\hat{F}_t$ including $\hat{F}_{PCA}$ and $\hat{F}^{0.99}_{QFA}$, and (v) AR plus $\hat{F}_t$ including $\hat{F}_{PCA}$, $\hat{F}^{0.9}_{QFA}$ and $\hat{F}^{0.99}_{QFA}$. Following Chudik, Kapetanios, and Pesaran (2018), the initial estimation period is 1960Q1 to 1989Q4 (120 periods), and the forecast evaluation period is split into great moderation pre-crisis (1990Q1 to 2007Q2) and financial crisis/recovery (2007Q3 to 2019Q2) sub-periods. A rolling window of 120 periods is used both to estimate the coefficients and generate the rolling forecasts. In particular, following Chudik et al. (2018), the number of mean factors in this model is estimated using PC_{p1} at each rolling window, where the maximum number of factors is set equal to 5.

The mean squared error (MSE) of these procedures, and their relative MSE (R-MSE) to the benchmark AR model are reported in Table 4 for the whole evaluation period and each of the relevant sub-samples. As can be observed, adding the upper tail QFA factors ranks better in terms of R-MSE than the AR and AR+$\hat{F}_{PCA}$ models for the three periods under considerations. The gains are not very sizable but still can be considered to be relevant, with reductions in R-MSE of between 3.5 and 8 percent.

A well-known shortcoming of point forecasts is that their uncertainty is generally unknown; hence it is difficult to quantify their precision at any given period of time. To address this issue, it has became customary among central banks to report density forecasts for important macroeconomic variables. In this respect, Adrian et al. (2019) argue that a simple way of producing such densities is via QR. Following their approach, we next evaluate the predictive power of the QFA factors for forecasting the densities of real GDP growth and inflation. In particular, we first predict the conditional quantiles of the target variable y_{t+h} by $\hat{q}_{\tau,t+h} = \hat{\alpha}_\tau + \sum_{j=0}^{p} \hat{\beta}_{\tau,j} y_{t-j} + \hat{\gamma}'_\tau \hat{F}_{\tau,t}$ for $\tau \in \{0.05, 0.25, 0.75, 0.95\}$, where $\hat{\alpha}_\tau, \hat{\beta}_{\tau,j}, \hat{\gamma}_\tau$ are estimated coefficients by running QR of y_{t+h} on $[1, y_t, \ldots, y_{t-p}, \hat{F}_{\tau,t}]$, and $\hat{F}_{\tau,t}$ is a vector of estimated quantile factors using the IQR algorithm.[10] Next, given the predicted quantiles: $[\hat{q}_{0.05,t+h}, \hat{q}_{0.25,t+h}, \hat{q}_{0.75,t+h}, \hat{q}_{0.95,t+h}]$, the predicted density of y_{t+h} is constructed as the density of a skewed t-distribution by matching the predicted quantiles.[11] Finally, the accuracy of the density forecast is measured by the predictive score, namely, the predicted density evaluated at the realized value of y_{t+h}. Higher predictive scores signify more accurate predictions. The out-of-sample density forecasts are constructed using rolling windows with the most recent 120 observations, and the evaluation period is 1990Q1 to 2019Q2. Moreover, we set $p = 3$, and consider as the benchmark model the one with $\hat{F}_{\tau,t} = 0$, where the quantiles of y_{t+h} are predicted only using its own lags.

The four panels in Figure 1 display the predictive scores of the one-quarter-ahead ($h = 1$) and one-year-ahead ($h = 4$) density forecasts for GDP growth (upper panels) and inflation (lower panels). In both instances, it can be seen that the predictive scores of the"AR + Quantile Factors" procedure are frequently above those of the "AR benchmark" model, and sometimes

[10] The number of factors at each of the four chosen quantiles is given in Table 3 (column 2).

[11] We refer to Adrian et al. (2019) for further details, and to Azzalini and Capitanio (2003) for the definition and properties of the skewed t-distribution.

by a large margin. In sum, these empirical results indicate that the QFA factors could indeed be very informative for density forecasting of highly relevant macroeconomic variables.

7 Conclusions

Inspired by the generalization of linear regressions to quantile regressions (QR), this paper proposes quantile factor models (QFM) as a new class of factor models in relation to the conventional approximate factor models (AFM). Compared to AFM, both factors and loadings in QFM are allowed to be quantile-dependent objects which affect other distributional characteristics of the data (volatility, higher moments, extreme values, etc.), and not (or not only) their mean. Using tools in the interface of QR, principal component analysis (PCA) and the theory of empirical processes, we propose an estimation procedure of the quantile-dependent objects in QFM, labelled Quantile Factor Analysis (QFA), which yields consistent and asymptotically normal estimators of factors and loadings at each quantile. In addition, we propose novel selection criteria to estimate consistently the number of factors at each quantile.

The previous theoretical findings receive support in finite samples from a range of Monte Carlo simulations, including the robustness of QFA when the idiosyncratic errors lack moments. Furthermore, as illustrated by our empirical application, these extra factors could be useful, among other issues, for forecasting exercises with factor-augmented regressions.

Among the research issues which have been left out of this paper, four topics stand out: (i) undertake further analysis on the role of QFA in the propagation of structural shocks in factor-augmented VAR (FAVAR) models, where recent developments in quantile VAR estimation, as in White, Kim, and Manganelli (2015), provide useful tools to address this issue; (ii) relax the independence assumption on the error terms which, in view of our previous simulation results with mildly dependent errors in QFM, seems feasible; (iii) extend our results for static QFM to dynamic QFM, where the set of quantile-dependent variables include lagged factors (see Forni, Hallin, Lippi, and Reichlin 2000 and Stock and Watson 2011); and finally (iv) provide theories about how to interpret QFA factors in different economic and financial setups.

Appendix A: Proofs of the Main Results

Definitions and Notations: Throughout the Appendix, $K_1, K_2, \ldots$ denote some positive constants that do not depend on N, T. For any random variable Y, define the Orlicz norm $\|Y\|_\psi$ as:

$$\|Y\|_\psi = \inf\left\{C > 0 : \mathbb{E}\psi\left(|Y|/C\right) \leq 1\right\},$$

where ψ is a non-decreasing, convex function with $\psi(0) = 0$. In particular, when $\psi(x) = e^{x^2} - 1$, the norm is written as $\|Y\|_{\psi_2}$. We use $\|\cdot\|_S$ to denote the spectral norm, and $C(\cdot, g, \mathcal{G})$ to denote the covering number of space $\mathcal{G}$ endowed with semimetric g. Moreover, define

$$\mathbb{W}_{NT}(\theta) = \mathbb{M}^*_{NT}(\theta) - \bar{\mathbb{M}}^*_{NT}(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\omega_{it}(\lambda_i, f_t) - \mathbb{E}[\omega_{it}(\lambda_i, f_t)]\right),$$

where $\omega_{it}(\lambda_i, f_t) = \rho_\tau(X_{it} - \lambda_i' f_t) - \rho_\tau(X_{it} - \lambda_{0i}' f_{0t})$.

A.1 Proof of Theorem 1

Lemma 1. *Under Assumption 1, we have $d(\hat{\theta}, \theta_0) = o_P(1)$ as $N, T \to \infty$.*

Proof. First, for any $\lambda_i \in \mathcal{A}$ and $f_t \in \mathcal{F}$, expanding $\mathbb{E}[\rho_\tau(X_{it} - \lambda_i' f_t) - \rho_\tau(X_{it} - \lambda_{0i} f_{0t})]$ around $c_{0,it} = \lambda_{0i} f_{0t}$, we have

$$\mathbb{E}[\rho_\tau(X_{it} - \lambda_i' f_t) - \rho_\tau(X_{it} - \lambda_{0i} f_{0t})] = 0.5 \cdot \mathsf{f}_{it}(c^*_{it}) \cdot (\lambda_i' f_t - \lambda_{0i} f_{0t})^2,$$

where c^*_{it} is between $\lambda_i' f_t$ and $c_{0,it}$. It then follows from Assumption 1(ii) that for all $\lambda_i \in \mathcal{A}$ and $f_t \in \mathcal{F}$,

$$\mathbb{E}[\rho_\tau(X_{it} - \lambda_i' f_t) - \rho_\tau(X_{it} - \lambda_{0i} f_{0t})] \gtrsim (\lambda_i' f_t - \lambda_{0i}' f_{0t})^2$$

since $|\lambda_i' f_t|$ and $|\lambda_{0i}' f_{0t}|$ are both bounded. Therefore, for any $\theta \in \Theta^r$, we have:

$$\bar{\mathbb{M}}^*_{NT}(\theta) \gtrsim d^2(\theta, \theta_0). \tag{A.1}$$

Second, by the definition of $\hat{\theta}$, we have $\mathbb{M}^*_{NT}(\hat{\theta}) = \mathbb{M}_{NT}(\hat{\theta}) - \mathbb{M}_{NT}(\theta_0) \leq 0$, or equivalently $\mathbb{W}_{NT}(\hat{\theta}) + \bar{\mathbb{M}}^*_{NT}(\hat{\theta}) \leq 0$. It then follows from (A.1) that

$$0 \leq d^2(\hat{\theta}, \theta_0) \lesssim \bar{\mathbb{M}}^*_{NT}(\hat{\theta}) \leq \sup_{\theta \in \Theta^r} |\mathbb{W}_{NT}(\theta)|.$$

