WARWICK

ECONOMICS

High Dimensional Latent Panel Quantile Regression with an Application to Asset Pricing

Alexandre Belloni, Mingli Chen, Oscar Hernan
Madrid Padilla, and Zixuan (Kevin) Wang

December 2019 No: 1230

Warwick Economics Research Papers

ISSN 2059-4283 (online)
ISSN 0083-7350 (print)

High Dimensional Latent Panel Quantile Regression with an Application to Asset Pricing *

Alexandre Belloni[1], Mingli Chen[2], Oscar Hernan Madrid Padilla[3], and Zixuan (Kevin) Wang[4]

[1]Fuqua Business School, Duke University
[2]Department of Economics, University of Warwick
[3]Department of Statistics, University of California, Los Angeles
[4]Harvard University

December 4, 2019

Abstract

We propose a generalization of the linear panel quantile regression model to accommodate both *sparse* and *dense* parts: sparse means 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 ℓ_1-penalised Quantile Regression, and for traditional latent factor estimator, such as PCA. We propose a new estimation procedure, based on the ADMM algorithm, that consists of combining the quantile loss function with ℓ_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.

Our proposed model has a "Characteristics + Latent Factors" 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.

Keywords: High-dimensional quantile regression; factor model; nuclear norm regularization; panel data; asset pricing; characteristic-based model

*We are grateful to conversations with Victor Chernozhukov, Iván Fernández-Val, Bryan Graham, Hiroaki Kaido, Anna Mikusheva, Whitney Newey, and Vasilis Syrgkanis.

1 Introduction

A central question in asset pricing is to explain the returns of stocks. According to the Arbitrage Pricing Theory, when the asset returns are generated by a linear factor model, there exists a stochastic discount factor linear in the factors that prices the returns (Ross (1976), Cochrane (2009)). However, from an empirical perspective, testing the theory is difficult because the factors are not directly observed.

To overcome this challenge, one approach uses characteristic-sorted portfolio returns to mimic the unknown factors. This approach for understanding expected returns can be captured by the following panel data model (Cochrane (2011)):

$$\mathbb{E}(Y_{i,t}|X_{i,t-1}) = X'_{i,t-1}\theta. \tag{1}$$

The drawback of this approach is that it requires a list of pre-specified characteristics which are chosen based on empirical experience and thus somewhat arbitrarily (Fama and French (1993)). In addition, the literature has documented a zoo of new characteristics, and the proliferation of characteristics in this "variable zoo" leads to the following questions: "which characteristics really provide independent information about average returns, and which are subsumed by others?" (Cochrane (2011)). Another approach uses statistical factor analysis, e.g. Principal Component Analysis (PCA), to extract latent factors from asset returns (Chamberlain and Rothschild (1983), Connor and Korajczyk (1988)):

$$\mathbb{E}(Y_{i,t}) = \lambda'_i g_t. \tag{2}$$

However, one main critique with this approach is that the latent factors estimated via PCA are purely statistical factors, thus lacks economic insight Campbell (2017).

We extend both modeling approaches and propose a "Characteristics + Latent Factors" framework. By incorporating the "Characteristics" documented, we improve the economic interpretability and explanatory power of the model. The characteristics can have a *sparse* structure, meaning although a large set of variables is available, only a much smaller subset of them might have predictive power. We also incorporate "Latent Factors", one benefit of having this part is that it might help alleviate the "omitted variable bias" problem (Giglio and Xiu (2018)). As in the literature typically these latent factors are estimated via PCA, which means all possible latent explanatory variables (those are not included in the model) might be important for prediction although their individual contribution might be small, we term this as the *dense* part. [1]

Hence, our framework allows for "Sparse + Dense" modeling with the time series and cross-section of asset returns. In addition, we focus on understanding the quantiles (hence the entire distribution) of returns rather than just the mean, in line with recent interest in quantile factor models (e.g. Ando and Bai (2019), Chen et al. (2018), Ma et al. (2019), Feng (2019), and Sagner (2019)). Specifically, we study the following high dimensional latent panel quantile regression model with $Y \in \mathbb{R}^{n\times T}$ and $X \in \mathbb{R}^{n\times T\times p}$ satisfying

$$F^{-1}_{Y_{i,t}|X_{i,t};\theta(\tau),\lambda_i(\tau),g_t(\tau)}(\tau) = X'_{i,t}\theta(\tau) + \lambda_i(\tau)'g_t(\tau), \quad i = 1\ldots,n, \ t = 1,\ldots,T, \tag{3}$$

[1] More about sparse modeling and dense modeling can be found in Giannone et al. (2017). See also Chernozhukov et al. (2017).

where i denotes subjects (or assets in our asset pricing setting), t denotes time, $\theta(\tau) \in \mathbb{R}^p$ is the vector of coefficients, $\lambda_i(\tau)$, $g_t(\tau) \in \mathbb{R}^{r_\tau}$ with $r_\tau \ll \min\{n, T\}$ (denote $\Pi_{i,t}(\tau) = \lambda_i(\tau)' g_t(\tau)$, then $\Pi(\tau) \in \mathbb{R}^{n \times T}$ is a low-rank matrix with unknown rank r_τ), $\tau \in [0, 1]$ is the quantile index, and $F_{Y_{i,t}|X_{i,t};\theta(\tau),\lambda_i(\tau),g_t(\tau)}$ (or $F_{Y_{i,t}|X_{i,t};\theta(\tau),\Pi_{i,t}(\tau)}$) is the cumulative distribution function of $Y_{i,t}$ conditioning on $X_{i,t}$, $\theta(\tau)$ and $\lambda_i(\tau)$, $g_t(\tau)$ (or $\Pi_{i,t}(\tau)$). Our framework also allows for the possibility of lagged dependent data. Thus, we model the quantile function at level τ as a linear combination of the predictors plus a low-rank matrix (or a factor structure). Here, we allow for the number of covariates p, and the time horizon T, to grow to infinity as n grows. Throughout the paper 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.

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 (Kozak et al. (2019); Feng et al. (2019); Han et al. (2018)). The second question is related to an classical literature starting from 1980s on statistical factor models of stock returns (Chamberlain and Rothschild (1983); Connor and Korajczyk (1988) and recently Lettau and Pelger (2018)). The third question extends the literature in late 1990s on stock return and firm characteristics (Daniel and Titman (1997, 1998)) and further asks whether the relationship is heterogenous 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 exhibits 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. She and Chen (2017)). As our framework is based on 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 Candès and Recht (2009) 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 ℓ_1 regularization on the vector of coefficients of the control variables, similarly to Belloni and Chernozhukov (2011) which mainly focused on the cross-sectional data setting. Note that with regards to sparsity, we must be cautious, specially when considering predictive models (She and Chen (2017)). 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 also propose a novel Alternating Direction Method of Multipliers (ADMM) algorithm (Boyd et al. (2011)) that allows us to estimate, at different quantile levels, both the vector of coefficients and the low-rank matrix. Our proposed algorithm can easily be adjusted to other nonlinear models with a low-rank matrix (with or without covariates).

We view our work as complementary to the low dimensional quantile regression with interactive fixed effects framework as of the recent work of Feng (2019), and the mean estimation setting in Moon and Weidner (2018). However, unlike Moon and Weidner (2018) and Feng (2019), we allow the number of covariates p to be large, perhaps $p \gg nT$. 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 Moon and Weidner (2018) and Feng (2019). On the theoretical side, allowing $p \gg nT$ requires a systematically different analysis as compared to Feng (2019), as it is known that ordinary quantile regression is inconsistent in high dimensional settings ($p \gg nT$), see Belloni and Chernozhukov (2011).

Related Literature. Our work contributes to the recent growing literature on panel quantile model. Abrevaya and Dahl (2008), Graham et al. (2018), Arellano and Bonhomme (2017), considered the fixed T asymptotic case. Kato et al. (2012) formally derived the asymptotic properties of the fixed effect quantile regression estimator under large T asymptotics, and Galvao and Kato (2016) further proposed fixed effects smoothed quantile regression estimator. Galvao (2011) works on dynamic panel. Koenker (2004) 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 Lamarche (2010), Galvao and Montes-Rojas (2010). We refer to Chapter 19 of Koenker et al. (2017) for a review.

Our work also contributes to the literature on nuclear norm penalisation, which has been widely studied in the machine learning and statistical learning literature, Fazel (2002), Recht et al. (2010); Koltchinskii et al. (2011); Rohde and Tsybakov (2011), Negahban and Wainwright (2011), Brahma et al. (2017). Recently, in the econometrics literature Athey et al. (2018) proposes a framework of matrix completion for estimating causal effects, Bai and Ng (2017) for estimating approximate factor model, Chernozhukov et al. (2018) considered the heterogeneous coefficients version of the linear panel data interactive fixed model where the main coefficients has a latent low-rank structure, Bai and Feng (2019) for robust principal component analysis, and Bai and Ng (2019) for imputing counterfactual outcome.

Finally, our results contribute to a growing literature on high dimensional quantile regression. Wang et al. (2012) considered quantile regression with concave penalties for ultra-high dimensional data; Zheng et al. (2015) proposed an adaptively weighted ℓ_1-penalty for globally concerned quantile regression. Screening procedures based on moment conditions motivated by the quantile models have been proposed and analyzed in He et al. (2013) and Wu and Yin (2015) in the high-dimensional regression setting. We refer to Koenker et al. (2017) for a review.

To sum-up, our paper makes the following contributions. First, we propose a new class 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 derive all the above-mentioned results. This leads to the use in our proofs of some novel techniques from high dimensional statistics/econometrics, spectral theory, empirical process, etc. Third, the proposed estimators inherit from quantile regression certain robustness properties to the presence of outliers and heavy-tailed distributions in the idiosyncratic component of a factor model. 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 Appendix.

Notation. For $m \in \mathbb{N}$, we write $[m] = \{1, \ldots, m\}$. For a vector $v \in \mathbb{R}^p$ we define its ℓ_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 ℓ_1 norm as $\|v\|_1 = \sum_{j=1}^p |v_j|$. We denote $\|v\|_{1,n,T} = \sum_j^p \hat{\sigma}_j |v_j|$ the ℓ_1-norm weighted by $\hat{\sigma}_j$'s (details can be found in eq (20)). 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 $\|A\|_2 = \sup_{x\,:\,\|x\|=1} \sqrt{x'A'Ax}$, its infinity norm by $\|A\|_\infty = \max\{|A_{i,j}| \,:\, i \in [n],\ j \in [T]\}$, its rank by $\text{rank}(A)$, and its nuclear norm by $\|A\|_* = \text{trace}(\sqrt{A'A})$ where A' is the transpose of A. The jth column A is denoted by $A_{\cdot,j}$. Furthermore, the multiplication of a tensor $X \in \mathbb{R}^{I_1\times\ldots\times I_m}$ with a vector $a \in \mathbb{R}^{I_m}$ is denoted by $Z := Xa \in \mathbb{R}^{I_1\times\ldots\times I_{m-1}}$, and, explicitly, $Z_{i_1,\ldots i_{m-1}} = \sum_{j=1}^{I_m} X_{i_1,\ldots,i_{m-1},j}\, a_j$. 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

$$F^{-1}_{Y_{i,t}|X_{i,t};\theta(\tau),\Pi_{i,t}(\tau)}(\tau) \;=\; X'_{i,t}\theta(\tau) + \Pi_{i,t}(\tau), \quad i \;=\; 1\ldots,n, \;\; t \;=\; 1,\ldots,T, \tag{4}$$

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 $F_{Y_{i,t}|X_{i,t};\theta(\tau),\Pi_{i,t}(\tau)}$ 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 τ 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 $\lambda_i(\tau),\, g_t(\tau) \in \mathbb{R}^{r_\tau}$, this immediately leads to the following setting

$$F^{-1}_{Y_{i,t}|X_{i,t};\theta(\tau),\Pi_{i,t}(\tau)}(\tau) \;=\; X'_{i,t}\theta(\tau) + \lambda_i(\tau)'g_t(\tau). \tag{5}$$

where we model the quantile function at level τ 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 (Bai (2009)), nonlinear panel data models (Chen (2014); Chen et al. (2014)).

Note, for eq (5), additional identification restrictions are needed for estimating $\lambda_i(\tau)$ and $g_t(\tau)$ (see Bai and Ng (2013)). In addition, in nonlinear panel data models, this create additional difficulties in estimation, as the latent factors and their loadings part are nonconvex. [2]

2.2 The Penalized Estimator, and its Convex Relaxation

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

$$\begin{array}{ll} \underset{\tilde\theta\in\mathbb{R}^p,\ \tilde\Pi\in\mathbb{R}^{n\times T}}{\text{minimize}} & \dfrac{1}{nT}\displaystyle\sum_{t=1}^{T}\sum_{i=1}^{n}\rho_\tau(Y_{i,t} - X'_{i,t}\tilde\theta - \tilde\Pi_{i,t}) \\ \text{subject to} & \text{rank}(\tilde\Pi) \le r_\tau, \\ & \|\tilde\theta\|_0 = s_\tau, \end{array} \tag{6}$$

[2] Different identification conditions might result in different estimation procedures for λ and f, see Bai and Li (2012) and Chen (2014).

where $\rho_\tau(t) = (\tau - 1\{t \leq 0\})t$ is the quantile loss function as in Koenker (2005), s_τ is a parameter that directly controls the sparsity of $\tilde{\theta}$, and r_τ controls the rank of the estimated latent matrix.

While the formulation in (6) seems appealing, as it enforces variable selection and low-rank matrix estimation simultaneously, (6) is a non-convex problem due to the constraints posed by the $\|\cdot\|_0$ and $\mathrm{rank}(\cdot)$ functions. We propose a convex relaxation of (6). Inspired by the seminal works of Tibshirani (1996) and Candès and Recht (2009), we formulate the problem

$$\begin{aligned} \min_{\tilde{\theta}\in\mathbb{R}^p,\ \tilde{\Pi}\in\mathbb{R}^{n\times T}} \quad & \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\rho_\tau(Y_{i,t} - X_{i,t}'\tilde{\theta} - \tilde{\Pi}_{i,t}) \\ \text{subject to} \quad & \|\tilde{\Pi}\|_* \leq \nu_2, \\ & \sum_{j=1}^{p} w_j|\tilde{\theta}_j| \leq \nu_1, \end{aligned} \tag{7}$$

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).

In principle, one can use any convex solver software to solve (7), since this is a convex optimization problem. However, for large scale problems a more careful implementation might be needed. Section 3.1 presents a scheme for solving (7) that is based on the ADMM algorithm ((Boyd et al., 2011)).

2.3 Summary of results

We now summarize our main results. For the model defined in (4):

- Under (4), $s_\tau \ll \min\{n, T\}$, 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

$$\|\hat{\theta}(\tau) - \theta(\tau)\| = O_{\mathbb{P}}\left(\sqrt{s_\tau}\max\{\sqrt{\log p}, \sqrt{\log n}, \sqrt{r_\tau}\}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{T}}\right)\right). \tag{8}$$

  and

$$\frac{1}{nT}\|\hat{\Pi}(\tau) - \Pi(\tau)\|_F^2 = O_{\mathbb{P}}\left(s_\tau \max\{\log p, \log n, r_\tau\}\left(\frac{1}{n} + \frac{1}{T}\right)\right), \tag{9}$$

  Importantly, the rates in (8) and (9), ignoring logarithmic factors, 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 β-mixing assumption, Theorem 1 provides a statistical guarantee for estimating $(\theta(\tau), \Pi(\tau))$. This result can be thought as a generalization of the statements in (8) and (9).

- 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 3.1), which is based on the ADMM algorithm (Boyd et al. (2011)).
- 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 Belloni and Chernozhukov (2011), 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

3.1 Estimation with High Dimensional Covariates

In this subsection, we describe the main steps of our proposed ADMM algorithm, details can be found in Section A. We start by introducing slack variables to the original problem (7). As a result, a problem equivalent to (7) is

$$\begin{aligned} \min_{\substack{\tilde{\theta},\tilde{\Pi},V \\ Z_\theta, Z_\Pi, W}} \quad & \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\rho_\tau(V_{i,t}) \,+\, \nu_1 \sum_{j=1}^{p} w_j |Z_{\theta_j}| \,+\, \nu_2 \|\tilde{\Pi}\|_* \\ \text{subject to} \quad & V = W, \quad W = Y - X\tilde{\theta} - Z_\Pi, \\ & Z_\Pi - \tilde{\Pi} = 0, \quad Z_\theta - \tilde{\theta} = 0. \end{aligned} \tag{10}$$

To solve (10), we propose a scaled version of the ADMM algorithm which relies on the following

Augmented Lagrangian

$$\begin{aligned}\mathcal{L}(\tilde{\theta}, \tilde{\Pi}, VZ_\theta, Z_\Pi, W, U_V, U_W, U_\Pi, U_\theta) \;=\; & \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\rho_\tau(V_{i,t}) \;+\; \nu_1\sum_{j=1}^{p} w_j|Z_{\theta_j}| \;+\; \nu_2\|\tilde{\Pi}\|_* \\ & + \frac{\eta}{2}\|V - W + U_V\|_F^2 \;+\; \frac{\eta}{2}\|W - Y + X\theta + Z_\Pi + U_W\|_F^2 \\ & + \frac{\eta}{2}\|Z_\Pi - \tilde{\Pi} + U_\Pi\|_F^2 \;+\; \frac{\eta}{2}\|Z_\theta - \tilde{\theta} + U_\theta\|_F^2,\end{aligned} \tag{11}$$

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

Notice that in (11), we have followed the usual construction of ADMM via introducing the scaled dual variables corresponding to the constraints in (10) – those are U_V, U_W, U_Π, and U_θ. Next, recall that ADMM proceeds by iteratively minimizing the Augmented Lagrangian in blocks with respected to the original variables, in our case $(V, \tilde{\theta}, \tilde{\Pi})$ and (W, Z_θ, Z_Π), and then updating the scaled dual variables (see Equations 3.5–3.7 in Boyd et al. (2011)). The explicit updates can be found in the Appendix. Here, we highlight the updates for Z_θ, $\tilde{\Pi}$, and V. For updating Z_θ at iteration $k+1$, we solve the problem

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

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

$$\tilde{\Pi}^{(k+1)} \leftarrow \underset{\tilde{\Pi} \in \mathbb{R}^{n\times T}}{\arg\min}\left\{\frac{\nu_2}{\eta}\|\tilde{\Pi}\|_* + \frac{1}{2}\|Z_\Pi^{(k)} - \tilde{\Pi} + U_\Pi^{(k)}\|_F^2\right\}, \tag{12}$$

via the singular value shrinkage operator, see Theorem 2.1 in Cai et al. (2010).

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

$$V^{(k+1)} \leftarrow \underset{V \in \mathbb{R}^{n\times T}}{\arg\min}\left\{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\rho_\tau(V_{i,t}) + \frac{\eta}{2}\|V - W^{(k)} + U_V^{(k)}\|_F^2\right\}, \tag{13}$$

which can be found in closed formula by Lemma 5.1 from Ali et al. (2016).

