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

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

December 2019

Warwick Economics Research Papers
High Dimensional Latent Panel Quantile Regression
with an Application to Asset Pricing *

Alexandre Belloni1, Mingli Chen2, Oscar Hernan Madrid Padilla3, and Zixuan (Kevin) Wang4

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 $\ell_1$-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 $\ell_1$ and nuclear norm regularization. We show, under general conditions, that our estimator can consistently estimate both the nonzero coefficients of the covariates and the latent low-rank matrix.

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)):

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

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)):

\[ \mathbb{E}(Y_{i,t}) = \lambda'_i g_t. \]  

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.  

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

\[ F_{Y_{i,t} | X_{i,t}, \theta(\tau), \lambda_i(\tau), g_t(\tau)}^{-1}(\tau) = X'_{i,t} \theta(\tau) + \lambda_i(\tau)' g_t(\tau), \quad i = 1, \ldots, n, \quad t = 1, \ldots, T. \]  

\[ \text{More about sparse modeling and dense modeling can be found in Giannone et al. (2017). See also Chernozhukov et al. (2017).} \]
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, $\lambda_i(\tau)$, $g_t(\tau) \in \mathbb{R}^r$ 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 $r_\tau$, $\tau \in [0, 1]$ 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 $\tau$ 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 $\ell_1$ 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 ℓ1-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 ℓ0 norm as $\|v\|_0 = \sum_{j=1}^p 1\{v_j \neq 0\}$, where 1{⋯} takes value 1 if the statement inside {⋯} is true, and zero otherwise; its ℓ1 norm as $\|v\|_1 = \sum_{j=1}^p |v_j|$. We denote $\|v\|_{1,n,T} = \sum_{j=1}^p \hat{\sigma}_j |v_j|$ the ℓ1-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 $\|A\|_2 = \sup_{x: \|x\|=1} \sqrt{x' A' A x}$, its infinity norm by $\|A\|_\infty = \max\{|A_{i,j}| : i \in [n], j \in [T]\}$, its rank by rank$(A)$, and its nuclear norm by $\|A\|_* = \text{trace}(\sqrt{A' A})$ where $A'$ is the transpose of $A$. The $j$th column of $A$ is denoted by $A_{\cdot,j}$. Furthermore, the multiplication of a tensor $X \in \mathbb{R}^{I_1 \times \cdots \times I_m}$ with a vector $a \in \mathbb{R}^{I_m}$ is denoted by $Z := Xa \in \mathbb{R}^{I_1 \times \cdots \times I_{m-1}}$, and, explicitly, $Z_{i_1,\ldots,i_{m-1}} = \sum_{j=1}^{I_m} X_{i_1,\ldots,i_{m-1},j} a_j$. We also use the notation $a \lor b = \max\{a,b\}$, $a \land b = \min\{a,b\}$, and $(a)_- = \max\{-a,0\}$. 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 $n$. 

5
2 The Estimator and Overview of Rate Results

2.1 Basic Setting

The setting of interest corresponds to a high dimension latent panel quantile regression model, where \( Y \in \mathbb{R}^{n \times T} \), and \( X \in \mathbb{R}^{n \times T \times p} \) satisfying
\[
F_{Y_{i,t}|X_{i,t};\theta(\tau),\Pi_{i,t}(\tau)}^{-1}(\tau) = X_{i,t}'\theta(\tau) + \Pi_{i,t}(\tau), \quad i = 1, \ldots, n, \quad t = 1, \ldots, T, \tag{4}
\]
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 \( \tau \) as a linear combination of the predictors plus a low-rank matrix. Here, we allow for the number of covariates \( p \), and the time horizon \( T \), to grow to infinity as \( n \) grows. Throughout the paper the quantile index \( \tau \in (0, 1) \) is fixed. We mainly focus on the case where \( p \) is large, possibly much larger than \( nT \), but for the true model \( \theta(\tau) \) is sparse and has only \( s_\tau \ll p \) non-zero components. Mathematically, \( s_\tau := \|\theta(\tau)\|_0 \).

When \( \Pi_{i,t}(\tau) = \lambda_i(\tau)'g_t(\tau) \), with \( \lambda_i(\tau), g_t(\tau) \in \mathbb{R}^{r_\tau} \), this immediately leads to the following setting
\[
F_{Y_{i,t}|X_{i,t};\theta(\tau),\Pi_{i,t}(\tau)}^{-1}(\tau) = X_{i,t}'\theta(\tau) + \lambda_i(\tau)'g_t(\tau). \tag{5}
\]
where we model the quantile function at level \( \tau \) as a linear combination of the covariates (as predictors) plus a latent factor structure. This is directly related to the panel data models with interactive fixed effects literature in econometrics, e.g. linear panel data model (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
\[
\begin{align*}
\text{minimize} & \quad \frac{1}{nT} \sum_{t=1}^{T} \sum_{i=1}^{n} \rho_\tau(Y_{i,t} - X_{i,t}'\theta - \Pi_{i,t}) \\
\text{subject to} & \quad \text{rank}(\Pi) \leq r_\tau, \\
& \quad \|\theta\|_0 = s_\tau,
\end{align*} \tag{6}
\]

\(^2\)Different identification conditions might result in different estimation procedures for \( \lambda \) 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), \( s_\tau \) is a parameter that directly controls the sparsity of \( \hat{\theta} \), and \( r_\tau \) 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 \( \text{rank}(\cdot) \) 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

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

(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

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

(8) and

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

(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 \( \beta \)-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

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

(10)

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

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

(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 \(U_V, U_W, U_{\Pi},\) and \(U_\theta\). Next, recall that ADMM proceeds by iteratively minimizing the Augmented Lagrangian in blocks with respected to the original variables, in our case \((V, \tilde{\theta}, \tilde{\Pi})\) and \((W, Z_{\theta}, Z_{\Pi})\), 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 \(Z_{\theta}, \tilde{\Pi},\) and \(V\). For updating \(Z_{\theta}\) at iteration \(k + 1\), we solve the problem

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

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

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

(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

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

(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 \(g_{il}(\tau)\) via the singular value decomposition of \(\tilde{\Pi}(\tau)\) and following equation

\[
\tilde{\Pi}(\tau)_{i,t} = \hat{\lambda}_i(\tau)' \hat{g}_{il}(\tau),
\]

(14)

where \(\hat{\lambda}_i(\tau)\) and \(\hat{g}_{il}(\tau)\) 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

\[
\min_{\tilde{\Pi} \in \mathbb{R}^{n \times T}} \left\{ \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \rho_{r}(Y_{i,t} - \tilde{\Pi}_{i,t}) + \nu_2 \|\tilde{\Pi}\|_* \right\},
\]

(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

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

(16)

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

(17)

and

\[
U_{\Pi}^{(k+1)} \leftarrow \tilde{\Pi}^{(k+1)} - Z_{\Pi}^{(k+1)} + U_{\Pi}^{(k)},
\]

(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 $T_\tau$ be the support of $\theta(\tau)$, thus

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

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

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 $\Pi$, $\{(Y_{i,t}, X_{i,t})_{t=1,\ldots,T} : i \in [n]\}$ is independent across $i$. Also, for each $i \in [n]$, the sequence $\{(Y_{i,t}, X_{i,t})_{t=1,\ldots,T} : i \in [n]\}$ is stationary and $\beta$-mixing with mixing coefficients satisfying $\sup_i \gamma_i(k) = O(k^{-\mu})$ for some $\mu > 2$. Moreover, there exists $\mu' \in (0, \mu)$, such that

$$npT \left( \left[ T^{1/(1+\mu')} \right] \right)^{-\mu'} \to 0. \quad (19)$$

Here, $\gamma_i(k) = \frac{1}{2} \sup_{l \geq 1} \sum_{j=1}^{L} \sum_{j'=1}^{L'} |\mathbb{P}(A_j \cap B_{j'}) - \mathbb{P}(A_j)\mathbb{P}(B_{j'})|$, where $\{A_j\}_{j=1}^L$ is a partition of $\sigma(\{X_{i,1}, Y_{i,1}\}, \ldots, \{X_{i,t}, Y_{i,t}\}, \{X_{i,t+k}, Y_{i,t+k}\})$, and $\{B_{j'}\}_{j'=1}^{L'}$ is a partition of $\sigma(\{X_{i,t+k}, Y_{i,t+k}\})$.

(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 $f_{Y_{i,t}|X_{i,t}; \theta(\tau), \Pi_{i,t}(\tau)}$ 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 $\frac{\partial}{\partial \theta} f_{Y_{i,t}|X_{i,t}; \theta(\tau), \Pi_{i,t}(\tau)}(y|x; \theta(\tau), \Pi_{i,t}(\tau))$ are both bounded by $\overline{f}$ and $\overline{f}'$, 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 $\beta$-mixing.

It is worth mentioning that the parameter $\mu$ in Assumption 1 controls the strength of the time dependence in the data. In the case that $\{(Y_{i,t}, X_{i,t})_{t \in [n], t \in [T]} : i \in [n]\}$ 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

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

and we require that

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

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 \( \{\hat{\sigma}^2_j\}_{j=1}^p \), we analyze the performance of the estimator

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

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

\[
\hat{Q}_\tau(\tilde{\theta}, \tilde{\Pi}) = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \rho_\tau(Y_{i,t} - X_{i,t}^\prime \tilde{\theta} - \tilde{\Pi}_{i,t}),
\]

with \( \rho_\tau \) as defined in Section 2.2.

As it can be 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

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

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

Assumption 3. For \( (\delta, \Delta) \in A_\tau \), let

\[
J^{1/2}_\tau(\delta, \Delta) := \sqrt{\frac{f}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \mathbb{E}\left( \left( X_{i,t}^\prime \delta + \Delta_{i,t} \right)^2 \right)}.
\]
Then there exists $m \geq 0$ such that

\[ 0 < \kappa_m := \inf_{(\delta, \Delta) \in A_T, \delta \neq 0} \frac{J_T^{1/2}(\delta, \Delta)}{\| \delta \|_{A_T} + \frac{\| \Delta \|_F}{\sqrt{nT}}}, \tag{23} \]

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

\[ 0 < q := 3 \frac{3^{3/2} f}{\sqrt{\mathcal{I}}(\delta, \Delta) \in A_T, \delta \neq 0} \frac{\left( \mathbb{E} \left( \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T (X_{i,t}^2 + \Delta_{i,t}) \right) \right)^{3/2}}{\mathbb{E} \left( \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T (|X_{i,t}^2 + \Delta_{i,t}|) \right)}, \tag{24} \]

with $f$ and $\mathcal{I}$ 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

\[ J_T(\delta, \Delta) = f \mathbb{E}(X_{i,t}^2) + \frac{f}{nT} \| \Delta \|_F^2. \]

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

\[ q \geq C \frac{\phi_n \sqrt{c_T(1 + s_\tau)}}{\sqrt{nd_T \kappa_0 f^{1/2}}}, \tag{25} \]