Thus, it remains to be shown that

$$\sup_{\theta \in \Theta^r} |\mathbb{W}_{NT}(\theta)| = o_P(1). \tag{A.2}$$

Choose K_1 large enough such that $\|\lambda_{0i}\|, \|f_{0t}\|, \|\lambda_i\|, \|f_t\| \leq K_1$ for all i, t for any $\theta \in \Theta^r$. Let $B_r(K_1)$ be a Euclidean ball in $\mathbb{R}^r$ with radius K_1. For any $\epsilon > 0$, let $\lambda_{(1)}, \ldots, \lambda_{(J)}$ be a maximal set of points in $B_r(K_1)$ such that $\|\lambda_{(j)} - \lambda_{(h)}\| > \epsilon/K_1$ for any $j \neq h$ where "maximal" signifies that no point can be added without violating the validity of the inequality. Similarly, let $f_{(1)}, \ldots, f_{(J)}$ be a maximal set of points in $B_r(K_1)$ such that $\|f_{(j)} - f_{(h)}\| > \epsilon/K_1$ for any $j \neq h$. It is well known that J, the packing number of $B_r(K_1)$, is equal to $K_2(K_1/\epsilon)^r$.

For any $\theta \in \Theta^r$, define $\theta^* = (\lambda_1^{*'}, \ldots, \lambda_N^{*'}, f_1^{*'}, \ldots, f_T^{*'})'$, where $\lambda_i^* = \{\lambda_{(j)} : j \leq J, \|\lambda_{(j)} - \lambda_i\| \leq \epsilon/K_1\}$ and $f_t^* = \{f_{(j)} : j \leq J, \|f_{(j)} - f_t\| \leq \epsilon/K_1\}$. Thus, we can write

$$\mathbb{W}_{NT}(\theta) = \mathbb{W}_{NT}(\theta^*) + \mathbb{W}_{NT}(\theta) - \mathbb{W}_{NT}(\theta^*).$$

Note that $|\rho_\tau(X_{it} - \lambda_i' f_t) - \rho_\tau(X_{it} - \lambda_i^{*'} f_t^*)| \leq 2|\lambda_i' f_t - \lambda_i^{*'} f_t^*| \leq 2\|\lambda_i\|\|f_t - f_t^*\| + 2\|f_t^*\|\|\lambda_i - \lambda_i^*\| \leq 4\epsilon$.

Thus,

$$\sup_{\theta\in\Theta^r} |\mathbb{W}_{NT}(\theta) - \mathbb{W}_{NT}(\theta^*)| \leq 4\epsilon. \tag{A.3}$$

Also, note that $|\omega_{it}(\lambda_i^*, f_t^*)| = |\rho_\tau(X_{it} - \lambda_i^{*\prime} f_t^*) - \rho_\tau(X_{it} - \lambda_{0i}' f_{0t})| \leq 2|\lambda_i^{*\prime} f_t^* - \lambda_{0i}' f_{0t}|$. Then, by Hoeffding's inequality we have

$$P[|\sqrt{NT}\mathbb{W}_{NT}(\theta^*)| > c] \leq 2e^{-\frac{c^2}{2\cdot d^2(\theta^*,\theta_0)}},$$

and by Lemma 2.2.1 of van der Vaart and Wellner (1996) it follows that $\|\mathbb{W}_{NT}(\theta^*)\|_{\psi_2} \lesssim d(\theta^*,\theta_0)/\sqrt{NT}$. Since θ^* can take at most $J^{N+T} \lesssim (K_1/\epsilon)^{r(N+T)}$ different values, and $d(\theta^*,\theta_0) \lesssim K_1$, it follows from Lemma 2.2.2 of van der Vaart and Wellner (1996) that

$$\mathbb{E}\left[\sup_{\theta\in\Theta^r} |\mathbb{W}_{NT}(\theta^*)|\right] \leq \left\|\sup_{\theta\in\Theta^r} |\mathbb{W}_{NT}(\theta^*)|\right\|_{\psi_2} \lesssim \sqrt{\log(K_1/\epsilon)}\sqrt{r(N+T)}/\sqrt{NT} \lesssim \sqrt{\log(K_1/\epsilon)}/L_{NT}. \tag{A.4}$$

Finally, by Markov's inequality and (A.3), for any $\delta > 0$,

$$\begin{aligned} P\left[\sup_{\theta\in\Theta^r} |\mathbb{W}_{NT}(\theta)| > \delta\right] &\leq P\left[\sup_{\theta\in\Theta^r} |\mathbb{W}_{NT}(\theta^*)| > \delta/2\right] + P\left[\sup_{\theta\in\Theta^r} |\mathbb{W}_{NT}(\theta) - \mathbb{W}_{NT}(\theta^*)| > \delta/2\right] \\ &\leq 2/\delta \cdot \mathbb{E}\left[\sup_{\theta\in\Theta^r} |\mathbb{W}_{NT}(\theta^*)|\right] + P\left[4\epsilon > \delta/2\right]. \end{aligned}$$

Thus, (A.2) follows from (A.4) since ϵ is arbitrary. This concludes the proof. □

Next, define $\Theta^r(\delta) = \{\theta \in \Theta^r : d(\theta,\theta_0) \leq \delta\}$.

Lemma 2. *Under Assumption 1 and for sufficiently small $\delta > 0$, for any $\theta \in \Theta^r(\delta)$, we have*

$$\|\Lambda - \Lambda_0 \mathsf{S}\|/\sqrt{N} + \|F - F_0\mathsf{S}\|/\sqrt{T} \leq K_3\delta,$$

where $\mathsf{S} = \mathsf{sgn}(F'F_0/T)$.

Proof. First, let $\mathsf{U} \in \mathbb{R}^{r\times r}$ be a diagonal matrix whose diagonal elements are either 1 or −1. Since $F'F/T = F_0'F_0/T = \mathbb{I}_r$ and $\|\Lambda_0\|/\sqrt{N} \leq K_4$ by Assumption 1(i), we have

$$\begin{aligned} & \|\Lambda - \Lambda_0\mathsf{U}\|/\sqrt{N} = \|(\Lambda - \Lambda_0\mathsf{U})F'\|/\sqrt{NT} = \|\Lambda F' - \Lambda_0 F_0' + \Lambda_0 F_0' - \Lambda_0 \mathsf{U} F'\|/\sqrt{NT} \\ \leq\ & \|\Lambda F' - \Lambda_0 F_0'\|/\sqrt{NT} + \|\Lambda_0\|/\sqrt{N}\cdot\|F - F_0\mathsf{U}\|/\sqrt{T} \\ \leq\ & d(\theta,\theta_0) + K_4\|F - F_0\mathsf{U}\|/\sqrt{T}. \end{aligned}$$

Thus, for $\theta \in \Theta^r(\delta)$,

$$\|\Lambda - \Lambda_0\mathsf{U}\|/\sqrt{N} + \|F - F_0\mathsf{U}\|/\sqrt{T} \leq \delta + (1 + K_4)\|F - F_0\mathsf{U}\|/\sqrt{T}. \tag{A.5}$$

Second,

$$\begin{aligned}\|F - F_0\mathsf{U}\|/\sqrt{T} &= \|F_0\mathsf{U} - F(F'F_0\mathsf{U}/T) + F(F'F_0\mathsf{U}/T) - F\|/\sqrt{T} \\ &\leq \|F_0\mathsf{U} - F(F'F_0\mathsf{U}/T)\|/\sqrt{T} + \|F(F'F_0\mathsf{U}/T) - F\|/\sqrt{T} = \|M_F F_0\|/\sqrt{T} + \|F'F_0/T - \mathsf{U}\|, \quad \text{(A.6)}\end{aligned}$$

where $P_A = A(A'A)^{-1}A'$ and $M_A = \mathbb{I} - P_A$.