Remark 1. *After estimating $\Pi(\tau)$, we can estimate $\lambda_i(\tau)$ and $g_t(\tau)$ via the singular value decomposition of $\hat{\Pi}(\tau)$ and following equation*

$$\hat{\Pi}(\tau)_{i,t} \;=\; \hat{\lambda}_i(\tau)'\hat{g}_t(\tau), \tag{14}$$

where $\hat{\lambda}_i(\tau)$ and $\hat{g}_t(\tau)$ 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 2.

3.2 Estimation without Covariates

Note, when there are no covariates, our proposed ADMM can be simplified. In this case, we face the following problem

$$\min_{\tilde{\Pi}\in\mathbb{R}^{n\times T}} \left\{ \frac{1}{nT}\sum_{i=1}^{n}\sum_{j=1}^{T}\rho_\tau(Y_{i,t}-\tilde{\Pi}_{i,t}) + \nu_2\|\tilde{\Pi}\|_* \right\}. \tag{15}$$

This can be thought as a convex relaxation of the estimator studied in Chen et al. (2018). Problem (15) is also related to the setting of robust estimation of a latent low-rank matrix, e.g. Elsener and van de Geer (2018). However, our approach can also be used to estimate different quantile levels. As for solving (15), we can proceed by doing the iterative updates

$$\tilde{\Pi}^{(k+1)} \leftarrow \arg\min_{\tilde{\Pi}} \left\{ \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\rho_\tau(Y_{i,t}-\tilde{\Pi}_{i,t}) + \frac{\eta}{2}\|\tilde{\Pi}-Z_\Pi^{(k)}+U_\Pi^{(k)}\|_F^2 \right\}, \tag{16}$$

$$Z_\Pi^{(k+1)} \leftarrow \arg\min_{Z_\Pi} \left\{ \frac{\eta}{2}\|\tilde{\Pi}^{(k+1)}-Z_\Pi+U_\Pi^{(k)}\|_F^2 + \nu_2\|Z_\Pi\|_* \right\}, \tag{17}$$

and

$$U_\Pi^{(k+1)} \leftarrow \Pi^{(k+1)} - Z_\Pi^{(k+1)} + U_\Pi^{(k)}, \tag{18}$$

where $\eta > 0$ is the penalty parameter (Boyd et al. (2011)). The minimization in (16) is similar to (13), whereas (17) can be done similarly as in (12).

Although our proposed estimation procedure can be applied to settings (i) with low dimensional covariates, or (ii) without covariates, 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. Our analysis combines tools from high dimensional quantile regression theory (e.g. Belloni and Chernozhukov (2011)), spectral theory (e.g. Vu (2007) and Chatterjee (2015)), and empirical process theory (e.g. Yu (1994) and van der Vaart and Wellner (1996)).

4.1 Main Result

We show that our proposed estimator is consistent in a broad range of models, and in some cases attains minimax rates, as in Candès and Plan (2011).

Before arriving at our main theorem, we start by stating some modeling assumptions. For a fixed $\tau > 0$, we assume that (4) holds. We also let T_τ 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 $r_\tau \;=\; \text{rank}(\Pi(\tau))$.

Throughout, we treat $\Pi(\tau)$ as 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. *The following holds:*

(i) Conditional on Π, $\{(Y_{i,t}, X_{i,t}\}_{t=1,\ldots,T}$ is independent across i. Also, for each $i \in [n]$, the sequence $\{(Y_{i,t}, X_{i,t})\}_{t=1,\ldots,T}$ is stationary and β-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

$$npT\left(\left\lfloor T^{1/(1+\mu')}\right\rfloor\right)^{-\mu} \;\to\; 0. \tag{19}$$

Here, $\gamma_i(k) = \frac{1}{2}\sup\limits_{l\geq 1}\sum_{j=1}^{L}\sum_{j'=1}^{L'}|\mathbb{P}(A_j \cap B_{j'}) - \mathbb{P}(A_j)\mathbb{P}(B_{j'})|$ with $\{A_j\}_{j=1}^{L}$ is a partition of $\sigma(\{X_{i,1}, Y_{i,1}\}, \ldots, \{X_{i,l}, Y_{i,l}\})$, and $\{B_{j'}\}_{j'=1}^{L'}$ partition of $\sigma(\{X_{i,l+k}, Y_{i,l+k}\}, \{X_{i,l+k+1}, Y_{i,l+k+1}\}\ldots)$.

(ii) There exists $\underline{f} > 0$ satisfying

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

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 $\Pi_{i,t}(\tau)$. Furthermore, $f_{Y_{i,t}|X_{i,t};\theta(\tau),\Pi_{i,t}(\tau)}(y|x; \theta(\tau), \Pi_{i,t}(\tau))$ and $\frac{\partial}{\partial y} f_{Y_{i,t}|X_{i,t};\theta(\tau),\Pi_{i,t}(\tau)}(y|x; \theta(\tau), \Pi_{i,t}(\tau))$ are both bounded by $\bar{f}$ and $\bar{f}'$, respectively, uniformly in y and x in the support of $X_{i,t}$.

Note that Assumption 1 is a generalization of the sampling and smoothness assumption of Belloni and Chernozhukov (2011). Furthermore, we highlight that similar to Belloni and Chernozhukov (2011), our framework is rich enough that avoids imposing Gaussian or homoscedastic modeling constraints. However, unlike Belloni and Chernozhukov (2011), we consider panel data with weak correlation across time. In particular, we refer readers to Yu (1994) for thorough discussions on β-mixing.

It is worth mentioning that the parameter μ in Assumption 1 controls the strength of the time dependence in the data. 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 (19).

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

$$\hat{\sigma}_j^2 \;=\; \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} X_{i,t,j}^2, \quad \forall j \in [p], \tag{20}$$

and we require that

$$\mathbb{P}\left(\max_{1\le j\le p} |\hat{\sigma}_j^2 - 1| \;\le\; \frac{1}{4}\right) \;\ge\; 1-\gamma \;\;\to\;\; 1, \quad \textit{as} \quad n\to\infty.$$

Assumption 2 appeared as Condition D.3 in Belloni and Chernozhukov (2011). It is met by general models on the covariates, see for instance Design 2 in Belloni and Chernozhukov (2011).

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

$$(\hat{\theta}(\tau), \hat{\Pi}(\tau)) \;=\; \underset{(\tilde{\theta},\tilde{\Pi})}{\arg\min}\left\{\hat{Q}_\tau(\tilde{\theta},\tilde{\Pi}) \;+\; \nu_1\|\tilde{\theta}\|_{1,n,T} + \nu_2\|\tilde{\Pi}\|_*\right\}, \tag{21}$$

where $\nu_2 \;>\; 0$ is a tuning parameter, $\|\tilde{\theta}\|_{1,n,T} \;:=\; \sum_{j=1}^p \hat{\sigma}_j|\tilde{\theta}_j|$, and

$$\hat{Q}_\tau(\tilde{\theta},\tilde{\Pi}) \;=\; \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\rho_\tau(Y_{i,t} - X_{i,t}'\tilde{\theta} - \tilde{\Pi}_{i,t}),$$

with ρ_τ as defined in Section 2.2.

As it can been seen in Lemma 7 from Appendix B.1, $(\hat{\theta}(\tau)-\theta(\tau), \hat{\Pi}(\tau)-\Pi(\tau))$ belongs to a restricted set, which in our framework is defined as

$$A_\tau = \left\{(\delta,\Delta)\in\mathbb{R}^p\times\mathbb{R}^{n\times T} \;:\; \|\delta_{T_\tau^c}\|_1 + \frac{\|\Delta\|_*}{\sqrt{nT}\sqrt{\log(\max\{n,pc_T\})}} \le C_0\left(\|\delta_{T_\tau}\|_1 + \frac{\sqrt{r_\tau}\|\Delta\|_F}{\sqrt{nT}\sqrt{\log(\max\{n,pc_T\})}}\right)\right\}, \tag{22}$$

for an appropriate positive constant C_0.

Similar in spirit to other high dimensional settings such as those in Candès and Tao (2007), Bickel et al. (2009), Belloni and Chernozhukov (2011) and Dalalyan et al. (2017), we impose an identifiability condition involving the restricted set which is expressed next and will be used in order to attain our main results. Before that, we introduce some notation.

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

Assumption 3. *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

$$0 < \kappa_m := \inf_{(\delta,\Delta)\in A_\tau, \delta\neq 0} \frac{J_\tau^{1/2}(\delta,\Delta)}{\|\delta_{T_\tau \cup \overline{T}_\tau(\delta,m)}\| + \frac{\|\Delta\|_F}{\sqrt{nT}}}, \tag{23}$$

where $c_T = \lceil T^{1/(1+\mu')} \rceil$ for μ' as defined in Assumption 2. Moreover, we assume that the following holds

$$0 < q := \frac{3}{8}\frac{\underline{f}^{3/2}}{\overline{f}'} \inf_{(\delta,\Delta)\in A_\tau, \delta\neq 0} \frac{\left(\mathbb{E}\left(\frac{1}{nT}\sum_{i=1}^n \sum_{t=1}^T (X_{i,t}'\delta + \Delta_{i,t})^2\right)\right)^{3/2}}{\mathbb{E}\left(\frac{1}{nT}\sum_{i=1}^n \sum_{t=1}^T |X_{i,t}'\delta + \Delta_{i,t}|^3\right)}, \tag{24}$$

with $\underline{f}$ and $\overline{f}'$ as in Assumption 1.

Few comments are in order. First, if $\Delta = 0$ then (23) and (24) become the restricted identifiability and nonlinearity conditions as of Belloni and Chernozhukov (2011). Second, the denominator of (23) 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

$$J_\tau(\delta,\Delta) = \underline{f}\,\mathbb{E}((\delta' X_{i,t})^2) + \frac{\underline{f}}{nT}\|\Delta\|_F^2.$$

Hence, $\|\Delta\|_F/(\sqrt{nT})$ appears also in the numerator of (23) and its presence in the denominator of (23) is not restrictive.

We now state our result for estimating $\theta(\tau)$ and $\Pi(\tau)$.

Theorem 1. *Suppose that Assumptions 1-3 hold and that*

$$q \geq C\,\frac{\phi_n\sqrt{c_T(1+s_\tau)\log p}(\sqrt{n}+\sqrt{d_T})}{\sqrt{nd_T}\kappa_0 \underline{f}^{1/2}}, \tag{25}$$

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

$$\|\hat{\theta}(\tau) - \theta(\tau)\| = O_{\mathbb{P}}\left(\frac{\phi_n\left(1+\sqrt{\frac{s_\tau}{m}}\right)}{\kappa_m}\frac{\sqrt{c_T(1+s_\tau)\,\max\{\log(pc_T\vee n), r_\tau\}}}{\kappa_0 \underline{f}^{1/2}}\left(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{d_T}}\right)\right), \tag{26}$$

and

$$\frac{1}{nT}\|\hat{\Pi}(\tau) - \Pi(\tau)\|_F^2 = O_{\mathbb{P}}\left(\frac{\phi_n^2 c_T(1+s_\tau)\max\{\log(pc_T\vee n), r_\tau\}}{\kappa_0^4\,\underline{f}}\left(\frac{1}{n}+\frac{1}{d_T}\right)\right), \tag{27}$$

for choices of the tuning parameters satisfying

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

and

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

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

Theorem 1 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 $\sqrt{s_\tau r_\tau}/\min\{\sqrt{n}, \sqrt{T}\}$, if we ignore all the other factors. Hence, we can consistently estimate $\theta(\tau)$ provided that $\max\{s_\tau, r_\tau\} << \min\{n, T\}$. This is similar to the low-rank condition in Negahban and Wainwright (2011). In the low dimensional case $s_\tau = O(1)$, the rate $\sqrt{r_\tau}/\min\{\sqrt{n}, \sqrt{T}\}$ matches that of Theorem 1 in Moon and Weidner (2018). However, we mainly focus on a loss function that is robust to outliers, and our assumptions also allow for weak dependence across time. Furthermore, the same applies to our rate on the mean squared error for estimating $\Pi(\tau)$, which also matches that in Theorem 1 of Moon and Weidner (2018).

Interestingly, it is expected that the rate in Theorem 1 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(σ^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 Candès and Plan (2011), we have the following lower bound for estimating $\Pi(\tau)$:

$$\inf_{\hat{\Pi}} \sup_{\Pi(\tau)\,:\,\text{rank}(\Pi(\tau))\leq r_\tau} \mathbb{E}\left(\frac{\|\hat{\Pi}(\tau) - \Pi(\tau)\|_F^2}{nT}\right) \;\geq\; \frac{r_\tau \sigma^2}{n}. \tag{28}$$

Notably, the lower bound in (28) matches the rate implied by Theorem 1, ignoring other factors depending on s_τ, κ_0, κ_m, p and ϕ_n. However, we highlight that the upper bound (27) in Theorem 1 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 loadings of the latent matrix $\Pi(\tau)$. This is expressed in Corollary 2 below and is immediate consequence of Theorem 1 and Theorem 3 in Yu et al. (2014).

Corollary 2. *Suppose that the all the conditions of Theorem 1 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 $\lambda(\tau), \hat{\lambda}(\tau), \tilde{\lambda}(\tau), \tilde{\hat{\lambda}}(\tau) \in \mathbb{R}^{n\times r_\tau}$ be matrices with orthonormal columns satisfying*

$$\Pi(\tau) \;=\; \sum_{j=1}^{r_\tau} \sigma_j\, \tilde{\lambda}_{\cdot,j}(\tau) g_{\cdot,j}(\tau)' \;=\; \sum_{j=1}^{r_\tau} \lambda_{\cdot,j}(\tau) g_{\cdot,j}(\tau)',$$

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

$$v_1 \;:=\; \min_{O\in\mathbb{O}_{r_\tau}} \|\hat{g}(\tau)O - g(\tau)\|_F \;=\; O_{\mathbb{P}}\left(\frac{(\sigma_1(\tau) + \sqrt{r_\tau}\,\text{Err})\text{Err}}{(\sigma_{r_\tau-1}(\tau))^2 - (\sigma_{r_\tau}(\tau))^2}\right), \tag{29}$$

and

$$v_2 =: \frac{\|\hat{\lambda}(\tau) - \lambda(\tau)\|_F^2}{nT} = O_{\mathbb{P}}\left(\frac{r_\tau \, \phi_n^2 c_T (1+s_\tau) \max\{\log(pc_T \vee n), r_\tau\}}{\kappa_0^4 \, \underline{f}} \left(\frac{1}{n} + \frac{1}{d_T}\right) + \frac{\sigma_1^2}{nT} \frac{(\sigma_1(\tau) + \sqrt{r_\tau}\,\mathrm{Err})^2 \mathrm{Err}^2}{((\sigma_{r_\tau - 1}(\tau))^2 - (\sigma_{r_\tau}(\tau))^2)^2} \right). \tag{30}$$

Here, $\mathbb{O}_{r_\tau}$ is the group of $r_\tau \times r_\tau$ orthonormal matrices, and

$$\mathrm{Err} := \frac{\phi_n c_T \sqrt{(1+s_\tau) \max\{\log(pc_T \vee n), r_\tau\}}}{\kappa_0^2 \, \underline{f}^{1/2}} \left(\sqrt{n} + \sqrt{d_T}\right).$$

A particularly interesting instance of Corollary 2 is when

$$(\sigma_1(\tau))^2, \, (\sigma_{r_\tau - 1}(\tau))^2 - (\sigma_{r_\tau}(\tau))^2 \asymp nT,$$

a natural setting if the entries of $\Pi(\tau)$ are $O(1)$. Then the upper bound (29) becomes

$$v_1 = O_{\mathbb{P}}\left(\frac{\phi_n \sqrt{c_T (1+s_\tau) \, \max\{\log(pc_T \vee n), r_\tau\}}}{\kappa_0^2 \underline{f}^{1/2}} \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{d_T}}\right)\right),$$

whereas (30) is now

$$v_2 = O_{\mathbb{P}}\left(\frac{r_\tau \, \phi_n^2 c_T (1+s_\tau) \max\{\log(pc_T \vee n), r_\tau\}}{\kappa_0^4 \, \underline{f}} \left(\frac{1}{n} + \frac{1}{d_T}\right)\right).$$

The conclusion of Corollary 2 allows us to provide an upper on the estimation of factors ($g(\tau)$) and loadings ($\lambda(\tau)$) of the latent matrix $\Pi(\tau)$. In particular this justifies the heuristic discussed in (14).

4.2 Correlated predictors

We conclude our theory section by studying the case where the vector of coefficients $\theta(\tau)$ can be dense, in the sense that the number of non-zero entries can be large, perhaps even larger than nT. To make estimation feasible, we impose the condition that $X\theta(\tau)$ can be perturbed into a low-rank matrix, a scenario that can happen when the covariates are highly correlated to each other.

We view our setting below as an extension of the linear model in Chernozhukov et al. (2018) to the quantile framework and reduced rank regression. The specific condition is stated next.

Assumption 4. *With probability approaching one, it holds that* $\mathrm{rank}(X\theta(\tau) + \xi) = O(r_\tau)$*, and*

$$\frac{\|\xi\|_*}{\sqrt{nT}} = O_{\mathbb{P}}\left(\frac{c_T \phi_n \sqrt{r_\tau}(\sqrt{n} + \sqrt{d_T})}{\sqrt{nT}\,\underline{f}} \right),$$

with c_T as defined in Theorem 1. Furthermore, $\|X\theta(\tau) + \Pi(\tau)\| = O_{\mathbb{P}}(1)$.

Notice that in Assumption 4, ξ is an approximation error. In the case $\xi = 0$, the condition implies that $\mathrm{rank}(X\theta(\tau)) = O(r_\tau)$ with probability close to one.

Next, exploiting Assumption 4, we show that (21) provides consistent estimation of the quantile function, namely, of $X\theta(\tau) + \Pi(\tau)$.

Theorem 3. *Suppose that Assumptions 1–2 and 4 hold. Let $(\hat{\theta}(\tau), \hat{\Pi}(\tau))$ be the solution to (21) with the additional constraint that $\|\tilde{\Pi}\|_\infty \leq C$, for a large enough positive constant C.*

Then

$$\frac{1}{nT}\|\hat{\Pi}(\tau) - \Pi(\tau)\|_F^2 \;=\; O_{\mathbb{P}}\left(\frac{(\overline{f}')^2\phi_n^2 c_T r_\tau}{\underline{f}^4}\left(\frac{1}{n} + \frac{1}{d_T}\right)\right),$$

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

$$\nu_1 \asymp \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\|X_{i,t}\|_\infty,$$

and

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

Interestingly, unlike Theorem 1, Theorem 3 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 4 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 ν_1 increases significantly with respect to the choice in Theorem 1, while keeping $\nu_2 \asymp c_T(\sqrt{n}+\sqrt{d_T})/(nT)$.