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

\[ \| \hat{\theta}(\tau) - \theta(\tau) \| = O_p \left( \frac{\phi_n (1 + \sqrt{\frac{s_\tau}{m}})}{\kappa_m} \sqrt{c_T(1 + s_\tau)} \max\{\log(p c_T \vee n), r_x\} \left( \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{d_T}} \right) \right), \tag{26} \]

and

\[ \frac{1}{nT} \| \hat{\Pi}(\tau) - \Pi(\tau) \|_F^2 = O_p \left( \frac{\phi_n^2 c_T(1 + s_\tau)}{\kappa_0^2 f} \max\{\log(p c_T \vee n), r_x\} \left( \frac{1}{n} + \frac{1}{d_T} \right) \right), \tag{27} \]

for choices of the tuning parameters satisfying

\[ \nu_1 \asymp \sqrt{\frac{c_T \log(\max\{n, p c_T\})}{n d_T}} \left( \sqrt{n} + \sqrt{d_T} \right), \]

and

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

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

13
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 \(\max\{s_\tau, r_\tau\} \ll \min\{n, T\}\). 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(\(\sigma^2\)). The latter implies that
\[
\Pr(|e_{1,1}| > z) \leq C_1 \exp\left(-\frac{z^2}{2\sigma^2}\right),
\]
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 \(\Pi(\tau)\):
\[
\inf_{\Pi} \sup_{\Pi(\tau) : \text{rank}(\Pi(\tau)) \leq r_\tau} \mathbb{E} \left(\frac{\|\hat{\Pi}(\tau) - \Pi(\tau)\|_F^2}{nT}\right) \geq \frac{r_\tau \sigma^2}{n}.
\]
(28)
Notably, the lower bound in (28) matches the rate implied by Theorem 1, ignoring other factors depending on \(s_\tau, \kappa_0, \kappa_m, p\) and \(\phi_m\). 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 \(\lambda(\tau), \hat{\lambda}(\tau), \hat{\lambda}(\tau) \in \mathbb{R}^{n \times r_\tau}\) be matrices with orthonormal columns satisfying
\[
\Pi(\tau) = \sum_{j=1}^{r_\tau} \sigma_j \hat{\lambda}_j(\tau)g_{.,j}(\tau)' = \sum_{j=1}^{r_\tau} \lambda_{.,j}(\tau)g_{.,j}(\tau)',
\]
and \(\hat{\Pi}(\tau)\hat{g}_{.,j}(\tau) = \hat{\sigma}_j(\tau)\hat{\lambda}_j(\tau) = \hat{\lambda}_j(\tau)\) for \(j = 1, \ldots, r_\tau\). Then
\[
v_1 := \min_{\mathcal{O} \in \mathcal{O}_{r_\tau}} \|\hat{g}(\tau)O - g(\tau)\|_F = O_P\left(\frac{(\sigma_1(\tau) + \sqrt{r_\tau} Err)Err}{(\sigma_{r_\tau - 1}(\tau))^2 - (\sigma_{r_\tau}(\tau))^2}\right),
\]
(29)
\[ v_2 = \frac{\| \hat{\lambda}(\tau) - \lambda(\tau) \|^2_2}{nT} = \mathbb{O}_P \left( \frac{\phi_0 \sqrt{c_T(1 + s_T) \max \{ \log(p c_T \vee n), r_T \} \left( \frac{1}{n} + \frac{1}{d_T} \right)}}{\kappa_0^2 f^{1/2}} \right) + \]

\[ \frac{\sigma^2}{nT} \frac{(\sigma_1(\tau) + \sqrt{r_T} \text{Err})^2 \text{Err}^2}{(\sigma_{r_T - 1}(\tau))^2 - (\sigma_{r_T}(\tau))^2}. \]

(30)

Here, \( \mathbb{O}_r \) is the group of \( r \times r \) orthonormal matrices, and

\[ \text{Err} := \phi_0 c_T \sqrt{(1 + s_T) \max \{ \log(p c_T \vee n), r_T \}} \left( \sqrt{n} + \sqrt{d_T} \right). \]

A particularly interesting instance of Corollary 2 is when

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

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

\[ v_1 = \mathbb{O}_P \left( \frac{\phi_0 \sqrt{c_T(1 + s_T) \max \{ \log(p c_T \vee n), r_T \} \left( \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{d_T}} \right)}}{\kappa_0 f^{1/2}} \right), \]

whereas (30) is now

\[ v_2 = \mathbb{O}_P \left( \frac{r_T \phi_0^2 c_T(1 + s_T) \max \{ \log(p c_T \vee n), r_T \} \left( \frac{1}{n} + \frac{1}{d_T} \right)}{\kappa_0^4 f} \right). \]

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 \( \text{rank}(X\theta(\tau) + \xi) = O(r_T) \), and

\[ \frac{\| \xi \|_*}{\sqrt{nT}} = \mathbb{O}_P \left( \frac{c_T \phi_n \sqrt{r_T(\sqrt{n} + \sqrt{d_T})}}{\sqrt{nT} f} \right), \]

with \( c_T \) as defined in Theorem 1. Furthermore, \( \| X\theta(\tau) + \Pi(\tau) \| = \mathbb{O}_P(1) \).
Notice that in Assumption 4, $\xi$ is an approximation error. In the case $\xi = 0$, the condition implies that\[ \text{rank}(X\theta(\tau)) = O(r_\tau) \]with probability close to one.

Next, exploiting Assumption 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 $\|\hat{\Pi}\|_\infty \leq C$, for a large enough positive constant $C$.

Then\[ \frac{1}{nT}\|\hat{\Pi}(\tau) - \Pi(\tau)\|_F^2 = O_P\left(\left(\frac{T}{f}\right)^2 \phi_n^2 \tau \left(\frac{1}{n} + \frac{1}{d_T}\right)\right), \]
where $\{\phi_n\}$ is a sequence with $\phi_n/(\sqrt{T} \log(1 + c_T)) \to \infty$, and for choices

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

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 $\nu_1$ 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 $f$, $f'$, and $\phi_n$. When the data are independent, the rate becomes of the order $r_\tau (n^{-1} + T^{-1})$. In such framework, our result matches the minimax rate of estimation in Candès and Plan (2010) for estimating an $n \times T$ matrix of rank $r_\tau$, 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 $\hat{\Pi}$. 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 ($\ell_1$-NN-QR) with extensive numerical simulations focusing on the median case, namely the case when $\tau = 0.5$. As benchmarks, we consider the $\ell_1$-penalized quantile regression studied in Belloni and Chernozhukov (2009), and similarly we refer to
this procedure as $\ell_1$-QR. We also compare with the mean case, which we denote it as $\ell_1$-NN-LS as it combines the $\ell_2$-loss function with $\ell_1$ and nuclear norm regularization. We consider different generative scenarios. For each scenario we randomly generate 100 different data sets and compute the estimates of the methods for a grid of values of $\nu_1$ and $\nu_2$. Specifically, these tuning parameters are taken to satisfy $\nu_1 \in \{10^{-4}, 10^{-4.5}, \ldots, 10^{-8}\}$ and $\nu_2 \in \{10^{-3}, 10^{-4}, \ldots, 10^{-9}\}$. Given any choice of tuning parameters, we evaluate the performance of each competing method, averaging over the 100 data sets, and report values that correspond to the best performance. These are referred as optimal tuning parameters and can be thought of as oracle choices.

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

$$
\text{BIC}(\nu_1, \nu_2) = \sum_{i=1}^{n} \sum_{t=1}^{T} \rho_\tau(Y_{i,t} - X_{i,t}'\hat{\theta}(\nu_1, \nu_2) - \hat{\Pi}(\nu_1, \nu_2)) + \frac{\log(nT)}{2} (c_1 \cdot \hat{s}_\tau(\nu_1, \nu_2) + (1 + n + T) \cdot \hat{r}_\tau(\nu_1, \nu_2)),
$$

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 $\nu_1$ and $\nu_2$ 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 $\nu_1$ and $\nu_2$, and report the average over the 100 Monte Carlo simulations. We refer to this as BIC-$\ell_1$-NN-QR.

As performance measure we use a scaled version (see Tables 1-2) of the squared distance between the true vector of coefficients $\theta$ and the corresponding estimate. We also consider a different metric, the “Quantile error” (Koenker and Machado (1999)):

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

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

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

17
Design 1. (Location shift model) The data is generated from the model

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

(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

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

(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 \( \theta \) and \( \Pi \) 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 \( \Pi \) as

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

(35)

