We propose a generalization of the linear panel quantile regression model to accommodate both sparse and dense parts: sparse means that while the number of covariates available is large, potentially only a much smaller number of them have a nonzero impact on each conditional quantile of the response variable; while the dense part is represented by a low-rank matrix that can be approximated by latent factors and their loadings. Such a structure poses problems for traditional sparse estimators, such as the $\ell_1$-penalized Quantile Regression, and for traditional latent factor estimators such as PCA. We propose a new estimation procedure, based on the ADMM algorithm, that consists of combining the quantile loss function with $\ell_1$ and nuclear norm regularization. We show, under general conditions, that our estimator can consistently estimate both the nonzero coefficients of the covariates and the latent low-rank matrix. This is done in a challenging setting that allows for temporal dependence, heavy-tail distributions, and the presence of latent factors.

Our proposed model has a “Characteristics + Latent Factors” Quantile Asset Pricing Model interpretation: we apply our model and estimator with a large-dimensional panel of financial data and find that (i) characteristics have sparser predictive power once latent factors were controlled (ii) the factors and coefficients at upper and lower quantiles are different from the median.

1. Introduction. A central question in asset pricing is to explain why certain assets pay higher returns than others. The Arbitrage Pricing Theory ([69]) and the Fama-French three factors model ([37]) explains the asset return variations by a linear combination of common risk factors. Assets with similar exposure to a common factor shall rise and fall together ([29]). However, empirical evidence appears to indicate that the firm characteristics, rather than common factors, can also explain the variations in stock returns ([33]), which suggests a characteristic-based model.

We generalize both modeling approaches and propose a “Characteristics + Latent Factors” quantile asset pricing framework. By incorporating the “Characteristics”, we improve the economic interpretability and explanatory power of the model. On the other hand, the finance literature has documented a zoo of new characteristics, and the proliferation of characteristics in this “variable zoo” leads to a concern about which characteristics really provide independent information about returns ([30]). Our model addresses this issue by imposing a sparse structure, meaning although a large set of characteristics is available, only a much smaller subset of them might have predictive power. We also incorporate “Latent Factors” to
capture the common variations in asset returns. One additional benefit of having this part is that it might help alleviate the "omitted variable bias" problem ([45]). As in the literature, typically, these latent factors are estimated via principal component analysis, which means all possible latent explanatory variables might be important for prediction although their individual contribution might be small, we term this as the dense part. Hence, our framework allows for "Sparse + Dense" modeling with large scale panel data that consist of a large number of asset returns that are allowed to be weakly correlated across time. In addition, we focus on understanding the quantiles (hence the entire distribution) of returns rather than just the mean, in line with the recent interest in quantile factor models (e.g. [3], [25], [61], [39], and [71]). Our quantile asset pricing framework also inherits micro-foundation from the seminal quantile preference framework([62, 70, 46]) and, in particular, the dynamic quantile preference framework of [35].

Specifically, with \( Y_{i,t} \) as the excess return of asset \( i \) in period \( t \), \( X_{i,t} \) as a \( p \)-dimensional vector of observable characteristics such as return volatility and trading volume, we study the following high dimensional latent panel quantile regression model:

\[
F_{Y_{i,t}}^{-1}(\tau; \theta(\tau), \lambda_\tau, g_\tau) = X_{i,t}'\theta(\tau) + \lambda_\tau g_\tau, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T,
\]

where \( \theta(\tau) \in \mathbb{R}^p \) is the vector of coefficients, \( g_\tau \) is an \( r_\tau \)-dimensional vector of unobservable factor returns, \( \lambda_\tau \) represents the factor loadings which captures the sensitivity of asset \( i \) on the \( r_\tau \) factors, \( \tau \in [0, 1] \) is the quantile index. We allow for the possibility of quantile dependence of sensitivity to risk factors as such evidence has been reported in the literature (e.g., [3]). For notation simplicity, we often denote \( \Pi_{i,t}(\tau) = \lambda_\tau g_\tau \), then \( \Pi(\tau) \) is a low-rank matrix with unknown rank \( r_\tau \). Thus, with \( F_{Y_{i,t}}^{-1}|X_{i,t}, \theta(\tau), \lambda_\tau, g_\tau \) is the cumulative distribution function of \( Y_{i,t} \) conditioning on \( X_{i,t}, \theta(\tau) \) and \( \lambda_\tau, g_\tau \), we model the quantiles of returns (instead of expected returns) as a linear combination of the characteristics and latent factors. Our framework allows for the possibility of lagged dependent data. Here, we allow for the number of characteristics \( p \), and the time horizon \( T \), to grow to infinity as \( n \) grows. Throughout, we focus on the case where \( p \) is large, possibly much larger than \( nT \), but for the true model \( \theta(\tau) \) is sparse and has only \( s_\tau \ll p \) non-zero components.

Our framework is flexible enough that allows us to jointly answer the following three questions in asset pricing: (i) Which characteristics are important to explain the time series and cross-section of stock returns, after controlling for the factors? (ii) How much would the latent factors explain stock returns after controlling for firm characteristics? (iii) Does the relationship of stock returns and firm characteristics change across quantiles? The first question is related to the recent literature on variable selection in asset pricing using machine learning ([58, 40, 49]). The second question is related to an classical literature starting from 1980s on statistical factor models of stock returns ([22, 31] and recently [60]). The third question extends the literature in late 1990s on stock return and firm characteristics ([33, 34]) and further asks whether the relationship is heterogeneous across quantiles.

There are several key features of considering prediction problem at the panel quantile model in this setting. First, stock returns are known to be asymmetric and exhibit heavy tail, thus modeling different quantiles of return provides extra information in addition to models of first and second moments. Second, quantile regression provides a richer characterization of the data, allowing heterogeneous relationship between stock returns and firm characteristics across the entire return distribution. Third, the latent factors might also be different at different quantiles of stock returns. Finally, quantile regression is more robust to the presence of outliers relative to other widely used mean-based approaches. Using a robust method is crucial when estimating low-rank structures (see e.g. [72]). As our framework is based on

---

1 More about sparse modeling and dense modeling can be found in [44]. See also [27].
modeling the quantiles of the response variable, we do not put assumptions directly on the moments of the dependent variable.

Our main goal is to consistently estimate both the sparse part and the low-rank matrix. Recovery of a low-rank matrix, when there are additional high dimensional covariates, in a nonlinear model can be very challenging. The rank constraint will result in the optimization problem NP-hard. In addition, estimation in high dimensional regression is known to be a challenging task, which in our frameworks becomes even more difficult due to the additional latent structure. We address the former challenge via nuclear norm regularization which is similar to [20] in the matrix completion setting. Without covariates, the estimation can be done via solving a convex problem, and similarly there are strong statistical guarantees of recovery of the underlying low-rank structure. We address the latter challenge by imposing \( \ell_1 \) regularization on the vector of coefficients of the control variables, similarly to [13] which mainly focused on the cross-sectional data setting. Note that with regards to sparsity, we must be cautious, specially when considering predictive models ([72]). Furthermore, we explore the performance of our procedure under settings where the vector of coefficients can be dense (due to the low-rank matrix).

We view our work as complementary to the low dimensional quantile regression with interactive fixed effects framework as of the very recent work of [39], and the mean estimation setting in [64]. However, unlike [64] and [39], we allow the number of covariates \( p \) to be large, perhaps \( p \geq nT \). This comes with different significant challenges. On the computational side, it requires us to develop novel estimation algorithms, which turns out can also be used for the contexts in [64] and [39]. On the theoretical side, allowing \( p \gg nT \) requires a systematically different analysis as compared to [39], as it is known that ordinary quantile regression is inconsistent in high dimensional settings \( (p \gg nT) \), see [13].

\textit{Related Literature.} Our work contributes to the recent growing literature on panel quantile model. [1], [47], [4], considered the fixed \( T \) asymptotic case. [51] formally derived the asymptotic properties of the fixed effect quantile regression estimator under large \( T \) asymptotics, and [42] further proposed fixed effects smoothed quantile regression estimator. [41] works on dynamic panel. [53] proposed a penalized estimation method where the individual effects are treated as pure location shift parameters common to all quantiles, for other related literature see [59], [43]. We refer to Chapter 19 of [56] for a review. Furthermore, our framework can be viewed as a generalization of the model in [3] which considered panel quantile model with independent errors and low dimensional covariates.

Our work also contributes to the literature on nuclear norm penalisation, which has been widely studied in the machine learning and statistical learning literature, [38], [67, 57, 68], [65], [16]. Recently, in the econometrics literature [5] proposes a framework of matrix completion for estimating causal effects, [10] for estimating approximate factor model, [28] considered the heterogeneous coefficients version of the linear panel data interactive fixed model where the main coefficients has a latent low-rank structure, [7] for robust principal component analysis, and [11] for imputing counterfactual outcome.

Finally, our results contribute to a growing literature on high dimensional quantile regression. [75] considered quantile regression with concave penalties for ultra-high dimensional data; [80] proposed an adaptively weighted \( \ell_1 \) penalty for globally concerned quantile regression. Screening procedures based on moment conditions motivated by the quantile models have been proposed and analyzed in [50] and [77] in the high-dimensional regression setting. We refer to [56] for a review.

To sum-up, our paper makes the following contributions. First, we propose a new class of models that consist of both high dimensional regressors and latent factor structures. We provide a scalable estimation procedure, and show that the resulting estimator is consistent under suitable regularity conditions. Second, the high dimensional and non-smooth objective
function require innovative strategies to deriving the above-mentioned results. In particular, our paper allows for serial dependence and this flexibility is important for the panel data case. It is well known that dealing with data dependence is a non-trivial problem, and there are additional challenges for high dimensional models even for those without incorporating the latent factors: the extension of lasso-based methods with relaxing the i.i.d. assumption and other restrictive assumptions (for instance, Gaussianity), is just beginning to occur, e.g. for time series data some recent development along this line can be found in [76]. Those lead to the novel use in our proofs of some techniques from the high dimensional statistics and econometrics literature, such as the localization argument from [13], and the new loss function introduced in [66]; from spectral theory, namely, properties of nuclear norm studied in [36], and concentration results by [23]; and from empirical process theory [78, 74]. We also generalize the sampling and smoothness assumption of [13] by considering panel data with weak correlation across time. In particular, we refer readers to [78] for thorough discussions on $\beta$-mixing.

On the theoretical side, we also present multiple results that entirely differ from those in [13]. In addition to allowing time dependence and a latent factor structure, we can consistently estimate the conditional quantiles without requiring a minimum eigenvalue condition on the behavior of the design matrix, which is typically required in the literature (e.g. [13]).

Relative to approaches that incorporate latent factor structure but relies on the squared loss, the proposed estimators inherit from quantile regression certain robustness properties to the presence of outliers and heavy-tailed distributions. Finally, we apply our proposed model and estimator to a large-dimensional panel of financial data in the US stock market and find that different return quantiles have different selected firm characteristics and that the number of latent factors can be also be different.

Outline. The rest of the paper is organized as follows. Section 2 introduces the high dimensional latent quantile regression model, and provides an overview of the main theoretical results. Section 3 presents the estimator and our proposed ADMM algorithm. Section 4 discusses the statistical properties of the proposed estimator. Section 5 provides simulation results. Section 6 consists of the empirical results of our model applied to a real data set. The proofs of the main results are in the Supplementary Material.