Third, we have

$$\begin{aligned}\frac{1}{\sqrt{NT}}\|(\Lambda F' - \Lambda_0 F_0')M_F\| &\leq \sqrt{\operatorname{rank}[(\Lambda F' - \Lambda_0 F_0')M_F]} \cdot \|M_F\|_S \cdot \|\Lambda F' - \Lambda_0 F_0'\|_S/\sqrt{NT} \\ &\lesssim \|\Lambda F' - \Lambda_0 F_0'\|/\sqrt{NT} = d(\theta, \theta_0), \quad \text{(A.7)}\end{aligned}$$

and since

$$\begin{aligned}\|(\Lambda F' - \Lambda_0 F_0')M_F\|/\sqrt{NT} = \|\Lambda_0 F_0' M_F\|/\sqrt{NT} &= \sqrt{\operatorname{Tr}\left[(\Lambda_0'\Lambda_0/N) \cdot (F_0' M_F F_0/T)\right]} \\ &\geq \sqrt{\sigma_{Nr}}\sqrt{\operatorname{Tr}\left(F_0' M_F F_0/T\right)} = \sqrt{\sigma_{Nr}}\|M_F F_0\|/\sqrt{T}, \quad \text{(A.8)}\end{aligned}$$

it follows from (A.7) and (A.8) that

$$\|M_F F_0\|/\sqrt{T} \lesssim \sqrt{\frac{1}{\sigma_{Nr}}d(\theta, \theta_0)}. \quad \text{(A.9)}$$

Similarly, it can be shown that

$$\|M_{F_0} F\|/\sqrt{T} \lesssim \sqrt{\frac{1}{\rho_{\min}(\Lambda'\Lambda/N)}d(\theta, \theta_0)}, \quad \text{(A.10)}$$

where $\rho_{\min}$ denotes the minimum eigenvalue.

Fourth, we have

$$\frac{1}{\sqrt{NT}}\|(\Lambda F' - \Lambda_0 F_0')P_F\| \leq \frac{1}{\sqrt{NT}}\|\Lambda F' - \Lambda_0 F_0'\| \cdot \|P_F\| = \sqrt{r}d(\theta, \theta_0),$$

so

$$\frac{1}{\sqrt{NT}}\|(\Lambda F' - \Lambda_0 F_0')P_F\| = \frac{1}{\sqrt{NT}}\|\Lambda F' - \Lambda_0(F_0'F/T)F'\| = \frac{1}{\sqrt{N}}\|\Lambda - \Lambda_0(F_0'F/T)\| \leq \sqrt{r}d(\theta, \theta_0). \quad \text{(A.11)}$$

Likewise, it can be shown that

$$\frac{1}{\sqrt{N}}\|\Lambda_0 - \Lambda(F'F_0/T)\| \leq \sqrt{r}d(\theta, \theta_0). \quad \text{(A.12)}$$

Fifth, define $R_T = F'F_0/T$. Note that $FR_T = FF'F_0/T = P_F F_0$, thus

$$\begin{aligned}\mathbb{I}_r = F_0'F_0/T &= R_T'(F'F/T)R_T + F_0'F_0/T - R_T'(F'F/T)R_T \\ &= R_T'R_T + F_0'F_0/T - F_0'FR_T/T + F_0'FR_T/T - R_T'(F'F/T)R_T \\ &= R_T'R_T + F_0'(F_0 - FR_T)/T = R_T'R_T + F_0'M_F F_0/T, \end{aligned} \quad (A.13)$$

because

$$F_0'FR_T/T - R_T'(F'F/T)R_T = (F_0 - FR_T)'FR_T/T = F_0'M_F FR_T/T = 0.$$

In addition,

$$\begin{aligned}\Lambda_0'\Lambda_0/N &= R_T'(\Lambda'\Lambda/N)R_T + (\Lambda_0'\Lambda_0/N - R_T'(\Lambda'\Lambda/N)R_T) \\ &= R_T'(\Lambda'\Lambda/N)R_T + \Lambda_0'(\Lambda_0 - \Lambda R_T)/N + (\Lambda_0 - \Lambda R_T)'\Lambda R_T/N. \end{aligned} \quad (A.14)$$

Similarly to the proof of (A.13), we have

$$\mathbb{I}_r = R_T R_T' + F'(F - F_0 R_T')/T = R_T R_T' + F'M_{F_0}F/T. \quad (A.15)$$

From (A.14), it holds that

$$\begin{aligned} & \Lambda_0'\Lambda_0/N = R_T'(\Lambda'\Lambda/N)(R_T')^{-1}R_T'R_T + \Lambda_0'(\Lambda_0 - \Lambda R_T)/N + (\Lambda_0 - \Lambda R_T)'\Lambda R_T/N \\ = \; & R_T'(\Lambda'\Lambda/N)(R_T')^{-1} + R_T'(\Lambda'\Lambda/N)(R_T')^{-1}(R_T'R_T - \mathbb{I}_r) + \Lambda_0'(\Lambda_0 - \Lambda R_T)/N + (\Lambda_0 - \Lambda R_T)'\Lambda R_T/N, \end{aligned}$$

and then it follows from the above equation and (A.13) that

$$(\Lambda_0'\Lambda_0/N + D_{NT})\, R_T' = R_T'(\Lambda'\Lambda/N), \quad (A.16)$$

where

$$D_{NT} = R_T'(\Lambda'\Lambda/N)(R_T')^{-1}F_0'M_F F_0/T - \Lambda_0'(\Lambda_0 - \Lambda R_T)/N - (\Lambda_0 - \Lambda R_T)'\Lambda R_T/N.$$

From (A.9) (A.11) and (A.12), we have that $\|D_{NT}\| \lesssim d(\theta, \theta_0)$. Hence, by the Bauer-Fike Theorem (see Theorem 7.2.2 of Golub and Van Loan 2013), it holds that

$$|\rho_{\min}[\Lambda'\Lambda/N] - \rho_{\min}[\Lambda_0'\Lambda_0/N]| \le \|D_{NT}\|_S \le \|D_{NT}\| \lesssim d(\theta, \theta_0). \quad (A.17)$$

Moreover, by Assumption 1(i) and the perturbation theory for eigenvectors (see Section 6.12 of Franklin 2012),

$$\|R_T'V_T\mathsf{S} - \mathbb{I}_r\| = \|R_T'V_T - \mathsf{S}\| \lesssim d(\theta, \theta_0), \quad (A.18)$$

where $V_T = \operatorname{diag}\left((R_{T,1}R_{T,1}')^{-1/2}, \ldots, (R_{T,r}R_{T,r}')^{-1/2}\right)$, and $R_{T,j}'$ is the jth column of R_T'.

Furthermore, (A.10) and (A.17) imply that $\rho_{\min}[\Lambda'\Lambda/N]$ is bounded below by a positive constant, and that

$$\|M_{F_0}F\|/\sqrt{T} \lesssim d(\theta, \theta_0). \quad (A.19)$$

Note that the triangular inequality implies that

$$\|R_T' - \mathsf{S}\| \leq \|R_T'V_T - \mathsf{S}\| + \|R_T'V_T - R_T\| \leq \|R_T'V_T - \mathsf{S}\| + \|R_T\| \cdot \|V_T - \mathbb{I}_r\|. \tag{A.20}$$

From (A.15) and (A.19), we get

$$\|V_T - \mathbb{I}_r\| \lesssim \|R_T R_T' - \mathbb{I}_r\| = \|M_{F_0}F\|^2/T \lesssim d^2(\theta, \theta_0). \tag{A.21}$$

For small enough $d(\theta, \theta_0)$, it then follows from (A.18) (A.20) and (A.21) that

$$\|F'F_0/T - \mathsf{S}\| = \|R_T - \mathsf{S}\| \lesssim d(\theta, \theta_0). \tag{A.22}$$

Finally, setting $\mathsf{U} = \mathsf{S}$, it follows from (A.6) (A.9) and (A.22) that for sufficiently small $d(\theta, \theta_0)$

$$\|F - F_0\mathsf{S}\|/\sqrt{T} \lesssim d(\theta, \theta_0). \tag{A.23}$$

Then the desired result follows from (A.5) and (A.23). □

Lemma 3. *Under Assumption 1, for sufficiently small δ, we have*

$$\mathbb{E}\left[\sup_{\theta \in \Theta^r(\delta)} |\mathbb{W}_{NT}(\theta)|\right] \lesssim \frac{\delta}{L_{NT}}.$$

Proof. In the proof of Lemma 1 we have shown that for $\theta_a, \theta_b \in \Theta^r$,

$$\left\| \sqrt{NT}|\mathbb{W}_{NT}(\theta_a) - \mathbb{W}_{NT}(\theta_b)| \right\|_{\psi_2} \lesssim d(\theta_a, \theta_b). \tag{A.24}$$

Since the process $\mathbb{W}_{NT}(\theta)$ is separable, it follows from Theorem 2.2.4 of van der Vaart and Wellner (1996) that

$$\mathbb{E}\left[\sup_{\Theta^r(\delta)} \sqrt{NT}\,|\mathbb{W}_{NT}(\theta)|\right] \lesssim \left\|\sup_{\Theta^r(\delta)} \sqrt{NT}\,|\mathbb{W}_{NT}(\theta)|\right\|_{\psi_2} \lesssim \int_0^\delta \sqrt{\log D(\epsilon, d, \Theta^r(\delta))}d\epsilon.$$