where

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

(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 \( \Pi \) 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-\( \ell_1 \)-NN-QR, performs very closely to the ideally tuned one \( \ell_1 \)-NN-QR. In the more challenging settings of Designs 3 and 4, we noticed that BIC-\( \ell_1 \)-NN-QR performs reasonably well compared to \( \ell_1 \)-NN-QR.
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 $\ell_2$ 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>$n$</th>
<th>$p$</th>
<th>$T$</th>
<th>Design 1</th>
<th>Design 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ell_1$-NN-QR</td>
<td>300</td>
<td>30</td>
<td>300</td>
<td>0.86 0.03</td>
<td>0.69 0.03</td>
</tr>
<tr>
<td>BIC-$\ell_1$-NN-QR</td>
<td>300</td>
<td>30</td>
<td>300</td>
<td>0.86 0.03</td>
<td>0.89 0.03</td>
</tr>
<tr>
<td>$\ell_1$-NN-LS</td>
<td>300</td>
<td>30</td>
<td>300</td>
<td>2.58 0.05</td>
<td>1.03 0.04</td>
</tr>
<tr>
<td>$\ell_1$-QR</td>
<td>300</td>
<td>30</td>
<td>300</td>
<td>8.18 4.19</td>
<td>6.81 4.18</td>
</tr>
<tr>
<td>$\ell_1$-NN-QR</td>
<td>300</td>
<td>30</td>
<td>100</td>
<td>2.99 0.04</td>
<td>2.39 0.04</td>
</tr>
<tr>
<td>BIC-$\ell_1$-NN-QR</td>
<td>300</td>
<td>30</td>
<td>100</td>
<td>2.99 0.04</td>
<td>2.39 0.04</td>
</tr>
<tr>
<td>$\ell_1$-NN-LS</td>
<td>300</td>
<td>30</td>
<td>100</td>
<td>8.06 0.12</td>
<td>3.11 0.08</td>
</tr>
<tr>
<td>$\ell_1$-QR</td>
<td>300</td>
<td>30</td>
<td>100</td>
<td>41.0 4.19</td>
<td>26.0 4.19</td>
</tr>
<tr>
<td>$\ell_1$-NN-QR</td>
<td>300</td>
<td>5</td>
<td>300</td>
<td>0.22 0.003</td>
<td>0.39 0.03</td>
</tr>
<tr>
<td>BIC-$\ell_1$-NN-QR</td>
<td>300</td>
<td>5</td>
<td>300</td>
<td>0.22 0.003</td>
<td>0.48 0.03</td>
</tr>
<tr>
<td>$\ell_1$-NN-LS</td>
<td>300</td>
<td>5</td>
<td>300</td>
<td>0.37 0.01</td>
<td>0.69 0.03</td>
</tr>
<tr>
<td>$\ell_1$-QR</td>
<td>300</td>
<td>5</td>
<td>300</td>
<td>2.6 4.19</td>
<td>3.27 4.19</td>
</tr>
<tr>
<td>$\ell_1$-NN-QR</td>
<td>300</td>
<td>5</td>
<td>100</td>
<td>0.50 0.008</td>
<td>0.80 0.03</td>
</tr>
<tr>
<td>BIC-$\ell_1$-NN-QR</td>
<td>300</td>
<td>5</td>
<td>100</td>
<td>0.53 0.009</td>
<td>0.97 0.03</td>
</tr>
<tr>
<td>$\ell_1$-NN-LS</td>
<td>300</td>
<td>5</td>
<td>100</td>
<td>1.12 0.02</td>
<td>1.46 0.04</td>
</tr>
<tr>
<td>$\ell_1$-QR</td>
<td>300</td>
<td>5</td>
<td>100</td>
<td>7.87 4.19</td>
<td>8.56 4.19</td>
</tr>
<tr>
<td>$\ell_1$-NN-QR</td>
<td>100</td>
<td>30</td>
<td>300</td>
<td>2.97 0.04</td>
<td>2.26 0.04</td>
</tr>
<tr>
<td>BIC-$\ell_1$-NN-QR</td>
<td>100</td>
<td>30</td>
<td>300</td>
<td>2.97 0.04</td>
<td>2.81 0.04</td>
</tr>
<tr>
<td>$\ell_1$-NN-LS</td>
<td>100</td>
<td>30</td>
<td>300</td>
<td>9.77 0.12</td>
<td>3.39 0.06</td>
</tr>
<tr>
<td>$\ell_1$-QR</td>
<td>100</td>
<td>30</td>
<td>300</td>
<td>40.0 4.23</td>
<td>24.0 4.23</td>
</tr>
<tr>
<td>$\ell_1$-NN-QR</td>
<td>100</td>
<td>30</td>
<td>100</td>
<td>2.3 0.04</td>
<td>10.0 0.03</td>
</tr>
<tr>
<td>BIC-$\ell_1$-NN-QR</td>
<td>100</td>
<td>30</td>
<td>100</td>
<td>2.3 0.06</td>
<td>11.0 0.04</td>
</tr>
<tr>
<td>$\ell_1$-NN-LS</td>
<td>100</td>
<td>30</td>
<td>100</td>
<td>8.4 0.79</td>
<td>13.0 0.16</td>
</tr>
<tr>
<td>$\ell_1$-QR</td>
<td>100</td>
<td>30</td>
<td>100</td>
<td>229.0 4.23</td>
<td>177.0 4.23</td>
</tr>
<tr>
<td>$\ell_1$-NN-QR</td>
<td>100</td>
<td>5</td>
<td>300</td>
<td>0.64 0.008</td>
<td>0.89 0.03</td>
</tr>
<tr>
<td>BIC-$\ell_1$-NN-QR</td>
<td>100</td>
<td>5</td>
<td>300</td>
<td>0.65 0.008</td>
<td>1.32 0.03</td>
</tr>
<tr>
<td>$\ell_1$-NN-LS</td>
<td>100</td>
<td>5</td>
<td>300</td>
<td>1.25 0.02</td>
<td>1.67 0.05</td>
</tr>
<tr>
<td>$\ell_1$-QR</td>
<td>100</td>
<td>5</td>
<td>300</td>
<td>8.79 4.23</td>
<td>8.61 4.23</td>
</tr>
<tr>
<td>$\ell_1$-NN-QR</td>
<td>100</td>
<td>5</td>
<td>100</td>
<td>1.82 0.009</td>
<td>3.30 0.03</td>
</tr>
<tr>
<td>BIC-$\ell_1$-NN-QR</td>
<td>100</td>
<td>5</td>
<td>100</td>
<td>1.85 0.01</td>
<td>3.74 0.04</td>
</tr>
<tr>
<td>$\ell_1$-NN-LS</td>
<td>100</td>
<td>5</td>
<td>100</td>
<td>4.06 0.03</td>
<td>4.45 0.14</td>
</tr>
<tr>
<td>$\ell_1$-QR</td>
<td>100</td>
<td>5</td>
<td>100</td>
<td>32.0 4.23</td>
<td>27.0 4.23</td>
</tr>
</tbody>
</table>
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 \(\ell_2\) 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>(n)</th>
<th>(p)</th>
<th>(T)</th>
<th>Design 3 (\frac{|\hat{\theta}(\tau) - \theta(\tau)|^2}{10^{-4}})</th>
<th>Quantile error</th>
<th>Design 4 (\frac{|\hat{\theta}(\tau) - \theta(\tau)|^2}{10^{-4}})</th>
<th>Quantile error</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\ell_1)-NN-QR</td>
<td>300</td>
<td>30</td>
<td>300</td>
<td>2.17 0.17</td>
<td>1.58 0.11</td>
<td>4.61 0.13</td>
<td>0.13</td>
</tr>
<tr>
<td>BIC-(\ell_1)-NN-QR</td>
<td>300</td>
<td>30</td>
<td>300</td>
<td>2.17 0.29</td>
<td>2.81 0.26</td>
<td>4.61 0.16</td>
<td>0.16</td>
</tr>
<tr>
<td>(\ell_1)-NN-LS</td>
<td>300</td>
<td>30</td>
<td>300</td>
<td>3.33 0.19</td>
<td>3.54 0.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\ell_1)-QR</td>
<td>300</td>
<td>30</td>
<td>300</td>
<td>6.59 1.09</td>
<td>12.0 1.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\ell_1)-NN-QR</td>
<td>300</td>
<td>30</td>
<td>100</td>
<td>10.0 0.18</td>
<td>4.61 0.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIC-(\ell_1)-NN-QR</td>
<td>300</td>
<td>30</td>
<td>100</td>
<td>11.0 0.26</td>
<td>9.21 0.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\ell_1)-NN-LS</td>
<td>300</td>
<td>30</td>
<td>100</td>
<td>11.0 0.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\ell_1)-QR</td>
<td>300</td>
<td>30</td>
<td>100</td>
<td>27.0 11.10</td>
<td>47.2 1.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\ell_1)-NN-QR</td>
<td>300</td>
<td>5</td>
<td>300</td>
<td>1.13 0.17</td>
<td>0.27 0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIC-(\ell_1)-NN-QR</td>
<td>300</td>
<td>5</td>
<td>300</td>
<td>1.56 0.33</td>
<td>0.74 0.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\ell_1)-NN-LS</td>
<td>300</td>
<td>5</td>
<td>300</td>
<td>1.58 0.19</td>
<td>0.49 0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\ell_1)-QR</td>
<td>300</td>
<td>5</td>
<td>300</td>
<td>2.47 1.10</td>
<td>5.52 1.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\ell_1)-NN-QR</td>
<td>300</td>
<td>5</td>
<td>100</td>
<td>3.04 0.19</td>
<td>0.69 0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIC-(\ell_1)-NN-QR</td>
<td>300</td>
<td>5</td>
<td>100</td>
<td>4.37 0.27</td>
<td>1.11 0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\ell_1)-NN-LS</td>
<td>300</td>
<td>5</td>
<td>100</td>
<td>4.43 0.27</td>
<td>1.12 0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\ell_1)-QR</td>
<td>300</td>
<td>5</td>
<td>100</td>
<td>7.65 1.11</td>
<td>13.4 1.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\ell_1)-NN-QR</td>
<td>100</td>
<td>30</td>
<td>300</td>
<td>11.0 0.18</td>
<td>7.06 0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIC-(\ell_1)-NN-QR</td>
<td>100</td>
<td>30</td>
<td>300</td>
<td>12.0 0.27</td>
<td>7.29 0.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\ell_1)-NN-LS</td>
<td>100</td>
<td>30</td>
<td>300</td>
<td>12.0 0.34</td>
<td>11.2 0.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\ell_1)-QR</td>
<td>100</td>
<td>30</td>
<td>300</td>
<td>26.0 1.10</td>
<td>51.0 1.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\ell_1)-NN-QR</td>
<td>100</td>
<td>30</td>
<td>100</td>
<td>6.12 0.17</td>
<td>32.1 0.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIC-(\ell_1)-NN-QR</td>
<td>100</td>
<td>30</td>
<td>100</td>
<td>6.64 0.22</td>
<td>35.4 0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\ell_1)-NN-LS</td>
<td>100</td>
<td>30</td>
<td>100</td>
<td>8.63 1.08</td>
<td>82.0 0.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\ell_1)-QR</td>
<td>100</td>
<td>30</td>
<td>100</td>
<td>16.5 1.12</td>
<td>267.6 1.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\ell_1)-NN-QR</td>
<td>100</td>
<td>5</td>
<td>300</td>
<td>2.99 0.18</td>
<td>0.86 0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIC-(\ell_1)-NN-QR</td>
<td>100</td>
<td>5</td>
<td>300</td>
<td>4.19 0.26</td>
<td>1.43 0.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\ell_1)-NN-LS</td>
<td>100</td>
<td>5</td>
<td>300</td>
<td>5.27 0.33</td>
<td>1.45 0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\ell_1)-QR</td>
<td>100</td>
<td>5</td>
<td>300</td>
<td>8.59 1.10</td>
<td>21.0 1.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\ell_1)-NN-QR</td>
<td>100</td>
<td>5</td>
<td>100</td>
<td>12.3 0.16</td>
<td>2.15 0.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIC-(\ell_1)-NN-QR</td>
<td>100</td>
<td>5</td>
<td>100</td>
<td>13.1 0.20</td>
<td>2.43 0.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\ell_1)-NN-LS</td>
<td>100</td>
<td>5</td>
<td>100</td>
<td>15.0 1.09</td>
<td>3.61 0.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\ell_1)-QR</td>
<td>100</td>
<td>5</td>
<td>100</td>
<td>24.7 1.10</td>
<td>45.7 1.09</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
6 Empirical Performance of the “Characteristics + Latent Factor” Model in Asset Pricing

Data Description

We use data from CRSP and Compustat to construct 24 firm level characteristics that are documented to explain the cross section and time series of stock returns in the finance and accounting literature. The characteristics we choose include well-known drivers of stock returns such as beta, size, book-to-market, momentum, volatility, liquidity, investment and profitability. Table 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 each asset can potentially be explained by $p$ observed asset characteristics (sparse part) and $r$ latent factors (dense part). The asset characteristics are the covariates in our model. Our model imposes a sparse structure on the $p$ characteristics so that only the characteristics having the strongest explanatory powers are selected by the model. The part that’s unexplained by the firm characteristics are captured by latent factors.

Suppose we have $n$ stock returns ($R_1, \ldots, R_n$), and $p$ observed firm characteristics ($X_1, \ldots, X_p$) over $T$ periods. The return quantile at level $\tau$ of portfolio $i$ in time $t$ is assumed to be the following:

$$F_{R_{i,t}|X_{i,t-1};\theta(\tau),\lambda_i(\tau),g_t(\tau)}^{-1}(\tau) = X_{i,t-1,1}\theta_1(\tau) + \ldots + X_{i,t-1,k}\theta_k(\tau) + \ldots + X_{i,t-1,p}\theta_p(\tau) + \lambda_i(\tau) g_t(\tau),$$

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 $\theta_k$ captures the extent to which assets with higher/lower characteristic $X_{i,t,k}$ delivers higher average return. The term $g_t$ contains the $r$ latent factors in period $t$ which captures systematic risks in the market, and $\lambda_i$ contains portfolio $i$’s loading on these factors (i.e. exposure to risk).