Notation. For $m \in \mathbb{N}$, we write $[m] = \{1, \ldots, m\}$. For a vector $v \in \mathbb{R}^p$ we define its $\ell_0$ norm as $\|v\|_0 = \sum_{j=1}^p 1\{v_j \neq 0\}$, where $1\{\cdot\}$ takes value 1 if the statement inside $\{\} is true, and zero otherwise; its $\ell_1$ norm as $\|v\|_1 = \sum_{j=1}^p |v_j|$ We denote $\|v\|_{1,n,T} = \sum_{j=1}^p \hat{\sigma}_j |v_j|$ the $\ell_1$-norm weighted by $\hat{\sigma}_j$’s (defined in eq(14)). The Euclidean norm is denoted by $\| \cdot \|$, thus $\|v\| = \sqrt{\sum_{j=1}^p v_j^2}$. If $A \in \mathbb{R}^{n \times T}$ is a matrix, its Frobenius norm is denoted by $\|A\|_F = \sqrt{\sum_{i=1}^n \sum_{t=1}^T A_{i,t}^2}$, its spectral norm by $\|A\|_2 = \sup_{x : \|x\|=1} \sqrt{x' A' A x}$, its infinity norm by $\|A\|_{\infty} = \max \{ |A_{i,j}| : i \in [n], j \in [T] \}$, its rank by rank($A$), and its nuclear norm by $\|A\|_* = \text{trace}(\sqrt{A' A})$ where $A'$ is the transpose of $A$. The $j$th column $A_j$ is denoted by $A_{:j}$. Furthermore, the multiplication of a tensor $X \in \mathbb{R}^{I_1 \times \cdots \times I_m}$ with a vector $\theta \in \mathbb{R}^{I_m}$ is denoted by $Z := X \theta \in \mathbb{R}^{I_1 \times \cdots \times I_{m-1}}$, and, explicitly, $Z_{i_1,\ldots,i_{m-1}} = \sum_{j=1}^{I_m} X_{i_1,\ldots,i_{m-1},j} \theta_j$. We also use the notation $a \vee b = \max\{a,b\}$, $a \wedge b = \min\{a,b\}$, $(a)_+ = \max\{-a,0\}$. For a sequence of random variables $\{z_j\}_{j=1}^\infty$, we denote by $\sigma(z_1, z_2, \ldots)$ the sigma algebra generated by $\{z_j\}_{j=1}^\infty$. Finally, for sequences $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$ we write $a_n \asymp b_n$ if there exists positive constants $c_1$ and $c_2$ such that $c_1 b_n \leq a_n \leq c_2 b_n$ for sufficiently large $n$.

2. The Estimator and Overview of Rate Results.
2.1. Basic Setting. The setting of interest corresponds to a high dimension latent panel quantile regression model, where \( Y \in \mathbb{R}^{n \times T} \), and \( X \in \mathbb{R}^{n \times T \times p} \) satisfying

\[
F_{Y_{i,t} | X_{i,t}; \theta(\tau), \Pi_{i,t}(\tau)}^{-1}(\tau) = X_{i,t}' \theta(\tau) + \Pi_{i,t}(\tau), \quad i = 1, \ldots, n, \ t = 1, \ldots, T,
\]

where \( i \) denotes subjects, \( t \) denotes time, \( \theta(\tau) \in \mathbb{R}^p \) is the vector of coefficients, \( \Pi(\tau) \in \mathbb{R}^{n \times T} \) is a low-rank matrix with unknown rank \( r_\tau \ll \min\{n, T\} \), \( \tau \in [0, 1] \) is the quantile index, and \( F_{Y_{i,t} | X_{i,t}; \theta(\tau), \Pi_i(\tau)} \) is the cumulative distribution function of \( Y_{i,t} \) conditioning on \( X_{i,t}, \theta(\tau) \) and \( \Pi_{i,t}(\tau) \). Thus, we model the quantile function at level \( \tau \) as a linear combination of the predictors plus a low-rank matrix. Here, we allow for the number of covariates \( p \) and the time horizon \( T \), to grow to infinity as \( n \) grows. Throughout the paper the quantile index \( \tau \in (0, 1) \) is fixed. We mainly focus on the case where \( p \) is large, possibly much larger than \( nT \), but for the true model \( \theta(\tau) \) is sparse and has only \( s_\tau \ll p \) non-zero components.

Mathematically, \( s_\tau := \|\theta(\tau)\|_0 \).

When \( \Pi_{i,t}(\tau) = \lambda_i(\tau)' g_t(\tau) \), with \( \lambda_i(\tau), \ g_t(\tau) \in \mathbb{R}^r \), this immediately leads to the following setting

\[
F_{Y_{i,t} | X_{i,t}; \theta(\tau), \Pi_{i,t}(\tau)}^{-1}(\tau) = X_{i,t}' \theta(\tau) + \lambda_i(\tau)' g_t(\tau),
\]

where we model the quantile function at level \( \tau \) as a linear combination of the covariates (as predictors) plus a latent factor structure. This is directly related to the panel data models with interactive fixed effects literature in econometrics, e.g. linear panel data model ([6]), nonlinear panel data models ([24, 26]).

Note, for eq (3), additional identification restrictions are needed for estimating \( \lambda_i(\tau) \) and \( g_t(\tau) \) (see [9]). In addition, in nonlinear panel data models, this creates additional difficult in estimation, as the latent factors and their loadings part induce a nonconvex quantile regression problem. However, we deal with this in the following subsection via using a nuclear norm constraint. \(^2\)

2.2. Estimator. In this subsection, we describe the high dimensional latent quantile estimator. With the sparsity and low-rank constraints in mind, a natural formulation for the estimation of \( (\theta(\tau), \Pi(\tau)) \) is

\[
\begin{align*}
\min_{\theta \in \mathbb{R}^p, \ \Pi \in \mathbb{R}^{n \times T}} & \quad \frac{1}{nT} \sum_{t=1}^{T} \sum_{i=1}^{n} \rho_\tau(Y_{i,t} - X_{i,t}' \tilde{\theta} - \tilde{\Pi}_{i,t}) \\
\text{subject to} & \quad \|\theta\|_0 = s_\tau,
\end{align*}
\]

where \( \rho_\tau(t) = (\tau - 1\{t \leq 0\})t \) is the quantile loss function as in [54], \( s_\tau \) is a parameter that directly controls the sparsity of \( \tilde{\theta} \), and \( r_\tau \) controls the rank of the estimated latent matrix.

While the formulation in (4) seems appealing, as it enforces variable selection and low-rank matrix estimation simultaneously, (4) is a non-convex problem due to the constraints posed by the \( \| \cdot \|_0 \) and \( \text{rank}(\cdot) \) functions. We propose a convex relaxation of (4). Inspired by the seminal works of [73] and [20], we formulate the problem as the following

\[
\begin{align*}
\min_{\theta \in \mathbb{R}^p, \ \Pi \in \mathbb{R}^{n \times T}} & \quad \left\{ \frac{1}{nT} \sum_{t=1}^{T} \sum_{i=1}^{n} \rho_\tau(Y_{i,t} - X_{i,t}' \tilde{\theta} - \tilde{\Pi}_{i,t}) + \nu_1 \sum_{j=1}^{p} |w_j| |\tilde{\theta}_j| + \nu_2 \|\tilde{\Pi}\|_* \right\}
\end{align*}
\]

\(^2\)Different identification conditions might result in different estimation procedures for \( \lambda \) and \( f \), see [8] and [24].
where \( \nu_1 > 0 \) and \( \nu_2 > 0 \) are tuning parameters, and \( w_1, \ldots, w_p \) are user specified weights (more on this in Section 4). Notice that \( \| \cdot \|_* \) is the nuclear norm defined on Page 4. The nuclear norm regularization works on the singular value of a matrix, the intuition is that via penalization with the nuclear norm, the resulting problem will be convex. Just as \( \ell_1 \)-minimization is the tightest convex relaxation of the combinatorial \( \ell_0 \)-minimization problem, nuclear-norm minimization is the tightest convex relaxation of the NP-hard rank minimization problem, see [18].

In principle, one can use any convex solver software to solve (5), since this is a convex optimization problem. However, for large scale problems a more careful implementation might be needed. Section 3 presents a scheme for solving (5) that is based on the ADMM algorithm ([15]).

2.3. Summary of results. We now summarize our main results. For the model defined in (2):

- Under (2), \( s_\tau \ll \min \{ n, T \} \), an assumption that implicitly requires \( r_\tau \ll \min \{ n, T \} \), and other regularity conditions defined in Section 4, we show that our estimator \((\hat{\theta}(\tau), \hat{\Pi}(\tau))\) defined in Section 3 is consistent for \((\theta(\tau), \Pi(\tau))\). Specifically, for the independent data case (across \( i \) and \( t \)), under suitable regularity conditions that can be found in Section 4, we have

\[
\| \hat{\theta}(\tau) - \theta(\tau) \| = O_P \left( \max \{ \sqrt{\log p}, \sqrt{\log n} \} \left( s_\tau + r_\tau \right) \left( \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{T}} \right) \right).
\]

and

\[
\frac{1}{nT} \| \hat{\Pi}(\tau) - \Pi(\tau) \|^2 = O_P \left( \max \{ \log p, \log n \} (s_\tau + r_\tau) \left( \frac{1}{n} + \frac{1}{T} \right) \right),
\]

Importantly, the rates in (6) and (7), up to logarithmic factor, match those in previous works. However, our setting allows for modeling at different quantile levels. We also complement our results by allowing for the possibility of lagged dependent data. Specifically, under a \( \beta \)-mixing assumption, Theorem 4.2 provides a statistical guarantee for estimating \((\theta(\tau), \Pi(\tau))\). This result can be thought as a generalization of the statements in (6) and (7).

Let \( q_{i,t} = \mathcal{X}_{i,t}' \theta_{i,t}(\tau) + \Pi_{i,t}(\tau) \) for \( t \in \{ 1, \ldots, T \} \) and \( i \in \{ 1, \ldots, n \} \) be the conditional quantiles. We show that, under weaker conditions than the ones needed for Theorem 4.2, our estimates \( \{ \hat{q}_{i,t} \} \) satisfy

\[
\frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \min \{ |q_{i,t} - \hat{q}_{i,t}|, (q_{i,t} - \hat{q}_{i,t})^2 \} = O_P \left( \left( \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{T}} \right) \left( \| \Pi(\tau) \|_* + \sqrt{\log \max \{ n, p \}} \| \beta(\tau) \|_1 \right) \right),
\]

for the independent data case (across \( i \) and \( t \)). This is a particular instance of Theorem 4.1 which allows the possibility of time dependence.

- An important aspect of our analysis is that we contrast the performance of our estimator in settings where the possibility of a dense \( \theta(\tau) \) provided that the features are highly correlated. We show that there exist choices of the tuning parameters for our estimator that lead to consistent estimation.

- For estimation, we provide an efficient algorithm (details can be found in Section 3), which is based on the ADMM algorithm ([15]).

- Section 6 provides thorough examples on financial data that illustrate the flexibility and interpretability of our approach.
Although our theoretical analysis builds on the work by [13], there are multiple challenges that we must face in order to prove the consistency of our estimator. First, the construction of the restricted set now involves the nuclear norm penalty. This requires us to define a new restricted set that captures the contributions of the low-rank matrix. Second, when bounding the empirical processes that naturally arise in our proof, we have to simultaneously deal with the sparse and dense components. Furthermore, throughout our proofs, we have to carefully handle the weak dependence assumption that can be found in Section 4.

3. High Dimensional Latent Panel Quantile Regression. In this subsection, we describe the main steps of our proposed ADMM algorithm, details can be found in Section S1. We start by introducing slack variables to the original problem (5). As a result, a problem equivalent to (5) is

$$
\begin{align*}
\min_{Z_{\theta}, Z_{\Pi}, W} & \quad \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \rho_{r}(V_{i,t}) + \nu_1 \sum_{j=1}^{p} w_j |Z_{\theta_j}| + \nu_2 \|\Pi\|_* \\
\text{subject to} & \quad V = W, \quad W = Y - X\hat{\theta} - Z_{\Pi}, \\
& \quad Z_{\Pi} - \Pi = 0, \quad Z_{\theta} - \hat{\theta} = 0.
\end{align*}
$$

(8)

To solve (8), we propose a scaled version of the ADMM algorithm which relies on the following Augmented Lagrangian

$$
\mathcal{L}(\hat{\theta}, \hat{\Pi}, V Z_{\theta}, Z_{\Pi}, W, U_V, U_W, U_{\Pi}, U_{\theta}) = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \rho_{r}(V_{i,t}) + \nu_1 \sum_{j=1}^{p} w_j |Z_{\theta_j}| + \nu_2 \|\Pi\|_* \\
+ \frac{\eta}{2} \|V - W + U_V\|_F^2 + \frac{\eta}{2} \|W - Y + X\theta + Z_{\Pi} + U_W\|_F^2 \\
+ \frac{\eta}{2} \|Z_{\Pi} - \Pi + U_{\Pi}\|_F^2 + \frac{\eta}{2} \|Z_{\theta} - \hat{\theta} + U_{\theta}\|_F^2,
$$

(9)

where $\eta > 0$ is a penalty parameter.