Thus, it remains to be shown that

$$\int_0^\delta \sqrt{\log D(\epsilon, d, \Theta^r(\delta))}d\epsilon = O(\sqrt{N+T}\delta). \tag{A.25}$$

To prove (A.25), first note that Lemma 2 implies that

$$\Theta^r(\delta) \subset \bigcup_{\mathsf{U} \in \mathcal{S}} \Theta^r(\delta; \mathsf{U})$$

where $\mathcal{S} = \{\mathsf{U} \in \mathbb{R}^{r \times r} : \mathsf{U} = \text{diag}(u_1, \ldots, u_r), u_j \in \{-1, 1\} \text{ for } j = 1, \ldots, r\}$ is the set of all the $r \times r$ sign matrices, and $\Theta^r(\delta; \mathsf{U}) = \{\theta \in \Theta^r : \|\Lambda - \Lambda_0\mathsf{U}\|/\sqrt{N} + \|F - F_0\mathsf{U}\|/\sqrt{T} \leq K_3\delta\}$. Since there are 2^r elements in $\mathcal{S}$ and r is fixed, it suffices to show that $\int_0^\delta \sqrt{\log D(\epsilon, d, \Theta^r(\delta; \mathsf{U}))}d\epsilon = O(\sqrt{N+T}\delta)$ for each

$\mathsf{U} \in \mathcal{S}$. Without loss of generality, we focus on the case $\mathsf{U} = \mathbb{I}_r$.

Second, for any $\theta_a, \theta_b \in \Theta^r$,

$$d(\theta_a, \theta_b) = \frac{1}{\sqrt{NT}} \|\Lambda_a F_a' - \Lambda_b F_b'\| = \frac{1}{\sqrt{NT}} \|\Lambda_a F_a' - \Lambda_b F_a' + \Lambda_b F_a' - \Lambda_b F_b'\|$$
$$\leq \frac{1}{\sqrt{N}} \|\Lambda_a - \Lambda_b\| + \frac{\|\Lambda_b\|}{\sqrt{N}} \cdot \frac{\|F_a - F_b\|}{\sqrt{T}} \leq K_5 \left(\frac{\|\Lambda_a - \Lambda_b\|}{\sqrt{N}} + \frac{\|F_a - F_b\|}{\sqrt{T}} \right),$$

where $K_5 \geq 1$. Now define

$$d^*(\theta_a, \theta_b) = 2K_5 \sqrt{\frac{\|\Lambda_a - \Lambda_b\|^2}{N} + \frac{\|F_a - F_b\|^2}{T}}.$$

It follows from $\sqrt{a} + \sqrt{b} \leq 2\sqrt{a+b}$ that $d(\theta_a, \theta_b) \leq d^*(\theta_a, \theta_b)$. Moreover, it follows from $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ that $\Theta^r(\delta; \mathbb{I}_r) \subset \Theta^{r*}(\delta) = \{\theta \in \Theta^r : d^*(\theta, \theta_0) \leq K_6 \delta\}$ where $K_6 = 2K_3K_5$. It then follows that[12]

$$D(\epsilon, d, \Theta^r(\delta; \mathbb{I}_r)) \leq D(\epsilon, d^*, \Theta^r(\delta; \mathbb{I}_r)) \leq D(\epsilon/2, d^*, \Theta^{r*}(\delta)) \leq C(\epsilon/4, d^*, \Theta^{r*}(\delta)). \tag{A.26}$$

Next, we find an upper bound for $C(\epsilon/4, d^*, \Theta^{r*}(\delta))$ as follows. Let $\eta = \epsilon/4$, and $\theta_1^*, \ldots, \theta_J^*$ be a maximal set in $\Theta^{r*}(\delta)$ such that $d^*(\theta_j^*, \theta_l^*) > \eta$ for any $j \neq l$. Define $B(\theta, c) = \{\gamma \in \Theta^r : d^*(\gamma, \theta) \leq c\}$. Then, the balls $B(\theta_1^*, \eta), \ldots, B(\theta_J^*, \eta)$ cover $\Theta^{r*}(\delta)$, and thus $C(\epsilon/4, d^*, \Theta^{r*}(\delta)) \leq J$. Moreover, the balls $B(\theta_1^*, \eta/4), \ldots, B(\theta_J^*, \eta/4)$ are disjoint and

$$\bigcup_{j=1}^{J} B(\theta_j^*, \eta/4) \subset \Theta^{r*}(\delta + \eta/4).$$

Note that the volume of a ball defined by the metric d^* with radius c is the volume of an ellipsoid, which is equal to $h_M \cdot c^M$, where h_M is a constant that depends on N, T and r, but not on c. Therefore, we have

$$J \cdot h_M \cdot (\eta/4)^M \leq h_M \cdot (K_6(\delta + \eta/4))^M,$$

which implies

$$J \leq \left(\frac{K_6(4\delta + \eta)}{\eta} \right)^M = \left(\frac{K_6(16\delta + \epsilon)}{\epsilon} \right)^M \leq \left(\frac{K_7 \delta}{\epsilon} \right)^M \tag{A.27}$$

for $\epsilon \leq \delta$. It then follows from (A.26) and (A.27) that

$$\int_0^\delta \sqrt{\log D(\epsilon, d, \Theta^r(\delta; \mathbb{I}_r))} d\epsilon \;\; \leq \;\; \int_0^\delta \sqrt{\log C(\epsilon/4, d^*, \Theta^{r*}(\delta))} d\epsilon \;\; \leq \;\; \sqrt{(N+T)r} \int_0^\delta \sqrt{\log(K_7 \delta/\epsilon)} d\epsilon.$$

It is easy to show that $\int_0^\delta \sqrt{\log(K_7\delta/\epsilon)} d\epsilon = O(\delta)$ and thus (A.25) follows. This concludes the proof of Lemma 3. □

Proof of Theorem 1:

[12] Let (T, d) be a semi-metric space. Then for any $S \subset T$, it can be shown that $D(\epsilon, d, S) \leq D(\epsilon/2, d, T) \leq C(\epsilon/4, d, T)$.

Proof. The parameter space Θ^r can be partitioned into shells $S_j = \{\theta \in \Theta^r : 2^{j-1} < L_{NT} \cdot d(\theta, \theta_0) \leq 2^j\}$. If $L_{NT} \cdot d(\hat{\theta}, \theta_0)$ is larger than 2^V for a given integer V, then $\hat{\theta}$ is in one of the shells S_j with $j > V$. In such a case the infimum of the mapping $\theta \mapsto \mathbb{M}^*_{NT}(\theta) = \mathbb{M}_{NT}(\theta) - \mathbb{M}_{NT}(\theta_0)$ over this shell is nonpositive by the definition of $\hat{\theta}$. From this we conclude that, for every $\eta > 0$,

$$P\left[L_{NT} \cdot d(\hat{\theta}, \theta_0) > 2^V\right] \leq \sum_{j>V, 2^{j-1} \leq \eta L_{NT}} P\left[\inf_{\theta \in S_j} \mathbb{M}^*_{NT}(\theta) \leq 0\right] + P[d(\hat{\theta}, \theta_0) > \eta].$$

For arbitrarily small $\eta > 0$, the second probability on the RHS of the above equation converges to 0 as $N, T \to \infty$ by Lemma 1.

Next, note that by (A.1), for each θ in S_j we have

$$-\bar{\mathbb{M}}^*_{NT}(\theta) \lesssim -d^2_{NT}(\theta, \theta_0) \leq -\frac{2^{2j-2}}{L^2_{NT}}.$$

Thus, $\inf_{\theta \in S_j} \mathbb{M}^*_{NT}(\theta) \leq 0$ implies that

$$\inf_{\theta \in S_j} \mathbb{W}_{NT}(\theta) \leq -\frac{2^{2j-2}}{L^2_{NT}},$$

and therefore

$$\sum_{j>V, 2^{j-1} \leq \eta L_{NT}} P\left[\inf_{\theta \in S_j} \mathbb{M}^*_{NT}(\theta) \leq 0\right] \leq \sum_{j>V, 2^{j-1} \leq \eta L_{NT}} P\left[\sup_{\theta \in S_j} |\mathbb{W}_{NT}(\theta)| \geq \frac{2^{2j-2}}{L^2_{NT}}\right].$$

By Lemma 3 and Markov's inequality, we have

$$P\left[\sup_{\theta \in S_j} |\mathbb{W}_{NT}(\theta)| \geq \frac{2^{2j-2}}{L^2_{NT}}\right] \lesssim \frac{L^2_{NT}}{2^{2j}} \cdot \mathbb{E}\left[\sup_{\theta \in S_j} |\mathbb{W}_{NT}(\theta)|\right] \lesssim \frac{L^2_{NT}}{2^{2j}} \cdot \frac{2^j}{L^2_{NT}} = 2^{-j},$$

which implies that

$$\sum_{j>V, 2^{j-1} \leq \eta L_{NT}} P\left[\inf_{\theta \in S_j} \mathbb{M}^*_{NT}(\theta) \leq 0\right] \lesssim \sum_{j>V} 2^{-j}.$$

The RHS of the previous expression convergences to 0 as $V \to \infty$, implying that $L_{NT} \cdot d(\hat{\theta}, \theta_0) = O_P(1)$, or $d(\hat{\theta}, \theta_0) = O_P(1/L_{NT})$. The desired result then follows from Lemma 2. □

A.2 Proof of Theorem 2

With a little abuse of notation, for $\theta_a \in \Theta^k$ and $\theta_b \in \Theta^r$, let

$$d(\theta_a, \theta_b) = \|\Lambda_a F'_a - \Lambda_b F'_b\| / \sqrt{NT}.$$

In this case, d is not a metric because θ_a and θ_b belong to different spaces. However, the following proofs will not be affected if we replace $d(\theta_a, \theta_b)$ by a different notation (say $\bar{d}(\theta_a, \theta_b)$).