As for the convergence rate in Theorem 3 for estimating $\Pi(\tau)$, this is of the order $r_\tau c_T(n^{-1} + d_T^{-1})$, if we ignore $\underline{f}$, $\overline{f}'$, and ϕ_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 Candès and Plan (2010) for estimating an $n \times T$ matrix of rank r_τ, provided that $n \asymp T$, see our discussion in Section 4.1.

Finally, notice that we have added an additional tuning parameter C that controls the magnitude of the possible estimate $\tilde{\Pi}$. This is done for technical reasons. We expect that the same upper bound holds without this additional constraint.

5 Simulation

In this section, we evaluate the performance of our proposed approach (ℓ_1-NN-QR) with extensive numerical simulations focusing on the median case, namely the case when $\tau = 0.5$. As benchmarks, we consider the ℓ_1-penalized quantile regression studied in Belloni and Chernozhukov (2009), and similarly we refer to

this procedure as ℓ_1-QR. We also compare with the mean case, which we denote it as ℓ_1-NN-LS as it combines the ℓ_2-loss function with ℓ_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 ν_1 and ν_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 (ν_1, ν_2), our method produces a score $(\hat{\theta}(\tau), \hat{\Pi}(\tau))$. Specifically, denote $\hat{s}_\tau = |\{j : \hat{\theta}_j(\tau) \neq 0\}|$ and $\hat{r}_\tau = \text{rank}(\hat{\Pi}(\tau))$,

$$\begin{aligned} \text{BIC}(\nu_1, \nu_2) \quad = \quad & \sum_{i=1}^{n} \sum_{t=1}^{T} \rho_\tau(Y_{i,t} - X'_{i,t}\hat{\theta}(\nu_1, \nu_2) - \hat{\Pi}(\nu_1, \nu_2)) + \\ & \frac{\log(nT)}{2} \left(c_1 \cdot \hat{s}_\tau(\nu_1, \nu_2) + (1 + n + T) \cdot \hat{r}_\tau(\nu_1, \nu_2) \right), \end{aligned} \tag{31}$$

where $c_1 > 0$ is a constant. The intuition here is that the first term in the right hand side of (31) 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 ν_1 and ν_2 is estimated by $\hat{s}_\tau$ for the vector of coefficients, and $(1+n+T) \cdot \hat{r}_\tau$ 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 $(1 + n + T) \cdot \hat{r}_\tau$. 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 $\text{BIC}(\nu_1, \nu_2)$, over the different choices of ν_1 and ν_2, and report the average over the 100 Monte Carlo simulations. We refer to this as BIC-ℓ_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 θ and the corresponding estimate. We also consider a different metric, the "Quantile error" (Koenker and Machado (1999)):

$$\frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} (F^{-1}_{Y_{i,t}|X_{i,t};\theta(\tau),\Pi(\tau)}(0.5) - \hat{F}^{-1}_{Y_{i,t}|X_{i,t};\theta(\tau),\Pi(\tau)}(0.5))^2, \tag{32}$$

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 $T \in \{100, 500\}$. 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

$$Y_{i,t} = X'_{i,t}\theta + \Pi_{i,t} + \epsilon_{i,t}, \tag{33}$$

where $\sqrt{3}\epsilon_{i,t} \overset{i.i.d.}{\sim} t(3)$, $i = 1, \ldots, n$ 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 (33), 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

$$Y_{i,t} = X'_{i,t}\theta + \Pi_{i,t} + (X'_{i,t}\theta)\epsilon_{i,t}, \tag{34}$$

where $\epsilon_{i,t} \overset{i.i.d.}{\sim} N(0,1)$, $i = 1, \ldots, n$ and $t = 1, \ldots, T$. The parameters in θ and Π in (34) are taken to be the same as in (33). 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 Π as

$$\Pi_{i,t} = \sum_{k=1}^{5} c_k u_k v_k^T, \tag{35}$$

where

$$c_k \sim U[0, 1/4], \quad u_k = \frac{\tilde{u}_k}{\|\tilde{u}_k\|}, \quad \tilde{u}_k \sim N(0, I_n), \quad v_k = \frac{\tilde{v}_k}{\|\tilde{v}_k\|}, \quad \tilde{v}_k \sim N(0, I_n), \quad k = 1, \ldots, 5. \tag{36}$$

Design 4. (Location-scale shift model with random factors) This is a combination of Designs 2 and 3. Specifically, we generate data as in (34) but with Π satisfying (35) and (36).

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 (31). In particular, under Designs 1 and 2, the data driven version of our estimator, BIC-ℓ_1-NN-QR, performs very closely to the ideally tuned one ℓ_1-NN-QR. In the more challenging settings of Designs 3 and 4, we noticed that BIC-ℓ_1-NN-QR performs reasonably well compared to ℓ_1-NN-QR.

Table 1: 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 ℓ_2 distance for estimating $\theta(\tau)$, and the Quantile error defined in (32). For each method we report the average, over 100 Monte Carlo simulations, of the two performance measures.

				Design 1		Design 2	
Method	n	p	T	$\frac{\|\hat{\theta}(\tau)-\theta(\tau)\|^2}{10^{-4}}$	Quantile error	$\frac{\|\hat{\theta}(\tau)-\theta(\tau)\|^2}{10^{-4}}$	Quantile error
ℓ_1-NN-QR	300	30	300	**0.86**	**0.03**	**0.69**	**0.03**
BIC-ℓ_1-NN-QR	300	30	300	**0.86**	**0.03**	0.89	**0.03**
ℓ_1-NN-LS	300	30	300	2.58	0.05	1.03	0.04
ℓ_1-QR	300	30	300	8.18	4.19	6.81	4.18
ℓ_1-NN-QR	300	30	100	**2.99**	**0.04**	**2.39**	**0.04**
BIC-ℓ_1-NN-QR	300	30	100	**2.99**	**0.04**	**2.39**	**0.04**
ℓ_1-NN-LS	300	30	100	8.06	0.12	3.11	0.08
ℓ_1-QR	300	30	100	41.0	4.19	26.0	4.19
ℓ_1-NN-QR	300	5	300	**0.22**	**0.003**	**0.39**	**0.03**
BIC-ℓ_1-NN-QR	300	5	300	**0.22**	**0.003**	0.48	**0.03**
ℓ_1-NN-LS	300	5	300	0.37	0.01	0.69	**0.03**
ℓ_1-QR	300	5	300	2.6	4.19	3.27	4.19
ℓ_1-NN-QR	300	5	100	**0.50**	**0.008**	**0.80**	**0.03**
BIC-ℓ_1-NN-QR	300	5	100	0.53	0.009	0.97	**0.03**
ℓ_1-NN-LS	300	5	100	1.12	0.02	1.46	0.04
ℓ_1-QR	300	5	100	7.87	4.19	8.56	4.19
ℓ_1-NN-QR	100	30	300	**2.97**	**0.04**	**2.26**	**0.04**
BIC-ℓ_1-NN-QR	100	30	300	**2.97**	**0.04**	2.81	**0.04**
ℓ_1-NN-LS	100	30	300	9.77	0.12	3.39	0.06
ℓ_1-QR	100	30	300	40.0	4.23	24.0	4.23
ℓ_1-NN-QR	100	30	100	**2.3**	**0.04**	**10.0**	**0.03**
BIC-ℓ_1-NN-QR	100	30	100	**2.3**	0.06	11.0	0.04
ℓ_1-NN-LS	100	30	100	8.4	0.79	13.0	0.16
ℓ_1-QR	100	30	100	229.0	4.23	177.0	4.23
ℓ_1-NN-QR	100	5	300	**0.64**	**0.008**	**0.89**	**0.03**
BIC-ℓ_1-NN-QR	100	5	300	0.65	**0.008**	1.32	**0.03**
ℓ_1-NN-LS	100	5	300	1.25	0.02	1.67	0.05
ℓ_1-QR	100	5	300	8.79	4.23	8.61	4.23
ℓ_1-NN-QR	100	5	100	**1.82**	**0.009**	**3.30**	**0.03**
BIC-ℓ_1-NN-QR	100	5	100	1.85	0.01	3.74	0.04
ℓ_1-NN-LS	100	5	100	4.06	0.03	4.45	0.14
ℓ_1-QR	100	5	100	32.0	4.23	27.0	4.23

Table 2: 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 ℓ_2 distance for estimating $\theta(\tau)$, and the Quantile error defined in (32). For each method we report the average, over 100 Monte Carlo simulations, of the two performance measures.

				Design 3		Design 4	
Method	n	p	T	$\frac{\|\hat{\theta}(\tau)-\theta(\tau)\|^2}{10^{-4}}$	Quantile error	$\frac{\|\hat{\theta}(\tau)-\theta(\tau)\|^2}{10^{-4}}$	Quantile error
ℓ_1-NN-QR	300	30	300	**2.17**	**0.17**	**1.58**	**0.11**
BIC-ℓ_1-NN-QR	300	30	300	2.17	0.29	2.81	0.26
ℓ_1-NN-LS	300	30	300	3.33	0.19	3.54	0.12
ℓ_1-QR	300	30	300	6.59	1.09	12.0	1.01
ℓ_1-NN-QR	300	30	100	**10.0**	0.18	**4.61**	**0.13**
BIC-ℓ_1-NN-QR	300	30	100	11.0	0.26	**4.61**	0.16
ℓ_1-NN-LS	300	30	100	11.0	0.25	9.21	0.17
ℓ_1-QR	300	30	100	27.0	11.10	47.2	1.10
ℓ_1-NN-QR	300	5	300	**1.13**	**0.17**	**0.27**	**0.03**
BIC-ℓ_1-NN-QR	300	5	300	1.56	0.33	0.74	0.13
ℓ_1-NN-LS	300	5	300	1.58	0.19	0.49	0.05
ℓ_1-QR	300	5	300	2.47	1.10	5.52	1.09
ℓ_1-NN-QR	300	5	100	**3.04**	**0.19**	**0.69**	0.05
BIC-ℓ_1-NN-QR	300	5	100	4.37	0.27	1.11	0.15
ℓ_1-NN-LS	300	5	100	4.43	0.27	1.12	**0.05**
ℓ_1-QR	300	5	100	7.65	1.11	13.4	1.10
ℓ_1-NN-QR	100	30	300	**11.0**	**0.18**	**7.06**	**0.15**
BIC-ℓ_1-NN-QR	100	30	300	12.0	0.27	7.29	0.24
ℓ_1-NN-LS	100	30	300	12.0	0.34	11.2	0.18
ℓ_1-QR	100	30	300	26.0	1.10	51.0	1.12
ℓ_1-NN-QR	100	30	100	**6.12**	**0.17**	**32.1**	0.10
BIC-ℓ_1-NN-QR	100	30	100	6.64	0.22	35.4	0.15
ℓ_1-NN-LS	100	30	100	8.63	1.08	82.0	0.84
ℓ_1-QR	100	30	100	16.5	1.12	267.6	1.13
ℓ_1-NN-QR	100	5	300	**2.99**	**0.18**	**0.86**	**0.05**
BIC-ℓ_1-NN-QR	100	5	300	4.19	0.26	1.43	0.14
ℓ_1-NN-LS	100	5	300	5.27	0.33	1.45	**0.05**
ℓ_1-QR	100	5	300	8.59	1.10	21.0	1.09
ℓ_1-NN-QR	100	5	100	**12.3**	**0.16**	**2.15**	**0.04**
BIC-ℓ_1-NN-QR	100	5	100	13.1	0.20	2.43	0.07
ℓ_1-NN-LS	100	5	100	15.0	1.09	3.61	0.07
ℓ_1-QR	100	5	100	24.7	1.10	45.7	1.09

6 Empirical Performance of the "Characteristics + Latent Factor" Model in Asset Pricing

Data Description

We use data from CRSP and Compustat to constrtuct 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 7 in the Appendix lists details of the characteristics used and the methods to construct the data. We follow the procedures of Green et al. (2017) 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 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 Daniel and Titman (1997, 1998). 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.

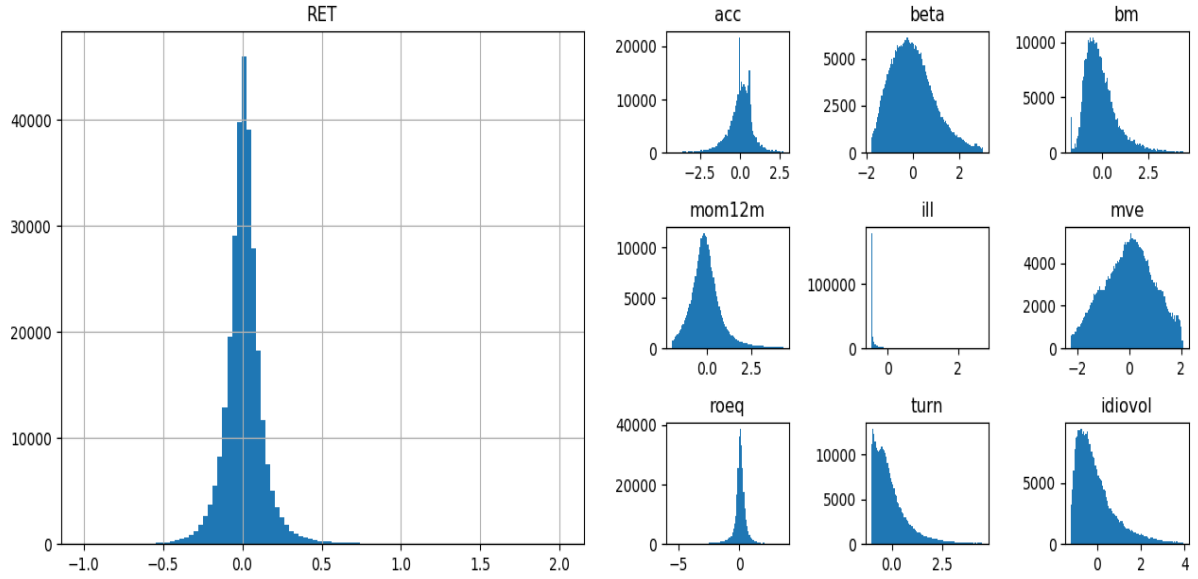


Figure 1: Histograms of monthly stock returns (left) and firm characteristics (right).

A "Characteristic + Latent Factor" Asset Pricing Model

We apply our model to fit the cross section and time series of stock returns ((Lettau and Pelger, 2018)). 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,...,R_n), and p observed firm characteristics (X_1,...,X_p) over T periods. The return quantile at level τ of portfolio i in time t is assumed to be the following:

$$F^{-1}_{R_{i,t}|X_{i,t-1};\theta(\tau),\lambda_i(\tau),g_t(\tau)}(\tau) = X_{i,t-1,1}\theta_1(\tau) + ... + X_{i,t-1,k}\theta_k(\tau) + ... + X_{i,t-1,p}\theta_p(\tau) + \underset{(1\times r_\tau)}{\lambda_i(\tau)} \underset{(r_\tau \times 1)}{g_t(\tau)}$$

where $X_{i,t-1,k}$ is the k-th characteristic (for example, the book-to-market ratio) of asset i in time $t-1$. The coefficient θ_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_τ latent factors in period t which captures systematic risks in the market, and λ_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 (Daniel and Titman (1997, 1998)). The formulation of our model accommodates both the factor view and the characteristics view. The sparse part is similar to Daniel and Titman (1997, 1998), in which stock returns 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(14). 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 ν_1 and ν_2, and we use our proposed BIC to select the optimal tuning parameters. The details of the information criteria can be found in equation (31).

The tuning parameter ν_1 governs the sparsity of the coefficient vector θ. The larger ν_1 is, the larger the shrinkage effect on θ. Figure 2 illustrate the effect of this shrinkage. With ν_2 fixed, as the value of ν_1 increases, more coefficients in the estimated θ vector shrink to zero. From a statistical point of view, the "effective characteristics" that can explain stock returns are those with non-zero coefficient θ at relatively large values of ν_1.

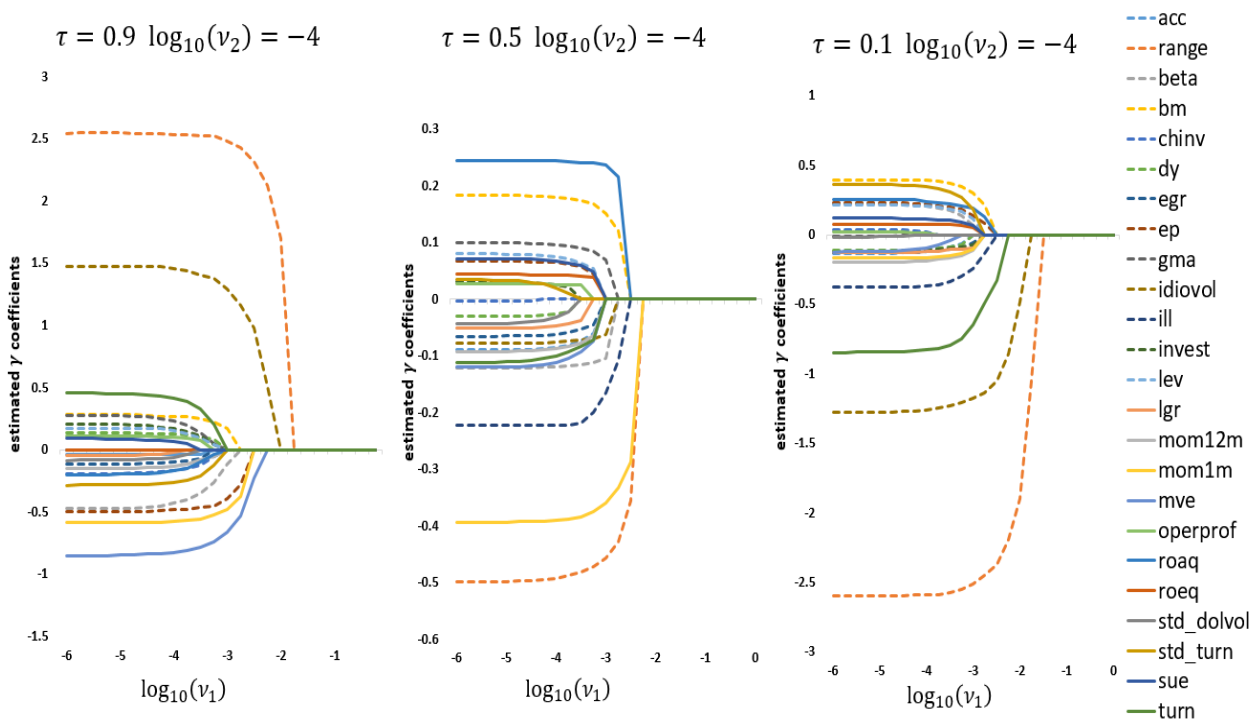


Figure 2: Estimated Coefficients as a Function of ν_1
The figure plots the estimated coefficient θ when the tuning parameter ν_1 changes, for $\tau = \{0.1, 0.5, 0.9\}$. The parameter ν_2 is fixed at $\log_{10}(\nu_2) = -4$.