Notice that in (9), we have followed the usual construction of ADMM via introducing the scaled dual variables corresponding to the constraints in (8) – those are $U_V$, $U_W$, $U_{\Pi}$, and $U_{\theta}$. Next, recall that ADMM proceeds by iteratively minimizing the Augmented Lagrangian in blocks with respected to the original variables, in our case $(V, \hat{\theta}, \hat{\Pi})$ and $(W, Z_{\theta}, Z_{\Pi})$, and then updating the scaled dual variables (see Equations 3.5–3.7 in [15]). The explicit updates can be found in the Supplementary Material. Here, we highlight the updates for $Z_{\theta}$, $Z_{\Pi}$, and $V$. For updating $Z_{\theta}$ at iteration $k + 1$, we solve the problem

$$
Z_{\theta}^{(k+1)} \leftarrow \arg\min_{Z_{\theta} \in \mathbb{R}^p} \left\{ \frac{1}{2} \|Z_{\theta} - \hat{\theta}^{(k+1)} + U_{\theta}^{(k)}\|_F^2 + \frac{\nu_1}{\eta} \sum_{j=1}^{p} w_j |(Z_{\theta})_j| \right\}.
$$

This can be solved in closed form exploiting the well known thresholding operator, see the details in Section S2.2. As for updating $\hat{\Pi}$, we solve

$$
\hat{\Pi}^{(k+1)} \leftarrow \arg\min_{\Pi \in \mathbb{R}^{n \times T}} \left\{ \frac{\nu_2}{\eta} \|\Pi\|_* + \frac{1}{2} \|Z_{\Pi}^{(k)} - \hat{\Pi} + U_{\Pi}^{(k)}\|_F^2 \right\},
$$

(10)

via the singular value shrinkage operator, see Theorem 2.1 in [17].

Furthermore, we update $V$, at iteration $k + 1$, via

$$
V^{(k+1)} \leftarrow \arg\min_{V \in \mathbb{R}^{n \times T}} \left\{ \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \rho_{r}(V_{i,t}) + \frac{\eta}{2} \|V - W^{(k)} + U_{V}^{(k)}\|_F^2 \right\},
$$

(11)

which can be found in closed formula by Lemma 5.1 from [2].
Remark 1. After estimating $\Pi(\tau)$, we can estimate $\lambda_i(\tau)$ and $g_i(\tau)$ via the singular value decomposition of $\hat{\Pi}(\tau)$ and following equation

$$\hat{\Pi}(\tau)_{i,t} = \lambda_i(\tau)'g_i(\tau),$$

where $\lambda_i(\tau)$ and $g_i(\tau)$ are of dimension $\hat{r}_\tau$. This immediately leads to factors and loadings estimated that can be used to obtain insights about the structure of the data. A formal identification statement is given in Corollary 4.3.

Finally, it is immediate to modify the proposed ADMM to the case when there are no covariates ($\beta(\tau) = 0$), see Section S1.1 in the Supplementary Material. Hence, our proposed estimation procedure can be applied to settings (i) with low dimensional covariates, or (ii) without covariates. However, in what follows, we focus on the high dimensional covariates setting.

4. Theory. The purpose of this section is to provide statistical guarantees for the estimator developed in the previous section. We focus on estimating the quantile function, allowing for the high dimensional scenario where $p$ and $T$ can grow as $n$ grows.

4.1. Estimating the quantiles. We show that our proposed estimator is consistent for estimating the conditional quantiles.

Throughout, we treat $\Pi(\tau)$ as fixed parameters. As for the data generation process, our next condition requires that the observations are independent across $i$, and weakly dependent across time.

Assumption 1. We assume that

$$Y_{i,t} | X_{i,t}; \theta(\tau), \Pi_{i,t}(\tau) = X_{i,t}'\theta(\tau) + \Pi_{i,t}(\tau) + \epsilon_{i,t},$$

where $P(\epsilon_{i,t} \leq 0 | X_{i,t}'\theta(\tau) + \Pi_{i,t}(\tau)) = \tau$. Furthermore, the following holds:

(i) There exists a function $G : [0,1]^d \rightarrow [g_1, g_2]$ for positive constants $g_1$ and $g_2$ such that, conditional on $\Pi$ and $\{X_{i,t}\}_{i \in [n]}$, $\epsilon_{i,t} = \epsilon_{i,t}G(X_{i,t})$ where $\{\epsilon_{i,t}\}_{t=1,...,T}$ are independent across $i$. Also, for each $i \in [n]$, the sequence $\{\epsilon_{i,t}\}_{t=1,...,T}$ is stationary and $\beta$-mixing with mixing coefficients satisfying $\sup_i \gamma_i(k) = O(k^{-\mu})$ for some $\mu > 2$. Moreover, there exists $\mu' \in (0, \mu)$, such that

$$npT\left[\left\lceil T^{1/(1+\mu')}\right\rceil\right]^{-\mu} \rightarrow 0.$$  

Here, 

$$\gamma_i(k) = \frac{1}{2} \sup_{l \geq 1} \left\{ \sum_{j=1}^{L} \sum_{l'=1}^{L'} |P(A_{j} \cap B_{j'}) - P(A_{j})P(B_{j'})| \right\} \text{ with } \{A_j\}_{j=1}^{L} \text{ partition of } \sigma(\{\epsilon_{i,1}\}, \ldots, \{\epsilon_{i,l}\}), \text{ and } \{B_{j'}\}_{j'=1}^{L'} \text{ partition of } \sigma(\{\epsilon_{i,l+k}\}, \{\epsilon_{i,l+k+1}\}, \ldots),$$

(ii) There exists $f > 0$ satisfying 

$$\inf_{1 \leq i \leq n, 1 \leq t \leq T, x \in X_t} f_{Y_{i,t} | X_{i,t}; \theta(\tau), \Pi_{i,t}(\tau)}(x'\theta(\tau) + \Pi_{i,t}(\tau) + \delta_{i,t} | x; \theta(\tau), \Pi_{i,t}(\tau)) > f,$$

for some $L > 0$, where $f_{Y_{i,t} | X_{i,t}; \theta, \Pi_{i,t}}$ is the probability density function associated with $Y_{i,t}$ when conditioning on $X_{i,t}$, and with parameters $\theta(\tau)$ and $\Pi_{i,t}(\tau)$. 

Note that Assumption 1 is related to the sampling and smoothness assumption of [13]. Furthermore, we highlight that similar to [13], our framework is rich enough that avoids imposing Gaussian modeling constraints. However, unlike [13], we consider panel data with weak correlation across time. In particular, we refer readers to [78] for thorough discussions on $\beta$-mixing. Another difference with [13] is that we do not require any smoothness assumption on the condition density function of the response given the covariates.

It is worth mentioning that the parameter $\mu$ in Assumption 1 controls the strength of the time dependence in the data while parameter $\mu'$ will relate to the tuning parameters and will impact the rates. As we decrease the value of $\mu'$ condition (13) is easier to satisfy but the tuning parameters increase and slow down the rates of convergence. Furthermore, in the case that $\{(Y_{i,t}, X_{i,t})\}_{i \in [n], t \in [T]}$ are independent our theoretical results will hold without imposing (13).

Next, we require that along each dimension the second moment of the covariates is one. We also assume that the second moments can be reasonably well estimated by their empirical counterparts.

**ASSUMPTION 2.** We assume $E(X_{i,t,j}^2) = 1$ for all $i \in [n], t \in [T], j \in [p]$. Then

\[
\hat{\sigma}_j^2 = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} X_{i,t,j}^2, \quad \forall j \in [p],
\]

and we require that

\[
P\left( \max_{1 \leq j \leq p} |\hat{\sigma}_j^2 - 1| \leq \frac{1}{4} \right) \geq 1 - \gamma \rightarrow 1, \quad as \quad n \rightarrow \infty.
\]

Assumption 2 appeared as Condition D.3 in [13]. It is met by general models on the covariates, see for instance Design 2 in [13].

Using the empirical second order moments $\{\hat{\sigma}_j^2\}_{j=1}^p$, we analyze the performance of the constrained estimator

\[
(\hat{\theta}(\tau), \hat{\Pi}(\tau)) = \arg \min_{(\tilde{\theta}, \tilde{\Pi})} \left\{ \hat{Q}_\tau(\tilde{\theta}, \tilde{\Pi}) + \nu_1 \|\tilde{\theta}\|_{1,n,T} + \nu_2 \|\tilde{\Pi}\|_* \right\},
\]

where $\nu_1, \nu_2 > 0$ are tuning parameters and $\|\tilde{\theta}\|_{1,n,T} := \sum_{j=1}^{p} \hat{\sigma}_j |\tilde{\theta}_j|$, $\hat{Q}_\tau(\tilde{\theta}, \tilde{\Pi}) = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \rho_\tau(Y_{i,t} - X_{i,t}^t \tilde{\theta} - \tilde{\Pi}_{i,t})$, with $\rho_\tau$ as defined in Section 2.2.

Our main result of this subsection is provided next.

**THEOREM 4.1.** Let $(\hat{\theta}(\tau), \hat{\Pi}(\tau))$ be the estimator defined in (15). Let us write

\[
q_{i,t} = X_{i,t}^t \theta_{i,t}(\tau) + \Pi_{i,t}(\tau), \quad \hat{q}_{i,t} = X_{i,t}^t \hat{\theta}_{i,t}(\tau) + \hat{\Pi}_{i,t}(\tau),
\]

for the conditional quantiles and their estimates. Then under Assumptions 1–2, we have that for any sequence $\{m_n\}$, $m_n \rightarrow \infty$, it holds that

\[
\frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \min \{ |q_{i,t} - \hat{q}_{i,t}|, (q_{i,t} - \hat{q}_{i,t})^2 \} = O_P \left( m_n \sqrt{c} \left( \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{d_T}} \right) \left( \|\Pi(\tau)\|_* + \sqrt{\log(\max\{n, pC_T\})} \right) \cdot \|\beta(\tau)\|_1 \right).
\]
with probability approaching one provided that $T, n \to \infty$, and the tuning parameters satisfy

$$\nu_1 \asymp \sqrt{\frac{c_T \log(\max\{n, p c_T\})}{n c_T (\sqrt{n} + \sqrt{d_T})}},$$

and

$$\nu_2 \asymp \frac{c_T}{n c_T} \left(\sqrt{n} + \sqrt{d_T}\right),$$

where $c_T = \lceil T^{1/(1+\mu')} \rceil$, $d_T = \lfloor T/(2c_T) \rfloor$.

Theorem 4.1 shows that we can consistently estimate the conditional quantiles at a rate that combines the sparse and dense signal magnitudes. This result differs from [13] in several ways. First, [13] work with cross sectional data and focus on the estimation of the vector of parameter coefficients instead of the conditional quantiles. Therefore the assumptions for Theorem 4.1 and those in [13] are very different. For example, [13] required a minimum eigenvalue condition on the behavior of the design matrix, whereas we can avoid this. Furthermore, unlike [13], Theorem 4.1 allows for panel data with dependence across time. This translates into a technical challenge to arrive at Theorem 4.1, since it is not possible to use the analysis from [13] which relies heavily on independence. In fact, the minimum eigenvalue condition in [13] (Condition D.4) can be difficult to be verified in practice without the independence data assumption, and hence Theorem 4.1 has a different setting than those results in [13]. Also, our setting and estimator involve the latent factors which make both the estimator and theory more complex than the framework in [13]. Lastly, while some of the ideas in the proof of Theorem 4.1 come from [66], all of the technical steps involved in the proof of Theorem 4.1 are novel.