For sufficiently small δ, define $\Theta^k(\delta) = \{\theta^k \in \Theta^k : d(\theta^k, \theta_0) \leq \delta\}$. Let $F^{k,r}$ denote the first r columns of F^k, and let $F^{k,-r}$ denote the remaining $k-r$ columns of F^k. $\Lambda^{k,r}$ and $\Lambda^{k,-r}$ are defined similarly.

Lemma 4. *Suppose that Assumption 1 holds and $r < k < \infty$. Then for any $\theta^k \in \Theta^k(\delta)$ and sufficiently small δ, we have*

$$\|F^{k,r} - F_0\mathsf{S}\|/\sqrt{T} \lesssim \delta, \quad \|\Lambda^{k,r} - \Lambda_0\mathsf{S}\|/\sqrt{N} \lesssim \delta, \quad \|\Lambda^{k,-r}\|/\sqrt{N} \lesssim \delta.$$

where $\mathsf{S} = \mathsf{sgn}((F^{k,r})'F_0/T)$.

Proof. The proof is similar to the proof of Lemma 2 and it is therefore omitted. □

For any $\theta^k \in \Theta^k$, write

$$\mathbb{M}^*_{NT}(\theta^k) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}[\rho_\tau(X_{it} - \lambda_i^{k'} f_t^k) - \rho_\tau(X_{it} - \lambda_{0i}' f_{0t})],$$

$$\bar{\mathbb{M}}^*_{NT}(\theta^k) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}[\rho_\tau(X_{it} - \lambda_i^{k'} f_t^k) - \rho_\tau(X_{it} - \lambda_{0i}' f_{0t})],$$

$$\mathbb{W}_{NT}(\theta^k) = \mathbb{M}^*_{NT}(\theta^k) - \bar{\mathbb{M}}^*_{NT}(\theta^k).$$

Lemma 5. *Suppose that Assumption 1 holds and $r < k < \infty$. For sufficiently small δ, we have:*

$$\mathbb{E}\left[\sup_{\theta^k \in \Theta^k(\delta)} \left|\mathbb{W}_{NT}(\theta^k)\right|\right] \lesssim \frac{\delta}{L_{NT}}.$$

Proof. The proof is similar to the proof of Lemma 3 and it is therefore omitted. □

Proof of Theorem 2:

Proof. First, similar to the proof of Lemma 1, we can show that $d(\hat{\theta}^k, \theta_0) = o_P(1)$. Second, like in the proof of Theorem 1, it follows from the previous lemma that

$$d(\hat{\theta}^k, \theta_0) = O_P(L_{NT}^{-1}). \tag{A.28}$$

Third, similar to the proof of Lemma 2 it can be shown that

$$|\hat{\sigma}^k_{N,j} - \sigma_j| = o_P(1) \text{ for } j = 1, \ldots, r. \tag{A.29}$$

Fourth, by Lemma 4 and (A.28),

$$\sum_{j=r+1}^{k} \hat{\sigma}^k_{N,j} = \|\hat{\Lambda}^{k,-r}\|^2/N \lesssim d(\hat{\theta}^k, \theta_0)^2 = O_P(L_{NT}^{-2}). \tag{A.30}$$

Finally, by (A.29) and (A.30), we have

$$P[\hat{r}_{\text{rank}} \neq r] = P[\hat{r}_{\text{rank}} < r] + P[\hat{r}_{\text{rank}} > r] \leq P[\hat{\sigma}^k_{N,r} \leq P_{NT}] + P[\hat{\sigma}^k_{N,r+1} > P_{NT}] = o(1). \tag{A.31}$$

It then follows that $P[\hat{r}_{\text{rank}} = r] \to 1$. $\square$

A.3 Proof of Theorem 3

Proof. Following the proof of Bai and Ng (2002), it suffices to show that for some $C > 0$,

$$\mathbb{M}_{NT}(\hat{\theta}^l) - \mathbb{M}_{NT}(\hat{\theta}^r) > C + o_P(1) \text{ for } l < r, \tag{A.32}$$

and

$$\mathbb{M}_{NT}(\hat{\theta}^l) - \mathbb{M}_{NT}(\hat{\theta}^r) = O_P(1/L_{NT}^2) \text{ for } l > r. \tag{A.33}$$

Case 1: Consider $l < r$. Adding and subtracting terms, we have

$$\mathbb{M}_{NT}(\hat{\theta}^l) - \mathbb{M}_{NT}(\hat{\theta}^r) = \mathbb{M}^*_{NT}(\hat{\theta}^l) - \mathbb{M}^*_{NT}(\hat{\theta}^r) = \bar{\mathbb{M}}^*_{NT}(\hat{\theta}^l) + \mathbb{W}_{NT}(\hat{\theta}^l) - \mathbb{M}^*_{NT}(\hat{\theta}^r).$$

As in the proof of Lemma 1, it can be shown that $|\mathbb{W}_{NT}(\hat{\theta}^l)| = o_P(1)$. Next, since $|\rho_\tau(X_{it} - \lambda_i' f_t) - \rho_\tau(X_{it} - \lambda_{0i}' f_{0t})| \lesssim |\lambda_i' f_t - \lambda_{0i}' f_{0t}|$, it follows from Lemma 1 that

$$\left|\mathbb{M}^*_{NT}(\hat{\theta}^r)\right| \lesssim \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} |\hat{\lambda}_i^{r\prime} \hat{f}_t^r - \lambda_{0i}' f_{0t}| \le d(\theta_0, \hat{\theta}^r) = o_P(1).$$

Thus, it remains to be shown that $\bar{\mathbb{M}}^*_{NT}(\hat{\theta}^l) \ge C$.

Note that, by Taylor expansion and Assumption 1(ii), it can be shown that $\bar{\mathbb{M}}^*_{NT}(\hat{\theta}^l) \gtrsim d^2(\hat{\theta}^l, \theta_0)$. Next, similar to (A.9) we can show that $\|M_{\hat{F}^l} F_0\|/\sqrt{T} \lesssim d(\hat{\theta}^l, \theta_0)$. It then follows that

$$\bar{\mathbb{M}}^*_{NT}(\hat{\theta}^l) \gtrsim \|M_{\hat{F}^l} F_0\|^2/T. \tag{A.34}$$

Note that

$$\|M_{\hat{F}^l} F_0\|^2/T = \text{Trace}\left[\mathbb{I}_r - F_0'\hat{F}^l\hat{F}^{l\prime}F_0/T^2\right] \ge \rho_{max}\left[\mathbb{I}_r - F_0'\hat{F}^l\hat{F}^{l\prime}F_0/T^2\right]. \tag{A.35}$$

By Lemma A.5 of Ahn and Horenstein (2013), we have

$$\rho_{max}\left[\mathbb{I}_r - F_0'\hat{F}^l\hat{F}^{l\prime}F_0/T^2\right] + \rho_{min}\left[F_0'\hat{F}^l\hat{F}^{l\prime}F_0/T^2\right] \ge \rho_{min}\left[\mathbb{I}_r\right]. \tag{A.36}$$

Since $F_0'\hat{F}^l\hat{F}^{l\prime}F_0$ is a $r \times r$ symmetric matrix with rank less or equal to l, we have $\rho_{min}\left[F_0'\hat{F}^l\hat{F}^{l\prime}F_0/T^2\right] = 0$, and the above inequality implies that

$$\rho_{max}\left[\mathbb{I}_r - F_0'\hat{F}^l\hat{F}^{l\prime}F_0/T^2\right] \ge 1. \tag{A.37}$$

Thus, (A.32) follows from (A.34) to (A.37).