There is a discussion in academic research on “factor versus characteristics” in late 1990s and early 2000s. The factor/risk based view argues that an asset has higher expected returns because of its exposure to risk factors (e.g. Fama-French 3 factors) which represent some unobserved systematic risk. An asset’s exposure to risk factors are measured by factor loadings. The characteristics view claims that stocks have higher expected returns simply because they have certain characteristics (e.g. higher book-to-market ratios, smaller market capitalization), which might be independent of systematic risk (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 $\nu_1$ and $\nu_2$, 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 $\nu_1$ governs the sparsity of the coefficient vector $\theta$. The larger $\nu_1$ is, the larger the shrinkage effect on $\theta$. Figure 2 illustrate the effect of this shrinkage. With $\nu_2$ fixed, as the value of $\nu_1$ increases, more coefficients in the estimated $\theta$ vector shrink to zero. From a statistical point of view, the “effective characteristics” that can explain stock returns are those with non-zero coefficient $\theta$ at relatively large values of $\nu_1$.

The parameter $\nu_2$ is fixed at $\log_{10}(\nu_2) = -4$. Figure 2: Estimated Coefficients as a Function of $\nu_1$
The figure plots the estimated coefficient $\theta$ when the tuning parameter $\nu_1$ changes, for $\tau = \{0.1, 0.5, 0.9\}$. The parameter $\nu_2$ is fixed at $\log_{10}(\nu_2) = -4$.

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

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.

<table>
<thead>
<tr>
<th>log₁₀(ν₂)</th>
<th>τ = 0.1</th>
<th>τ = 0.5</th>
<th>τ = 0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>-6.0</td>
<td>228</td>
<td>228</td>
<td>228</td>
</tr>
<tr>
<td>-5.5</td>
<td>228</td>
<td>228</td>
<td>228</td>
</tr>
<tr>
<td>-5.0</td>
<td>228</td>
<td>228</td>
<td>228</td>
</tr>
<tr>
<td>-4.5</td>
<td>164</td>
<td>228</td>
<td>168</td>
</tr>
<tr>
<td>-4.0</td>
<td>1</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>-3.5</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>-3.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Note: Estimated under different values of turning parameter ν₂, when ν₁ = 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 ν₁ and ν₂ for different quantiles. The tuning parameters are selected via BIC based on (31) as discussed in Section 5. For every ν₁ and ν₂, we get the estimates \( \theta(ν₁, ν₂) \) and \( \Pi(ν₁, ν₂) \) and the number of factors \( r = \text{rank}(\Pi(ν₁, ν₂)) \). The \( \theta \) vector is sparse with non-zero coefficients on selected characteristics. The 10% quantile of returns has only 1 latent factor, and 3 selected characteristics. The median of returns has 7 latent factors and 2 selected characteristics. The 90% quantile of returns has 2 latent factors and 7 selected characteristics. Range is the only characteristic selected across all 3 quantiles. Idiosyncratic volatility is selected at 10% and 90% quantiles, with opposite signs. 1-month momentum is selected at 50% and 90% percentiles, with negative sign suggesting reversal in returns.

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.

<table>
<thead>
<tr>
<th></th>
<th>Fixed $\nu_1$ and $\nu_2$</th>
<th></th>
<th></th>
<th></th>
<th>Optimal $\nu_1$ and $\nu_2$ (BIC)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\tau = 0.1$</td>
<td>$\tau = 0.5$</td>
<td>$\tau = 0.9$</td>
<td>$\tau = 0.1$</td>
<td>$\tau = 0.5$</td>
<td>$\tau = 0.9$</td>
</tr>
<tr>
<td>acc</td>
<td>-0.089</td>
<td>-0.074</td>
<td>-0.041</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>range</td>
<td>-2.574</td>
<td>-0.481</td>
<td>2.526</td>
<td>-2.372</td>
<td>-0.356</td>
<td>2.429</td>
</tr>
<tr>
<td>beta</td>
<td>0.174</td>
<td>-0.116</td>
<td>-0.406</td>
<td>0</td>
<td>0</td>
<td>-0.115</td>
</tr>
<tr>
<td>bm</td>
<td>0.371</td>
<td>0.175</td>
<td>0.263</td>
<td>0</td>
<td>0</td>
<td>0.168</td>
</tr>
<tr>
<td>chinv</td>
<td>0</td>
<td>0</td>
<td>-0.152</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>dy</td>
<td>-0.086</td>
<td>0</td>
<td>0.119</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>egr</td>
<td>-0.106</td>
<td>-0.053</td>
<td>-0.091</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ep</td>
<td>0.199</td>
<td>0.057</td>
<td>-0.479</td>
<td>0</td>
<td>0</td>
<td>-0.391</td>
</tr>
<tr>
<td>gma</td>
<td>0</td>
<td>0.091</td>
<td>0.201</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>idiovol</td>
<td>-1.229</td>
<td>-0.071</td>
<td>1.438</td>
<td>-1.055</td>
<td>0</td>
<td>1.286</td>
</tr>
<tr>
<td>ill</td>
<td>-0.334</td>
<td>-0.218</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>invest</td>
<td>-0.097</td>
<td>0</td>
<td>0.146</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>lev</td>
<td>0.183</td>
<td>0.063</td>
<td>0.129</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>lgr</td>
<td>-0.106</td>
<td>-0.037</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mom12m</td>
<td>-0.166</td>
<td>-0.077</td>
<td>-0.117</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mom1m</td>
<td>-0.150</td>
<td>-0.384</td>
<td>-0.571</td>
<td>0</td>
<td>-0.286</td>
<td>-0.477</td>
</tr>
<tr>
<td>mve</td>
<td>-0.038</td>
<td>-0.093</td>
<td>-0.811</td>
<td>0</td>
<td>0</td>
<td>-0.667</td>
</tr>
<tr>
<td>operprof</td>
<td>0</td>
<td>0.025</td>
<td>0.088</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>roaq</td>
<td>0.221</td>
<td>0.242</td>
<td>-0.147</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>roeq</td>
<td>0.073</td>
<td>0.041</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>std_dolvol</td>
<td>0</td>
<td>0</td>
<td>-0.039</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>std_turn</td>
<td>0.310</td>
<td>0</td>
<td>-0.247</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>sue</td>
<td>0.105</td>
<td>0.061</td>
<td>0.045</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>turn</td>
<td>-0.796</td>
<td>-0.083</td>
<td>0.386</td>
<td>-0.330</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>$\tau = 0.1$</th>
<th>$\tau = 0.5$</th>
<th>$\tau = 0.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>optimal $r$</td>
<td>1</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>optimal $\nu_1$</td>
<td>$10^{-2.5}$</td>
<td>$10^{-2.5}$</td>
<td>$10^{-2.75}$</td>
</tr>
<tr>
<td>optimal $\nu_2$</td>
<td>$10^{-4}$</td>
<td>$10^{-4}$</td>
<td>$10^{-4}$</td>
</tr>
</tbody>
</table>

Note: This table reports the selected optimal tuning parameter $\nu_1$ and $\nu_2$ that minimize the objective function in equation (31) for different quantiles.

**Interpretation of Latent Factors**

Table 6 below reports the variance in the matrix $\Pi$ 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 $\Pi$, with second PC explaining 13.8% and third PC explaining 6.8%.

Table 6: Percentage of $\Pi$ explained by PC

<table>
<thead>
<tr>
<th></th>
<th>$\tau = 0.1$</th>
<th>$\tau = 0.5$</th>
<th>$\tau = 0.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC1</td>
<td>100.00%</td>
<td>73.82%</td>
<td>99.68%</td>
</tr>
<tr>
<td>PC2</td>
<td>13.71%</td>
<td>0.32%</td>
<td>0.32%</td>
</tr>
<tr>
<td>PC3</td>
<td>6.78%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PC4</td>
<td>4.12%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PC5</td>
<td>1.11%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PC6</td>
<td>0.45%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PC7</td>
<td>0.01%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

Note: Variance of matrix $\Pi$ 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 $\Pi$ against S&P500 index monthly return for quantiles 10% (left), 50% (middle), and 90% (right).

References


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


Maryam Fazel. Matrix rank minimization with applications. 2002.


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


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


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

$$V^{(k+1)} \leftarrow P_+ \left( W^{(k)} - U^{(k)}_V - \frac{\tau}{nT\eta} \mathbf{11}' + P_- \left( W^{(k)} - U^{(k)}_V - \frac{\tau}{nT\eta} \mathbf{11}' \right) \right)$$

$$\tilde{\theta}^{(k+1)} \leftarrow \arg\min_{\theta} \left\{ \frac{\eta}{2} n \sum_{i=1}^{T} \left( W_{i,t}^{(k)} - Y_{i,t} + X_{i,t}^{(k)} + (Z_{\Pi}^{(k)})_{i,t} + (U_{W}^{(k)})_{i,t} \right)^2 + \frac{\eta}{2} \|Z_{\theta}^{(k)} - \theta + \nu^{(k)} \|_2^2 \right\}$$

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

$$Z_{\theta}^{(k+1)} \leftarrow \arg\min_{Z_{\theta}} \left\{ \frac{1}{2} \|\tilde{\theta}^{(k+1)} - U_{\theta}^{(k)} - Z_{\theta} \|_F^2 + \frac{\nu_1}{\eta} \sum_{j=1}^{p} \sum_{t=1}^{T} \|Z_{\theta})_{j}\|_F \right\}$$

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

$$(42)$$

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

$$(43)$$

$$U_{V}^{(k+1)} \leftarrow V^{(k+1)} - W^{(k+1)} + U_{V}^{(k)} ,$$

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

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

$$U_{\theta}^{(k+1)} \leftarrow Z_{\theta}^{(k+1)} - \tilde{\theta}^{(k+1)} + U_{\theta}^{(k)} ,$$

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

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

$$\tilde{\theta}^{(k+1)} \leftarrow \left[ \sum_{i=1}^{n} \sum_{t=1}^{T} X_{i,t} X_{i,t}^T + I_p \right]^{-1} \left[ - \sum_{i=1}^{n} \sum_{t=1}^{T} X_{i,t} A_{i,t} + Z_{\theta}^{(k)} + U_{\theta}^{(k)} \right] ,$$

where

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

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

$$\tilde{\Pi}^{(k+1)} \leftarrow P \text{ diag} \left( \max \left\{ 0, \nu_j - \frac{\nu_2}{\eta} \right\} \right) Q' ,$$

$$32$$
where
\[ Z^{(k)}_{\Pi} + U^{(k)}_{\Pi} = P \, \text{diag}(\{v_j\}_{1 \leq j \leq l}) Q' \]

Furthermore, for \( Z_{\theta} \),
\[ Z^{(k+1)}_{\theta,j} \leftarrow \text{sign}(\tilde{\theta}^{(k+1)}_j - U^{(k)}_{\theta,j}) \left[ |\tilde{\theta}^{(k+1)}_j - U^{(k)}_{\theta,j}| - \frac{\nu_1 w_j}{\eta} \right] . \]

Finally, defining
\[ \tilde{A} = -Y + X\tilde{\theta}^{(k+1)} + U^{(k)}_W, \quad \tilde{B} = -V^{(k+1)} - U^{(k)}_V, \quad \tilde{C} = -\tilde{\Pi}^{(k+1)} + U^{(k)}_{\Pi}, \]
the remaining updates are
\[ Z^{(k+1)}_{\Pi} \leftarrow -\tilde{A} - 2\tilde{C} + \tilde{B}, \]
and
\[ W^{(k+1)} \leftarrow -\tilde{A} - \tilde{C} - 2Z^{(k+1)}_{\Pi}. \]

\section*{B Proofs of the Main Results in the Paper}

\subsection*{B.1 Auxiliary lemmas for proof of Theorem 1}

Throughout, we use the notation
\[ Q_{\tau}(\tilde{\theta}, \tilde{\Pi}) = \mathbb{E} (\tilde{Q}_{\tau}(\tilde{\theta}, \tilde{\Pi})). \]

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

- \( \{(Y_{i,t}, X_{i,t})\}_{i \in [n], t \in [T]} \) is independent of \( \{(Y_{i,t}, X_{i,t})\}_{i \in [n], t \in [T]} \);
- for a fixed \( t \) the random vectors \( \{(Y_{i,t}, X_{i,t})\}_{i \in [n]} \) are independent;
- for a fixed \( i \):
  \[ \mathcal{L}(\{(Y_{i,t}, X_{i,t})\}_{t \in H_l}) = \mathcal{L}(\{(Y_{i,t}, X_{i,t})\}_{t \in H_1}) = \mathcal{L}(\{(Y_{i,t}, X_{i,t})\}_{t \in H_1}) \quad \forall l \in [dT], \]
  and the blocks \( \{(Y_{i,t}, X_{i,t})\}_{t \in H_1}, \ldots, \{(Y_{i,t}, X_{i,t})\}_{t \in H_{dT}} \) are independent.