On another note, the rate of $T$ in Theorem 4.1 must be such that $T \to \infty$ and $T$ satisfies (13). The parameter $\mu'$ in (13) controls the terms $c_T = \lceil T^{1/(1+\mu')} \rceil$ and $d_T = \lfloor T/(2c_T) \rfloor$ both of which determine the tuning parameters and imply specific rates of convergence. Clearly, $c_T \cdot d_T \asymp T$. Moreover, the parameter $c_T$ measures the strength of the dependence in the data, whereas $d_T$ can be interpreted as the effective number of independent samples across time. In particular, when $c_T \asymp 1$, we have that $\{\epsilon_{i,t}\}$ are basically independent across time, and so the final rate depends on $T$. As a different example, since we would like $d_T \gg c_T$, with $\mu' = 2$ we have $c_T \asymp T^{1/3}$ and $d_T \asymp T^{2/3}$ so that $\nu_1 \sim (T^{-1/6} + \sqrt{T^{1/3}/n}) \log(npT)$ and $\nu_2 \sim \sqrt{T^{2/3}/n} + \frac{1}{n c_T}$. In this case condition (13) requires the parameter $\mu$ that governs the mixing speed to be such that $npT \ll T^{\mu/3}$.

### 4.2. Estimating the coefficients and latent factors

We show that our proposed estimator is consistent for estimating the vector of coefficients and latent factors separately in a broad range of models, and in some cases attains minimax rates, as in [19].

In order to obtain our main result of this subsection, we first provide some additional notation and assumptions. For a fixed $\tau > 0$, we assume that (2) holds. We also let $T_\tau$ be the support of $\theta(\tau)$, thus

$$T_\tau = \{j \in [p] : \theta_j(\tau) \neq 0\},$$

and we write $s_\tau = |T_\tau|$, and $r_\tau = \text{rank}(\Pi(\tau))$.

**Assumption 3.** Conditional on $\Pi$, $\{(X_{i,t}, \epsilon_{i,t})\}_{t=1,...,T}$ are independent across $i$. Also, for each $i \in [n]$, the sequence $\{(X_{i,t}, \epsilon_{i,t})\}_{t=1,...,T}$ is stationary and $\beta$-mixing with mixing coefficients satisfying $\sup_t \gamma_t(k) = O(k^{-\mu})$ for some $\mu > 2$. Moreover, there exists $\mu' \in (0, \mu)$, such that (13) holds. In addition, there exists $f > 0$ satisfying

$$\inf_{1 \leq i \leq n, 1 \leq t \leq T, x \in \mathcal{X}} \tilde{f}_{Y_i,1|X_{i,t},\theta(\tau),\Pi_{i,t}(\tau)}(x' \theta(\tau) + \Pi_{i,t}(\tau)|x) > f.$$
where \( f_{Y_{i,t} \mid X_{i,t} = \theta_i, \Pi_{i,t}} \) is the probability density function associated with \( Y_{i,t} \) when conditioning on \( X_{i,t} \) and with parameters \( \theta(\tau) \) and \( \Pi_{i,t}(\tau) \). Furthermore, \( f_{Y_{i,t} \mid X_{i,t} = \theta(\tau), \Pi_{i,t}(\tau)}(y \mid x; \theta(\tau), \Pi_{i,t}(\tau)) \) and \( \frac{\partial}{\partial \theta} f_{Y_{i,t} \mid X_{i,t} = \theta(\tau), \Pi_{i,t}(\tau)}(y \mid x; \theta(\tau), \Pi_{i,t}(\tau)) \) are both bounded by \( f \) and \( f' \), respectively, uniformly in \( y \) and \( x \) in the support of \( X_{i,t} \).

As it can been seen in Lemma S4 from the Supplementary Material, the error of our estimator defined in (15), \( \theta(\tau) - \bar{\theta}(\tau), \Pi(\tau) - \bar{\Pi}(\tau) \), belongs to a restricted set, which in our framework is defined as

\[
A_{\tau} = \left\{ (\delta, \Delta) \in \mathbb{R}^p \times \mathbb{R}^{n \times T} : \|\delta_T\|_1 + \frac{\|\Delta\|_{\infty}}{\sqrt{nT} \sqrt{\log(\max\{n, pc\})}} \leq C_0 \left( \|\delta_T\|_1 + \frac{\|\Delta\|_{\infty}}{\sqrt{nT} \sqrt{\log(\max\{n, pc\})}} \right) \right\}
\]

for an appropriate positive constant \( C_0 \).

Similar in spirit to other high dimensional settings such as those in [21], [14], [13] and [32], we impose an identifiability condition involving the restricted set which is expressed next and will be used in order to attain our main results. This is the key difference with the analysis in Section 4.1. Before arriving at our next condition, we introduce some notation.

For \( m \geq 0 \), we denote by \( T_{\tau}(\delta, m) \subset \{1, \ldots, p\} \setminus T_{\tau} \) the support of the \( m \) largest components, excluding entries in \( T_{\tau} \), of the vector \( (|\delta_1|, \ldots, |\delta_p|)^T \). We also use the convention \( \overline{T}_{\tau}(\delta, 0) = \emptyset \).

**Assumption 4.** For \( (\delta, \Delta) \in A_{\tau} \), let

\[
J_{\tau}^{1/2}(\delta, \Delta) := \sqrt{\frac{f}{nT} \sum_{i=1}^n \sum_{t=1}^T \mathbb{E} \left( \left( X'_{i,t} \delta + \Delta_{i,t} \right)^2 \right)}.
\]

Then there exists \( m \geq 0 \) such that

\[
0 < \kappa_m := \inf_{(\delta, \Delta) \in A_{\tau}, \delta \neq 0} \frac{J_{\tau}^{1/2}(\delta, \Delta)}{\|\delta_T \cup \overline{T}_{\tau}(\delta, m)\| + \|\Delta\|_{\infty}},
\]

where \( c_{\tau} = \left[ T^{1/(1+\mu')} \right] \) for \( \mu' \) as defined in Assumption 2. Moreover, we assert that the following holds

\[
0 < q := \frac{3}{8} \frac{f^{3/2}}{f'} \inf_{(\delta, \Delta) \in A_{\tau}, \delta \neq 0} \frac{\left( \mathbb{E} \left( \left( \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T (X'_{i,t} \delta + \Delta_{i,t})^2 \right) \right) \right)^{3/2}}{\mathbb{E} \left( \left( \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T |X'_{i,t} \delta + \Delta_{i,t}|^3 \right) \right)},
\]

with \( f \) and \( f' \) as in Assumption 1.

Few comments are in order. First, if \( \Delta = 0 \) then (17) and (18) become the restricted identifiability and nonlinearity conditions as of [13]. Second, the denominator of (17) contains the term \( \|\Delta\|_{F} / \sqrt{nT} \). To see why this is reasonable, consider the case where \( \mathbb{E}(X_{i,t}) = 0 \), and \( X_{i,t} \) are i.i.d.. Then

\[
J_{\tau}(\delta, \Delta) = \frac{f}{nT} \mathbb{E}((s'X_{i,t})^2) + \frac{f}{nT} \|\Delta\|^2_{F}.
\]

Hence, \( \|\Delta\|_{F} / \sqrt{nT} \) appears also in the numerator of (17) and it is not restrictive its presence in the denominator of (17).

We now state our result for estimating \( \theta(\tau) \) and \( \Pi(\tau) \).
Let \( (\hat{\theta}(\tau), \hat{\Pi}(\tau)) \) be the estimator defined in (15). Suppose that Assumptions 2–4 hold and that

\[
q \geq C \phi_n \sqrt{ct \log(pcT \lor n)} \left( \sqrt{s_\tau + 1} + \sqrt{r_\tau / \log(pcT \lor n)} \right) (\sqrt{n} + \sqrt{d_T}) \frac{1}{\kappa_0 f^{1/2}}
\]

for a large enough constant \( C \), and \( \{\phi_n\} \) is a sequence with \( \phi_n / (\sqrt{T} \log(cT + 1)) \to \infty \). Then

\[
||\hat{\theta}(\tau) - \theta(\tau)|| = O_P \left( \phi_n \left( 1 + \sqrt{\frac{\sigma}{m}} \right) \sqrt{ct \log(pcT \lor n)} \left( \sqrt{1 + s_\tau} + \sqrt{\frac{r_\tau}{\log(pcT \lor n)}} \right) \frac{1}{\kappa_0 f^{1/2}} \right)
\]

and

\[
\frac{1}{nT} ||\hat{\Pi}(\tau) - \Pi(\tau)||^2_F = O_P \left( \phi_n^2 cT \log(pcT \lor n) \left( 1 + s_\tau + \frac{r_\tau}{\log(pcT \lor n)} \right) \frac{1}{\kappa_0^4 f^4} \left( \frac{1}{n} + \frac{1}{d_T} \right) \right)
\]

for choices of the tuning parameters as in Theorem 4.1.

Theorem 4.2 gives an upper bound on the performance of \( (\hat{\theta}(\tau), \hat{\Pi}(\tau)) \) for estimating the vector of coefficients \( \theta(\tau) \) and the latent matrix \( \Pi(\tau) \). For simplicity, consider the case of i.i.d data. Then the convergence rate of our estimation of \( \theta(\tau) \), under the Euclidean norm, is in the order of \( (\sqrt{s_\tau} + \sqrt{r_\tau}) / \min\{\sqrt{n}, \sqrt{d_T}\} \), if we ignore all the other factors. Hence, we can consistently estimate \( \theta(\tau) \) provided that \( \max\{s_\tau, r_\tau\} \ll \min\{n, T\} \). This is similar to the low-rank condition in [65]. In the low dimensional case \( s_\tau = O(1) \), the rate \( \sqrt{r_\tau} / \min\{\sqrt{n}, \sqrt{d_T}\} \) matches that of Theorem 1 in [64]. However, unlike [64], our estimator is based on a loss function that is robust to outliers, and our assumptions also allow for weak dependence across time, making our framework potentially more general. Furthermore, the same applies to our rate on the mean squared error for estimating \( \hat{\Pi}(\tau) \), which also matches that in Theorem 1 of [64].

With regards to the novelty of Theorem 4.2, we highlight that, while its proof is similar in spirit to that of Theorem 2 in [13], there are some significant differences. First, the construction of the restricted set used in Theorem 4.2 involves two different penalties which makes challenging to disentangle the behavior of \( \beta(\tau) \) and \( \Pi \), whereas [13] only had to deal with one penalty. Second, the empirical processes in [13] all involved independent data whereas as the proof of Theorem 4.2 handles the time dependence of our model and the latent factors.

Interestingly, it is expected that the rate in Theorem 4.2 is optimal. To elaborate on this point, consider the simple case where \( n = T, \theta = 0, \tau = 0.5 \), and \( e_{i,t} := Y_{i,t} - \Pi_{i,t}(\tau) \) are mean zero i.i.d. sub-Gaussian(\( \sigma^2 \)). The latter implies that

\[
\mathbb{P}(|e_{1,1}| > z) \leq C_1 \exp \left( \frac{-z^2}{2\sigma^2} \right)
\]

for a positive constant \( C_1 \), and for all \( z > 0 \). Then by Theorem 2.3 in [19], we have the following lower bound for estimating \( \Pi(\tau) \):

\[
\inf_{\Pi} \sup_{\Pi(\tau) \leq \text{rank}(\Pi(\tau)) \leq r_\tau} \mathbb{E} \left[ \left( \frac{||\hat{\Pi}(\tau) - \Pi(\tau)||^2_F}{nT} \right) \right] \geq \frac{r_\tau \sigma^2}{n}.
\]

Notably, the lower bound in (22) matches the rate implied by Theorem 4.2, ignoring other factors depending on \( s_\tau, \kappa_0, \kappa_m, p \) and \( \phi_n \). However, we highlight that the upper bound
(21) in Theorem 4.2 holds without the perhaps restrictive condition that the errors are sub-
Gaussian.