Table 3 reports the relationship between tuning parameter ν_2 and rank of estimated Π at different quantiles. It shows that the tuning parameter ν_2 governs the rank of matrix Π, and that as ν_2 increases, we penalize more on the rank of matrix Π through its nuclear norm.

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 (mom1m, mom12m), accurals (acc), book equity growth (egr), leverage (lev), and standardized unexpected earnings (sue). However, some characteristics have heterogenous effects on future returns at different quantiles. For example, at the 10% quantile,

Table 3: The estimated rank of Π.

$\log_{10}(\nu_2)$	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.9$
-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 ν_2, when $\nu_1 = 10^{-5}$ is fixed. The results are reported for quantiles 10%, 50% and 90%.

high beta stocks have high future returns, which is consistent with results found via the CAPM; while at 50% and 90% quantile, high beta stocks have low future returns, which conforms the "low beta anomaly" phenomenon. Volatility (measured by both range and idiosyncratic volatility) is positively correlated with future returns at 90% quantile, but negatively correlated with future returns at 10% and 50% percentile. The result suggests that quantile models can capture a wider picture of the heterogenous relationship between asset returns and firm characteristics at different parts of the distribution (Koenker (2000)).

Table 5 reports the selected optimal tuning parameters ν_1 and ν_2 for different quantiles. The tuning parameters are selected via BIC based on (31) as discussed in Section 5. For every ν_1 and ν_2, we get the estimates $\widetilde{\theta}(\nu_1, \nu_2)$ and $\widetilde{\Pi}(\nu_1, \nu_2)$ and the number of factors $r = \text{rank}(\widetilde{\Pi}(\nu_1, \nu_2))$. The θ 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.

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.

Table 4: Sparse Part Coefficients at Different Quantiles.

	Fixed ν_1 and ν_2			Optimal ν_1 and ν_2 (BIC)		
	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.9$
acc	-0.089	-0.074	-0.041	0	0	0
range	-2.574	-0.481	2.526	-2.372	-0.356	2.429
beta	0.174	-0.116	-0.406	0	0	-0.115
bm	0.371	0.175	0.263	0	0	0.168
chinv	0	0	-0.152	0	0	0
dy	-0.086	0	0.119	0	0	0
egr	-0.106	-0.053	-0.091	0	0	0
ep	0.199	0.057	-0.479	0	0	-0.391
gma	0	0.091	0.201	0	0	0
idiovol	-1.229	-0.071	1.438	-1.055	0	1.286
ill	-0.334	-0.218	0	0	0	0
invest	-0.097	0	0.146	0	0	0
lev	0.183	0.063	0.129	0	0	0
lgr	-0.106	-0.037	0	0	0	0
mom12m	-0.166	-0.077	-0.117	0	0	0
mom1m	-0.150	-0.384	-0.571	0	-0.286	-0.477
mve	-0.038	-0.093	-0.811	0	0	-0.667
operprof	0	0.025	0.088	0	0	0
roaq	0.221	0.242	-0.147	0	0	0
roeq	0.073	0.041	0	0	0	0
std_dolvol	0	0	-0.039	0	0	0
std_turn	0.310	0	-0.247	0	0	0
sue	0.105	0.061	0.045	0	0	0
turn	-0.796	-0.083	0.386	-0.330	0	0

Note: The left panel reports the estimated coefficient vector θ 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 θ under the when the turning parameters are optimal, as selected by BIC (indicated in Table 5).

Table 5: Selected Optimal Tuning Parameters and Number of Factors

	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.9$
optimal r	1	7	2
optimal ν_1	$10^{-2.5}$	$10^{-2.5}$	$10^{-2.75}$
optimal ν_2	10^{-4}	10^{-4}	10^{-4}

Note: This table reports the selected optimal tuning parameter ν_1 and ν_2 that minimize the objective function in equation (31) for different 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%.

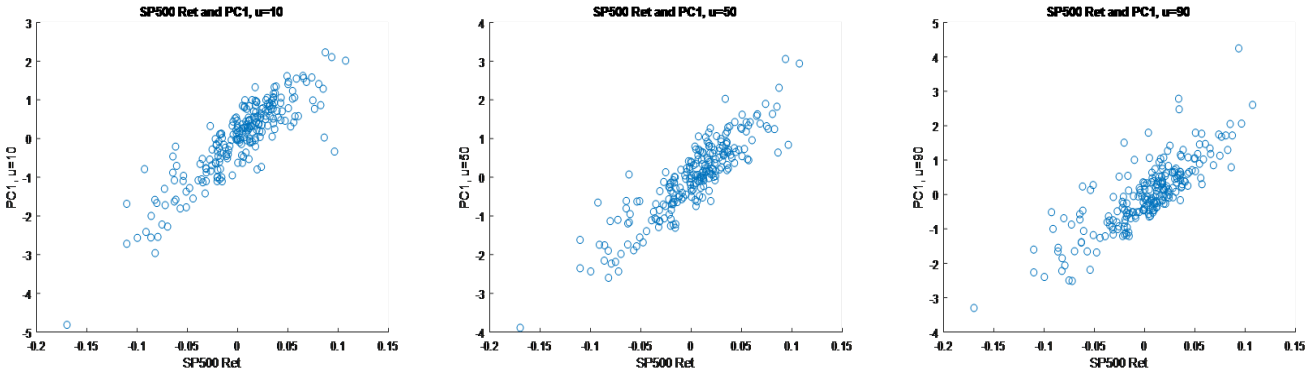
Table 6: Percentage of Π explained by PC

	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.9$
PC1	100.00%	73.82%	99.68%
PC2		13.71%	0.32%
PC3		6.78%	
PC4		4.12%	
PC5		1.11%	
PC6		0.45%	
PC7		0.01%	
Total	100.00%	100.00%	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.

Figure 3: The S&P 500 Index Return and the First PC at Different Quantiles.
This figure plots the first PC of matrix Π against S&P500 index monthly return for quantiles 10% (left), 50% (middle), and 90% (right).



References

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

Alnur Ali, Zico Kolter, and Ryan Tibshirani. The multiple quantile graphical model. In Advances in Neural Information Processing Systems, pages 3747–3755, 2016.

Tomohiro Ando and Jushan Bai. Quantile co-movement in financial markets: A panel quantile model with unobserved heterogeneity. Journal of the American Statistical Association, pages 1–31, 2019.

Manuel Arellano and Stéphane Bonhomme. Quantile selection models with an application to understanding changes in wage inequality. Econometrica, 85(1):1–28, 2017.

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. 2018.

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

Jushan Bai and Junlong Feng. Robust principal components analysis with non-sparse errors. arXiv preprint arXiv:1902.08735, 2019.

Jushan Bai and Kunpeng Li. Statistical analysis of factor models of high dimension. The Annals of Statistics, 40(1):436–465, 2012.

Jushan Bai and Serena Ng. Principal components estimation and identification of static factors. Journal of Econometrics, 176(1):18–29, 2013.

Jushan Bai and Serena Ng. Principal components and regularized estimation of factor models. arXiv preprint arXiv:1708.08137, 2017.

Jushan Bai and Serena Ng. Matrix completion, counterfactuals, and factor analysis of missing data. arXiv preprint arXiv:1910.06677, 2019.

Alexandre Belloni and Victor Chernozhukov. On the computational complexity of mcmc-based estimators in large samples. The Annals of Statistics, 37(4):2011–2055, 2009.

Alexandre Belloni and Victor Chernozhukov. ℓ_1-penalized quantile regression in high-dimensional sparse models. The Annals of Statistics, 39(1):82–130, 2011.

Peter Bickel, Ya'acov Ritov, and Alexandre Tsybakov. Simultaneous analysis of lasso and dantzig selector. The Annals of Statistics, 37(4):1705–1732, 2009.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine learning, 3(1):1–122, 2011.

Pratik Prabhanjan Brahma, Yiyuan She, Shijie Li, Jiade Li, and Dapeng Wu. Reinforced robust principal component pursuit. IEEE transactions on neural networks and learning systems, 29(5):1525–1538, 2017.

Jian-Feng Cai, Emmanuel Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization, 20(4):1956–1982, 2010.

John Campbell. Financial decisions and markets: a course in asset pricing. Princeton University Press, 2017.

Emmanuel Candès and Yaniv Plan. Matrix completion with noise. Proceedings of the IEEE, 98(6):925–936, 2010.

Emmanuel Candès and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. IEEE Transactions on Information Theory, 57(4):2342–2359, 2011.

Emmanuel Candès and Benjamin Recht. Exact matrix completion via convex optimization. Foundations of Computational mathematics, 9(6):717, 2009.

Emmanuel Candès and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. The annals of Statistics, 35(6):2313–2351, 2007.

Gary Chamberlain and Michael Rothschild. Arbitrage, factor structure, and mean-variance analysis on large asset markets. Econometrica (pre-1986), 51(5):1281, 1983.

Sourav Chatterjee. Matrix estimation by universal singular value thresholding. The Annals of Statistics, 43 (1):177–214, 2015.

Liang Chen, Juan Dolado, and Jesús Gonzalo. Quantile factor models. 2018.

Mingli Chen. Estimation of nonlinear panel models with multiple unobserved effects. Warwick Economics Research Paper Series No. 1120, 2014.

Mingli Chen, Iván Fernández-Val, and Martin Weidner. Nonlinear panel models with interactive effects. arXiv preprint arXiv:1412.5647, 2014.

Victor Chernozhukov, Christian Hansen, and Yuan Liao. A lava attack on the recovery of sums of dense and sparse signals. The Annals of Statistics, 45(1):39–76, 2017.

Victor Chernozhukov, Christian Hansen, Yuan Liao, and Yinchu Zhu. Inference for heterogeneous effects using low-rank estimations. arXiv preprint arXiv:1812.08089, 2018.

John H Cochrane. Asset pricing: Revised edition. Princeton university press, 2009.

John H Cochrane. Presidential address: Discount rates. The Journal of finance, 66(4):1047–1108, 2011.

Gregory Connor and Robert A Korajczyk. Risk and return in an equilibrium apt: Application of a new test methodology. Journal of financial economics, 21(2):255–289, 1988.

Arnak Dalalyan, Mohamed Hebiri, and Johannes Lederer. On the prediction performance of the lasso. Bernoulli, 23(1):552–581, 2017.

Kent Daniel and Sheridan Titman. Evidence on the characteristics of cross sectional variation in stock returns. the Journal of Finance, 52(1):1–33, 1997.

Kent Daniel and Sheridan Titman. Characteristics or covariances. Journal of Portfolio Management, 24(4): 24–33, 1998.

Andreas Elsener and Sara van de Geer. Robust low-rank matrix estimation. The Annals of Statistics, 46 (6B):3481–3509, 2018.

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

Maryam Fazel. Matrix rank minimization with applications. 2002.

Guanhao Feng, Stefano Giglio, and Dacheng Xiu. Taming the factor zoo: A test of new factors. Technical report, National Bureau of Economic Research, 2019.

Junlong Feng. Regularized quantile regression with interactive fixed effects. arXiv preprint arXiv:1911.00166, 2019.

Antonio Galvao. Quantile regression for dynamic panel data with fixed effects. Journal of Econometrics, 164(1):142–157, 2011.

Antonio Galvao and Kengo Kato. Smoothed quantile regression for panel data. Journal of econometrics, 193(1):92–112, 2016.

Antonio Galvao and Gabriel V Montes-Rojas. Penalized quantile regression for dynamic panel data. Journal of Statistical Planning and Inference, 140(11):3476–3497, 2010.

Domenico Giannone, Michele Lenza, and Giorgio Primiceri. Economic predictions with big data: The illusion of sparsity. 2017.

Stefano Giglio and Dacheng Xiu. Asset pricing with omitted factors. Chicago Booth Research Paper, (16-21), 2018.

Bryan S Graham, Jinyong Hahn, Alexandre Poirier, and James L Powell. A quantile correlated random coefficients panel data model. Journal of Econometrics, 206(2):305–335, 2018.

Jeremiah Green, John Hand, and Frank Zhang. The characteristics that provide independent information about average us monthly stock returns. The Review of Financial Studies, 30(12):4389–4436, 2017.

Yufeng Han, Ai He, David Rapach, and Guofu Zhou. What firm characteristics drive us stock returns? Available at SSRN 3185335, 2018.

Xuming He, Lan Wang, and Hyokyoung Grace Hong. Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. The Annals of Statistics, 41(1):342–369, 2013.

Kengo Kato, Antonio F Galvao Jr, and Gabriel V Montes-Rojas. Asymptotics for panel quantile regression models with individual effects. Journal of Econometrics, 170(1):76–91, 2012.

Roger Koenker. Galton, edgeworth, frisch, and prospects for quantile regression in econometrics. Journal of Econometrics, 95(2):347–374, 2000.

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

Roger Koenker. Quantile regression. Cambridge University Press, New York, 2005.

Roger Koenker and Jose AF Machado. Goodness of fit and related inference processes for quantile regression. Journal of the american statistical association, 94(448):1296–1310, 1999.

Roger Koenker, Victor Chernozhukov, Xuming He, and Limin Peng. Handbook of quantile regression. CRC press, 2017.

Vladimir Koltchinskii, Karim Lounici, and Alexandre Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. The Annals of Statistics, 39(5):2302–2329, 2011.

Serhiy Kozak, Stefan Nagel, and Shrihari Santosh. Shrinking the cross-section. Journal of Financial Economics, 2019.

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

Martin Lettau and Markus Pelger. Estimating latent asset-pricing factors. Technical report, National Bureau of Economic Research, 2018.

Shujie Ma, Oliver Linton, and Jiti Gao. Estimation and inference in semiparametric quantile factor models. 2019.

Hyungsik Roger Moon and Martin Weidner. Nuclear norm regularized estimation of panel regression models. arXiv preprint arXiv:1810.10987, 2018.

Sahand Negahban and Martin Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. The Annals of Statistics, 39(2):1069–1097, 2011.

Benjamin Recht, Maryam Fazel, and Pablo Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM review, 52(3):471–501, 2010.

Angelika Rohde and Alexandre Tsybakov. Estimation of high-dimensional low-rank matrices. The Annals of Statistics, 39(2):887–930, 2011.

Stephen Ross. The arbitrage theory of capital asset pricing. Journal of Economic Theory, 13(3):341–360, 1976.

Andrés G Sagner. Three essays on quantile factor analysis. PhD thesis, Boston University, 2019.

Yiyuan She and Kun Chen. Robust reduced-rank regression. Biometrika, 104(3):633–647, 2017.

Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288, 1996.

Aad W. van der Vaart and Jon A Wellner. Weak convergence and empirical processes: with applications to statistics. Springer, 1996.

Van Vu. Spectral norm of random matrices. Combinatorica, 27(6):721–736, 2007.

Lan Wang, Yichao Wu, and Runze Li. Quantile regression for analyzing heterogeneity in ultra-high dimension. Journal of the American Statistical Association, 107(497):214–222, 2012.

Yuanshan Wu and Guosheng Yin. Conditional quantile screening in ultrahigh-dimensional heterogeneous data. Biometrika, 102(1):65–76, 2015.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. The Annals of Probability, pages 94–116, 1994.

Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. Biometrika, 102(2):315–323, 2014.

Qi Zheng, Limin Peng, and Xuming He. Globally adaptive quantile regression with ultra-high dimensional data. Annals of statistics, 43(5):2225, 2015.

A Implementation Details of the Proposed ADMM Algorithm

Denoting by $P_+(\cdot)$ and $P_-(\cdot)$ the element-wise positive and negative part operators, the ADMM proceeds doing the iterative updates

$$V^{(k+1)} \leftarrow P_+\left(W^{(k)} - U_V^{(k)} - \frac{\tau}{nT\eta}\mathbf{1}\mathbf{1}'\right) + P_-\left(W^{(k)} - U_V^{(k)} - \frac{\tau}{nT\eta}\mathbf{1}\mathbf{1}'\right) \tag{37}$$

$$\tilde{\theta}^{(k+1)} \leftarrow \arg\min_{\theta}\left\{\frac{\eta}{2}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(W_{i,t}^{(k)} - Y_{i,t} + X_{i,t}'\theta + (Z_\Pi^{(k)})_{i,t} + (U_W^{(k)})_{i,t}\right)^2 + \frac{\eta}{2}\|Z_\theta^{(k)} - \theta + U_\theta^{(k)}\|^2\right\} \tag{38}$$

$$\tilde{\Pi}^{(k+1)} \leftarrow \arg\min_{\tilde{\Pi}}\left\{\frac{1}{2}\|Z_\Pi^{(k)} - \tilde{\Pi} + U_\Pi^{(k)}\|_F^2 + \frac{\nu_2}{\eta}\|\tilde{\Pi}\|_*\right\} \tag{39}$$

$$\tag{40}$$

$$Z_\theta^{(k+1)} \leftarrow \arg\min_{Z_\theta}\left\{\frac{1}{2}\|\tilde{\theta}^{(k+1)} - U_\theta^{(k)} - Z_\theta\|^2 + \frac{\nu_1}{\eta}\sum_{j=1}^{p} w_j|(Z_\theta)_j|\right\} \tag{41}$$

$$(Z_\Pi^{(k+1)}, W^{(k+1)}) \leftarrow \arg\min_{Z_\Pi, W}\left\{\frac{\eta}{2}\|W - Y + X\tilde{\theta}^{(k+1)} + Z_\Pi + U_W^{(k)}\|_F^2 + \frac{\eta}{2}\|V^{(k+1)} - W + U_V^{(k)}\|_F^2\right. \tag{42}$$

$$\left. + \frac{\eta}{2}\|Z_\Pi - \tilde{\Pi}^{(k+1)} + U_\Pi^{(k)}\|_F^2\right\} \tag{43}$$

$$U_V^{(k+1)} \leftarrow V^{(k+1)} - W^{(k+1)} + U_V^{(k)}, \quad U_W^{(k+1)} \leftarrow W^{(k+1)} - Y + X\tilde{\theta}^{(k+1)} + Z_\Pi^{(k+1)} + U_W^{(k)},$$
$$U_\Pi^{(k+1)} \leftarrow Z_\Pi^{(k+1)} - \tilde{\Pi}^{(k+1)} + U_\Pi^{(k)}, \quad U_\theta^{(k+1)} \leftarrow Z_\theta^{(k+1)} - \tilde{\theta}^{(k+1)} + U_\theta^{(k)},$$

where $\eta > 0$ is the penalty, see Boyd et al. (2011).