Here, we define
\[ \Lambda := \{H_1, H'_1, \ldots, H_{dT}, H'_{dT}, R\} \]
with
\[ H_j = \{t : 1 + 2(j - 1)c_T \leq t \leq (2j - 1)c_T\}, \]
\[ H'_j = \{t : 1 + (2j - 1)c_T \leq t \leq 2jc_T\}, \quad j = 1, \ldots, d_T, \]
and \( R = \{t : 2c_Td_T + 1 \leq t \leq T\}. \)
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 $\tilde{a}_{i,t} = \tau - 1\{\tilde{Y}_{i,t} \leq \tilde{X}_{i,t}'\theta(\tau) + \Pi_{i,t}(\tau)\}$.

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

$$\mathbb{P} \left( \max_{j=1,\ldots,p} \frac{1}{n^T} \sum_{i=1}^{n} \sum_{t=1}^{T} \frac{X_{i,t,j}a_{i,t}}{\sigma_j} \geq \eta \frac{|X|}{9} \right) \leq \frac{16}{n} + 8npT \left( \frac{1}{c_T} \right) \mu.$$ 

**Proof.** Notice that

$$\mathbb{P} \left( \max_{j=1,\ldots,p} \frac{1}{n^T} \sum_{i=1}^{n} \sum_{t=1}^{T} \frac{X_{i,t,j}a_{i,t}}{\sigma_j} \geq \eta \frac{|X|}{9} \right) \leq 2p \max_{j=1,\ldots,p} \mathbb{P} \left( \frac{1}{n^T} \sum_{i=1}^{n} \sum_{t=1}^{T} \frac{X_{i,t,j}a_{i,t}}{\sigma_j} \geq \eta \frac{|X|}{9} \right)$$

$$\leq 4p \max_{j=1,\ldots,p} \mathbb{P} \left( \frac{1}{n^T} \sum_{i=1}^{n} \sum_{t=1}^{T} \frac{X_{i,t,j}a_{i,t}}{\sigma_j} \geq \eta \frac{|X|}{9} \right) + 2p \max_{j=1,\ldots,p} \mathbb{P} \left( \frac{1}{n^T} \sum_{i=1}^{n} \sum_{t=1}^{T} \frac{X_{i,t,j}a_{i,t}}{\sigma_j} \geq \eta \frac{|X|}{9} \right) + 8npT \left( \frac{1}{c_T} \right) \mu.$$ 

(45)

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

$$\frac{1}{n^T} \sum_{i=1}^{n} \sum_{t=1}^{T} X_{i,t,j}^2 \leq 3c_T \sigma_j^2,$$

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

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

(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),

$$
P \left( \sup_{\Delta \in \mathcal{G}} \frac{1}{nT} \left| \sum_{i=1}^{n} \sum_{t=1}^{T} \Delta_{i,t} a_{i,t} \right| \geq \eta \right) 
\leq P \left( \sup_{\Delta \in \mathcal{G}} \frac{1}{nT} \left| \sum_{i=1}^{n} \sum_{t=1}^{T} \Delta_{i,t} a_{i,t} \right| \geq \frac{\eta}{3} \right) 
+ P \left( \sup_{\Delta \in \mathcal{G}} \frac{1}{nT} \left| \sum_{i=1}^{n} \sum_{t=1}^{T} \Delta_{i,t} \tilde{a}_{i,t} \right| \geq \frac{\eta}{3} \right) + 2nT \left( \frac{1}{c_T} \right)^{\mu}
$$

$$
\leq 2c_T \max_{m \in [c_T]} P \left( \sup_{\Delta \in \mathcal{G}} \frac{1}{ndT} \left| \sum_{i=1}^{n} \sum_{t=1}^{T} \Delta_{i,(2c_T t + m)} \tilde{a}_{i,(2c_T t + m)} \right| \geq \frac{\eta}{9} \right) 
+ 2c_T \max_{m \in [n]} P \left( \sup_{\Delta \in \mathcal{G}} \frac{1}{ndT} \left| \sum_{i=1}^{n} \sum_{t=1}^{T} \Delta_{i,(2c_T d_T + m)} \tilde{a}_{i,(2c_T d_T + m)} \right| \geq \frac{\eta}{9} \right) + 2nT \left( \frac{1}{c_T} \right)^{\mu}
$$

(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 $m$

$$
\sup_{\Delta \in \mathcal{G}} \frac{1}{ndT} \left| \sum_{i=1}^{n} \sum_{t=0}^{d_T-1} \Delta_{i,(2c_T t + m)} \tilde{a}_{i,(2c_T t + m)} \right| 
\leq \sup_{\Delta \in \mathcal{G}} \frac{1}{ndT} \left\| \left\{ \tilde{a}_{i,(2c_T t + m)} \right\}_{i \in [n], t \in [d_T]} \right\|_2 \| \Delta \|_*
$$

(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 $1 - c_1 \exp(-c_2 \max\{n, d_T\})$ by Theorem 3.4 from Chatterjee (2015).

On the other hand,

$$
\sup_{\Delta \in \mathcal{G}} \frac{1}{ndT} \left| \sum_{i=1}^{n} \Delta_{i,(2c_T d_T + m)} \tilde{a}_{i,(2c_T d_T + m)} \right| 
\leq \sup_{\Delta \in \mathcal{G}} \frac{\sqrt{n} \| \Delta_{i,(2c_T d_T + m)} \|_{*}}{ndT}
$$

(49)

$$
\leq \frac{\sqrt{n} \| \Delta \|_*}{ndT},
$$

35
with probability at least 

\[ 1 - c_1 \exp(-c_2 \max\{n, d_T\}) \]

\textit{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 

\[ c_T/T \leq 1/3 \]

\textit{Lemma 6. For every } \tilde{\Pi}, \bar{\Pi} \in \mathbb{R}^{n \times T}, \text{ we have that}

\[ ||\tilde{\Pi} - \bar{\Pi}||_* + ||\tilde{\Pi}||_* - ||\bar{\Pi}||_* \leq 6\sqrt{\text{rank}(\tilde{\Pi})||\tilde{\Pi} - \bar{\Pi}||_F} \]

\textit{Proof. This follows directly from Lemma 2.3 in Elsener and van de Geer (2018).}

\textit{Lemma 7. Assume that 1–3 hold. Then, with probability approaching one,}

\[ \frac{3}{4} ||\theta||_1 \leq ||\theta||_{1,n,T} \leq \frac{5}{4} ||\theta||_1, \] (50)

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

\textit{Moreover, for } c_0 \in (0, 1) \text{ letting}

\[ \nu_1 = \frac{9}{1 - c_0} \sqrt{\frac{c_T \log(\max\{n, p_{cT}\})}{nd_T}}(\sqrt{n} + \sqrt{d_T}), \]

\textit{and}

\[ \nu_2 = \frac{200c_T}{nT} (\sqrt{n} + \sqrt{d_T}), \]

\textit{we have that}

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

\textit{with probability approaching one, where}

\[ A_\tau = \left\{ (\delta, \Delta) : \|\delta_T\|_1 + \frac{||\Delta||_*}{\sqrt{nT}\sqrt{\log(\max\{n, p_{cT}\})}} \leq C_0 \left( \|\delta_T\|_1 + \frac{\sqrt{\tau}||\Delta||_F}{\sqrt{nT}\sqrt{\log(\max\{n, p_{cT}\})}} \right) \right\}, \]

\textit{and } C_0 \text{ is a positive constant that depends on } \tau \text{ and } c_0.

\textit{Proof. By Lemma 4, Lemma 5, Lemma 6, and Assumption 2, we have that
\[ 0 \leq \tilde{Q}(\theta(\tau), \Pi(\tau)) - \tilde{Q}(\hat{\theta}(\tau), \hat{\Pi}(\tau)) + \nu_1 \left( \|\theta(\tau)\|_{1,n} - \|\hat{\theta}(\tau)\|_{1,n,T} \right) + \nu_2 \left( \|\Pi(\tau)\|_* - \|\hat{\Pi}(\tau)\|_* \right) \]

\[ \leq \max_{1 \leq j \leq p} \left| \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} X_{i,t,j} a_{i,t} \right| \left[ \sum_{k=1}^{p} \hat{\sigma}_k |\theta_k(\tau) - \hat{\theta}_k(\tau)| \right] + \nu_1 \left( \|\theta(\tau)\|_{1,n,T} - \|\hat{\theta}(\tau)\|_{1,n,T} \right) + \nu_2 \left( \|\Pi(\tau)\|_* - \|\hat{\Pi}(\tau)\|_* \right) \]

\[ \leq 9 \sqrt{c_T \log(\max\{n, pc_T\})} \left[ \sum_{k=1}^{p} \hat{\sigma}_k |\theta_k(\tau) - \hat{\theta}_k(\tau)| \right] + \nu_1 \left( \|\theta(\tau)\|_{1,n,T} - \|\hat{\theta}(\tau)\|_{1,n,T} \right) \]

\[ + \left( \frac{200 c_T}{nT} \left( \sqrt{n} + \sqrt{d_T} \right) \|\Pi(\tau) - \hat{\Pi}(\tau)\|_* \right) + \left( \frac{200 c_T}{nT} \left( \sqrt{n} + \sqrt{d_T} \right) \left( \|\Pi(\tau)\|_* - \|\hat{\Pi}(\tau)\|_* \right) \right) \]

\[ \leq 9 \sqrt{c_T \log(\max\{n, pc_T\})} \left[ \sum_{k=1}^{p} \hat{\sigma}_k |\theta_k(\tau) - \hat{\theta}_k(\tau)| \right] + \nu_1 \left( \|\theta(\tau)\|_{1,n,T} - \|\hat{\theta}(\tau)\|_{1,n,T} \right) \]

\[ + \left( \frac{1200 c_T \sqrt{c_T}}{nT} \left( \sqrt{n} + \sqrt{d_T} \right) \|\Pi(\tau) - \hat{\Pi}(\tau)\|_F \right) \]

\[ - \left( \frac{100 c_T}{nT} \left( \sqrt{n} + \sqrt{d_T} \right) \|\Pi(\tau) - \hat{\Pi}(\tau)\|_* \right) \]