We conclude this section with a result regarding the estimation of the factors and load-
ings of the latent matrix \( \Pi(\tau) \). This is expressed in Corollary 4.3 below and is immediate consequence of Theorem 4.2 and Theorem 3 in [79].

**COROLLARY 4.3.** Suppose that the all the conditions of Theorem 4.2 hold. Let \( \sigma_1(\tau) \geq \sigma_2(\tau) \geq \ldots \geq \sigma_{r_\tau}(\tau) > 0 \) be the singular values of \( \Pi(\tau) \), and \( \hat{\sigma}_1(\tau) \geq \ldots \geq \hat{\sigma}_{\min(n,T)}(\tau) \) the singular values of \( \hat{\Pi}(\tau) \). Let \( g(\tau), \hat{g}(\tau) \in \mathbb{R}^{T \times r_\tau} \) and \( \lambda(\tau), \hat{\lambda}(\tau), \tilde{\lambda}(\tau) \in \mathbb{R}^{n \times r_\tau} \) be matrices with orthonormal columns satisfying

\[
\Pi(\tau) = \sum_{j=1}^{r_\tau} \sigma_j \hat{\lambda}_j(\tau) g_{j,\cdot}(\tau)' = \sum_{j=1}^{r_\tau} \lambda_j(\tau) g_{j,\cdot}(\tau)',
\]

and \( \hat{\Pi}(\tau) \hat{g}_{j,\cdot}(\tau) = \hat{\sigma}_j(\tau) \hat{\lambda}_j(\tau) = \hat{\lambda}_j(\tau) \) for \( j = 1, \ldots, r_\tau \).

Then

\[ \min_{\hat{O} \in \hat{O}_{r_\tau}} \| \hat{g}(\tau) O - g(\tau) \|_F = O_P \left( \frac{(\sigma_1(\tau) + \sqrt{\tau_{\text{Err}}}) Err}{(\sigma_{r_\tau-1}(\tau))^2 - (\sigma_{r_\tau}(\tau))^2} \right), \]

and

\[ \frac{\| \hat{\lambda}(\tau) - \lambda(\tau) \|_F^2}{nT} = O_P \left( \frac{\sigma_1^2(\tau) + \sqrt{\tau_{\text{Err}}}) Err^2}{(\sigma_{r_\tau-1}(\tau))^2 - (\sigma_{r_\tau}(\tau))^2} \right). \]

Here, \( \hat{O}_{r_\tau} \) is the group of \( r_\tau \times r_\tau \) orthonormal matrices, and

\[ \text{Err} := \frac{\phi_n c_T \sqrt{\log(p c_T \vee n)}((1 + s_\tau + r_\tau/\log(p c_T \vee n))}{\kappa_0^2 f^{1/2}} \left( \sqrt{n} + \sqrt{d_T} \right). \]

A particularly interesting instance of Corollary 4.3 is when

\[ (\sigma_1(\tau))^2, (\sigma_{r_\tau-1}(\tau))^2 - (\sigma_{r_\tau}(\tau))^2 \approx nT, \]

a natural setting if the entries of \( \Pi(\tau) \) are \( O(1) \). Then the upper bound (23) becomes

\[ \min_{\hat{O} \in \hat{O}_{r_\tau}} \| \hat{g}(\tau) O - g(\tau) \|_F = O_P \left( \frac{\phi_n c_T \sqrt{\log(p c_T \vee n)}((1 + s_\tau + \sqrt{r_\tau/\log(p c_T \vee n)})}{\kappa_0^2 f^{1/2}} \left( \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{d_T}} \right) \right), \]

whereas (24) is now

\[ \frac{\| \hat{\lambda}(\tau) - \lambda(\tau) \|_F^2}{nT} = O_P \left( \frac{r_\tau \phi_n c_T \log(p c_T \vee n)(1 + s_\tau + r_\tau/\log(p c_T \vee n))}{\kappa_0^2 f^{1/2}} \left( \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{d_T}} \right) \right). \]

The conclusion of Corollary 4.3 allows us to provide an upper bound on the estimation of factors \( g(\tau) \) and loadings \( \lambda(\tau) \) of the latent matrix \( \Pi(\tau) \). Notice that we are not claiming that we provide consistent estimation of the number latent factors, as Theorem 4.2 only guarantees consistent estimation of \( \Pi(\tau) \). However, other authors, e.g. [63], have observed that estimation can be possible even if \( r_\tau \) is unknown.
4.3. Nearly low rank quantiles. We conclude our theory section by studying the case where the matrix $X\theta(\tau)$ has nearly low rank. We make this formal by imposing the condition that $X\theta(\tau)$ can be perturbed into a low-rank matrix. While the result in this subsection differs in the choice of tuning parameters from those in Theorem 4.2, our result here suggests that even in this low-rank setting there exists a choice of tuning parameters that allows our estimator to provide consistent estimation.

We view our setting below as an extension of the linear model in [28] to the quantile framework and reduced rank regression. The specific condition is stated next.

**Assumption 5.** With probability approaching one, it holds that $\text{rank}(X\theta(\tau) + \xi) = O(r_\tau)$, and

$$\frac{\|\xi\|}{\sqrt{nT}} = O_P\left(\frac{c_T \phi_n \sqrt{T}(\sqrt{n} + \sqrt{dT})}{\sqrt{nTf}}\right),$$

with $c_T$ as defined in Theorem 4.2. Furthermore, $\|X\theta(\tau) + \Pi(\tau)\|_\infty = O_P(1)$.

Notice that in Assumption 5, $\xi$ is an approximation error. In the case $\xi = 0$, the condition implies that $\text{rank}(X\theta(\tau)) = O(r_\tau)$ with probability close to one.

Next, exploiting Assumption 5, we show that (15) provides consistent estimation of the quantile function, namely, of $X\theta(\tau) + \Pi(\tau)$.

**Theorem 4.4.** Suppose that Assumptions 2–5 hold. Let $(\hat{\theta}(\tau), \hat{\Pi}(\tau))$ be the solution to (15) with the additional constraint that $\|\hat{\Pi}\|_\infty \leq C$, for a large enough positive constant $C$. Then

$$\frac{1}{nT} \|\hat{\Pi}(\tau) - \Pi(\tau) - X\theta(\tau)\|_F^2 = O_P\left(\frac{(\tilde{T})^2 \phi_n^2 c_T r_\tau}{f^4} \left(\frac{1}{n} + 1 + \frac{1}{dT}\right)\right),$$

where $\{\phi_n\}$ is a sequence with $\phi_n/(\sqrt{T} \log(1 + c_T)) \to \infty$, and for choices

$$\nu_1 \asymp \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \max_{j=1,\ldots,p} \frac{|X_{i,t,j}|}{\hat{\sigma}_{j}},$$

and

$$\nu_2 \asymp \frac{c_T}{nT} \left(\sqrt{n} + \sqrt{dT}\right).$$

Interestingly, unlike Theorem 4.2, Theorem 4.4 does not show that we can estimate $\theta(\tau)$ and $\Pi(\tau)$ separately. Instead, we show that $\hat{\Pi}(\tau)$, the estimated matrix of latent factors, captures the overall contribution of both $\theta(\tau)$ and $\Pi(\tau)$. This is expected since Assumption 5 states that, with high probability, $X\theta(\tau)$ has rank of the same order as of $\Pi(\tau)$. Notably, $\hat{\Pi}(\tau)$ is able to estimate $X\theta(\tau) + \Pi(\tau)$ via requiring that the value of $\nu_1$ increases significantly with respect to the choice in Theorem 4.2, while keeping $\nu_2 \asymp c_T/(\sqrt{n} + \sqrt{dT})/(nT)$.

As for the convergence rate in Theorem 4.4 for estimating $\Pi(\tau)$, this is of the order $r_\tau c_T (n^{-1} + d_T^{-1})$, if we ignore $f_\tau, \tilde{T}$, and $\phi_n$. When the data are independent, the rate becomes of the order $r_\tau (n^{-1} + T^{-1})$. In such framework, our result matches the minimax rate of estimation in [18] for estimating an $n \times T$ matrix of rank $r_\tau$, provided that $n \asymp T$, see our discussion in Section 4.2.
5. Simulation. In this section, we evaluate the performance of our proposed approach ($\ell_1$-NN-QR) with extensive numerical simulations focusing on the median case, namely the case when $\tau = 0.5$. As benchmarks, we consider the $\ell_1$-penalized quantile regression studied in [12], and similarly we refer to this procedure as $\ell_1$-QR. We also compare with the mean case, which we denote as $\ell_1$-NN-LS as it combines the $\ell_2$-loss function with $\ell_1$ and nuclear norm regularization. We consider different generative scenarios. For each scenario we randomly generate 100 different data sets and compute the estimates of the methods for a grid of values of $\nu_1$ and $\nu_2$. Specifically, these tuning parameters are taken to satisfy $\nu_1 \in \{10^{-4}, 10^{-4.5}, \ldots, 10^{-8}\}$ and $\nu_2 \in \{10^{-3}, 10^{-4}, \ldots, 10^{-9}\}$. Given any choice of tuning parameters, we evaluate the performance of each competing method, averaging over the 100 data sets, and report values that correspond to the best performance. These are referred as optimal tuning parameters and can be thought of as oracle choices.

We also propose a modified Bayesian Information Criterion (BIC) to select the best pair of tuning parameters. Given a pair $(\nu_1, \nu_2)$, our method produces a score $(\hat{\theta}(\tau), \hat{\Pi}(\tau))$. Specifically, denote $s_{\tau} = |\{j : \hat{\theta}_j(\tau) \neq 0\}|$ and $r_{\tau} = \text{rank}(\hat{\Pi}(\tau))$,

$$
\text{BIC}(\nu_1, \nu_2) = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \log(nT) \left( c_1 \cdot s_{\tau}(\nu_1, \nu_2) + (1 + n + T) \cdot r_{\tau}(\nu_1, \nu_2) \right),
$$

where $c_1 > 0$ is a constant. The intuition here is that the first term in the right hand side of (25) corresponds to the fit to the data. The second term includes the factor $\log(nT)/2$ to emulate the usual penalization in BIC. The number of parameters in the model with choices $\nu_1$ and $\nu_2$ is estimated by $s_{\tau}$ for the vector of coefficients, and $(1 + n + T) \cdot r_{\tau}$ for the latent matrix. The latter is reasonable since $\hat{\Pi}(\tau)$ is potentially a low rank matrix and we simply count the number of parameters in its singular value decomposition. As for the extra quantity $c_1$, we have included this term to balance the dominating contribution of the $(1 + n + T) \cdot r_{\tau}$. We find that in practice $c_1 = \log^2(nT)$ gives reasonable performance in both simulated and real data. This is the choice that we use in our experiments. Then for each of data set under each design, we calculate the minimum value of BIC$(\nu_1, \nu_2)$, over the different choices of $\nu_1$ and $\nu_2$, and report the average over the 100 Monte Carlo simulations. We refer to this as BIC-$\ell_1$-NN-QR.

As performance measure we use a scaled version (see Tables 1-2) of the squared distance between the true vector of coefficients $\theta$ and the corresponding estimate. We also consider a different metric, the “Quantile error” ([55]):

$$
\frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} (F_{Y_{i,t} | X_{i,t}; \theta(\tau), \Pi(\tau)}(0.5) - \hat{F}_{Y_{i,t} | X_{i,t}; \hat{\theta}(\tau), \hat{\Pi}(\tau)}(0.5))^2,
$$

which measures the average squared error between the quantile functions at the samples and their respective estimates. Since our simulations consider models with symmetric mean zero error, the above metric corresponds to the mean squared error for estimating the conditional expectation.