Case 2: Now consider $l > r$. Adding and subtracting terms we can write

$$\mathbb{M}_{NT}(\hat{\theta}^l) - \mathbb{M}_{NT}(\hat{\theta}^r) = \mathbb{W}_{NT}(\hat{\theta}^l) - \mathbb{W}_{NT}(\hat{\theta}^r) + \bar{\mathbb{M}}^*_{NT}(\hat{\theta}^l) - \bar{\mathbb{M}}^*_{NT}(\hat{\theta}^r). \tag{A.38}$$

First, similar to the proof of Theorem 2, we can show that for sufficiently small δ,

$$\mathbb{E}\left[\sup_{d(\theta^l,\theta_0)\leq\delta}\left|\mathbb{W}_{NT}(\theta^l)\right|\right]\lesssim\frac{\delta}{L_{NT}},$$

and $d(\hat{\theta}^l,\theta_0)=O_P(1/L_{NT})$. It then follows that

$$\mathbb{W}_{NT}(\hat{\theta}^l)=O_P(1/L_{NT}^2). \tag{A.39}$$

Second, similar to the proof of Lemma 3 and Theorem 1 we can show that

$$\mathbb{W}_{NT}(\hat{\theta}^r)=O_P(1/L_{NT}^2). \tag{A.40}$$

Finally, consider $\bar{\mathbb{M}}^*_{NT}(\hat{\theta}^l)-\bar{\mathbb{M}}^*_{NT}(\hat{\theta}^r)$. Using a Taylor expansion and the assumption that $|\mathsf{f}_{it}(\cdot)|$ is uniformly (in i and t) bounded above, it can be shown that

$$\left|\bar{\mathbb{M}}^*_{NT}(\hat{\theta}^l)\right|\lesssim d^2(\hat{\theta}^l,\theta_0)\quad\text{and}\quad\left|\bar{\mathbb{M}}^*_{NT}(\hat{\theta}^r)\right|\lesssim d^2(\hat{\theta}^r,\theta_0).$$

It then follows from $d(\hat{\theta}^l,\theta_0)=O_P(1/L_{NT})$ and $d(\hat{\theta}^r,\theta_0)=O_P(1/L_{NT})$ that

$$\bar{\mathbb{M}}^*_{NT}(\hat{\theta}^l)-\bar{\mathbb{M}}^*_{NT}(\hat{\theta}^r)=O_P(1/L_{NT}^2). \tag{A.41}$$

Thus, (A.33) follows from (A.38), (A.39), (A.40) and (A.41). This concludes the proof. □

Appendix B: Tables and Figures

Table 1: AFM with Heavy-tailed Idiosyncratic Errors

(N,T)	Aver. $\hat{r}_{rank}$	$P[\hat{r}_{rank}=r]$	$f_1,\hat{F}_{KEN}$	$f_2,\hat{F}_{KEN}$	$f_3,\hat{F}_{KEN}$	$f_1,\hat{F}^{0.5}_{QFA}$	$f_2,\hat{F}^{0.5}_{QFA}$	$f_3,\hat{F}^{0.5}_{QFA}$
				$u_{it}\sim$ **i.i.d** t_3				
(50,50)	2.428	0.533	0.966	0.945	0.933	0.977	0.963	0.954
(50,100)	2.553	0.617	0.973	0.949	0.938	0.982	0.966	0.958
(50,200)	2.635	0.670	0.975	0.952	0.940	0.983	0.967	0.958
(100,50)	2.497	0.589	0.984	0.972	0.966	0.989	0.982	0.978
(100,100)	2.849	0.864	0.986	0.975	0.969	0.991	0.984	0.980
(100,200)	2.937	0.937	0.988	0.977	0.970	0.992	0.985	0.981
(200,50)	2.538	0.624	0.992	0.986	0.983	0.995	0.991	0.989
(200,100)	2.894	0.904	0.993	0.988	0.985	0.996	0.992	0.990
(200,200)	2.995	0.995	0.994	0.988	0.985	0.996	0.993	0.991
				$u_{it}\sim$ **i.i.d** t_2				
(50,50)	2.472	0.562	0.887	0.816	0.769	0.974	0.956	0.946
(50,100)	2.560	0.623	0.907	0.832	0.798	0.979	0.960	0.950
(50,200)	2.642	0.676	0.912	0.840	0.801	0.981	0.963	0.953
(100,50)	2.505	0.593	0.937	0.883	0.845	0.988	0.979	0.974
(100,100)	2.854	0.869	0.946	0.898	0.877	0.990	0.982	0.977
(100,200)	2.940	0.940	0.949	0.904	0.877	0.991	0.983	0.978
(200,50)	2.548	0.632	0.961	0.923	0.906	0.994	0.989	0.985
(200,100)	2.897	0.907	0.968	0.941	0.927	0.995	0.991	0.989
(200,200)	2.995	0.995	0.972	0.945	0.928	0.996	0.992	0.990
				$u_{it}\sim$ **i.i.d** t_1				
(50,50)	2.706	0.330	0.049	0.020	0.015	0.956	0.923	0.902
(50,100)	2.518	0.563	0.049	0.017	0.010	0.967	0.939	0.925
(50,200)	2.456	0.566	0.034	0.011	0.007	0.971	0.944	0.930
(100,50)	2.570	0.568	0.024	0.011	0.007	0.981	0.968	0.959
(100,100)	2.867	0.877	0.027	0.009	0.004	0.986	0.972	0.968
(100,200)	2.942	0.943	0.027	0.006	0.004	0.988	0.976	0.970
(200,50)	2.525	0.602	0.012	0.004	-0.001	0.991	0.983	0.980
(200,100)	2.903	0.911	0.013	0.003	0.001	0.994	0.988	0.985
(200,200)	2.994	0.994	0.011	0.003	0.001	0.994	0.989	0.986

Note: Simulation results of 1000 repetitions. The DGP considered in this Table is: $X_{it}=\sum_{j=1}^{3}\lambda_{ji}f_{jt}+u_{it}$, where $f_{1t}=0.8f_{1,t-1}+\epsilon_{1t},\ f_{2t}=0.5f_{2,t-1}+\epsilon_{2t},\ f_{3t}=0.2f_{3,t-1}+\epsilon_{3t},\ \lambda_{ji},\epsilon_{jt}\sim i.i.d\ \mathcal{N}(0,1),\ u_{it}\sim i.i.d\ t_\nu$ with $\nu=3$ (upper panel), $\nu=2$ (middle panel) and $\nu=1$ (lower panel). $\hat{F}^{0.5}_{QFA}$ and $\hat{F}_{KEN}$ denote the estimates of the factors by QFA at $\tau=0.5$ and the method proposed by He et al. (2020), respectively. Columns 2-3 report the average estimated number of factors and the frequency of choosing the right number of factors using our rank-minimization estimator; columns 4-6 report the average adjusted R^2 of regressing each of the true factors on $\hat{F}_{KEN}$; and the last three columns report the average adjusted R^2 of regressing each of the true factors on $\hat{F}_{KEN}$.