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

\[ 0 \leq \left[ \sum_{j=1}^{p} \left( (1 - c_0) \hat{\sigma}_j |\hat{\theta}_j(\tau) - \theta_j(\tau)| + \hat{\sigma}_j |\theta_j(\tau) - \hat{\theta}_j(\tau)| \right) \right] \]

\[ + \left[ \frac{3C_1 \sqrt{c_T} \|\Pi(\tau) - \hat{\Pi}(\tau)\|_F}{\sqrt{nT \log(\max\{n, pc_T\})}} - \frac{C_2 \|\Pi(\tau) - \hat{\Pi}(\tau)\|_*}{\sqrt{nT \log(\max\{n, pc_T\})}} \right], \]

37
and the claim follows.

\[\square\]

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

\[Q_r(\theta(\tau) + \delta, \Pi(\tau) + \Delta) - Q_r(\theta(\tau), \Pi(\tau)) \geq \min \left\{ \frac{J_r^{1/2}(\delta, \Delta)}{4}, qJ_r^{1/2}(\delta, \Delta) \right\}.\]

Proof. Let

\[v_{A_r} = \sup_v \left\{ v : Q_r(\theta(\tau) + \delta, \Pi(\tau) + \Delta) - Q_r(\theta(\tau), \Pi(\tau)) \geq \frac{(J_r^{1/2}(\delta, \Delta))^2}{4}, \forall (\delta, \Delta) \in A_r, \quad J_r^{1/2}(\delta, \Delta) \leq v \right\}.\]

Then by the convexity of \(Q_r(\cdot)\) and the definition of \(v_{A_r}\), we have that

\[Q_r(\theta(\tau) + \delta, \Pi(\tau) + \Delta) - Q_r(\theta(\tau), \Pi(\tau)) \geq \frac{(J_r^{1/2}(\delta, \Delta))^2}{4} \wedge \left\{ \frac{J_r^{1/2}(\delta, \Delta)}{4} \right\} \wedge \left\{ \frac{J_r^{1/2}(\delta, \Delta) v_{A_r}^2}{4} \right\} \geq \frac{(J_r^{1/2}(\delta, \Delta))^2}{4} \wedge qJ_r^{1/2}(\delta, \Delta),\]

where in last inequality we have used the fact that \(v_{A_r} \geq 4q\). To see why this is true, notice that there exists \(z_{X_{i,t}} \in [0, z]\) such that

\[Q_r(\theta(\tau) + \delta, \Pi(\tau) + \Delta) - Q_r(\theta(\tau), \Pi(\tau)) \geq \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T \mathbb{E} \left( (X_{i,t}^t \delta + \Delta_{i,t})^2 - \frac{f^p}{6nT} \sum_{i=1}^n \sum_{t=1}^T \mathbb{E} \left( |X_{i,t}^t \delta + \Delta_{i,t}|^3 \right) \right).\]

Hence, if \((\delta, \Delta) \in A_r\) with \(J_r^{1/2}(\delta, \Delta) \leq 4q\) then

\[\sqrt{\frac{f}{nT} \sum_{i=1}^n \sum_{t=1}^T \mathbb{E} \left( (X_{i,t}^t \delta + \Delta_{i,t})^2 \right)} \leq \frac{3}{2} \frac{f^{3/2}}{f^p} \inf_{(\delta, \Delta) \in A_r, \delta \neq 0} \left( \frac{\mathbb{E} \left( \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T (X_{i,t}^t \delta + \Delta_{i,t})^2 \right)^{3/2}}{\mathbb{E} \left( \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T |X_{i,t}^t \delta + \Delta_{i,t}|^3 \right)} \right).\]
combined with (51) implies

\[ Q_\tau(\theta(\tau) + \delta, \Pi(\tau) + \Delta) - Q_\tau(\theta(\tau), \Pi(\tau)) \geq \left( \frac{J^{1/2}_\tau(\delta, \Delta)}{4} \right)^2. \]

\[ \square \]

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

\[ \|\delta\|_{1,n,T} \leq \frac{2(C_0 + 1)\sqrt{s_\tau + 1}}{\kappa_0} \max \left\{ \frac{\sqrt{T_\tau}}{\log(n \vee pct)}, 1 \right\} J^{1/2}_\tau(\delta, \Delta), \]

and

\[ \|\Delta\|_* \leq (C_0 + 1)\sqrt{s_\tau + 1} \sqrt{nT \log(\max\{pct, n\})} \kappa_0^{-1} \max \left\{ \frac{\sqrt{T_\tau}}{\log(n \vee pct)}, 1 \right\} J^{1/2}_\tau(\delta, \Delta), \]

with \(C_0\) as in Lemma 7.

**Proof.** By Cauchy-Schwartz’s inequality, and the definition of \(A_\tau\) and \(J^{1/2}_\tau(\delta, \Delta)\) we have

\[ \|\delta\|_{1,n,T} \leq \frac{2}{3} \left( \|\delta_{T_\tau}\|_1 + \|\delta_{T_\eta}\|_1 \right) \]

\[ \leq \frac{2}{3} \|\delta_{T_\tau}\|_1 + \frac{3C_0}{4} \left( \|\delta_{T_\tau}\|_1 + \frac{\sqrt{T_\tau} \|\Delta\|_F}{\sqrt{nT \log(n \vee pct)}} \right) \]

\[ \leq 2(C_0 + 1)\sqrt{s_\tau} \|\delta_{T_\tau}\|_2 + 2C_0 \left( \frac{\sqrt{T_\tau} \|\Delta\|_F}{\sqrt{nT \log(n \vee pct)}} \right) \]

\[ \leq 2(C_0 + 1)\sqrt{s_\tau + 1} \left( \|\delta_{T_\tau}\|_2 + \frac{\sqrt{T_\tau}}{\sqrt{nT \log(n \vee pct)}} \|\Delta\|_F \right) \]

\[ \leq 2(C_0 + 1)\sqrt{s_\tau + 1} \max \left\{ \frac{\sqrt{T_\tau}}{\log(n \vee pct)}, 1 \right\} J^{1/2}_\tau(\delta, \Delta). \]

On the other hand, by the triangle inequality, the construction of the set \(A_\tau\), and Cauchy-Schwartz’s inequality

\[ \|\Delta\|_* \leq C_0 \sqrt{nT \log(n \vee pct)} \left( \|\delta_{T_\tau}\|_1 + \frac{\sqrt{T_\tau} \|\Delta\|_F}{\sqrt{nT \log(n \vee pct)}} \right) \]

\[ \leq \sqrt{s_\tau + 1}(C_0 + 1) \sqrt{nT \log(n \vee pct)} \left( \|\delta_{T_\tau}\|_2 + \frac{\sqrt{T_\tau} \|\Delta\|_F}{\sqrt{nT \log(n \vee pct)}} \right) \]

\[ \leq \sqrt{s_\tau + 1}(C_0 + 1) \sqrt{nT \log(n \vee pct)} \max \left\{ \frac{\sqrt{T_\tau}}{\log(n \vee pct)}, 1 \right\} \frac{J^{1/2}_\tau(\delta, \Delta)}{\kappa_0}. \]

\[ \square \]

**Lemma 10.** Let

\[ \epsilon(\eta) = \sup_{(\delta, \Delta) \in A_\tau : J^{1/2}_\tau(\delta, \Delta) \leq \eta} \left| Q_\tau(\theta(\tau) + \delta, \Pi(\tau) + \Delta) - Q_\tau(\theta(\tau), \Pi(\tau)) - Q_\tau(\theta(\tau) + \delta, \Pi(\tau) + \Delta) + Q_\tau(\theta(\tau), \Pi(\tau)) \right|. \]
and \( \{\phi_n\} \) a sequence with \( \phi_n/(\sqrt{T}\log(c_T + 1)) \to \infty \). Then for all \( \eta > 0 \)
\[
\epsilon(\eta) \leq C_0\eta c_T\phi_n\sqrt{(1 + s_T)\max\{\log(p c_T \wedge n), r_T\} (\sqrt{n} + \sqrt{d_T})},
\]
for some constant \( C_0 > 0 \), with probability at least \( 1 - \alpha_n \). Here, the sequence \( \{\alpha_n\} \) is independent of \( \eta \), and \( \alpha_n \to 0 \).

**Proof.** Let \( \Omega_1 \) be the event max\(_j\leq p |\hat{s}_j - 1| \leq 1/4 \). Then, by Assumption, \( P(\Omega_1) \geq 1 - \gamma \). Next let \( \kappa > 0 \), and \( f = (\delta, \Delta) \in A_r \) and write
\[
\mathcal{F} = \{((\delta, \Delta) \in A_r : J^{1/2}_r(\delta, \Delta) \leq \eta}\}.
\]
Then notice that by Lemmas 4.1 and 4.2 from Yu (1994),
\[
\mathbb{P}\left(\epsilon(\eta)\sqrt{nT} \geq \kappa\right) \leq 2\mathbb{P}\left(\sup_{f \in \mathcal{F}} \frac{1}{\sqrt{nT}} \left| \sum_{i=1}^{n} \sum_{t \in H_i} Z_{i,t}(f) \right| \geq \frac{\kappa}{3}\right) + \mathbb{P}\left(\sup_{f \in \mathcal{F}} \frac{1}{\sqrt{nT}} \left| \sum_{i=1}^{n} \sum_{t \in \tau} Z_{i,t}(f) \right| \geq \frac{\kappa}{3}\right) + 2nT \left(\frac{1}{c_T}\right)^\mu, \tag{52}
\]
where
\[
Z_{i,t}(f) = \rho_T(\tilde{Y}_{i,t} - \tilde{X}_{i,t}^t(\theta(\tau) + \delta) - (\Pi_{i,t}(\tau) + \Delta_{i,t})) - \rho_T(\tilde{Y}_{i,t} - \tilde{X}_{i,t}^t(\theta(\tau) - \Pi_{i,t}(\tau))) - \mathbb{E}\left(\rho_T(\tilde{Y}_{i,t} - \tilde{X}_{i,t}^t(\theta(\tau) + \delta) - (\Pi_{i,t}(\tau) + \Delta_{i,t})) - \rho_T(\tilde{Y}_{i,t} - \tilde{X}_{i,t}^t(\theta(\tau) - \Pi_{i,t}(\tau)))\right).
\]
Next we proceed to bound each term in (52). To that end, notice that
\[
\text{Var}\left(\sum_{i=1}^{n} \sum_{t \in H_i} Z_{i,t}(f) \right) \leq \sum_{i=1}^{n} \sum_{t \in H_i} \mathbb{E}\left(\left(\frac{1}{\sqrt{c_T}} \sum_{t \in \tau} Z_{i,t}(f) \right)^2\right) \leq \sum_{i=1}^{n} \sum_{t \in H_i} \sum_{l=1}^{d_T} \mathbb{E}\left(\left(\tilde{X}_{i,l}\delta + \Delta_{i,l}\right)^2\right) \leq \frac{nT}{f} \left(\frac{J^{1/2}_r(\delta, \Delta)}{\delta}\right)^2.
\]
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)
\[
\mathbb{P}\left(\sup_{f \in \mathcal{F}} \frac{1}{\sqrt{nT}} \left| \sum_{i=1}^{n} \sum_{t \in H_i} Z_{i,t}(f) \right| \geq \kappa\right) \leq \mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \sum_{t \in H_i} \varepsilon_{i,t} \left(\sum_{t \in H_i} Z_{i,t}(f) \right) \right| \geq \kappa\right) \leq \frac{P(A^n(\eta) \geq \frac{n}{2}(1 - \eta\Omega_1)) + P(\Omega_1^c)}{1 - \frac{12\epsilon^2(\eta)}{L^2}}, \tag{53}
\]
where

\[ A^0(\eta) := \sup_{f \in F} \left| \frac{1}{\sqrt{ndT}} \sum_{i=1}^{n} \sum_{l=1}^{d_f} \varepsilon_{i,l} \left( \sum_{t \in H_l} \rho_r (Y_{i,t} - \hat{X}_{i,t}^t(\theta(\tau) + \delta) - (\Pi_{i,t}(\tau) + \Delta_{i,t})) - \rho_r (\hat{Y}_{i,t} - \hat{X}_{i,t}^t(\theta(\tau) - \Pi_{i,t}(\tau)) \right) \right| \].