Next, we provide a detailed description of each of the generative models that we consider in our experiments. In each model design the dimensions of the problem are given by $n \in \{100, 500\}$, $p \in \{5, 30\}$ and $T \in \{100, 500\}$, and we also consider the instance $n = p = T = 50$. The covariates $\{X_{i,t}\}$ are i.i.d $N(0, I_p)$.

Design 1. (Location shift model) The data is generated from the model

$$
Y_{i,t} = X_{i,t}' \theta + \Pi_{i,t} + \epsilon_{i,t},
$$
where \( \sqrt{3} \epsilon_{i,t} \stackrel{i.i.d.}{\sim} t(3) \), \( i = 1, \ldots, n \) and \( t = 1, \ldots, T \), with \( t(3) \) the Student’s t-distribution with 3 degrees of freedom. The scaling factor \( \sqrt{3} \) simply ensures that the errors have variance 1. In (27), we take the vector \( \theta \in \mathbb{R}^p \) to satisfy

\[
\theta_j = \begin{cases} 
1 & \text{if } j \in \{1, \ldots, \min\{10, p\} \} \\
0 & \text{otherwise.} 
\end{cases}
\]

We also construct \( \Pi \in \mathbb{R}^{n \times T} \) to be rank one, defined as \( \Pi_{i,t} = 5i (\cos(4\pi t/T)) / n \).

**Design 2. (Location-scale shift model)** We consider the model

\[
Y_{i,t} = X'_{i,t} \theta + \Pi_{i,t} + (X'_{i,t} \theta)^t \epsilon_{i,t},
\]

where \( \epsilon_{i,t} \stackrel{i.i.d.}{\sim} N(0,1) \), \( i = 1, \ldots, n \) and \( t = 1, \ldots, T \). The parameters in \( \theta \) and \( \Pi \) in (28) are taken to be the same as in (27). The only difference now is that we have the extra parameter \( \theta \in \mathbb{R}^p \), which we define as \( \theta_j = j/(2p) \) for \( j \in \{1, \ldots, p\} \).

**Design 3. (Location shift model with random factors)** This is the same as Design 1 with the difference that we now generate \( \Pi \) as

\[
\Pi_{i,t} = \sum_{k=1}^{5} c_k u_k v_k^T,
\]

where

\[
c_k \sim U[0, 1/4], \quad u_k = \tilde{u}_k / \| \tilde{u}_k \|, \quad \tilde{u}_k \sim N(0, I_n), \quad v_k = \tilde{v}_k / \| \tilde{v}_k \|, \quad \tilde{v}_k \sim N(0, I_n), \quad k = 1, \ldots, 5.
\]

**Design 4. (Location-scale shift model with random factors)** This is a combination of Designs 2 and 3. Specifically, we generate data as in (28) but with \( \Pi \) satisfying (29) and (30).

The results in Tables 1-2 show a clear advantage of our proposed method against the benchmarks across the four designs we consider. This is true for estimating the vector of coefficients, and under the measure of quantile error. Importantly, our approach is not only the best under the optimal choice of tuning parameters but it remains competitive with the BIC type of criteria defined with the score (25). In particular, under Designs 1 and 2, the data driven version of our estimator, BIC-\( \ell_1 \)-NN-QR, performs very closely to the ideally tuned one \( \ell_1 \)-NN-QR. In the more challenging settings of Designs 3 and 4, we noticed that BIC-\( \ell_1 \)-NN-QR performs reasonably well compared to \( \ell_1 \)-NN-QR.

6. Empirical Performance of the “Characteristics + Latent Factor” Model in Asset Pricing.

**Data Description.** We use data from CRSP and Compustat to construct 24 firm level characteristics that are documented to explain the cross section and time series of stock returns in the finance and accounting literature. The characteristics we choose include well-known drivers of stock returns such as beta, size, book-to-market, momentum, volatility, liquidity, investment and profitability. Table S1 in the Supplementary Material lists details of the characteristics used and the methods to construct the data. We follow the procedures of [48] to construct the characteristics of interest. The characteristics used in our model are standardized to have zero mean and unit variance. Figure 1 plots the histogram of monthly stock returns and 9 standardized firm characteristics. Each of them have different distribution
For Designs 1-2 described in the main text, under different values of \((n,p,T)\), we compare the performance of different methods. The metrics use are the scaled \(\ell_2\) distance for estimating \(\theta(\tau)\), and the Quantile error defined in (26). For each method we report the average, over 100 Monte Carlo simulations, of the two performance measures.

| Method          | Design 1 | Design 2 |
|-----------------|----------|----------|
|                 | \(n\) | \(p\) | \(T\) | \(\|\hat{\theta}(\tau) - \theta(\tau)\|_2^{\infty}/\sqrt{T}\) | Quantile error | \(\|\hat{\theta}(\tau) - \theta(\tau)\|_2^{\infty}/\sqrt{T}\) | Quantile error |
| \(\ell_1\)-NN-QR | 300 | 30 | 300 | 0.86 | 0.03 | 0.69 | 0.03 |
| \(\ell_1\)-NN-QR | 300 | 30 | 300 | 0.86 | 0.03 | 0.89 | 0.03 |
| \(\ell_1\)-NN-LS | 300 | 30 | 300 | 2.58 | 0.05 | 1.03 | 0.04 |
| \(\ell_1\)-QR | 300 | 30 | 300 | 8.18 | 4.19 | 6.81 | 4.18 |
| \(\ell_1\)-NN-QR | 300 | 30 | 100 | 2.39 | 0.04 | 2.39 | 0.04 |
| \(\ell_1\)-NN-QR | 300 | 30 | 100 | 0.86 | 0.03 | 0.69 | 0.03 |
| \(\ell_1\)-NN-LS | 300 | 30 | 100 | 8.06 | 0.12 | 3.11 | 0.08 |
| \(\ell_1\)-QR | 300 | 30 | 100 | 41.0 | 4.19 | 26.0 | 4.19 |
| \(\ell_1\)-NN-QR | 300 | 5 | 300 | 0.22 | 0.003 | 0.39 | 0.03 |
| \(\ell_1\)-NN-QR | 300 | 5 | 300 | 0.22 | 0.003 | 0.89 | 0.03 |
| \(\ell_1\)-NN-QR | 300 | 5 | 300 | 0.37 | 0.01 | 0.69 | 0.03 |
| \(\ell_1\)-QR | 300 | 5 | 300 | 2.6 | 4.19 | 3.27 | 4.19 |
| \(\ell_1\)-NN-QR | 100 | 30 | 300 | 0.50 | 0.008 | 0.80 | 0.03 |
| \(\ell_1\)-NN-QR | 100 | 30 | 300 | 0.53 | 0.009 | 0.97 | 0.03 |
| \(\ell_1\)-NN-LS | 100 | 30 | 300 | 9.77 | 0.02 | 3.39 | 0.06 |
| \(\ell_1\)-QR | 100 | 30 | 300 | 40.0 | 4.23 | 24.0 | 4.23 |
| \(\ell_1\)-NN-QR | 100 | 5 | 300 | 2.3 | 0.04 | 10.0 | 0.03 |
| \(\ell_1\)-NN-QR | 100 | 5 | 300 | 2.3 | 0.06 | 11.0 | 0.04 |
| \(\ell_1\)-NN-LS | 100 | 5 | 300 | 8.4 | 0.02 | 13.0 | 0.16 |
| \(\ell_1\)-QR | 100 | 5 | 300 | 229.0 | 4.23 | 177.0 | 4.23 |
| \(\ell_1\)-NN-QR | 100 | 5 | 100 | 0.64 | 0.008 | 0.89 | 0.03 |
| \(\ell_1\)-NN-QR | 100 | 5 | 100 | 0.65 | 0.008 | 1.32 | 0.03 |
| \(\ell_1\)-NN-LS | 100 | 5 | 100 | 1.25 | 0.02 | 1.67 | 0.05 |
| \(\ell_1\)-QR | 100 | 5 | 100 | 8.79 | 4.23 | 8.61 | 4.23 |
| \(\ell_1\)-NN-QR | 100 | 100 | 5 | 100 | 1.82 | 0.009 | 3.30 | 0.03 |
| \(\ell_1\)-NN-QR | 100 | 100 | 5 | 100 | 1.85 | 0.01 | 3.74 | 0.04 |
| \(\ell_1\)-NN-LS | 100 | 100 | 5 | 100 | 4.06 | 0.03 | 4.45 | 0.14 |
| \(\ell_1\)-QR | 100 | 100 | 5 | 100 | 32.0 | 4.23 | 27.0 | 4.23 |
| \(\ell_1\)-NN-QR | 50 | 50 | 50 | 136.0 | 0.09 | 388.0 | 0.21 |
| \(\ell_1\)-NN-QR | 50 | 50 | 50 | 241.0 | 0.33 | 522.0 | 0.59 |
| \(\ell_1\)-NN-LS | 50 | 50 | 50 | 2010.0 | 0.17 | 982.0 | 0.45 |
| \(\ell_1\)-QR | 50 | 50 | 50 | 1258.0 | 4.27 | 2000.6 | 4.33 |

patterns, suggesting the potential nonlinear relationship between returns and firm characteristics, which can be potentially captured by our quantile model.

Our empirical design is closely related to the characteristics model proposed by [33, 34]. To avoid any “data snooping” issue cause by grouping, we conduct the empirical analysis at individual stock level. Specifically, we use the sample period from January 2000 to December 2018, and estimate our model using monthly returns (228 months) from 1306 firms that have non-missing values during this period.
For Designs 3-4 described in the main text, under different values of \((n, p, T)\), we compare the performance of different methods. The metrics use are the scaled \(\ell_2\) distance for estimating \(\theta(\tau)\), and the Quantile error defined in (26). For each method we report the average, over 100 Monte Carlo simulations, of the two performance measures.