The update for $\tilde{\theta}$ is

$$\tilde{\theta}^{(k+1)} \leftarrow \left[\sum_{i=1}^{n}\sum_{t=1}^{T} X_{i,t}X_{i,t}' + I_p\right]^{-1}\left[-\sum_{i=1}^{n}\sum_{t=1}^{T} X_{i,t}A_{i,t} + Z_\theta^{(k)} + U_\theta^{(k)}\right],$$

where

$$A := W^{(k)} + Z_\Pi^{(k)} + U_W^{(k)} - Y.$$

The update for $\tilde{\Pi}$ is

$$\tilde{\Pi}^{(k+1)} \leftarrow P\,\mathrm{diag}\left(\max\left\{0, v_j - \frac{\nu_2}{\eta}\right\}_{1\le j\le l}\right)Q',$$

where

$$Z_{\Pi}^{(k)} + U_{\Pi}^{(k)} \;=\; P\,\mathrm{diag}(\{v_j\}_{1\le j\le l})Q'.$$

Furthermore, for Z_θ,

$$Z_{\theta,j}^{(k+1)} \;\leftarrow\; \mathrm{sign}(\tilde{\theta}_j^{(k+1)} - U_{\theta,j}^{(k)})\left[|\tilde{\theta}_j^{(k+1)} - U_{\theta,j}^{(k)}| - \frac{\nu_1 w_j}{\eta}\right].$$

Finally, defining

$$\tilde{A} \;=\; -Y + X\tilde{\theta}^{(k+1)} + U_W^{(k)},\;\; \tilde{B} \;=\; -V^{(k+1)} - U_V^{(k)},\;\; \tilde{C} \;=\; -\tilde{\Pi}^{(k+1)} + U_{\Pi}^{(k)},$$

the remaining updates are

$$Z_{\Pi}^{(k+1)} \;\leftarrow\; \frac{-\tilde{A} - 2\tilde{C} + \tilde{B}}{3},$$

and

$$W^{(k+1)} \;\leftarrow\; -\tilde{A} - \tilde{C} - 2Z_{\Pi}^{(k+1)}.$$

B Proofs of the Main Results in the Paper

B.1 Auxiliary lemmas for proof of Theorem 1

Throughout, we use the notation

$$Q_\tau(\tilde{\theta}, \tilde{\Pi}) \;=\; \mathbb{E}(\hat{Q}_\tau(\tilde{\theta}, \tilde{\Pi})).$$

Moreover, as in Yu (1994), we define the sequence $\{(\tilde{Y}_{i,t}, \tilde{X}_{i,t})\}_{i\in[n],t\in[T]}$ such that

- $\{(\widetilde{Y}_{i,t}, \widetilde{X}_{i,t})\}_{i\in[n],t\in[T]}$ is independent of $\{(Y_{i,t}, X_{i,t})\}_{i\in[n],t\in[T]}$;
- for a fixed t the random vectors $\{(\widetilde{Y}_{i,t}, \widetilde{X}_{i,t})\}_{i\in[n]}$ are independent;
- for a fixed i:

$$\mathcal{L}(\{(\widetilde{Y}_{i,t}, \widetilde{X}_{i,t})\}_{t\in H_l}) \;=\; \mathcal{L}(\{(Y_{i,t}, X_{i,t})\}_{t\in H_l}) \;=\; \mathcal{L}(\{(Y_{i,t}, X_{i,t})\}_{t\in H_1}) \;\;\forall l \;\in [d_T],$$

  and the blocks $\{(\widetilde{Y}_{i,t}, \widetilde{X}_{i,t})\}_{t\in H_1}, \ldots, \{(\widetilde{Y}_{i,t}, \widetilde{X}_{i,t})\}_{t\in H_{d_T}}$ are independent.

Here, we define $\Lambda \;:=\; \{H_1, H_1', \ldots, H_{d_T}, H_{d_T}', R\}$ with

$$\begin{aligned} H_j \;&=\; \{t \,:\, 1 + 2(j-1)c_T \le t \le (2j-1)c_T\}\,, \\ H_j' \;&=\; \{t \,:\, 1 + (2j-1)c_T \le t \le 2jc_T\}\,,\; j = 1, \ldots, d_T, \\ &\quad\text{and}\;\; R \;=\; \{t \,:\, 2c_T d_T + 1 \;\le t \le T\}. \end{aligned} \tag{44}$$

We also use the symbol $\mathcal{L}(\cdot)$ to denote the distribution of a sequence of random variables.

Next, define the scores $a_{i,t} = \tau - 1\{Y_{i,t} \leq X'_{i,t}\theta(\tau) + \Pi_{i,t}(\tau)\}$, and $\tilde{a}_{i,t} = \tau - 1\{\widetilde{Y}_{i,t} \leq \widetilde{X}'_{i,t}\theta(\tau) + \Pi_{i,t}(\tau)\}$.

Lemma 4. *Under Assumptions 1–3, we have*

$$\mathbb{P}\left(\max_{j=1,\ldots,p} \frac{1}{nT}\left|\sum_{i=1}^{n}\sum_{t=1}^{T}\frac{X_{i,t,j}a_{i,t}}{\hat{\sigma}_j}\right| \geq 9\sqrt{\frac{c_T\log(\max\{n,pc_T\})}{nd_T}}\right) \leq \frac{16}{n} + 8npT\left(\frac{1}{c_T}\right)^{\mu}.$$

Proof. Notice that

$$\begin{aligned}
&\mathbb{P}\left(\max_{j=1,\ldots,p} \frac{1}{nT}\left|\sum_{i=1}^{n}\sum_{t=1}^{T}\frac{X_{i,t,j}a_{i,t}}{\hat{\sigma}_j}\right| \geq \eta|X\right) \\
&\leq 2p\max_{j=1,\ldots,p}\mathbb{P}\left(\frac{1}{nd_T}\left|\sum_{i=1}^{n}\sum_{l=1}^{d_T}\left(\frac{1}{c_T}\sum_{t=2(l-1)+1}^{2(l-1)+c_T}\frac{X_{i,t,j}a_{it}}{\hat{\sigma}_j}\right)\right| \geq \frac{\eta}{9}|X\right) + \\
&p\max_{j=1,\ldots,p}\mathbb{P}\left(\frac{1}{nd_T}\left|\sum_{i=1}^{n}\frac{1}{c_T}\left(\sum_{t\in R}\frac{X_{i,t,j}}{\hat{\sigma}_j}\right)\right| \geq \frac{\eta}{9}|X\right) \\
&\leq 4p\max_{j=1,\ldots,p}\mathbb{P}\left(\max_{m=1,\ldots,c_T}\frac{1}{nd_T}\left|\sum_{i=1}^{n}\sum_{l=0}^{d_T-1}\frac{X_{i,(2lc_T+m),j}\tilde{a}_{i,t}}{\hat{\sigma}_j}\right| \geq \frac{\eta}{9}|X\right) + \\
&2p\max_{j=1,\ldots,p}\mathbb{P}\left(\max_{m=1,\ldots,|R|}\frac{1}{nd_T}\left|\sum_{i=1}^{n}\frac{X_{i,(2d_Tc_T+m)j}\tilde{a}_{i,t}}{\hat{\sigma}_j}\right| \geq \frac{\eta}{9}|X\right) + 8npT\left(\frac{1}{c_T}\right)^{\mu} \\
&\leq 4pc_T\max_{j\in[p],m\in[c_T]}\mathbb{P}\left(\frac{1}{nd_T}\left|\sum_{i=1}^{n}\sum_{l=0}^{d_T-1}\frac{X_{i,(2lc_T+m),j}\tilde{a}_{i,t}}{\hat{\sigma}_j}\right| \geq \frac{\eta}{9}|X\right) + \\
&2pc_T\max_{j\in[p],m\in[|R|]}\mathbb{P}\left(\frac{1}{nd_T}\left|\sum_{i=1}^{n}\frac{X_{i,(2d_Tc_T+m),j}\tilde{a}_{i,t}}{\hat{\sigma}_j}\right| \geq \frac{\eta}{9}|X\right) + 8npT\left(\frac{1}{c_T}\right)^{\mu}
\end{aligned} \tag{45}$$

where the first inequality follows from union bound, and the second by Lemmas 4.1 and 4.2 from Yu (1994). Therefore, since

$$\frac{1}{nd_T}\sum_{i=1}^{n}\sum_{l=0}^{d_T-1}X^2_{i,(2lc_T+m),j} \leq 3c_T\hat{\sigma}_j^2,$$

and with a similar argument for the second term in the last inequality of (45), we obtain the result by Hoeffding's inequality and integrating over X. □

Lemma 5. *Supposes that Assumptions 1–3 hold, and let*

$$\mathcal{G} = \left\{\Delta \in \mathbb{R}^{n\times T} : \|\Delta\|_* \leq 1\right\}. \tag{46}$$

Then there exists positive constants c_1 and c_2 such that

$$\sup_{\Delta\in\mathcal{G}} \frac{1}{nT}\left|\sum_{i=1}^{n}\sum_{t=1}^{T}\Delta_{i,t}a_{i,t}\right| \;\leq\; \frac{100c_T}{nT}\left(\sqrt{n}+\sqrt{d_T}\right),$$

with probability at least

$$1 \,-\, 2nT\left(\frac{1}{c_T}\right)^{\mu} \,-\, 2c_1\exp(-c_2\max\{n,d_T\}+\log c_T),$$

for some positive constants c_1 and c_2.

Proof. Notice that by Lemma 4.3 from Yu (1994),

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\Delta\in\mathcal{G}}\left[\frac{1}{nT}\left|\sum_{i=1}^{n}\sum_{l=1}^{d_T}\sum_{t\in H_l}\Delta_{i,t}a_{i,t}\right| + \frac{1}{nT}\left|\sum_{i=1}^{n}\sum_{l=1}^{d_T}\sum_{t\in H_l'}\Delta_{i,t}a_{i,t}\right| + \frac{1}{nT}\left|\sum_{i=1}^{n}\sum_{t\in R}\Delta_{i,t}a_{i,t}\right|\right] \geq \eta\right)\\
&\leq \mathbb{P}\left(\sup_{\Delta\in\mathcal{G}}\frac{1}{nT}\left|\sum_{i=1}^{n}\sum_{l=1}^{d_T}\sum_{t\in H_l}\Delta_{i,t}\tilde{a}_{i,t}\right| \geq \frac{\eta}{3}\right) + \mathbb{P}\left(\sup_{\Delta\in\mathcal{G}}\frac{1}{nT}\left|\sum_{i=1}^{n}\sum_{l=1}^{d_T}\sum_{t\in H_l'}\Delta_{i,t'}\tilde{a}_{i,t}\right| \geq \frac{\eta}{3}\right) +\\
&\mathbb{P}\left(\sup_{\Delta\in\mathcal{G}}\frac{1}{nT}\left|\sum_{i=1}^{n}\sum_{t\in R}\Delta_{i,t}\tilde{a}_{i,t}\right| \geq \frac{\eta}{3}\right) + 2nT\left(\frac{1}{c_T}\right)^{\mu}\\
&\leq 2c_T\max_{m\in[c_T]}\mathbb{P}\left(\sup_{\Delta\in\mathcal{G}}\frac{1}{nd_T}\left|\sum_{i=1}^{n}\sum_{l=0}^{d_T-1}\Delta_{i,\,(2c_Tl+m)}\tilde{a}_{i,\,(2c_Tl+m)}\right| \geq \frac{\eta}{9}\right) +\\
&2c_T\max_{m\in[|R|]}\mathbb{P}\left(\sup_{\Delta\in\mathcal{G}}\frac{1}{nd_T}\left|\sum_{i=1}^{n}\Delta_{i,\,(2c_Td_T+m)}\tilde{a}_{i,\,(2c_Td_T+m)}\right| \geq \frac{\eta}{9}\right) + 2nT\left(\frac{1}{c_T}\right)^{\mu}
\end{aligned} \tag{47}$$

We now proceed to bound each of the terms in the upper bound of (47). For the first term, notice that for a fixed m

$$\begin{aligned}
\sup_{\Delta\in\mathcal{G}}\frac{1}{nd_T}\left|\sum_{i=1}^{n}\sum_{l=0}^{d_T-1}\Delta_{i,\,(2c_Tl+m)}\tilde{a}_{i,\,(2c_Tl+m)}\right| \;&\leq\; \sup_{\Delta\in\mathcal{G}}\frac{1}{nd_T}\left\|\left\{\tilde{a}_{i,\,(2c_Tl+m)}\right\}_{i\in[n],l\in[d_T]}\right\|_2\|\Delta\|_*\\
&\leq\; \frac{3}{nd_T}\left(\sqrt{n}+\sqrt{d_T}\right),
\end{aligned} \tag{48}$$

where the first inequality holds by the duality between the nuclear norm and spectral norm, and the second inequality happens with probability at least $1-c_1\exp\left(-c_2\max\{n,d_T\}\right)$ by Theorem 3.4 from Chatterjee (2015).

On the other hand,

$$\begin{aligned}
\sup_{\Delta\in\mathcal{G}}\frac{1}{nd_T}\left|\sum_{i=1}^{n}\Delta_{i,\,(2c_Td_T+m)}\tilde{a}_{i,\,(2c_Td_T+m)}\right| \;&\leq\; \sup_{\Delta\in\mathcal{G}}\frac{\sqrt{n}\|\{\Delta_{i,\,(2c_Td_T+m)}\}_{i\in[n]}\|}{nd_T}\\
&\leq\; \frac{\sqrt{n}\|\Delta\|_*}{nd_T},
\end{aligned} \tag{49}$$

with probability at least $1 - c_1 \exp\left(-c_2 \max\{n, d_T\}\right)$, also by Theorem 3.4 from Chatterjee (2015).

The claim follows by combining (47), (48), and (49), taking $\eta = 30(\sqrt{n} + \sqrt{d_T})/\sqrt{nd_T}$, and the fact that $c_T/T \leq 1/3$.

□

Lemma 6. *For every $\tilde{\Pi}, \check{\Pi} \in \mathbb{R}^{n \times T}$, we have that*

$$\|\check{\Pi} - \tilde{\Pi}\|_* + \|\check{\Pi}\|_* - \|\tilde{\Pi}\|_* \leq 6\sqrt{\operatorname{rank}(\check{\Pi})}\|\check{\Pi} - \tilde{\Pi}\|_F$$

Proof. This follows directly from Lemma 2.3 in Elsener and van de Geer (2018). □

Lemma 7. *Assume that 1–3 hold. Then, with probability approaching one,*

$$\frac{3}{4}\|\theta\|_1 \leq \|\theta\|_{1,n,T} \leq \frac{5}{4}\|\theta\|_1, \tag{50}$$

for all $\theta \in \mathbb{R}^p$.

Moreover, for $c_0 \in (0, 1)$ letting

$$\nu_1 = \frac{9}{1 - c_0}\sqrt{\frac{c_T \log(\max\{n, pc_T\})}{nd_T}}(\sqrt{n} + \sqrt{d_T}),$$

and

$$\nu_2 = \frac{200c_T}{nT}\left(\sqrt{n} + \sqrt{d_T}\right),$$

we have that

$$(\hat{\theta}(\tau) - \theta(\tau), \hat{\Pi}(\tau) - \Pi(\tau)) \in A_\tau,$$

with probability approaching one, where

$$A_\tau = \left\{(\delta, \Delta) \,:\, \|\delta_{T_\tau^c}\|_1 + \frac{\|\Delta\|_*}{\sqrt{nT}\sqrt{\log(\max\{n, pc_T\})}} \leq C_0\left(\|\delta_{T_\tau}\|_1 + \frac{\sqrt{r_\tau}\|\Delta\|_F}{\sqrt{nT}\sqrt{\log(\max\{n, pc_T\})}}\right)\right\},$$

and C_0 is a positive constant that depends on τ and c_0.

Proof. By Lemma 4, Lemma 5, Lemma 6, and Assumption 2, we have that

$$
\begin{aligned}
0 \;\leq\;& \hat{Q}(\theta(\tau),\Pi(\tau)) \,-\, \hat{Q}(\hat{\theta}(\tau),\hat{\Pi}(\tau)) \,+\, \nu_1\left(\|\theta(\tau)\|_{1,n} - \|\hat{\theta}(\tau)\|_{1,n,T}\right) \,+\, \nu_2(\|\Pi(\tau)\|_* - \|\hat{\Pi}(\tau)\|_*) \\
\leq\;& \max_{1\leq j\leq p}\left|\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\frac{X_{i,t,j}a_{i,t}}{\hat{\sigma}_j}\right|\left[\sum_{k=1}^{p}\hat{\sigma}_k|\theta_k(\tau)-\hat{\theta}_k(\tau)|\right] \,+\, \nu_1\left(\|\theta(\tau)\|_{1,n,T} - \|\hat{\theta}(\tau)\|_{1,n,T}\right) + \\
& \left|\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}a_{i,t}(\Pi_{i,t}(\tau)-\hat{\Pi}_{i,t}(\tau))\right| \,+\, \nu_2(\|\Pi(\tau)\|_* - \|\hat{\Pi}(\tau)\|_*) \\
\leq\;& 9\sqrt{\frac{c_T\log(\max\{n,pc_T\})}{nd_T}}\left[\sum_{k=1}^{p}\hat{\sigma}_k|\theta_k(\tau)-\hat{\theta}_k(\tau)|\right] \,+\, \nu_1\left(\|\theta(\tau)\|_{1,n,T} - \|\hat{\theta}(\tau)\|_{1,n,T}\right) \\
& + \|\Pi(\tau)-\hat{\Pi}(\tau)\|_*\left(\sup_{\|\tilde{\Delta}\|_*\leq 1}\left|\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}a_{i,t}\tilde{\Delta}_{i,t}\right|\right) + \nu_2(\|\Pi(\tau)\|_* - \|\hat{\Pi}(\tau)\|_*) \\
\leq\;& 9\sqrt{\frac{c_T\log(\max\{n,pc_T\})}{nd_T}}\left[\sum_{k=1}^{p}\hat{\sigma}_k|\theta_k(\tau)-\hat{\theta}_k(\tau)|\right] \,+\, \nu_1\left(\|\theta(\tau)\|_{1,n,T} - \|\hat{\theta}(\tau)\|_{1,n,T}\right) \\
& + \frac{200c_T}{nT}\left(\sqrt{n}+\sqrt{d_T}\right)\|\Pi(\tau)-\hat{\Pi}(\tau)\|_* + \frac{200c_T}{nT}\left(\sqrt{n}+\sqrt{d_T}\right)\left(\|\Pi(\tau)\|_* - \|\hat{\Pi}(\tau)\|_*\right), \\
& - \frac{100c_T}{nT}\left(\sqrt{n}+\sqrt{d_T}\right)\|\Pi(\tau)-\hat{\Pi}(\tau)\|_* \\
\leq\;& 9\sqrt{\frac{c_T\log(\max\{n,pc_T\})}{nd_T}}\left[\sum_{k=1}^{p}\hat{\sigma}_k|\theta_k(\tau)-\hat{\theta}_k(\tau)|\right] \,+\, \nu_1\left(\|\theta(\tau)\|_{1,n,T} - \|\hat{\theta}(\tau)\|_{1,n,T}\right) \\
& + \frac{1200c_T\sqrt{r_\tau}}{nT}\left(\sqrt{n}+\sqrt{d_T}\right)\|\Pi(\tau)-\hat{\Pi}(\tau)\|_F \\
& - \frac{100c_T}{nT}\left(\sqrt{n}+\sqrt{d_T}\right)\|\Pi(\tau)-\hat{\Pi}(\tau)\|_*
\end{aligned}
$$

with probability at least

$$
1-\gamma-\frac{16}{n}-8npT\left(\frac{1}{c_T}\right)^{\mu}-2nT\left(\frac{1}{c_T}\right)^{\mu}-2c_1\exp(-c_2\max\{n,d_T\}+\log c_T).
$$

Therefore, with probability approaching one, for positive constants C_1 and C_2, we have

$$
\begin{aligned}
0 \;\leq\;& \left[\sum_{j=1}^{p}\left((1-c_0)\hat{\sigma}_j|\hat{\theta}_j(\tau)-\theta_j(\tau)| + \hat{\sigma}_j|\theta_j(\tau)| - \hat{\sigma}_j|\hat{\theta}_j(\tau)|\right)\right] \\
& + \left[\frac{3C_1\sqrt{r_\tau}\|\Pi(\tau)-\hat{\Pi}(\tau)\|_F}{\sqrt{nT}\sqrt{\log(\max\{n,pc_T\})}} - \frac{C_2\|\Pi(\tau)-\hat{\Pi}(\tau)\|_*}{\sqrt{nT}\sqrt{\log(\max\{n,pc_T\})}}\right],
\end{aligned}
$$

and the claim follows.