Next, note that

\[ \rho_r (\hat{Y}_{i,t} - \hat{X}_{i,t}^t(\theta(\tau) + \delta) - (\Pi_{i,t}(\tau) + \Delta_{i,t})) - \rho_r (\hat{Y}_{i,t} - \hat{X}_{i,t}^t(\theta(\tau) - \Pi_{i,t}(\tau)) = \tau \left( \hat{X}_{i,t}^t \delta + \Delta_{i,t} \right) + v_{i,t}(\delta, \Delta) + w_{i,t}(\delta, \Delta), \]

where

\[ |v_{i,t}(\delta, \Delta)| = \left| \hat{Y}_{i,t} - \hat{X}_{i,t}^t(\theta(\tau) + \delta) - (\Pi_{i,t}(\tau) + \Delta_{i,t}) - (Y_{i,t} - \hat{X}_{i,t}^t(\theta(\tau) + \delta) - \Pi_{i,t}(\tau)) \right| \]

and

\[ |w_{i,t}(\delta, \Delta)| = \left| \hat{Y}_{i,t} - \hat{X}_{i,t}^t(\theta(\tau) + \delta) - (\Pi_{i,t}(\tau)) - (\hat{Y}_{i,t} - \hat{X}_{i,t}^t(\theta(\tau) - \Pi_{i,t}(\tau)) \right| \]

Moreover, notice that by Lemma 9,

\[ \{(\delta, \Delta) \in A_\tau : J_r^{1/2}(\delta, \Delta) \leq \eta\} \subset \{(\delta, \Delta) \in A_\tau : \|\delta\|_{1,n,T \leq \eta \nu}\}, \]

where

\[ \nu := \frac{2(C_0 + 1)\sqrt{1 + s_\tau}}{\kappa_0} \max \left\{ \frac{\sqrt{T_\tau}}{\sqrt{\log(n \lor \text{pc}_T)}}, 1 \right\}. \]

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

\[ \|\Delta\|_* \leq \sqrt{1 + s_\tau (C_0 + 1)} J_r^{1/2}(\delta, \Delta) \sqrt{nT} \max\{\log(pc_T \lor n), r_\tau\}, \]

and so,

\[ \{(\delta, \Delta) \in A_\tau : J_r^{1/2}(\delta, \Delta) \leq \eta\} \subset \{(\delta, \Delta) \in A_\tau : \|\Delta\|_* \leq \sqrt{1 + s_\tau (C_0 + 1)} \eta \sqrt{nT} \max\{\log(pc_T \lor n), r_\tau\} / \kappa_0 \}. \]
Hence, defining

\[
B_0^1(\eta) = \sqrt{c_T} \sup_{\delta : \|\delta\|_{1,n,T} \leq \eta \nu} \left| \frac{1}{\sqrt{nT}} \sum_{i=1}^{n} \sum_{l=1}^{d_T} \varepsilon_{i,l} \left( \sum_{t \in H_i} \tilde{X}_{i,t} \delta \right) \right|,
\]

\[
B_0^2(\eta) = \sqrt{c_T} \sup_{\Delta : \|\Delta\|_* \leq \sqrt{T+\sigma_T(C_0+1)\eta \sqrt{n T \max\{\log(p c_T n \nu), r_T\} / \kappa_0}}} \left| \frac{1}{\sqrt{nT}} \sum_{i=1}^{n} \sum_{l=1}^{d_T} \varepsilon_{i,l} \left( \sum_{t \in H_i} \Delta_{i,t} \right) c_T \right|,
\]

\[
B_0^3(\eta) = \sqrt{c_T} \sup_{\delta : \|\delta\|_{1,n,T} \leq \eta \nu} \left| \frac{1}{\sqrt{nT}} \sum_{i=1}^{n} \sum_{l=1}^{d_T} \varepsilon_{i,l} \left( \sum_{t \in H_i} w_{i,l}(\delta, \Delta) \right) c_T \right|,
\]

By union bound we obtain that

\[
P(A_0^0(\eta) \geq \kappa | \Omega_1) \leq \sum_{j=1}^{4} P(B_0^j(\eta) \geq \kappa | \Omega_1), \tag{56}
\]

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

First, notice that

\[
B_0^1(\eta) \leq 2 c_T \max_{m \in [c_T]} \sup_{\delta : \|\delta\|_{1,n,T} \leq \eta \nu} \left| \frac{1}{\sqrt{nT}} \sum_{i=1}^{n} \sum_{l=0}^{d_T-1} \varepsilon_{i,l} \tilde{X}_{i,(2lc_T+m)} \delta \right|,
\]

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

\[
P(B_0^1(\eta) \geq \kappa | \Omega_1) \leq 2pc_T \exp \left( -\frac{\kappa^2}{4c_T^2(16\sqrt{2}\eta \nu)^2} \right). \tag{57}
\]

Next we proceed to bound \(B_0^3(\eta)\), by noticing that

\[
B_0^3(\eta) \leq \max_{m \in [c_T]} \sup_{\Delta : \|\Delta\|_* \leq \sqrt{T+\sigma_T(C_0+1)\eta \sqrt{n T \max\{\log(p c_T n \nu), r_T\} / \kappa_0}}} \left| \sqrt{c_T} \sum_{i=1}^{n} \sum_{l=0}^{d_T-1} \varepsilon_{i,l} v_{i,l}(\delta, \Delta) \right|.
\]

Towards that end we proceed to bound the moment generating function of \(B_0^3(\eta)\) and the use that to obtain

42
an upper bound on $B_3^0(\eta)$. Now fix $m \in [d_T]$ and notice that

\[
\mathbb{E} \left( \exp \left( \lambda \sup_{\Delta: \|\Delta\|_\| \leq \sqrt{1 + s_\tau(C_0 + 1)\eta} \sqrt{nT \max\{\log(p \cap T \wedge n), r_T\}} / \kappa_0} \left( \frac{\mathbb{E} \left[ \|\Delta\|_\| \right]}{\sqrt{nT}} \right) \right) \right) \leq \mathbb{E} \left( \exp \left( \lambda \frac{\sqrt{nT} \max\{\log(p \cap T \wedge n), r_T\}}{\kappa_0} \left( \frac{\mathbb{E} \left[ \|\Delta\|_\| \right]}{\sqrt{nT}} \right) \right) \right)
\]

\[
\exp \left( \frac{(s_\tau + 1)(C_0 + 1)^2 c_4 \lambda^2 c_T^2 \eta^2 \max\{\log(p \cap T \wedge n), r_T\}}{\kappa_0^2} \right),
\]

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),

\[
P(B_3^0(\eta) \geq \kappa|\Omega_1) \leq c_T \max_{m \in [d_T]} \mathbb{P} \left( \sup_{\Delta: \|\Delta\|_\| \leq \sqrt{1 + s_\tau(C_0 + 1)\eta} \sqrt{nT \max\{\log(p \cap T \wedge n), r_T\}} / \kappa_0} \left( \frac{\mathbb{E} \left[ \|\Delta\|_\| \right]}{\sqrt{nT}} \right) \right) \leq \inf_{\lambda > 0} \left[ \mathbb{E} \left( \exp \left( -\lambda \mathbb{E} \left[ \|\Delta\|_\| \right] \right) \right) \right]
\]

\[
\exp \left( \frac{(1 + s_\tau)(C_0 + 1)^2 c_4 \lambda^2 c_T^2 \eta^2 \max\{\log(p \cap T \wedge n), r_T\}}{\kappa_0^2} + \log c_T \right) \leq c_5 \exp \left( -\sqrt{1 + s_\tau(C_0 + 1)\eta} \sqrt{3 \max\{\log(p \cap T \wedge n), r_T\}} \left( \sqrt{n + d_T} \right) \right) + \log c_T \right),
\]

for a positive constant $c_5 > 0$.

Furthermore, we observe that

\[
B_2^0(\eta) \leq \max_{m \in [d_T]} \mathbb{E} \left( \sup_{\Delta: \|\Delta\|_\| \leq \sqrt{1 + s_\tau(C_0 + 1)\eta} \sqrt{nT \max\{\log(p \cap T \wedge n), r_T\}} / \kappa_0} \left( \frac{\mathbb{E} \left[ \|\Delta\|_\| \right]}{\sqrt{nT}} \right) \right) \leq \inf_{\lambda > 0} \left[ \mathbb{E} \left( \exp \left( -\lambda \mathbb{E} \left[ \|\Delta\|_\| \right] \right) \right) \right]
\]

\[
\exp \left( \frac{(1 + s_\tau)(C_0 + 1)^2 c_4 \lambda^2 c_T^2 \eta^2 \max\{\log(p \cap T \wedge n), r_T\}}{\kappa_0^2} + \log c_T \right) \leq c_5 \exp \left( -\sqrt{1 + s_\tau(C_0 + 1)\eta} \sqrt{3 \max\{\log(p \cap T \wedge n), r_T\}} \left( \sqrt{n + d_T} \right) \right) + \log c_T \right),
\]

(59)
Hence, with the same argument for bounding $B^0_2(\eta)$, we have

\[
P( B^0_2(\eta) \geq \kappa | \Omega_1 ) \leq c_5 \exp \left( - \frac{\kappa \kappa_0}{\sqrt{1 + s_T (C_0 + 1) \eta c_T \sqrt{3 \max\{\log(p c_T \lor n), r_T\}} \left( \sqrt{n} + \sqrt{d_T} \right) } + \log c_T \right). \tag{60}
\]

Finally, we proceed to bound $B^0_4(\eta)$. To that end, notice that

\[
B^0_4(\eta) \leq \max_{m \in [d_T]} \sup_{\delta : \| \delta \|_{1,n,T} \leq \eta \nu} \left| \frac{\sqrt{c_T}}{\sqrt{\nu n d_T}} \sum_{i=1}^{n} \sum_{l=0}^{d_T-1} \varepsilon_{i,l} w_i (2 t c_T + m) (\delta, \Delta) \right|,
\]

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

\[
P(B^0_4(\eta) \geq \kappa | \Omega_1 ) \leq 2 p c_T \exp \left( - \frac{\kappa^2}{4 c_T^2 (16 \sqrt{2} \eta \nu)^2} \right). \tag{61}
\]

Therefore, letting

\[
\kappa = \frac{\eta c_T \phi_n (1 + C_0)^2 \sqrt{1 + s_T} \max\{\log(p c_T \lor n), r_T\} (\sqrt{n} + \sqrt{d_T})}{\kappa_0 f^{1/2}},
\]