Table 2: Estimation of QFM

M1	$\tau = 0.25$				$\tau = 0.5$				$\tau = 0.75$			
(N,T)	$\hat{r}$	f_{1t}	f_{2t}	f_{3t}	$\hat{r}$	f_{1t}	f_{2t}	f_{3t}	$\hat{r}$	f_{1t}	f_{2t}	f_{3t}
(50,50)	2.19	0.917	0.686	0.340	1.88	0.960	0.808	0.010	2.19	0.917	0.694	0.348
(50,100)	2.32	0.936	0.700	0.413	1.90	0.968	0.863	0.001	2.33	0.939	0.705	0.413
(50,200)	2.51	0.934	0.732	0.521	1.94	0.972	0.892	0.001	2.49	0.933	0.722	0.511
(100,50)	2.25	0.943	0.751	0.403	1.91	0.980	0.856	0.004	2.25	0.944	0.740	0.402
(100,100)	2.68	0.973	0.850	0.680	1.97	0.985	0.948	0.002	2.69	0.974	0.868	0.678
(100,200)	2.86	0.981	0.915	0.797	2.00	0.987	0.971	0.000	2.88	0.981	0.910	0.811
(200,50)	2.29	0.962	0.774	0.456	1.89	0.989	0.853	0.009	2.26	0.956	0.758	0.445
(200,100)	2.75	0.985	0.884	0.766	1.98	0.993	0.967	0.001	2.75	0.984	0.894	0.766
(200,200)	2.99	0.992	0.980	0.933	2.00	0.993	0.987	0.000	2.99	0.991	0.981	0.937
M2	$\tau = 0.25$				$\tau = 0.5$				$\tau = 0.75$			
(N,T)	$\hat{r}$	f_{1t}	f_{2t}	f_{3t}	$\hat{r}$	f_{1t}	f_{2t}	f_{3t}	$\hat{r}$	f_{1t}	f_{2t}	f_{3t}
(50,50)	2.89	0.904	0.747	0.558	2.38	0.952	0.832	0.039	2.87	0.914	0.762	0.558
(50,100)	2.75	0.931	0.771	0.627	1.97	0.961	0.858	0.004	2.75	0.930	0.778	0.631
(50,200)	2.85	0.940	0.818	0.699	1.95	0.965	0.888	0.001	2.85	0.938	0.820	0.693
(100,50)	2.92	0.951	0.815	0.667	2.43	0.976	0.878	0.044	2.93	0.951	0.818	0.665
(100,100)	3.00	0.972	0.912	0.820	2.07	0.982	0.946	0.007	3.00	0.971	0.898	0.816
(100,200)	2.97	0.974	0.937	0.857	2.00	0.984	0.964	0.001	2.98	0.974	0.937	0.858
(200,50)	2.92	0.968	0.836	0.723	2.50	0.988	0.894	0.052	2.96	0.971	0.837	0.733
(200,100)	2.99	0.985	0.932	0.890	2.10	0.991	0.963	0.008	2.98	0.985	0.940	0.887
(200,200)	3.00	0.988	0.976	0.931	2.00	0.992	0.985	0.001	3.00	0.988	0.976	0.931
M3	$\tau = 0.25$				$\tau = 0.5$				$\tau = 0.75$			
(N,T)	$\hat{r}$	f_{1t}	f_{2t}	f_{3t}	$\hat{r}$	f_{1t}	f_{2t}	f_{3t}	$\hat{r}$	f_{1t}	f_{2t}	f_{3t}
(50,50)	2.47	0.903	0.686	0.373	2.14	0.953	0.809	0.025	2.45	0.912	0.692	0.369
(50,100)	2.52	0.931	0.722	0.420	1.98	0.959	0.848	0.005	2.53	0.934	0.721	0.421
(50,200)	2.68	0.936	0.766	0.491	1.95	0.961	0.882	0.002	2.66	0.942	0.755	0.483
(100,50)	2.47	0.939	0.758	0.467	2.05	0.976	0.854	0.017	2.45	0.948	0.746	0.475
(100,100)	2.86	0.975	0.891	0.680	2.04	0.981	0.946	0.005	2.85	0.974	0.892	0.673
(100,200)	2.97	0.980	0.943	0.742	2.00	0.983	0.963	0.000	2.95	0.979	0.934	0.735
(200,50)	2.48	0.958	0.774	0.554	1.98	0.986	0.865	0.018	2.47	0.961	0.774	0.554
(200,100)	2.89	0.987	0.927	0.798	2.00	0.990	0.965	0.002	2.89	0.988	0.926	0.799
(200,200)	3.00	0.990	0.979	0.866	2.00	0.992	0.984	0.000	3.00	0.990	0.979	0.864

Note: Simulation results from 1000 repetitions. The DGP considered in this Table is: $X_{it} = \lambda_{1i} f_{1t} + \lambda_{2i} f_{2t} + (\lambda_{3i} f_{3t}) \cdot e_{it}$, $f_{1t} = 0.8 f_{1,t-1} + \epsilon_{1t}$, $f_{2t} = 0.5 f_{2,t-1} + \epsilon_{2t}$, $f_{3t} = |g_t|$, $\lambda_{1i}, \lambda_{2i}, \epsilon_{1t}, \epsilon_{2t}, g_t \sim i.i.d\ \mathcal{N}(0,1)$, and $\lambda_{3i} \sim i.i.d\ U[1,2]$. $e_{it} = \beta e_{i,t-1} + v_{it} + \rho \cdot \sum_{j=i-J, j\neq i}^{i+J} v_{jt}$. **M1**: $v_{it} \sim i.i.d\ \mathcal{N}(0,1)$, $\beta = \rho = 0$; **M2**: $v_{it} \sim i.i.d\ t_3$, $\beta = \rho = 0$; **M3**: $v_{it} \sim i.i.d\ \mathcal{N}(0,1)$, $\beta = \rho = 0.2$, $J = 3$. For each $\tau \in \{0.25, 0.5, 0.75\}$, the first column reports the averages of the rank estimator $\hat{r}$, while the second to the fourth columns report the average adjusted R^2 in the regression of (each of) the true factors on the QFA factors $\hat{F}^{\tau}_{QFA}$, obtained from the IQR algorithm.

Table 3: Macro Forecasting: Comparison of $\hat{F}_{QFA}$ and $\hat{F}_{PCA}$

		Elements of $\hat{F}^{\tau}_{QFA}$					
τ	$\hat{r}_{rank}$	1	2	3	4	5	6
0.01	1	0.657					
0.05	2	0.733					
0.10	2	0.796	0.871				
0.25	4	0.952	0.932	0.939	0.890		
0.50	5	0.993	0.976	0.964	0.945	0.923	
0.75	5	0.906	0.945	0.943	0.903	0.882	
0.90	2	0.316	0.911				
0.95	1	0.261					
0.99	1	0.266					

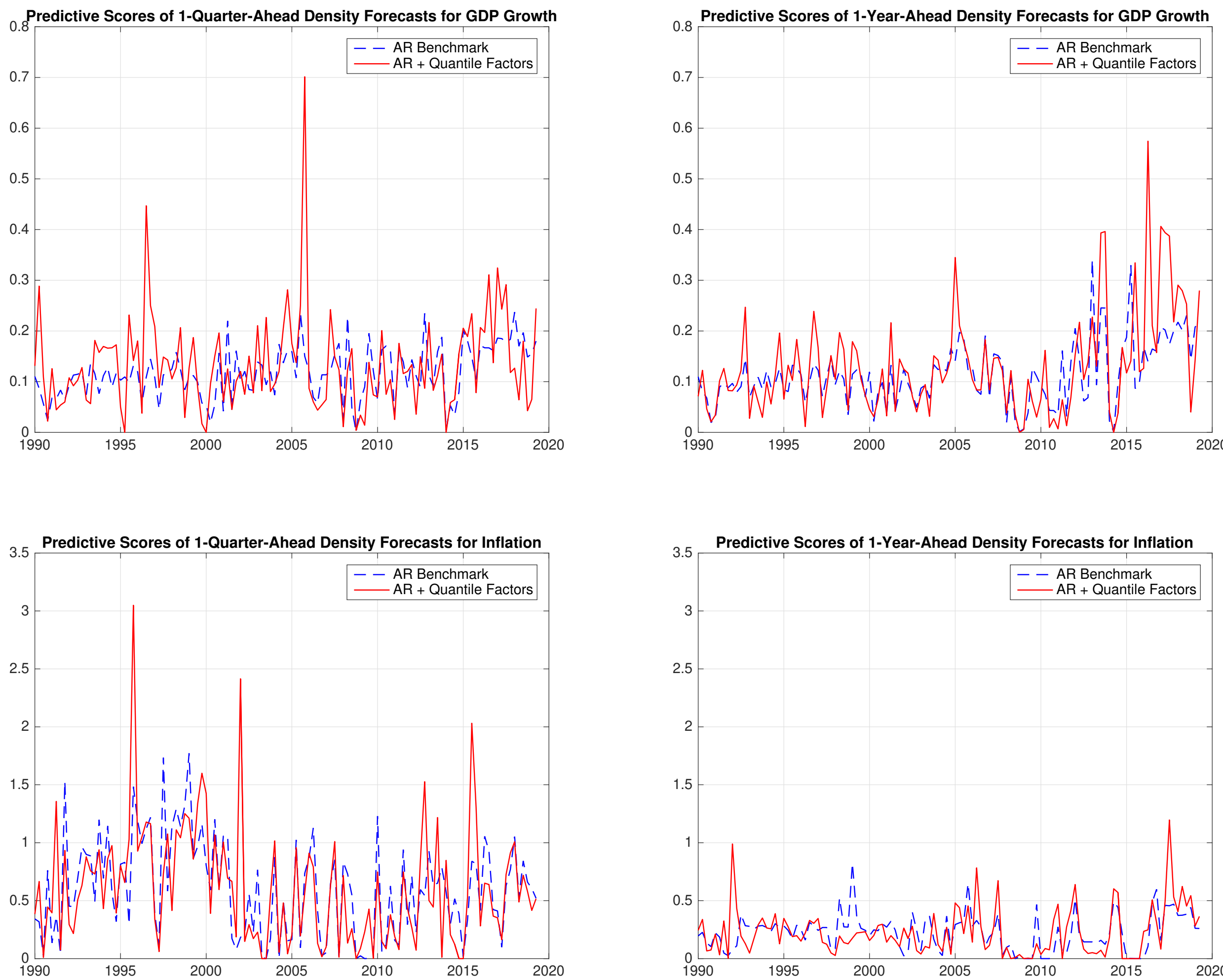
Note: The second column reports the estimated number of factors $\hat{r}_{rank}$ at each τ. The third to the last columns report the R^2 of regressing each column of $\hat{F}^{\tau}_{QFA}$ on $\hat{F}_{PCA}$. For $\hat{F}^{\tau}_{QFA}$ at each τ, the numbers of estimated factors is set to $\hat{r}_{rank}$, for $\hat{F}_{PCA}$, the number of estimated factors is set to 8.