□

Lemma 8. *Under Assumption 3, for all $(\delta, \Delta) \in A_\tau$, we have that*

$$Q_\tau(\theta(\tau)+\delta, \Pi(\tau)+\Delta) - Q_\tau(\theta(\tau), \Pi(\tau)) \;\geq\; \min\left\{ \frac{\left(J_\tau^{1/2}(\delta,\Delta)\right)^2}{4}, qJ_\tau^{1/2}(\delta,\Delta) \right\}.$$

Proof. Let

$$v_{A_\tau} \;=\; \sup_v \Bigg\{ v \;:\; Q_\tau(\theta(\tau)+\tilde{\delta}, \Pi(\tau)+\tilde{\Delta}) - Q_\tau(\theta(\tau), \Pi(\tau)) \geq \frac{\left(J_\tau^{1/2}(\tilde{\delta},\tilde{\Delta})\right)^2}{4},\; \forall (\tilde{\delta},\tilde{\Delta}) \in A_\tau, \\ J_\tau^{1/2}(\tilde{\delta},\tilde{\Delta}) \leq v \Bigg\}.$$

Then by the convexity of $Q_\tau(\cdot)$ and the definition of v_{A_u}, we have that

$$\begin{aligned}
&Q_\tau(\theta(\tau)+\tilde{\delta}, \Pi(\tau)+\tilde{\Delta}) - Q_\tau(\theta(\tau), \Pi(\tau)) \\
&\geq \frac{\left(J_\tau^{1/2}(\delta,\Delta)\right)^2}{4} \wedge \left\{ \frac{J_\tau^{1/2}(\delta,\Delta)}{v_{A_\tau}} \cdot \inf_{(\tilde{\delta},\tilde{\Delta})\in A_\tau,\, J_\tau^{1/2}(\tilde{\delta},\tilde{\Delta}) \geq v_{A_\tau}} Q_\tau(\theta(\tau)+\tilde{\delta}, \Pi(\tau)+\tilde{\Delta}) - Q_\tau(\theta(\tau), \Pi(\tau)) \right\} \\
&\geq \frac{\left(J_\tau^{1/2}(\delta,\Delta)\right)^2}{4} \wedge \left\{ \frac{J_\tau^{1/2}(\delta,\Delta)}{v_{A_\tau}} \frac{v_{A_\tau}^2}{4} \right\} \\
&\geq \frac{\left(J_\tau^{1/2}(\delta,\Delta)\right)^2}{4} \wedge qJ_\tau^{1/2}(\delta,\Delta),
\end{aligned}$$

where in last inequality we have used the fact that $v_{A_\tau} \;\geq\; 4q$. To see why this is true, notice that there exists $z_{X_{it},z} \in [0, z]$ such that

$$\begin{aligned}
&Q_\tau(\theta(\tau)+\delta, \Pi(\tau)+\Delta) - Q_\tau(\theta(\tau), \Pi(\tau)) \\
&= \frac{1}{nT}\sum_{i=1}^n \sum_{t=1}^T \mathbb{E}\left(\int_0^{X_{i,t}'\delta+\Delta_{it}} \left(F_{Y_{i,t}|X_{i,t},\Pi_{i,t}}(X_{i,t}'\theta(\tau)+\Pi_{i,t}+z) - F_{Y_{i,t}|X_{i,t},\Pi_{i,t}}(X_{i,t}'\theta(\tau)+\Pi_{i,t}(\tau)) \right) dz \right) \\
&= \frac{1}{nT}\sum_{i=1}^n \sum_{t=1}^T \mathbb{E}\Bigg(\int_0^{X_{i,t}'\delta+\Delta_{i,t}} \left(z f_{Y_{i,t}|X_{i,t},\Pi_{i,t}(\tau)}(X_{i,t}'\theta(\tau)+\Pi_{i,t}(\tau)) \right) + \\
&\quad \tfrac{z^2}{2} f'_{Y_{i,t}|X_{i,t},\Pi_{i,t}}(X_{i,t}'\theta(\tau)+\Pi_{i,t}(\tau)+z_{X_{i,t},z}) dz \Bigg) \\
&\geq \frac{\underline{f}}{nT}\sum_{i=1}^n \sum_{t=1}^T \mathbb{E}\left(\left(X_{i,t}'\delta+\Delta_{i,t}\right)^2 \right) - \frac{1}{6}\frac{\bar{f}'}{nT}\sum_{i=1}^n \sum_{t=1}^T \mathbb{E}\left(\left|X_{i,t}'\delta+\Delta_{i,t}\right|^3 \right).
\end{aligned} \tag{51}$$

Hence, if $(\delta, \Delta) \in A_\tau$ with $J_\tau^{1/2}(\delta,\Delta) \;\leq\; 4q$ then

$$\sqrt{\frac{\underline{f}}{nT}\sum_{i=1}^n \sum_{t=1}^T \mathbb{E}\left(\left(X_{i,t}'\delta+\Delta_{i,t}\right)^2 \right)} \;\leq\; \frac{3}{2}\frac{\underline{f}^{3/2}}{\bar{f}'} \cdot \inf_{(\delta,\Delta)\in A_\tau, \delta\neq 0} \frac{\left(\mathbb{E}\left(\frac{1}{nT}\sum_{i=1}^n \sum_{t=1}^T (X_{i,t}'\delta+\Delta_{i,t})^2 \right) \right)^{3/2}}{\mathbb{E}\left(\frac{1}{nT}\sum_{i=1}^n \sum_{t=1}^T |X_{i,t}'\delta+\Delta_{i,t}|^3 \right)}$$

combined with (51) implies

$$Q_\tau(\theta(\tau)+\delta,\Pi(\tau)+\Delta)-Q_\tau(\theta(\tau),\Pi(\tau)) \;\geq\; \frac{\left(J_\tau^{1/2}(\delta,\Delta)\right)^2}{4}.$$

□

Lemma 9. *Under Assumption 3, for all $(\delta,\Delta)\in A_\tau$, we have*

$$\|\delta\|_{1,n,T} \;\leq\; \frac{2(C_0+1)\sqrt{s_\tau+1}}{\kappa_0}\max\left\{\frac{\sqrt{r_\tau}}{\sqrt{\log(n\vee pc_T)}},1\right\}J_\tau^{1/2}(\delta,\Delta),$$

and

$$\|\Delta\|_* \;\leq\; (C_0+1)\sqrt{s_\tau+1}\sqrt{nT\log(\max\{pc_T,n\})}\kappa_0^{-1}\max\left\{\frac{\sqrt{r_\tau}}{\sqrt{\log(n\vee pc_T)}},1\right\}J_\tau^{1/2}(\delta,\Delta),$$

with C_0 as in Lemma 7.

Proof. By Cauchy-Schwartz's inequality, and the definition of A_τ and $J_\tau^{1/2}(\delta,\Delta)$ we have

$$\begin{aligned}
\|\delta\|_{1,n,T} &\leq \tfrac{5}{4}\left(\|\delta_{T_\tau}\|_1 + \|\delta_{T_\tau^c}\|_1\right)\\
&\leq \tfrac{5}{4}\|\delta_{T_\tau}\|_1 + \tfrac{5C_0}{4}\left(\|\delta_{T_\tau}\|_1 + \tfrac{\sqrt{r_\tau}\|\Delta\|_F}{\sqrt{nT\log(n\vee pc_T)}}\right)\\
&\leq 2(C_0+1)\sqrt{s_\tau}\|\delta_{T_\tau}\|_2 + 2C_0\left(\tfrac{\sqrt{r_\tau}\|\Delta\|_F}{\sqrt{nT\log(n\vee pc_T)}}\right)\\
&\leq 2(C_0+1)\sqrt{s_\tau+1}\left(\|\delta_{T_\tau}\|_2 + \tfrac{\sqrt{r_\tau}}{\sqrt{nT\log(n\vee pc_T)}}\|\Delta\|_F\right)\\
&\leq 2(C_0+1)\sqrt{s_\tau+1}\max\left\{\tfrac{\sqrt{r_\tau}}{\sqrt{\log(n\vee pc_T)}},1\right\}\tfrac{J_\tau^{1/2}(\delta,\Delta)}{\kappa_0}.
\end{aligned}$$

On the other hand, by the triangle inequality, the construction of the set A_τ, and Cauchy-Schwartz's inequality

$$\begin{aligned}
\|\Delta\|_* &\leq C_0\sqrt{nT\log(n\vee pc_T)}\left(\|\delta_{T_\tau}\|_1 + \tfrac{\sqrt{r_\tau}\|\Delta\|_F}{\sqrt{nT\log(n\vee pc_T)}}\right)\\
&\leq \sqrt{s_\tau+1}(C_0+1)\sqrt{nT\log(n\vee pc_T)}\left(\|\delta_{T_\tau}\|_2 + \tfrac{\sqrt{r_\tau}\|\Delta\|_F}{\sqrt{nT\log(n\vee pc_T)}}\right)\\
&\leq \sqrt{s_\tau+1}(C_0+1)\sqrt{nT\log(n\vee pc_T)}\max\left\{\tfrac{\sqrt{r_\tau}}{\sqrt{\log(n\vee pc_T)}},1\right\}\tfrac{J_\tau^{1/2}(\delta,\Delta)}{\kappa_0}.
\end{aligned}$$

□

Lemma 10. *Let*

$$\begin{aligned}
\epsilon(\eta) \;=\; \sup_{(\delta,\Delta)\in A_\tau\,:\,J_\tau^{1/2}(\delta,\Delta)\leq\eta} \Big|&\hat{Q}_\tau(\theta(\tau)+\delta,\Pi(\tau)+\Delta)-\hat{Q}_\tau(\theta,\Pi)-\\
&Q_\tau(\theta(\tau)+\delta,\Pi(\tau)+\Delta)+Q_\tau(\theta(\tau),\Pi(\tau))\Big|,
\end{aligned}$$

and $\{\phi_n\}$ *a sequence with* $\phi_n/(\sqrt{\underline{f}}\log(c_T+1)) \to \infty$. *Then for all* $\eta > 0$

$$\epsilon(\eta) \leq \frac{\tilde{C}_0 \eta c_T \phi_n \sqrt{(1+s_\tau)\max\{\log(pc_T \vee n), r_\tau\}}(\sqrt{n}+\sqrt{d_T})}{\sqrt{nT}\kappa_0 \underline{f}^{1/2}},$$

for some constant $\tilde{C}_0 > 0$, *with probability at least* $1-\alpha_n$. *Here, the sequence* $\{\alpha_n\}$ *is independent of* η, *and* $\alpha_n \to 0$.

Proof. Let Ω_1 be the event $\max_{j\leq p}|\hat{\sigma}_j - 1| \leq 1/4$. Then, by Assumption , $P(\Omega_1) \geq 1-\gamma$. Next let $\kappa > 0$, and $f = (\delta, \Delta) \in A_\tau$ and write

$$\mathcal{F} = \{(\delta,\Delta) \in A_\tau : \quad J_\tau^{1/2}(\delta,\Delta) \leq \eta\}.$$

Then notice that by Lemmas 4.1 and 4.2 from Yu (1994),

$$\begin{aligned}
\mathbb{P}\left(\epsilon(\eta)\sqrt{nT} \geq \kappa\right) &\leq 2\,\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{1}{\sqrt{nd_T}}\left|\sum_{i=1}^n\sum_{l=1}^{d_T}\sum_{t\in H_l}\frac{Z_{i,t}(f)}{\sqrt{c_T}}\right| \geq \frac{\kappa}{3}\right) + \\
&\quad \mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{1}{\sqrt{nd_T}}\left|\sum_{i=1}^n\sum_{t\in R}\frac{Z_{i,t}(f)}{\sqrt{c_T}}\right| \geq \frac{\kappa}{3}\right) + 2nT\left(\frac{1}{c_T}\right)^\mu, \\
&=: A_1 + A_2 + 2nT\left(\tfrac{1}{c_T}\right)^\mu,
\end{aligned} \tag{52}$$

where

$$\begin{aligned}
Z_{i,t}(f) &= \rho_\tau(\widetilde{Y}_{i,t} - \widetilde{X}'_{i,t}(\theta(\tau)+\delta) - (\Pi_{i,t}(\tau)+\Delta_{i,t})) - \rho_\tau(\widetilde{Y}_{i,t} - \widetilde{X}'_{i,t}\theta(\tau) - \Pi_{i,t}(\tau))) \\
&\quad -\mathbb{E}\left(\rho_\tau(\widetilde{Y}_{i,t} - \widetilde{X}'_{i,t}(\theta(\tau)+\delta) - (\Pi_{i,t}(\tau)+\Delta_{i,t})) - \rho_\tau(\widetilde{Y}_{i,t} - \widetilde{X}'_{i,t}\theta(\tau) - \Pi_{i,t}(\tau)))\right).
\end{aligned}$$

Next we proceed to bound each term in (52). To that end, notice that

$$\begin{aligned}
\mathrm{Var}\left(\sum_{i=1}^n\sum_{l=1}^{d_T}\sum_{t\in H_l}\frac{Z_{i,t}(f)}{\sqrt{c_T}}\right) &\leq \sum_{i=1}^n\sum_{l=1}^{d_T}\mathbb{E}\left(\left(\frac{1}{\sqrt{c_T}}\sum_{t\in H_l} Z_{i,t}(f)\right)^2\right) \\
&\leq \sum_{i=1}^n\sum_{l=1}^{d_T}\sum_{t\in H_l}\mathbb{E}\left(\left(\tilde{X}'_{i,t}\delta + \Delta_{i,t}\right)^2\right) \\
&\leq \frac{nT}{\underline{f}}\left(J_\tau^{1/2}(\delta,\Delta)\right)^2.
\end{aligned}$$

Let $\{\varepsilon_{i,l}\}_{i\in[n],\, l\in[d_T]}$ be i.i.d Rademacher variables independent of the data.

Therefore, by Lemma 2.3.7 in van der Vaart and Wellner (1996)

$$\begin{aligned}
\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{1}{\sqrt{nd_T}}\left|\sum_{i=1}^n\sum_{l=1}^{d_T}\sum_{t\in H_l}\frac{Z_{i,t}(f)}{\sqrt{c_T}}\right| \geq \kappa\right) &\leq \frac{\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{nd_T}}\sum_{i=1}^n\sum_{l=1}^{d_T}\varepsilon_{i,l}\left(\sum_{t\in H_l}\frac{Z_{i,t}(f)}{\sqrt{c_T}}\right)\right| \geq \frac{\kappa}{4}\right)}{1 - \frac{12}{nT\kappa^2}\sup_{f\in\mathcal{F}}\mathrm{Var}(\sum_{i=1}^n\sum_{l=1}^{d_T}\sum_{t\in H_l}\frac{Z_{i,t}(f)}{\sqrt{c_T}})} \\
&\leq \frac{\mathbb{P}\left(A^0(\eta) \geq \frac{\kappa}{12}|\Omega_1\right) + \mathbb{P}(\Omega_1^c)}{1 - \frac{72c_T\eta^2}{\underline{f}\kappa^2}},
\end{aligned} \tag{53}$$

where

$$A^0(\eta) := \sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{nd_T}}\sum_{i=1}^{n}\sum_{l=1}^{d_T}\varepsilon_{i,l}\left(\sum_{t\in H_l}\frac{\rho_\tau(\widetilde{Y}_{i,t}-\widetilde{X}'_{i,t}(\theta(\tau)+\delta)-(\Pi_{i,t}(\tau)+\Delta_{i,t}))-\rho_\tau(\widetilde{Y}_{i,t}-\widetilde{X}'_{i,t}\theta(\tau)-\Pi_{i,t}(\tau))}{\sqrt{c_T}}\right)\right|.$$

Next, note that

$$\begin{aligned}\rho_\tau(\widetilde{Y}_{i,t}-\widetilde{X}'_{i,t}(\theta(\tau)+\delta)-(\Pi_{i,t}(\tau)+\Delta_{i,t}))-\rho_\tau(\widetilde{Y}_{i,t}-\widetilde{X}'_{i,t}\theta(\tau)-\Pi_{i,t}(\tau)) &= \tau\left(\widetilde{X}'_{i,t}\delta+\Delta_{i,t}\right)\\ &\quad +v_{i,t}(\delta,\Delta)+w_{i,t}(\delta,\Delta),\end{aligned}$$

where

$$\begin{aligned}|v_{i,t}(\delta,\Delta)| &= \left|(\widetilde{Y}_{i,t}-\widetilde{X}'_{i,t}(\theta(\tau)+\delta)-(\Pi_{i,t}(\tau)+\Delta_{i,t}))_- - (Y_{i,t}-\widetilde{X}'_{i,t}(\theta(\tau)+\delta)-\Pi_{i,t}(\tau))_-\right|\\ &\le |\Delta_{i,t}|.\end{aligned} \tag{54}$$

and

$$\begin{aligned}|w_{i,t}(\delta,\Delta)| &= \left|(\widetilde{Y}_{i,t}-\widetilde{X}'_{i,t}(\theta(\tau)+\delta)-\Pi_{i,t}(\tau))_- - (\widetilde{Y}_{i,t}-\widetilde{X}'_{i,t}\theta(\tau)-\Pi_{i,t}(\tau))_-\right|\\ &\le |\widetilde{X}'_{i,t}\delta|.\end{aligned} \tag{55}$$

Moreover, notice that by Lemma 9,

$$\{(\delta,\Delta)\in A_\tau : J_\tau^{1/2}(\delta,\Delta)\le\eta\} \subset \{(\delta,\Delta)\in A_\tau : \|\delta\|_{1,n,T}\le\eta\upsilon\},$$

where

$$\upsilon := \frac{2(C_0+1)\sqrt{1+s_\tau}}{\kappa_0}\max\left\{\frac{\sqrt{r_\tau}}{\sqrt{\log(n\vee pc_T)}},1\right\}.$$