| Method       | \(n\) | \(p\) | \(T\) | \(\|\theta(\tau) - \hat{\theta}(\tau)\|_2^{10^{-14}}\) Quantile error | \(\|\theta(\tau) - \hat{\theta}(\tau)\|_2^{10^{-14}}\) Quantile error |
|--------------|-------|-------|------|-------------------------------------------------|-------------------------------------------------|
| \(\ell_1\)-NN-QR | 300   | 300   | 300  | 2.17 0.17                                       | 1.58 0.11                                       |
| BIC-\(\ell_1\)-NN-QR | 300   | 300   | 300  | 2.17 0.29                                       | 2.81 0.26                                       |
| \(\ell_1\)-NN-LS | 300   | 300   | 300  | 3.33 0.19                                       | 3.54 0.12                                       |
| \(\ell_1\)-QR | 300   | 300   | 300  | 6.59 1.09                                       | 12.0 1.01                                       |
| \(\ell_1\)-NN-QR | 300   | 300   | 100  | 10.0 0.18                                       | 4.61 0.13                                       |
| BIC-\(\ell_1\)-NN-QR | 300   | 300   | 100  | 11.0 0.26                                       | 4.61 0.16                                       |
| \(\ell_1\)-NN-LS | 300   | 300   | 100  | 11.0 0.25                                       | 9.21 0.17                                       |
| \(\ell_1\)-QR | 300   | 300   | 100  | 27.0 11.10                                      | 47.2 1.10                                       |
| \(\ell_1\)-NN-QR | 300   | 5     | 300  | 1.13 0.17                                       | 0.27 0.03                                       |
| BIC-\(\ell_1\)-NN-QR | 300   | 5     | 300  | 1.56 0.33                                       | 0.74 0.13                                       |
| \(\ell_1\)-NN-LS | 300   | 5     | 300  | 1.58 0.19                                       | 0.49 0.05                                       |
| \(\ell_1\)-QR | 300   | 5     | 300  | 2.47 1.10                                       | 5.52 1.09                                       |
| \(\ell_1\)-NN-QR | 300   | 5     | 100  | 3.04 0.19                                       | 0.69 0.05                                       |
| BIC-\(\ell_1\)-NN-QR | 300   | 5     | 100  | 4.37 0.27                                       | 1.11 0.15                                       |
| \(\ell_1\)-NN-LS | 300   | 5     | 100  | 4.43 0.27                                       | 1.12 0.05                                       |
| \(\ell_1\)-QR | 300   | 5     | 100  | 7.65 1.11                                       | 13.4 1.10                                       |
| \(\ell_1\)-NN-QR | 100   | 30    | 300  | 11.0 0.18                                       | 7.06 0.15                                       |
| BIC-\(\ell_1\)-NN-QR | 100   | 30    | 300  | 12.0 0.27                                       | 7.29 0.24                                       |
| \(\ell_1\)-NN-LS | 100   | 30    | 300  | 12.0 0.34                                       | 11.2 0.18                                       |
| \(\ell_1\)-QR | 100   | 30    | 300  | 26.0 1.10                                       | 51.0 1.12                                       |
| \(\ell_1\)-NN-QR | 100   | 30    | 100  | 6.12 0.17                                       | 32.1 0.10                                       |
| BIC-\(\ell_1\)-NN-QR | 100   | 30    | 100  | 6.64 0.22                                       | 35.4 0.15                                       |
| \(\ell_1\)-NN-LS | 100   | 30    | 100  | 8.63 1.08                                       | 82.0 0.84                                       |
| \(\ell_1\)-QR | 100   | 30    | 100  | 16.5 1.12                                       | 267.6 1.13                                      |
| \(\ell_1\)-NN-QR | 100   | 5     | 300  | 2.99 0.18                                       | 0.86 0.05                                       |
| BIC-\(\ell_1\)-NN-QR | 100   | 5     | 300  | 4.19 0.26                                       | 1.43 0.14                                       |
| \(\ell_1\)-NN-LS | 100   | 5     | 300  | 5.27 0.33                                       | 1.45 0.05                                       |
| \(\ell_1\)-QR | 100   | 5     | 300  | 8.59 1.10                                       | 21.0 1.09                                       |
| \(\ell_1\)-NN-QR | 100   | 5     | 100  | 12.3 0.16                                       | 2.15 0.04                                       |
| BIC-\(\ell_1\)-NN-QR | 100   | 5     | 100  | 13.1 0.20                                       | 2.43 0.07                                       |
| \(\ell_1\)-NN-LS | 100   | 5     | 100  | 15.0 1.09                                       | 3.61 0.07                                       |
| \(\ell_1\)-QR | 100   | 5     | 100  | 24.7 1.10                                       | 45.7 1.09                                       |
| \(\ell_1\)-NN-QR | 50    | 50    | 50   | 386.7 0.32                                      | 511 0.27                                        |
| BIC-\(\ell_1\)-NN-QR | 50    | 50    | 50   | 898.1 0.97                                      | 1128.1 0.74                                     |
| \(\ell_1\)-NN-LS | 50    | 50    | 50   | 605.2 0.44                                      | 967.3 0.49                                     |
| \(\ell_1\)-QR | 50    | 50    | 50   | 1035.8 1.15                                     | 1697.5 1.18                                     |

A “Characteristic + Latent Factor” Asset Pricing Model. We apply our model to fit the cross section and time series of stock returns ([60]). There are \(n\) assets (stocks), and the return of the each asset can potentially be explained by \(p\) observed asset characteristics (sparse part) and \(r\) latent factors (dense part). The asset characteristics are the covariates in our model. Our model imposes a sparse structure on the \(p\) characteristics so that only the characteristics having the strongest explanatory powers are selected by the model. The part that’s unexplained by the firm characteristics are captured by latent factors.
Suppose we have $n$ stock returns $(R_1, \ldots, R_n)$, and $p$ observed firm characteristics $(X_1, \ldots, X_p)$ over $T$ periods. The return quantile at level $\tau$ of portfolio $i$ in time $t$ is assumed to be the following:

$$F_{R_{i,t}|X_{i,t-1};\theta(\tau),\lambda_{i}(\tau),g_{t}(\tau)}^{-1}(\tau) = X_{i, t-1, 1}\theta_1(\tau) + \ldots + X_{i, t-1, p}\theta_p(\tau) + \lambda_{i}(\tau) g_{t}(\tau) \left(1 \times r_{\tau}\right) \left(r_{\tau} \times 1\right)$$

where $X_{i, t-1, k}$, e.g. $k = 1$ or $k = p$, is the $k$-th characteristic (e.g. the book-to-market ratio) of asset $i$ in time $t - 1$. The coefficient $\theta_k$ captures the extent to which assets with higher/lower characteristic $X_{i, t, k}$ delivers higher average return. The term $g_{t}$ contains the $r_{\tau}$ latent factors.
in period \( t \) which captures systematic risks in the market, and \( \lambda_i \) contains portfolio \( i \)’s loading on these factors (i.e. exposure to risk).

There is a discussion in academic research on “factor versus characteristics” in late 1990s and early 2000s. The factor/risk based view argues that an asset has higher expected returns because of its exposure to risk factors (e.g. Fama-French 3 factors) which represent some unobserved systematic risk. An asset’s exposure to risk factors are measured by factor loadings. The characteristics view claims that stocks have higher expected returns simply because they have certain characteristics (e.g. higher book-to-market ratios, smaller market capitalization), which might be independent of systematic risk ([33, 34]). The formulation of our model accommodates both the factor view and the characteristics view. The sparse part is similar to [33, 34], in which stock return are explained by firm characteristics. The dense part assumes a low-dimensional latent factor structure where the common variations in stock returns are driven by several “risk factors”.

**Empirical Results.** We first get the estimates \( \hat{\theta}(\tau) \) and \( \hat{\Pi}(\tau) \) at three different quantiles, \( \tau = \{0.1, 0.5, 0.9\} \) using our proposed ADMM algorithm. We then decompose \( \hat{\Pi}(\tau) \) into the products of its \( \hat{r}_\tau \) principal components \( \hat{g}(\tau) \) and their loadings \( \hat{\lambda}(\tau) \) via eq(12). The \( (i,k) \)-th element of \( \hat{\lambda}(\tau) \), denoted as \( \hat{\lambda}_{i,k}(\tau) \), can be interpreted as the exposure of asset \( i \) to the \( k \)-th latent factor (or in finance terminology, “quantity of risk”). And the \( (t,k) \)-th elements of \( \hat{g}(\tau) \), denoted as \( \hat{g}_{t,k}(\tau) \), can be interpreted as the compensation of the risk exposure to the \( k \)-th latent factor in time period \( t \) (or in finance terminology, “price of risk”). The model are estimated with different tuning parameters \( \nu_1 \) and \( \nu_2 \), and use our proposed BIC to select the optimal tuning parameters. The details of the information criteria can be found in equation (25).

The tuning parameter \( \nu_1 \) governs the sparsity of the coefficient vector \( \theta \). The larger \( \nu_1 \) is, the larger the shrinkage effect on \( \theta \). Figure 2 illustrate the effect of this shrinkage. With \( \nu_2 \) fixed, as the value of \( \nu_1 \) increases, more coefficients in the estimated \( \theta \) vector shrink to zero. From a statistical point of view, the “effective characteristics” that can explain stock returns are those with non-zero coefficient \( \theta \) at relatively large values of \( \nu_1 \).

Table 3 reports the relationship between tuning parameter \( \nu_2 \) and rank of estimated \( \Pi \) at different quantiles. It shows that the tuning parameter \( \nu_2 \) governs the rank of matrix \( \Pi \), and that as \( \nu_2 \) increases, we penalize more on the rank of matrix \( \Pi \) through its nuclear norm.

| \( \log_{10}(\nu_2) \) | \( \tau = 0.1 \) | \( \tau = 0.5 \) | \( \tau = 0.9 \) |
|-----------------|----------------|----------------|----------------|
| -6.0            | 228            | 228            | 228            |
| -5.5            | 228            | 228            | 228            |
| -5.0            | 228            | 228            | 228            |
| -4.5            | 164            | 228            | 168            |
| -4.0            | 1              | 7              | 2              |
| -3.5            | 1              | 1              | 1              |
| -3.0            | 0              | 0              | 0              |

Note: Estimated under different values of turning parameter \( \nu_2 \), when \( \nu_1 = 10^{-5} \) is fixed. The results are reported for quantiles 10%, 50% and 90%.

The left panel of Table 4 reports the estimated coefficients in the sparse part when we fix the tuning parameters at \( \log_{10}(\nu_1) = -3.5 \) and \( \log_{10}(\nu_2) = -4 \). The signs of some characteristics are the same across the quantiles, e.g. size (mve), book-to-market (bm), momentum
Table 4
Sparse Part Coefficients at Different Quantiles.

| Characteristic | acc | range | beta | bm | chinv | dy | egr | ep | gma | idiovol | ill | invest | lev | lgr | mom12m | mom1m | mve | operprof | roaq | roeq | std_dolvol | std_turn | sue | turn |
|----------------|-----|-------|------|----|-------|----|-----|----|-----|--------|-----|---------|----|-----|--------|--------|-----|--------|------|------|----------|----------|-----|------|
| \(\tau = 0.1\) | -0.089 | -2.574 | 0.174 | 0.371 | 0 | -0.086 | -0.106 | 0.199 | 0 | -1.229 | -0.097 | 0.183 | -0.106 | -0.166 | -0.150 | -0.038 | 0 | 0.221 | 0.073 | 0 | 0.105 | -0.796 |
| \(\tau = 0.5\) | -0.074 | -0.481 | -0.116 | 0.175 | 0 | 0 | -0.053 | -0.116 | 0.057 | -0.071 | 0.146 | 0.063 | -0.037 | -0.077 | -0.384 | -0.093 | 0.025 | 0.242 | 0.041 | 0 | 0.061 | -0.083 |
| \(\tau = 0.9\) | -0.041 | 2.526 | -0.406 | 0.263 | -0.152 | 0.119 | -0.091 | -0.479 | 0 | 1.438 | 0.146 | 0.129 | 0 | 0 | -0.571 | -0.811 | 0.088 | -0.147 | 0 | 0.045 | 0.386 | -0.330 | 0 |

Note: The left panel reports the estimated coefficient vector \(\theta\) in the sparse part for quantiles 10%, 50% and 90%, when the tuning parameters are fixed at \(\log_{10}(\nu_1) = -3.5\), \(\log_{10}(\nu_2) = -4\). The right panel reports the estimated coefficient vector \(\theta\) under the when the turning parameters are optimal, as selected by BIC (indicated in Table 5).

Table 5 reports the selected optimal tuning parameters \(\nu_1\) and \(\nu_2\) for different quantiles. The tuning parameters are selected via BIC based on (25) as discussed in Section 5. For every \(\nu_1\) and \(\nu_2\), we get the estimates \(\tilde{\theta}(\nu_1, \nu_2)\) and \(\tilde{\Pi}(\nu_1, \nu_2)\) and the number of factors \(r = \text{rank}(\tilde{\Pi}(\nu_1, \nu_2))\). The \(\theta\) vector is sparse with non-zero coefficients on selected characteristics. The 10% quantile of returns has only 1 latent factor, and 3 selected characteristics. The median of returns has 7 latent factors and 2 selected characteristics. The 90% quantile of returns has 2 latent factors and 7 selected characteristics. Range is the only characteristic selected across all 3 quantiles. Idiosyncratic volatility is selected at 10% and 90% quantiles.
with opposite signs. 1-month momentum is selected at 50% and 90% percentiles, with negative sign suggesting reversal in returns.