Table 4: Macro Forecasting: MSE of Different Methods

	Pre-Crisis		Crisis- Post-Crisis		Full	
	MSE	R. MSE	MSE	R. MSE	MSE	R. MSE
			Real GDP Growth			
AR Benchmark	4.526	1.000	5.456	1.000	4.904	1.000
AR + $\hat{F}_{PCA}$	4.282	0.946	5.373	0.985	4.725	0.964
AR + $\hat{F}_{PCA}$ + $\hat{F}^{90}_{QFA}$	4.155	**0.918**	5.331	0.977	4.634	**0.945**
AR + $\hat{F}_{PCA}$ + $\hat{F}^{99}_{QFA}$	4.354	0.962	5.270	**0.966**	4.728	0.964
AR + $\hat{F}_{PCA}$ + $\hat{F}^{90}_{QFA}$ + $\hat{F}^{99}_{QFA}$	4.191	0.926	5.456	1.000	4.688	0.960
			Inflation			
AR Benchmark	0.266	1.000	0.790	1.000	0.479	1.000
AR + $\hat{F}_{PCA}$	0.246	0.926	0.732	0.926	0.444	0.926
AR + $\hat{F}_{PCA}$ + $\hat{F}^{90}_{QFA}$	0.246	0.926	0.732	0.926	0.444	0.927
AR + $\hat{F}_{PCA}$ + $\hat{F}^{99}_{QFA}$	0.247	0.927	0.739	0.935	0.447	0.932
AR + $\hat{F}_{PCA}$ + $\hat{F}^{90}_{QFA}$ + $\hat{F}^{99}_{QFA}$	0.245	**0.922**	0.726	**0.919**	0.441	**0.920**

Note: This Table reports the MSE of five alternative 1-quarter-ahead forecasting methods for real GDP growth and inflation, and their relative MSE (R. MSE) compared with the AR benchmark model (the lowest R. MSE are shown in bold characters). The out-of-sample forecasting is implemented using rolling windows with 120 observations. The full forecasting evaluation period is from 1990Q1 to 2019Q2, the pre-crisis period is from 1990Q1 to 2007Q2, and the crisis plus post-crisis period is from 2007Q3 to 2019Q2.

Figure 1: Macro Forecasting: Predictive Scores of Density Forecasts for GDP Growth and Inflation



Note: The graphs plot the predictive scores of 1-quarter-ahead and 1-year-ahead density forecasts for real GDP growth and inflation. The evaluation period is from 1990Q1 to 2019Q2, and the out-of-sample forecasting is implemented using rolling windows with 120 observations. The predicted τ-quantiles are constructed using quantile regressions of the target variable on its owns lags and the estimated quantile factors at τ (denoted as $\hat{F}^{\tau}_{QFA}$). The predicted densities are constructed as the density functions of skewed t-distributions by matching the predicted quantiles of the target variable at $\tau \in \{0.05, 0.25, 0.75, 0.95\}$. The predictive scores are the predicted densities evaluated at the realized values of the target variable. Higher scores indicate more accurate forecasts. The dotted blue line plots the predictive scores of the benchmark AR model where only the lags of the target variable are used to predict the τ-quantiles, while the red line plots the predictive scores of the model where $\hat{F}^{\tau}_{QFA}$ is also used to predict the τ-quantiles.

References

Adrian, T., N. Boyarchenko, and D. Giannone (2019). Vulnerable growth. *American Economic Review 109*(4), 1263–89.

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

Amengual, D. and E. Sentana (2020). Is a normal copula the right copula? *Journal of Business & Economic Statistics 38*(2), 350–366.

Ando, T. and J. Bai (2020). Quantile co-movement in financial markets: A panel quantile model with unobserved heterogeneity. *Journal of the American Statistical Association 115*(529), 266–279.

Ang, A., R. J. Hodrick, Y. Xing, and X. Zhang (2006). The cross-section of volatility and expected returns. *The Journal of Finance 61*(1), 259–299.

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

Athey, S. and G. W. Imbens (2019). Machine learning methods that economists should know about. *Annual Review of Economics 11*(1), 685–725.

Azzalini, A. and A. Capitanio (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(2), 367–389.

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

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

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

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

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

Bai, J. and S. Ng (2019). Rank regularized estimation of approximate factor models. *Journal of Econometrics 212*(1), 78–96.

Barigozzi, M. and M. Hallin (2016). Generalized dynamic factor models and volatilities: recovering the market volatility shocks. *The Econometrics Journal 19*(1), 33–63.

Candès, E. J., X. Li, Y. Ma, and J. Wright (2011). Robust principal component analysis? *Journal of the ACM 58*(3), 1–37.

Chamberlain, G. and M. Rothschild (1983). Arbitrage, factor structure in arbitrage pricing models. *Econometrica 51*(5), 1281–1304.

Chen, L., J. J. Dolado, and J. Gonzalo (2020). Quantile factor models. *arXiv preprint arXiv:1911.02173*.

Chen, M., I. Fernández-Val, and M. Weidner (2020). Nonlinear factor models for network and panel data. *Journal of Econometrics (forthcoming)*.

Chen, X., L. P. Hansen, and J. Scheinkman (2009). Nonlinear principal components and long-run implications of multivariate diffusions. *The Annals of Statistics 37*(6B), 4279–4312.

Chudik, A., G. Kapetanios, and M. H. Pesaran (2018). A one covariate at a time, multiple testing approach to variable selection in high-dimensional linear regression models. *Econometrica 86*(4), 1479–1512.

Collins, M., S. Dasgupta, and R. E. Schapire (2001). A generalization of principal component analysis to the exponential family. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, Cambridge, MA, USA. MIT Press.

de Castro, L. and A. F. Galvao (2019). Dynamic quantile models of rational behavior. *Econometrica 87*(6), 1893–1939.

Fan, J., H. Liu, and W. Wang (2018). Large covariance estimation through elliptical factor models. *Annals of statistics 46*(4), 1383.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and statistics 82*(4), 540–554.

Franklin, J. N. (2012). *Matrix theory.* Courier Corporation.

Galvao, A. F. and K. Kato (2016). Smoothed quantile regression for panel data. *Journal of Econometrics 193*(1), 92–112.

Golub, G. H. and C. F. Van Loan (2013). *Matrix Computations*, Volume 3. JHU Press.

Gorodnichenko, Y. and S. Ng (2017). Level and volatility factors in macroeconomic data. *Journal of Monetary Economics 91*, 52–68.

He, Y., X. Kong, L. Yu, and X. Zhang (2020). Large-dimensional factor analysis without moment constraints. *Journal of Business & Economic Statistics* (forthcoming), 1–31.

Herskovic, B., B. Kelly, H. Lustig, and S. Van Nieuwerburgh (2016). The common factor in idiosyncratic volatility: Quantitative asset pricing implications. *Journal of Financial Economics 119*(2), 249–283.

Horowitz, J. (1998). Bootstrap methods for median regression models. *Econometrica 66*(6), 1327–1352.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology 24*(6), 417.

Koenker, R. (2005). *Quantile Regression.* Number 38. Cambridge University Press.

Ma, S., O. Linton, and J. Gao (2020). Estimation and inference in semiparametric quantile factor models. *Journal of Econometrics (forthcoming).*

McCracken, M. W. and S. Ng (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics 34*(4), 574–589.

Muller, H. G. (1984). Smooth optimum kernel estimators of densities, regression curves and modes. *The Annals of Statistics 12*(2), 766–774.

Pelger, M. and R. Xiong (2018). Interpretable proximate factors for large dimensions. *arXiv preprint arXiv:1805.03373*.

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica 74*(4), 967–1012.

Powell, J. L. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics 25*(3), 303–325.

Renault, E., T. Van Der Heijden, and B. Werker (2017). Arbitrage pricing theory for idiosyncratic variance factors. *Working paper, Brown University*.

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

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

Su, L. and X. Wang (2017). On time-varying factor models: Estimation and testing. *Journal of Econometrics 198*(1), 84–101.

van de Geer, S. A. (2002). On Hoeffding's inequality for dependent random variables. In *Empirical process techniques for dependent data*, pp. 161–169. Springer.

van der Vaart, A. and J. Wellner (1996). *Weak convergence and empirical processes.* Springer, New York.

Vidal, R., Y. Ma, and S. Sastry (2016). *Generalized Principal Component Analysis*, Volume 40. Springer.

White, H., T.-H. Kim, and S. Manganelli (2015). VAR for VaR: Measuring tail dependence using multivariate regression quantiles. *Journal of Econometrics 187*(1), 169–188.