Also by Lemma 9, for $(\delta,\Delta)\in A_\tau$

$$\|\Delta\|_* \le \frac{\sqrt{1+s_\tau}(C_0+1)J_\tau^{1/2}(\delta,\Delta)\sqrt{nT\max\{\log(pc_T\vee n),r_\tau\}}}{\kappa_0},$$

and so,

$$\begin{aligned}&\{(\delta,\Delta)\in A_\tau : J_\tau^{1/2}(\delta,\Delta)\le\eta\} \subset\\ &\left\{(\delta,\Delta)\in A_\tau : \|\Delta\|_*\le\sqrt{1+s_\tau}(C_0+1)\eta\sqrt{nT\max\{\log(pc_T\vee n),r_\tau\}}/\kappa_0\right\}.\end{aligned}$$

Hence, defining

$$
\begin{aligned}
B_1^0(\eta) &= \sqrt{c_T} \sup_{\delta\,:\,\|\delta\|_{1,n,T}\le \eta\upsilon} \left| \frac{1}{\sqrt{nd_T}} \sum_{i=1}^n \sum_{l=1}^{d_T} \varepsilon_{i,l} \left(\sum_{t\in H_l} \frac{\widetilde{X}_{i,t}'\delta}{c_T} \right) \right|, \\
B_2^0(\eta) &= \sqrt{c_T} \sup_{\Delta\,:\,\|\Delta\|_*\le \sqrt{1+s_\tau}(C_0+1)\eta\sqrt{nT\max\{\log(pc_T\vee n), r_\tau\}}/\kappa_0} \left| \frac{1}{\sqrt{nd_T}} \sum_{i=1}^n \sum_{l=1}^{d_T} \varepsilon_{i,l} \left(\sum_{t\in H_l} \frac{\Delta_{i,t}}{c_T} \right) \right| \\
B_3^0(\eta) &= \sqrt{c_T} \sup_{\Delta\,:\,\|\Delta\|_*\le \sqrt{1+s_\tau}(C_0+1)\eta\sqrt{nT\max\{\log(pc_T\vee n), r_\tau\}}/\kappa_0} \left| \frac{1}{\sqrt{nd_T}} \sum_{i=1}^n \sum_{l=1}^{d_T} \varepsilon_{i,l} \left(\sum_{t\in H_l} \frac{v_{i,t}(\delta,\Delta)}{c_T} \right) \right|, \\
B_4^0(\eta) &= \sqrt{c_T} \sup_{\delta\,:\,\|\delta\|_{1,n,T}\le \eta\upsilon} \left| \frac{1}{\sqrt{nd_T}} \sum_{i=1}^n \sum_{l=1}^{d_T} \varepsilon_{i,l} \left(\sum_{t\in H_l} \frac{w_{i,t}(\delta,\Delta)}{c_T} \right) \right|.
\end{aligned}
$$

By union bound we obtain that

$$
\mathbb{P}(A^0(\eta) \ge \kappa|\Omega_1) \le \sum_{j=1}^4 \mathbb{P}(B_j^0(\eta) \ge \kappa|\Omega_1), \tag{56}
$$

so we proceed to bound each term in the right hand side of the inequality above.

First, notice that

$$
B_1^0(\eta) \le 2c_T \max_{m\in[c_T]} \sup_{\delta\,:\,\|\delta\|_{1,n,T}\le \eta\upsilon} \left| \frac{1}{\sqrt{nT}} \sum_{i=1}^n \sum_{l=0}^{d_T-1} \varepsilon_{i,l} \widetilde{X}_{i,\,(2lc_T+m)}'\delta \right|,
$$

and hence by a union bound and the same argument on the proof of Lemma 5 in Belloni and Chernozhukov (2011), we have that

$$
\mathbb{P}(B_1^0(\eta) \ge \kappa|\Omega_1) \le 2pc_T \exp\left(-\frac{\kappa^2}{4c_T^2(16\sqrt{2}\eta\upsilon)^2} \right). \tag{57}
$$

Next we proceed to bound $B_3^0(\eta)$, by noticing that

$$
B_3^0(\eta) \le \max_{m\in[c_T]} \sup_{\Delta\,:\,\|\Delta\|_*\le \sqrt{1+s_\tau}(C_0+1)\eta\sqrt{nT\max\{\log(pc_T\vee n), r_\tau\}}/\kappa_0} \left| \frac{\sqrt{c_T}}{\sqrt{nd_T}} \sum_{i=1}^n \sum_{l=0}^{d_T-1} \varepsilon_{i,l} v_{i,\,(2lc_T+m)}(\delta,\Delta) \right|.
$$

Towards that end we proceed to bound the moment generating function of $B_3^0(\eta)$ and the use that to obtain

an upper bound on $B_3^0(\eta)$. Now fix $m \in [d_T]$ and notice that

$$
\begin{aligned}
&\mathbb{E}\left(\exp\left(\lambda \sup_{\Delta\,:\,\|\Delta\|_* \le \sqrt{1+s_\tau}(C_0+1)\eta\sqrt{nT\max\{\log(pc_T\vee n), r_\tau\}}/\kappa_0} \left|\frac{\sqrt{c_T}}{\sqrt{nd_T}}\sum_{i=1}^{n}\sum_{l=0}^{d_T-1}\varepsilon_{i,l}v_{i,\,(2lc_T+m)}(\delta,\Delta)\right|\right)\right)\\
&\le \mathbb{E}\left(\exp\left(\lambda \sup_{\Delta\,:\,\|\Delta\|_* \le \sqrt{1+s_\tau}(C_0+1)\eta\sqrt{nT\max\{\log(pc_T\vee n), r_\tau\}}/\kappa_0} \left|\frac{\sqrt{c_T}}{\sqrt{nd_T}}\sum_{i=1}^{n}\sum_{l=1}^{d_T}\varepsilon_{i,l}\,\Delta_{i,\,(2lc_T+m)}\right|\right)\right)\\
&\le \mathbb{E}\Bigg(\sup_{\Delta\,:\,\|\Delta\|_* \le \sqrt{1+s_\tau}(C_0+1)\eta\sqrt{nT\max\{\log(pc_T\vee n), r_\tau\}}/\kappa_0}\left(\exp\left(\frac{\lambda\sqrt{c_T}\|\Delta\|_*\,\mathbb{E}(\|\{\varepsilon_{il}\}\|_2)}{\sqrt{nd_T}}\right)\right.\\
&\qquad \left.\left.\exp\left(\lambda\frac{\sqrt{c_T}\|\Delta\|_*(\|\{\varepsilon_{i,l}\}\|_2 - \mathbb{E}(\|\{\varepsilon_{i,l}\}\|_2))}{\sqrt{nd_T}}\right)\right)\right)\\
&\le \exp\left(\lambda\frac{\sqrt{1+s_\tau}(C_0+1)c_4\eta c_T\sqrt{3\max\{\log(pc_T\vee n), r_\tau\}}\left(\sqrt{n}+\sqrt{d_T}\right)}{\kappa_0}\right)\cdot\\
&\exp\left(\frac{(s_\tau+1)(C_0+1)^2c_4\lambda^2c_T^2\eta^2\max\{\log(pc_T\vee n), r_\tau\}}{\kappa_0^2}\right),
\end{aligned}
\tag{58}
$$

for a positive constant $c_4 > 0$, and where the first inequality holds by Ledoux-Talagrand's contraction inequality, the second by the the duality of the spectral and nuclear norms and the triangle inequality, the third by Theorem 1.2 in Vu (2007) and by basic properties of sub-Gaussian random variables.

Therefore, by Markov's inequality and (58),

$$
\begin{aligned}
&\mathbb{P}\left(B_3^0(\eta) \ge \kappa|\Omega_1\right)\\
&\le c_T\max_{m\in[c_T]}\mathbb{P}\left(\sup_{\Delta\,:\,\|\Delta\|_* \le \sqrt{1+s_\tau}(C_0+1)\eta\sqrt{nT\max\{\log(pc_T\vee n), r_\tau\}}/\kappa_0}\left|\frac{\sqrt{c_T}}{\sqrt{nd_T}}\sum_{i=1}^{n}\sum_{l=0}^{d_T-1}\varepsilon_{i,l}v_{i,\,(2lc_T+m)}(\delta,\Delta)\right| \ge \kappa\right)\\
&\le \inf_{\lambda>0}\left[\exp\left(-\lambda\kappa\right)\exp\left(\lambda\frac{\sqrt{1+s_\tau}(C_0+1)c_4\eta c_T\sqrt{3\max\{\log(pc_T\vee n), r_\tau\}}\left(\sqrt{n}+\sqrt{d_T}\right)}{\kappa_0}\right)\cdot\right.\\
&\qquad\left.\exp\left(\frac{(1+s_\tau)(C_0+1)^2c_4\lambda^2c_T^2\eta^2\max\{\log(pc_T\vee n), r_\tau\}}{\kappa_0^2}+\log c_T\right)\right]\\
&\le c_5\exp\left(-\frac{\kappa\kappa_0}{\sqrt{1+s_\tau}(C_0+1)\eta c_T\sqrt{3\max\{\log(pc_T\vee n), r_\tau\}}\left(\sqrt{n}+\sqrt{d_T}\right)}+\log c_T\right),
\end{aligned}
\tag{59}
$$

for a positive constant $c_5 > 0$.

Furthermore, we observe that

$$
B_2^0(\eta) \le \max_{m\in[d_T]}\sup_{\Delta\,:\,\|\Delta\|_* \le \sqrt{1+s_\tau}(C_0+1)\eta\sqrt{nT\max\{\log(pc_T\vee n), r_\tau\}}/\kappa_0}\left|\frac{\sqrt{c_T}}{\sqrt{nd_T}}\sum_{i=1}^{n}\sum_{l=0}^{d_T-1}\varepsilon_{i,l}\Delta_{i,\,(2lc_T+m)}\right|.
$$

Hence, with the same argument for bounding $B_3^0(\eta)$, we have

$$\mathbb{P}\left(B_2^0(\eta) \geq \kappa|\Omega_1\right) \;\leq\; c_5 \exp\left(-\frac{\kappa\kappa_0}{\sqrt{1+s_\tau}(C_0+1)\eta c_T\sqrt{3\max\{\log(pc_T\vee n), r_\tau\}}\left(\sqrt{n}+\sqrt{d_T}\right)} + \log c_T\right). \tag{60}$$

Finally, we proceed to bound $B_4^0(\eta)$. To that end, notice that

$$B_4^0(\eta) \;\leq\; \max_{m\in[d_T]} \sup_{\delta\,:\,\|\delta\|_{1,n,T}\leq\eta\upsilon} \left|\frac{\sqrt{c_T}}{\sqrt{nd_T}}\sum_{i=1}^{n}\sum_{l=0}^{d_T-1}\varepsilon_{i,l}w_{i,\,(2lc_T+m)}(\delta,\Delta)\right|,$$

and by (55) and Ledoux-Talagrand's inequality, as in (57), we obtain

$$\mathbb{P}(B_4^0(\eta) \;\geq\; \kappa|\Omega_1) \;\leq\; 2pc_T\exp\left(-\frac{\kappa^2}{4c_T^2(16\sqrt{2}\eta\upsilon)^2}\right). \tag{61}$$

Therefore, letting

$$\kappa \;=\; \frac{\eta c_T\phi_n(1+C_0)^2\sqrt{(1+s_\tau)\max\{\log(pc_T\vee n), r_\tau\}}(\sqrt{n}+\sqrt{d_T})}{\kappa_0\underline{f}^{1/2}},$$

and repeating the argument above for bounding A_2 in (52), combining (52), (53), (56), (57), (59), (60) and (61), we obtain that

$$\begin{aligned}\mathbb{P}(\epsilon(\eta) \;\geq\; \tfrac{\kappa}{\sqrt{nT}}) \;\;\leq\;\; & 5\frac{\gamma+4\exp\left(\max\{\log(pc_T\vee n),r_\tau\}-C_1\frac{\phi_n^2\max\{\log(pc_T\vee n),r_\tau\}}{\underline{f}}(\sqrt{n}+\sqrt{d_T})^2\right)+2c_5\exp\left(-C_2\frac{2\phi_n}{\underline{f}^{1/2}}\right)}{1-\frac{3\kappa_0^2}{c_T\phi_n(s_\tau+1)\,(1+C_0)^2(\sqrt{n}+\sqrt{d_T})^2\max\{\log(pc_T\vee n),r_\tau\}}} \\ & + nT\left(\tfrac{1}{c_T}\right)^{\mu},\end{aligned}$$

for some positive constants C_1 and C_2. □

B.2 Proof of Theorem 1

Proof. Recall from Lemma 7, our choices of ν_1 and ν_2 are

$$\nu_1 \;=\; C_0'\sqrt{\frac{c_T\log(\max\{n, pc_T\})}{nd_T}}\left(\sqrt{n}+\sqrt{d_T}\right),$$

and

$$\nu_2 \;=\; \frac{200c_T}{nT}\left(\sqrt{n}+\sqrt{d_T}\right),$$

for $C_0' \;=\; 9/(1-c_0)$, and c_0 as in Lemma 7.

Let

$$\eta = \frac{8\phi_n(C_0'(1+C_0)+\tilde{C}_0+200(1+C_0))\sqrt{c_T(1+s_\tau)\ \max\{\log(pc_T\vee n), r_\tau\}}(\sqrt{n}+\sqrt{d_T})}{\sqrt{nd_T}\kappa_0 \underline{f}^{1/2}}, \quad (62)$$

for C_0 as in Lemma 7, and $\tilde{C}_0$ as in Lemma 10.

Throughout we assume that the following events happen:

- $\Omega_1 :=$ the event that $(\hat{\theta}(\tau)-\theta(\tau), \hat{\Pi}(\tau)-\Pi(\tau)) \in A_\tau$.
- $\Omega_2 :=$ the event for which the upper bound on $\epsilon(\eta)$ in Lemma 10 holds.

Suppose that

$$|J_\tau^{1/2}(\hat{\theta}(\tau)-\theta(\tau), \hat{\Pi}(\tau)-\Pi(\tau))| > \eta. \quad (63)$$

Then, by the convexity of A_τ, and of the objective $\hat{Q}$ with its constraint, we obtain that

$$\begin{aligned} 0 > \min_{(\delta,\Delta)\in A_\tau : |J_\tau^{1/2}(\delta,\Delta)|=\eta} & \hat{Q}_\tau(\theta(\tau)+\delta, \Pi(\tau)+\Delta) - \hat{Q}(\theta(\tau), \Pi(\tau)) + \nu_1 \left[\|\theta(\tau)+\delta\|_{1,n,T} - \|\theta(\tau)\|_{1,n,T}\right] \\ & + \nu_2 \left[\|\Pi(\tau)+\Delta\|_* - \|\Pi(\tau)\|_*\right] \end{aligned}$$

Moreover, by Lemma 9 and the triangle inequality,

$$\begin{aligned} \|\theta(\tau)\|_{1,n,T} - \|\theta(\tau)+\delta\|_{1,n,T} &\leq \|\delta_{T_\tau}\|_{1,n,T} \\ &\leq 2(1+C_0)\sqrt{1+s_\tau}\frac{J_\tau^{1/2}(\delta,\Delta)}{\kappa_0}\max\left\{\frac{\sqrt{r_\tau}}{\sqrt{\log(n\vee pc_T)}}, 1\right\}, \end{aligned}$$

and

$$\begin{aligned} \|\Pi(\tau)\|_* - \|\Pi(\tau)+\Delta\|_* &\leq \|\Delta\|_* \\ &\leq (1+C_0)\sqrt{1+s_\tau}\sqrt{nT\max\{\log(pc_T\vee n), r_\tau\}}\frac{J_\tau^{1/2}(\delta,\Delta)}{\kappa_0}. \end{aligned}$$

Therefore,

$$\begin{aligned} 0 > & \min_{(\delta,\Delta)\in A_\tau : |J_\tau^{1/2}(\delta,\Delta)|=\eta} \hat{Q}(\theta(\tau)+\delta, \Delta+\Pi(\tau)) - \\ & \hat{Q}(\theta(\tau), \Pi(\tau)) - 2\nu_1(1+C_0)\sqrt{1+s_\tau}\frac{J_\tau^{1/2}(\delta,\Delta)}{\kappa_0}\max\left\{\frac{\sqrt{r_\tau}}{\sqrt{\log(n\vee pc_T)}}, 1\right\}, \\ & -\nu_2(1+C_0)\sqrt{1+s_\tau}\sqrt{nT\max\{\log(pc_T\vee n), r_\tau\}}\frac{J_\tau^{1/2}(\delta,\Delta)}{\kappa_0} \\ = & \min_{(\delta,\Delta)\in A_\tau : |J_\tau^{1/2}(\delta,\Delta)|=\eta} \Bigg[\hat{Q}_\tau(\theta(\tau)+\delta, \Delta+\Pi(\tau)) - \hat{Q}(\theta(\tau), \Pi(\tau)) \\ & -Q(\theta(\tau)+\delta, \Delta+\Pi(\tau)) + Q(\theta(\tau), \Pi(\tau)) \\ & +Q(\theta(\tau)+\delta, \Delta+\Pi(\tau)) - Q(\theta(\tau), \Pi(\tau)) - \\ & 2C_0'(1+C_0)\sqrt{\frac{c_T(1+s_\tau)\ \max\{\log(pc_T\vee n), r_\tau\}}{nd_T}}(\sqrt{n}+\sqrt{d_T})\frac{J_\tau^{1/2}(\delta,\Delta)}{\kappa_0} - \\ & \frac{200c_T}{nT}\left(\sqrt{n}+\sqrt{d_T}\right)\sqrt{1+s_\tau}(1+C_0)\sqrt{nT\max\{\log(pc_T\vee n), r_\tau\}}\frac{J_\tau^{1/2}(\delta,\Delta)}{\kappa_0}\Bigg] \end{aligned}$$

$$
\begin{aligned}
\geq & \min_{(\delta,\Delta)\in A_\tau\,:\,|J_\tau^{1/2}(\delta,\Delta)|=\eta} Q(\theta(\tau)+\delta,\Delta+\Pi(\tau)) - Q(\theta(\tau),\Pi(\tau)) \\
& -[2C_0'(1+C_0)+200(C_0+1)]\sqrt{\tfrac{c_T(1+s_\tau)\,\max\{\log(pc_T\vee n),r_\tau\}}{nd_T}}(\sqrt{n}+\sqrt{d_T})\tfrac{J_\tau^{1/2}(\delta,\Delta)}{\kappa_0} \\
& -\frac{\tilde{C}_0\eta\phi_n\sqrt{c_T(1+s_\tau)\max\{\log(pc_T\vee n),r_\tau\}}(\sqrt{n}+\sqrt{d_T})}{\sqrt{nd_T}\kappa_0\underline{f}^{1/2}}
\end{aligned}
$$

$$
\begin{aligned}
\geq & \frac{\eta^2}{4}\wedge(\eta q) - [2C_0'(1+C_0)+200(C_0+1)]\sqrt{\frac{c_T(1+s_\tau)\,\max\{\log(pc_T\vee n),r_\tau\}}{nd_T}}(\sqrt{n}+\sqrt{d_T})\frac{\eta}{\kappa_0} - \\
& \tfrac{\tilde{C}_0\eta\phi_n\sqrt{c_Ts_\tau\,\max\{\log(pc_T\vee n),r_\tau\}}(\sqrt{n}+\sqrt{d_T})}{\sqrt{nd_T}\kappa_0\underline{f}^{1/2}} \\
\geq & \frac{\eta^2}{4} - \frac{2\eta\phi_n(C_0'(1+C_0)+\tilde{C}_0+200(C_0+1))\sqrt{c_T(1+s_\tau)\,\max\{\log(pc_T\vee n),r_\tau\}}(\sqrt{n}+\sqrt{d_T})}{\sqrt{nd_T}\kappa_0\underline{f}^{1/2}} \\
= & 0,
\end{aligned}
\tag{64}
$$

where the the second inequality follows from Lemma 10, the third from Lemma 8, the fourth from our choice of η and (25), and the equality also from our choice of η. Hence, (64) leads to a contradiction which shows that (63) cannot happen in the first place. As a result, by Assumption 3,

$$
\frac{\|\hat{\Pi}(\tau)-\Pi(\tau)\|_F}{\sqrt{nT}} \leq \frac{1}{\kappa_0}|J_\tau^{1/2}(\hat{\theta}(\tau)-\theta(\tau),\hat{\Pi}(\tau)-\Pi(\tau))| \leq \frac{\eta}{\kappa_0},
$$

which holds with probability approaching one.