### Table 5
Selected Optimal Tuning Parameters and Number of Factors

| τ    | optimal r | optimal ν₁ | optimal ν₂ |
|------|-----------|------------|------------|
| 0.1  | 1         | 10⁻⁶.⁵     | 10⁻⁴       |
| 0.5  | 7         | 10⁻².⁵     | 10⁻⁴       |
| 0.9  | 2         | 10⁻².⁷⁵    | 10⁻⁴       |

Note: This table reports the selected optimal tuning parameter ν₁ and ν₂ that minimize the objective function in equation (25) for different quantiles.

Overall, the empirical evidence suggests that both firm characteristics and latent risk factors have valuable information in explaining stock returns. In addition, we find that the selected characteristics and number of latent factors differ across the quantiles.

**Interpretation of Latent Factors.** Table 6 below reports the variance in the matrix Π explained by each Principal Component (PC) or latent factor. At upper and lower quantiles, the first PC dominates. At the median there are more latent factors accounting for the variations in Π, with second PC explaining 13.8% and third PC explaining 6.8%.

### Table 6
Percentage of Π explained by PC

| τ    | PC1   | PC2   | PC3   | PC4   | PC5   | PC6   | PC7   | Total |
|------|-------|-------|-------|-------|-------|-------|-------|-------|
| 0.1  | 100.00% | 13.71% | 6.78% | 4.12% | 1.11% | 0.45% | 0.01% | 100.00% |
| 0.5  | 73.82%   | 0.32%  | 0.45% | 0.01% |       |       |       | 100.00% |
| 0.9  | 99.68%   |       |       |       |       |       |       | 100.00% |

Note: Variance of matrix Π explained by each principal component for different quantiles.

We also found the first PC captures the market returns in all three quantiles: Figure 3 plots the first principal component against the monthly returns of S&P500 index, showing that they have strong positive correlations.

**Acknowledgements.** We are grateful to Victor Chernozhukov, Iván Fernández-Val, Bryan Graham, Hiroaki Kaido, Anna Mikusheva, Whitney Newey, Eric Renault, Jeremy Smith, and Vasilis Syrgkanis for helpful discussions.
Fig 3. The S&P 500 Index Return and the First PC at Different Quantiles. This figure plots the first PC of matrix \( \Pi \) against S&P500 index monthly return for quantiles 10% (left), 50% (middle), and 90% (right).

REFERENCES

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

[2] Ali, A., Kolter, Z. and Tibshirani, R. (2016). The multiple quantile graphical model. In *Advances in Neural Information Processing Systems* 3747–3755.

[3] Ando, T. and Baj, J. (2020). *Journal of the American Statistical Association* 115, 266–279.

[4] Arellano, M. and Bonhomme, S. (2017). Quantile selection models with an application to understanding changes in wage inequality. *Econometrica* 85, 1–28.

[5] Athey, S., Bayati, M., Doudchenko, N., Imbens, G. and Khosravi, K. (2018). Matrix completion methods for causal panel data models.

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

[7] Bai, J. and Feng, J. (2019). Robust Principal Components Analysis with Non-Sparse Errors. *arXiv preprint arXiv:1902.08735*.

[8] Bai, J. and Li, K. (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics* 40, 436–465.

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

[10] Bai, J. and Ng, S. (2017). Principal components and regularized estimation of factor models. *arXiv preprint arXiv:1708.08137*.

[11] Bai, J. and Ng, S. (2019). Matrix completion, counterfactuals, and factor analysis of missing data. *arXiv preprint arXiv:1910.06677*.

[12] Belloni, A. and Chernozhukov, V. (2009). On the computational complexity of MCMC-based estimators in large samples. *The Annals of Statistics* 37, 2011–2055.

[13] Belloni, A. and Chernozhukov, V. (2011). \( \ell_1 \)-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics* 39, 82–130.

[14] Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics* 37, 1705–1732.

[15] Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning* 3, 1–122.

[16] Brahma, P. P., She, Y., Li, S., Li, J. and Wu, D. (2017). Reinforced robust principal component pursuit. *IEEE transactions on neural networks and learning systems* 29, 1525–1538.

[17] Cai, J.-F., Candès, E. and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization* 20, 1956–1982.

[18] Candès, E. and Plan, Y. (2010). Matrix completion with noise. *Proceedings of the IEEE* 98, 925–936.

[19] Candès, E. and Plan, Y. (2011). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory* 57, 2342–2359.

[20] Candès, E. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics* 9, 717.

[21] Candès, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics* 35, 2313–2351.

[22] Chamberlain, G. and Rothschild, M. (1983). ARBITRAGE, FACTOR STRUCTURE, AND MEAN-VARIANCE ANALYSIS ON LARGE ASSET MARKETS. *Econometrica (pre-1986)* 51, 1281.
[23] Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. The Annals of Statistics 43 177–214.
[24] Chen, M. (2014). Estimation of nonlinear panel models with multiple unobserved effects. Warwick Economics Research Paper Series No. 1120.
[25] Chen, L., Dolado, J. and Gonzalo, J. (2018). Quantile factor models.
[26] Chen, M., Fernández-Val, I. and Weidner, M. (2014). Nonlinear panel models with interactive effects. arXiv preprint arXiv:1412.5647.
[27] Chernozhukov, V., Hansen, C. and Liao, Y. (2017). A laval attack on the recovery of sums of dense and sparse signals. The Annals of Statistics 45 39–76.
[28] Chernozhukov, V., Hansen, C., Liao, Y. and Zhu, Y. (2018). Inference For Heterogeneous Effects Using Low-Rank Estimations. arXiv preprint arXiv:1812.08089.
[29] Cochrane, J. H. (2009). Asset pricing: Revised edition. Princeton university press.
[30] Cochrane, J. H. (2011). Presidential address: Discount rates. The Journal of finance 66 1047–1108.
[31] Connor, G. and Korajczyk, R. A. (1988). Risk and return in an equilibrium APT: Application of a new test methodology. Journal of financial economics 21 255–289.
[32] Dalalyan, A., Hebiri, M. and Lederer, J. (2017). On the prediction performance of the lasso. Bernoulli 23 552–581.
[33] Daniel, K. and Titman, S. (1997). Evidence on the characteristics of cross sectional variation in stock returns. the Journal of Finance 52 1–33.
[34] Daniel, K. and Titman, S. (1998). Characteristics or covariances. Journal of Portfolio Management 24 24–33.
[35] de Castro, L. and Galvao, A. F. (2019). Dynamic quantile models of rational behavior. Econometrica 87 1893–1939.
[36] Elsener, A. and Van De Geer, S. (2018). Robust low-rank matrix estimation. The Annals of Statistics 46 3481–3509.
[37] Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. Journal of financial economics 33 3–56.
[38] Fazel, M. (2002). Matrix rank minimization with applications.
[39] Feng, J. (2019). Regularized Quantile Regression with Interactive Fixed Effects. arXiv preprint arXiv:1911.00166.
[40] Feng, G., Giglio, S. and Xiu, D. (2019). Taming the factor zoo: A test of new factors Technical Report, National Bureau of Economic Research.
[41] Galvao, A. (2011). Quantile regression for dynamic panel data with fixed effects. Journal of Econometrics 164 142–157.
[42] Galvao, A. and Kato, K. (2016). Smoothed quantile regression for panel data. Journal of econometrics 193 92–112.
[43] Galvao, A. and Montes-Rojas, G. V. (2010). Penalized quantile regression for dynamic panel data. Journal of Statistical Planning and Inference 140 3476–3497.
[44] Giannone, D., Lenza, M. and Primiceri, G. (2017). Economic predictions with big data: The illusion of sparsity.
[45] Giglio, S. and Xiu, D. (2018). Asset pricing with omitted factors. Chicago Booth Research Paper 16-21.
[46] Giovannetti, B. C. (2013). Asset pricing under quantile utility maximization. Review of Financial Economics 22 169–179.
[47] Graham, B. S., Hahn, J., Poirier, A. and Powell, J. L. (2018). A quantile correlated random coefficients panel data model. Journal of Econometrics 206 305–335.
[48] Green, J., Hand, J. and Zhang, F. (2017). The characteristics that provide independent information about average us monthly stock returns. The Review of Financial Studies 30 4389–4436.
[49] Han, Y., He, A., Rapach, D. and Zhou, G. (2018). What Firm Characteristics Drive US Stock Returns? Available at SSRN 3185335.
[50] He, X., Wang, L. and Hong, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. The Annals of Statistics 41 342–369.
[51] Kato, K., Galvao Jr, A. F. and Montes-Rojas, G. V. (2012). Asymptotics for panel quantile regression models with individual effects. Journal of Econometrics 170 76–91.
[52] Koenker, R. (2000). Galton, Edgeworth, Frisch, and prospects for quantile regression in econometrics. Journal of Econometrics 95 347–374.
[53] Koenker, R. (2004). Quantile regression for longitudinal data. Journal of Multivariate Analysis 91 74–89.
[54] Koenker, R. (2005). Quantile regression. Cambridge University Press, New York.
[55] Koenker, R. and Machado, J. A. (1999). Goodness of fit and related inference processes for quantile regression. Journal of the american statistical association 94 1296–1310.
Koeneker, R., Chernozhukov, V., He, X. and Peng, L. (2017). Handbook of quantile regression. CRC press.

Koltchinskii, V., Lounici, K. and Tsybakov, A. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. The Annals of Statistics 39 2302–2329.

Kozak, S., Nagel, S. and Santosh, S. (2019). Shrinking the cross-section. Journal of Financial Economics.

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

Lettau, M. and Pelger, M. (2018). Estimating latent asset-pricing factors Technical Report, National Bureau of Economic Research.

Ma, S., Linton, O. and Gao, J. (2019). Estimation and inference in semiparametric quantile factor models.

Manski, C. F. (1988). Ordinal utility models of decision making under uncertainty. Theory and Decision 25 79–104.

Moon, H. R. and Weidner, M. (2015). Linear regression for panel with unknown number of factors as interactive fixed effects. Econometrica 83 1543–1579.

Moon, H. R. and Weidner, M. (2018). Nuclear norm regularized estimation of panel regression models. arXiv preprint arXiv:1810.10987.

Negahban, S. and Wainwright, M. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. The Annals of Statistics 39 1069–1097.

Padilla, O. H. M. and Chatterjee, S. (2020). Risk Bounds for Quantile Trend Filtering. arXiv preprint arXiv:2007.07472.

Recht, B., Fazel, M. and Parrilo, P. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM review 52 471–501.

Rohde, A. and Tsybakov, A. (2011). Estimation of high-dimensional low-rank matrices. The Annals of Statistics 39 887–930.

Ross, S. (1976). The arbitrage theory of capital asset pricing. Journal of Economic Theory 13 341–360.

Rostek, M. (2010). Quantile maximization in decision theory. The Review of Economic Studies 77 339–371.

Sagner, A. G. (2019). Three essays on quantile factor analysis, PhD thesis, Boston University.

She, Y. and Chen, K. (2017). Robust reduced-rank regression. Biometrika 104 633–647.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58 267–288.

Van der Vaart, A. W. and Wellner, J. A. (1996). Weak convergence and empirical processes: with applications to statistics. Springer.

Wang, L., Wu, Y. and Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. Journal of the American Statistical Association 107 214–222.

Wong, K. C., Li, Z. and Tewari, A. (2020). Lasso guarantees for β-mixing heavy-tailed time series. The Annals of Statistics 48 1124–1142.

Wu, Y. and Yin, G. (2015). Conditional quantile screening in ultrahigh-dimensional heterogeneous data. Biometrika 102 65–76.

Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. The Annals of Probability 94–116.

Yu, Y., Wang, T. and Samworth, R. J. (2014). A useful variant of the Davis–Kahan theorem for statisticians. Biometrika 102 315–323.

Zheng, Q., Peng, L. and He, X. (2015). Globally adaptive quantile regression with ultra-high dimensional data. Annals of statistics 43 2225.