To conclude the proof, let $\hat{\delta} = \hat{\theta}-\theta$ and notice that

$$
\begin{aligned}
\|\hat{\delta}_{(T_\tau\cup\overline{T}_\tau(\hat{\delta},m))^c}\|^2 & \leq \sum_{k\geq m+1}\frac{\|\hat{\delta}_{T_\tau^c}\|_1^2}{k^2} \\
& \leq \frac{\|\hat{\delta}_{T_\tau^c}\|_1^2}{m} \\
& \leq \frac{4C_0}{m}\left[\|\hat{\delta}_{T_\tau}\|_1^2+\frac{r_\tau\|\Pi(\tau)-\hat{\Pi}(\tau)\|_F^2}{nT\log(pc_T\vee n)}\right] \\
& \leq \frac{4C_0}{m}\left[s_\tau\,\|\hat{\delta}_{T_\tau\cup\overline{T}_\tau(\hat{\delta},m)}\|^2+\frac{r_\tau\|\Pi(\tau)-\hat{\Pi}(\tau)\|_F^2}{nT\log(pc_T\vee n)}\right],
\end{aligned}
$$

which implies

$$
\begin{aligned}
\|\hat{\delta}\| & \leq \left(1+2C_0\sqrt{\tfrac{s_\tau}{m}}\right)\left(\|\hat{\delta}_{T_\tau\cup\overline{T}_\tau(\hat{\delta},m)}\|+\tfrac{\sqrt{r_\tau}\|\Pi(\tau)-\hat{\Pi}(\tau)\|_F}{\sqrt{nT\log(c_Tp\vee n)}}\right) \\
& \leq \tfrac{J_\tau^{1/2}(\hat{\delta},\hat{\Pi}(\tau)-\Pi(\tau))}{\kappa_m}\left(1+2C_0\sqrt{\tfrac{s_\tau}{m}}\right),
\end{aligned}
$$

and the result follows. □

B.3 Proof of Theorem 3

Lemma 11. *Suppose that Assumptions 1–2 and 4 hold. Let*

$$\nu_1 \;=\; \frac{2}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\|X_{i,t}\|_\infty.$$

and

$$\nu_2 \;=\; \frac{200c_T}{nT}\left(\sqrt{n}+\sqrt{d_T}\right).$$

We have that

$$(\hat{\Pi}(\tau)-\Pi(\tau)-X\theta(\tau)) \in A'_\tau,$$

with probability approaching one, where

$$A'_\tau = \left\{\Delta \in \mathbb{R}^{n\times T} \;:\; \|\Delta\|_* \le c_0\sqrt{r_\tau}\left(\|\Delta\|_F+\|\xi\|_*\right),\;\; \|\Delta\|_\infty \le c_1\right\},$$

and c_0 and c_1 are positive constants that depend on τ. Furthermore, $\hat{\theta}(\tau)=0$.

Proof. First, we observe that C in the statement of Theorem 3 can be take as $C=\|X'\theta(\tau)+\Pi(\tau)\|_\infty$. And so,

$$\left\|X\theta(\tau)+\Pi(\tau)-\hat{\Pi}(\tau)\right\|_\infty \;\le\; 2C \;=:\; c_1.$$

Next, notice that for any $\check{\Pi}\in\mathbb{R}^{n\times T}$ and $\check{\theta}\in\mathbb{R}^p$,

$$\begin{aligned}
\hat{Q}_\tau(0,\check{\Pi})-\hat{Q}_\tau(\check{\theta},\check{\Pi})-\nu_1\|\check{\theta}\|_{1,n,T} \;&\le\; \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}|X'_{i,t}\check{\theta}|-\nu_1\|\check{\theta}\|_1\\
&\le\; \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}|X'_{i,t}\check{\theta}|-\nu_1\|\check{\theta}\|_1\\
&\le\; \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\|X'_{i,t}\|_\infty\|\check{\theta}\|_1-\nu_1\|\check{\theta}\|_1\\
&<\; 0,
\end{aligned}$$

where the first inequality follows since ρ_τ is a contraction map. Therefore, $\hat{\theta}(\tau)=0$. Furthermore, by

Lemma 5, we have

$$
\begin{aligned}
0 \;\leq\;& \hat{Q}_\tau(0, X\theta(\tau)+\Pi(\tau)) - \hat{Q}_\tau(0,\hat{\Pi}(\tau)) \;+\; \nu_2\left(\|X\theta(\tau)+\Pi(\tau)\|_* - \|\hat{\Pi}(\tau)\|_*\right)\\
\leq\;& \left|\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} a_{i,t}\left((X\theta(\tau)+\Pi(\tau)) - \hat{\Pi}(\tau)\right)\right|\\
&+\nu_2\left(\|X\theta(\tau)+\Pi(\tau)\|_* - \|\hat{\Pi}(\tau)\|_*\right)\\
\leq\;& \|X\theta(\tau)+\Pi(\tau)-\hat{\Pi}(\tau)\|_*\left(\sup_{\|\tilde{\Delta}\|_*\leq 1}\left|\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} a_{i,t}\tilde{\Delta}_{i,t}\right|\right)\\
&+\nu_2\left(\|X\theta(\tau)+\Pi(\tau)\|_* - \|\hat{\Pi}(\tau)\|_*\right)\\
\leq\;& \frac{200c_T}{nT}\left(\sqrt{n}+\sqrt{d_T}\right)\left(\|X\theta(\tau)+\Pi(\tau)-\hat{\Pi}(\tau)\|_* + \|X\theta(\tau)+\Pi(\tau)\|_* - \|\hat{\Pi}(\tau)\|_*\right)\\
&-\frac{100c_T}{nT}\left(\sqrt{n}+\sqrt{d_T}\right)\|X\theta(\tau)+\Pi(\tau)-\hat{\Pi}(\tau)\|_*\\
\leq\;& \frac{200c_T}{nT}\left(\sqrt{n}+\sqrt{d_T}\right)\left(\|X\theta(\tau)+\Pi(\tau)+\xi-\hat{\Pi}(\tau)\|_* + \|X\theta(\tau)+\Pi(\tau)+\xi\|_* - \|\hat{\Pi}(\tau)\|_*\right)\\
&+\frac{400c_T}{nT}\left(\sqrt{n}+\sqrt{d_T}\right)\|\xi\|_* \;-\; \frac{100c_T}{nT}\left(\sqrt{n}+\sqrt{d_T}\right)\|X\theta(\tau)+\Pi(\tau)-\hat{\Pi}(\tau)\|_*\\
\leq\;& \frac{c_1\, c_T}{nT}\left(\sqrt{n}+\sqrt{d_T}\right)\sqrt{r_\tau}\|X\theta(\tau)+\Pi(\tau)+\xi-\hat{\Pi}(\tau)\|_F\\
&+\frac{400c_T}{nT}\left(\sqrt{n}+\sqrt{d_T}\right)\|\xi\|_* \;-\; \frac{100c_T}{nT}\left(\sqrt{n}+\sqrt{d_T}\right)\|X\theta(\tau)+\Pi(\tau)-\hat{\Pi}(\tau)\|_*
\end{aligned}
$$

for some positive constant c_1,

□

Lemma 12. *Let*

$$
\begin{aligned}
\epsilon'(\eta) \;=\; \sup_{(\delta,\Delta)\in A_\tau\,:\,J_\tau^{1/2}(\delta,\Delta)\leq\eta}\Big|&\hat{Q}_\tau(0, X\theta(\tau)+\Pi(\tau)+\Delta) - \hat{Q}_\tau(0, X\theta(\tau)+\Pi(\tau)) -\\
&Q_\tau(0, X\theta(\tau)+\Pi(\tau)+\Delta) + Q_\tau(0, X\theta(\tau)+\Pi(\tau))\Big|,
\end{aligned}
$$

and $\{\phi_n\}$ *a sequence with* $\phi_n/(\sqrt{\underline{f}}\log(c_T+1)) \to \infty$. *Then for all* $\eta > 0$

$$
\epsilon'(\eta) \;\leq\; \frac{\tilde{C}_0\eta c_T\phi_n\sqrt{r_\tau}(\sqrt{n}+\sqrt{d_T})}{\sqrt{nT}\underline{f}},
$$

for some constant $\tilde{C}_0 > 0$, *with probability at least* $1-\alpha_n$. *Here, the sequence* $\{\alpha_n\}$ *is independent of* η, *and* $\alpha_n \to 0$.

Proof. This follows similarly to the proof of Lemma 10.

□

Lemma 13. *Let*

$$A''_\tau \;=\; \left\{\Delta \in A'_\tau \;:\; q(\Delta) \geq 2\eta_0,\;\; \Delta \neq 0\right\},$$

with

$$\eta_0 \;=\; \frac{\tilde{C}_1 c_T \phi_n \sqrt{r_\tau}(\sqrt{n}+\sqrt{d_T})}{\sqrt{nT}\,\underline{f}},$$

for an appropriate constant $\tilde{C}_1 > 0$*, and*

$$q(\Delta) \;=\; \frac{3}{2}\frac{\underline{f}^{3/2}}{\bar{f}'}\,\frac{\left(\frac{1}{nT}\sum_{i=1}^n\sum_{t=1}^T(\Delta_{i,t})^2\right)^{3/2}}{\frac{1}{nT}\sum_{i=1}^n\sum_{t=1}^T|\Delta_{i,t}|^3}.$$

Under Assumptions 1-2 and 4, for any $\Delta \in A''_\tau$ *we have that*

$$Q_\tau(0, X\theta(\tau)+\Pi(\tau)+\Delta) - Q_\tau(0, X\theta(\tau)+\Pi(\tau)) \;\geq\; \min\left\{\frac{\underline{f}\|\Delta\|^2}{4nT}, \frac{2\eta\underline{f}^{1/2}\|\Delta\|}{\sqrt{nT}}\right\}.$$

Proof. This follows as the proof of Lemma 8. □

The proof of Theorem 3 proceeds by exploiting Lemmas 11 and 13. By Lemma 11, we have that

$$\hat{\Delta} \;:=\; \hat{\Pi}(\tau) - X\theta(\tau) - \Pi(\tau) \in A'_\tau, \tag{65}$$

with high probability. Therefore, we assume that (65) holds. Hence, if $\hat{\Delta} \notin A''_\tau$, then

$$\frac{1}{\sqrt{nT}}\|\hat{\Delta}\|_F \;<\; \frac{4\,\eta\,\left(\sum_{i=1}^n\sum_{t=1}^T|\hat{\Delta}_{i,t}|^3\right)}{3\,\left(\sum_{i=1}^n\sum_{t=1}^T\hat{\Delta}_{i,t}^2\right)}\frac{\overline{f}'}{\underline{f}^{3/2}} \;\leq\; \frac{4\overline{f}'\|\hat{\Delta}\|_\infty\,\eta}{3\underline{f}^{3/2}}. \tag{66}$$

If $\hat{\Delta} \in A''_\tau$, then we proceed as in the proof of Theorem 1 by exploiting Lemma 12, and treating $X\theta(\tau)+\Pi(\tau)$ as the latent factors matrix, the design matrix as the matrix zero, A''_τ as A_τ, and

$$\kappa_0 \;=\; \underline{f}^{1/2},$$

in Assumption 3. This leads to

$$\frac{1}{\sqrt{nT}}\|\hat{\Delta}\|_F \;\leq\; \eta, \tag{67}$$

and the claim in Theorem 3 follows combining (66) and (67).

B.4 Proof of Corollary 2

First notice that by Theorem 1 and Theorem 3 in Yu et al. (2014),

$$v := \max\left\{\min_{O\in\mathbb{O}_{r_\tau}}\|\hat{g}(\tau)O - g(\tau)\|_F, \min_{O\in\mathbb{O}_{r_\tau}}\|\tilde{\hat{\lambda}}(\tau)O - \tilde{\lambda}(\tau)\|_F\right\} = O_{\mathbb{P}}\left(\frac{(\sigma_1(\tau)+\sqrt{r_\tau}\,\mathrm{Err})\mathrm{Err}}{(\sigma_{r_\tau-1}(\tau))^2-(\sigma_{r_\tau}(\tau))^2}\right). \tag{68}$$

Furthermore,

$$\begin{aligned}
\frac{\|\hat{\lambda}(\tau)-\lambda(\tau)\|_F^2}{nT} &= \frac{1}{nT}\sum_{j=1}^{r_\tau}\|\lambda_{\cdot,j}(\tau)-\hat{\lambda}_{\cdot,j}(\tau)\|^2 \\
&\le \frac{2}{nT}\sum_{j=1}^{r_\tau}(\sigma_j-\hat{\sigma}_j)^2 + \frac{2}{nT}\sum_{j=1}^{r_\tau}\sigma_j^2\|\tilde{\hat{\lambda}}_j(\tau)-\tilde{\lambda}_j(\tau)\|^2 \\
&\le \frac{2}{nT}\sum_{j=1}^{r_\tau}(\sigma_j-\hat{\sigma}_j)^2 + \frac{2\,\sigma_1^2}{nT}\sum_{j=1}^{r_\tau}\|\tilde{\hat{\lambda}}_j(\tau)-\tilde{\lambda}_j(\tau)\|^2 \\
&\le \frac{2\,r_\tau}{nT}(\sigma_1-\hat{\sigma}_1)^2 + \frac{2\,\sigma_1^2}{nT}\|\tilde{\hat{\lambda}}(\tau)-\tilde{\lambda}(\tau)\|_F^2 \\
&\le \frac{2\,r_\tau}{nT}\|\Pi(\tau) - \hat{\Pi}(\tau)\|_F^2 + \frac{2\,\sigma_1^2}{nT}\|\tilde{\hat{\lambda}}(\tau)-\tilde{\lambda}(\tau)\|_F^2 \\
&= O_{\mathbb{P}}\left(\frac{r_\tau\phi_n^2 c_T(1+s_\tau)\max\{\log(pc_T\vee n), r_\tau\}}{\kappa_0^4\,\underline{f}}\left(\frac{1}{n}+\frac{1}{d_T}\right)\right) + \\
&\quad O_{\mathbb{P}}\left(\frac{\sigma_1^2}{nT}\frac{(\sigma_1(\tau)+\sqrt{r_\tau}\,\mathrm{Err})^2\mathrm{Err}^2}{((\sigma_{r_\tau-1}(\tau))^2-(\sigma_{r_\tau}(\tau))^2)^2}\right),
\end{aligned}$$

where the third inequality follows from Weyl's inequality, and the last one from (68).

Table 7: Firm Characteristics Construction.

Characteristics	Name	Construction
acc	Working capital accruals	Annual income before extraordinary items (ib) minus operating cash flows (oancf) divided by average total assets (at)
agr	Asset growth	Annual percent change in total assets (at)
beta	Beta	Estimated market beta from weekly returns and equal weighted market returns for 3 years
bm	Book-to-market	Book value of equity (ceq) divided by end of fiscal year-end market capitalization
chinv	Change in inventory	Change in inventory (inv) scaled by average total assets (at)
chmom	Change in 6-month momentum	Cumulative returns from months t-6 to t-1 minus months t-12 to t-7
dolvol	Dollar trading volume	Natural log of trading volume times price per share from month t-2
dy	Dividend to price	Total dividends (dvt) divided by market capitalization at fiscal year-end
egr	Earnings announcement return	Annual percent change in book value of equity (ceq)
ep	Earnings to price	Annual income before extraordinary items (ib) divided by end of fiscal year market cap
gma	Gross profitability	Revenues (revt) minus cost of goods sold (cogs) divided by lagged total assets (at)
idiovol	Idiosyncratic return volatility	Standard deviation of residuals of weekly returns on weekly equal weighted market returns for 3 years prior to month end
ill	Illiquidity (Amihud)	Average of daily (absolute return / dollar volume).
invest	Capital expenditures and inventory	Annual change in gross property, plant, and equipment (ppegt) + annual change in inventories (invt) all scaled by lagged total assets (at)
lev	Leverage	Annual change in gross property, plant, and equipment (ppegt) + annual change in inventories (invt) all scaled by lagged total assets (at)
lgr	Growth in long-term debt	Annual percent change in total liabilities (lt)
mom1m	1-month momentum	1-month cumulative return

Continuation of Table 7		
mom6m	6-month momentum	5-month cumulative returns ending one month before month end
mve	Size	Natural log of market capitalization at end of month t-1
operprof	Operating profitability	Revenue minus cost of goods sold - SG&A expense - interest expense divided by lagged common shareholders' equity
range	Range of stock price	Monthly average of daily price range: (high-low)/((high+low)/2) (alternative measure of volatility)
retvol	Return volatility	Standard deviation of daily returns from month t-1
roaq	Return on assets	Income before extraordinary items (ibq) divided by one quarter lagged total assets (atq)
roeq	Return on equity	Earnings before extraordinary items divided by lagged common shareholders' equity
sue	Unexpected quarterly earnings	Unexpected quarterly earnings divided by fiscal-quarter-end market cap. Unexpected earnings is I/B/E/S actual earnings minus median forecasted earnings if available, else it is the seasonally differenced quarterly earnings before extraordinary items from Compustat quarterly file
turn	Share turnover	Average monthly trading volume for most recent 3 months scaled by number of shares outstanding in current month

Note: Estimated under different values of turning parameter ν_2, when $\nu_1 = 10^{-5}$ is fixed. The results are reported for quantiles 10%, 50% and 90%.