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

\[
P(\epsilon(\eta) \geq \frac{\kappa}{\sqrt{n T}}) \leq 5^{\gamma + 4 \exp \left( \max\{\log(p c_T \lor n), r_T\} - C_1 \phi_n^2 \max\{\log(p c_T \lor n), r_T\} \frac{1}{\nu n d_T} \right) + 2 \log c_T} \exp \left( - C_2 \phi_n^2 2 \nu n^2 \max\{\log(p c_T \lor n), r_T\} \right) + n T \left( \frac{1}{c_T} \right)^\mu,
\]

for some positive constants $C_1$ and $C_2$.

\[\square\]

### B.2 Proof of Theorem 1

**Proof.** Recall from Lemma 7, our choices of $\nu_1$ and $\nu_2$ are

\[
\nu_1 = C_0' \sqrt{\frac{c_T \log(\max\{n, p c_T\})}{n d_T}} \left( \sqrt{n} + \sqrt{d_T} \right),
\]

and

\[
\nu_2 = \frac{200 c_T}{n T} \left( \sqrt{n} + \sqrt{d_T} \right),
\]

for $C_0' = 9 / (1 - c_0)$, and $c_0$ as in Lemma 7.
Let
\[
\eta = \frac{8\phi_n(C_0' + C_0) + \tilde{C}_0 + 200(1 + C_0) + \sqrt{r_T} \max \{ \log(p c_T \lor n), r_T \} \{ \sqrt{n} + \sqrt{d_T} \}}{\sqrt{n d_T} \kappa_0 f_{1/2}},
\]  
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
\[
|J_t^{1/2}(\hat{\theta}(\tau) - \theta(\tau), \hat{\Pi}(\tau) - \Pi(\tau))| > \eta.
\]
Then, by the convexity of \( A_\tau \), and of the objective \( \hat{Q} \) with its constraint, we obtain that
\[
0 > \min_{(\delta, \Delta) \in A_\tau : |J_t^{1/2}(\delta, \Delta)| = \eta} \hat{Q}(\theta(\tau) + \delta, \Pi(\tau) + \Delta) - \hat{Q}(\theta(\tau), \Pi(\tau)) + \nu_1 \max \{ \| \theta(\tau) + \delta \|_{1,n,T} - \| \theta(\tau) \|_{1,n,T} \}
\]
\[+ \nu_2 \max \{ \| \Pi(\tau) + \Delta \|_* - \| \Pi(\tau) \|_* \}, \]
Moreover, by Lemma 9 and the triangle inequality,
\[
\| \theta(\tau) \|_{1,n,T} - \| \theta(\tau) + \delta \|_{1,n,T} \leq \| \delta_T \|_{1,n,T} \leq 2(1 + C_0) \sqrt{1 + \frac{J_t^{1/2}(\delta, \Delta)}{\kappa_0}} \max \left\{ \frac{\sqrt{r_T}}{\log(n v p c_T)}, 1 \right\},
\]
and
\[
\| \Pi(\tau) \|_* - \| \Pi(\tau) + \Delta \|_* \leq \| \Delta \|_* \leq (1 + C_0) \sqrt{1 + \frac{\sqrt{r_T} \max \{ \log(p c_T \lor n), r_T \} J_t^{1/2}(\delta, \Delta)}{\kappa_0}}.
\]
Therefore,
\[
0 > \min_{(\delta, \Delta) \in A_\tau : |J_t^{1/2}(\delta, \Delta)| = \eta} \hat{Q}(\theta(\tau) + \delta, \Delta + \Pi(\tau)) - \hat{Q}(\theta(\tau), \Pi(\tau)) - 2\nu_1(1 + C_0) \sqrt{1 + \frac{J_t^{1/2}(\delta, \Delta)}{\kappa_0}} \max \left\{ \frac{\sqrt{r_T}}{\log(n v p c_T)}, 1 \right\},
\]
\[
- \nu_2(1 + C_0) \sqrt{1 + \frac{\sqrt{r_T} \max \{ \log(p c_T \lor n), r_T \} J_t^{1/2}(\delta, \Delta)}{\kappa_0}} + 2C_0' \sqrt{\frac{\sqrt{r_T} \max \{ \log(p c_T \lor n), r_T \}}{\kappa_0}} \sqrt{n} + \sqrt{45 d_T})
\]
\[
\frac{200 c_T}{\kappa_0} \left( \sqrt{n} + \sqrt{d_T} \right) \sqrt{1 + s_T(1 + C_0) \sqrt{n T} \max \{ \log(p c_T \lor n), r_T \}} \frac{J_t^{1/2}(\delta, \Delta)}{\kappa_0}.
\]
\[ \geq \min_{(\delta, \Delta) \in A_T} \frac{Q(\theta) + \delta \Delta + \Pi(\tau) - Q(\theta, \Pi(\tau))}{\sqrt{nT}} \]

\[ - [2C_0' (1 + C_0) + 200(C_0 + 1)] \sqrt{C_T (1 + s_T) \max\{\log(p^T \land n), r_T\}} (\sqrt{n} + \sqrt{d_T}) J^{1/2}_{\kappa_0} \]

\[ \geq \frac{\eta^2}{4} \land (\eta q) - \frac{[2C_0'(1 + C_0) + 200(C_0 + 1)]}{\sqrt{C_T (1 + s_T) \max\{\log(p^T \land n), r_T\}} (\sqrt{n} + \sqrt{d_T})} \]

\[ \geq \frac{\eta^2}{4} - \frac{C_0 \eta \phi_n (1 + C_0) + 200(C_0 + 1)}{\sqrt{C_T (1 + s_T) \max\{\log(p^T \land n), r_T\}} (\sqrt{n} + \sqrt{d_T})} \]

\[ = 0, \quad (64) \]

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

\[ \frac{||\hat{\Pi}(\tau) - \Pi(\tau)||_F}{\sqrt{nT}} \leq \frac{1}{\kappa_0} |J^{1/2}_{\tau}(\hat{\theta}(\tau) - \theta(\tau), \hat{\Pi}(\tau) - \Pi(\tau))| \leq \frac{\eta}{\kappa_0}, \]

which holds with probability approaching one.

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

\[ ||\hat{\delta}_{T_{\tau} \cup T_{\tau}(\hat{\delta}, m)}||^2 \leq \sum_{k \geq m+1} \frac{||\hat{\delta}_{T_{\tau}}||^2_k}{k^2} \]

\[ \leq \frac{4C_0}{m} \left[ \frac{||\hat{\delta}_{T_{\tau}}||^2_1 + r_T ||\Pi(\tau) - \hat{\Pi}(\tau)||^2_F}{nT \log(p^T \land n)} \right] \]

\[ \leq \frac{4C_0}{m} \left[ \frac{s_T ||\hat{\delta}_{T_{\tau} \cup T_{\tau}(\hat{\delta}, m)}||^2 + r_T ||\Pi(\tau) - \hat{\Pi}(\tau)||^2_F}{nT \log(p^T \land n)} \right], \]

which implies

\[ \|\hat{\delta}\| \leq (1 + 2C_0 \sqrt{\frac{s_T}{m}}) \left( ||\hat{\delta}_{T_{\tau} \cup T_{\tau}(\hat{\delta}, m)}|| + \frac{\sqrt{r_T} ||\Pi(\tau) - \hat{\Pi}(\tau)||_F}{\sqrt{nT \log(p^T \land n)}} \right) \]

\[ \leq J^{1/2}_{\kappa_0} (\hat{\Pi}(\tau) - \Pi(\tau)) \left( 1 + 2C_0 \sqrt{\frac{s_T}{m}} \right), \]

and the result follows. \( \square \)
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

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

and \( c_0 \) and \( c_1 \) are positive constants that depend on \( \tau \). 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,

\[ \left\| X\theta(\tau) + \Pi(\tau) - \hat{\Pi}(\tau) \right\|_{\infty} \leq 2C =: c_1. \]

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

\[ \hat{Q}_\tau(0, \hat{\Pi}) - \hat{Q}_\tau(\hat{\theta}, \hat{\Pi}) - \nu_1 \|\hat{\theta}\|_{1,n,T} \leq \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} |X'_{i,t} \hat{\theta}| - \nu_1 \|\hat{\theta}\|_1 \]

\[ \leq \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} |X'_{i,t} \hat{\theta}| - \nu_1 \|\hat{\theta}\|_1 \]

\[ \leq \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \|X'_{i,t}\|_{\infty} \|\hat{\theta}\|_1 - \nu_1 \|\hat{\theta}\|_1 \]

\[ < 0, \]

where the first inequality follows since \( \rho_\tau \) is a contraction map. Therefore, \( \hat{\theta}(\tau) = 0 \). Furthermore, by
Lemma 5, we have

\[
0 \leq \hat{Q}_\tau(0, X\theta(\tau) + \Pi(\tau)) - \hat{Q}_\tau(0, \hat{\Pi}(\tau)) + \nu_2 \left( \|X\theta(\tau) + \Pi(\tau)\|_* - \|\hat{\Pi}(\tau)\|_* \right)
\]

\[
\leq \left| \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} a_{i,t} \left( (X\theta(\tau) + \Pi(\tau)) - \hat{\Pi}(\tau) \right) \right|
\]

\[
+ \nu_2 \left( \|X\theta(\tau) + \Pi(\tau)\|_* - \|\hat{\Pi}(\tau)\|_* \right)
\]

\[
\leq \|X\theta(\tau) + \Pi(\tau) - \hat{\Pi}(\tau)\|_* \left( \sup_{\|\hat{\Delta}\|_* \leq 1} \left| \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} a_{i,t} \hat{\Delta}_{i,t} \right| \right)
\]

\[
+ \nu_2 \left( \|X\theta(\tau) + \Pi(\tau)\|_* - \|\hat{\Pi}(\tau)\|_* \right)
\]

\[
\leq \frac{200c_T}{nT} \left( \sqrt{n} + \sqrt{d_T} \right) \left( \|X\theta(\tau) + \Pi(\tau) - \hat{\Pi}(\tau)\|_* + \|X\theta(\tau) + \Pi(\tau)\|_* - \|\hat{\Pi}(\tau)\|_* \right)
\]

\[
- \frac{100c_T}{nT} \left( \sqrt{n} + \sqrt{d_T} \right) \|X\theta(\tau) + \Pi(\tau) - \hat{\Pi}(\tau)\|_*
\]

\[
\leq \frac{200c_T}{nT} \left( \sqrt{n} + \sqrt{d_T} \right) \left( \|X\theta(\tau) + \Pi(\tau) + \xi - \hat{\Pi}(\tau)\|_* + \|X\theta(\tau) + \Pi(\tau) + \xi\|_* - \|\hat{\Pi}(\tau)\|_* \right)
\]

\[
+ \frac{400c_T}{nT} \left( \sqrt{n} + \sqrt{d_T} \right) \|\xi\|_* - \frac{100c_T}{nT} \left( \sqrt{n} + \sqrt{d_T} \right) \|X\theta(\tau) + \Pi(\tau) - \hat{\Pi}(\tau)\|_*
\]

\[
\leq \frac{c_1 c_T}{nT} \left( \sqrt{n} + \sqrt{d_T} \right) \sqrt{T} \|X\theta(\tau) + \Pi(\tau) + \xi - \hat{\Pi}(\tau)\|_F
\]

\[
+ \frac{400c_T}{nT} \left( \sqrt{n} + \sqrt{d_T} \right) \|\xi\|_* - \frac{100c_T}{nT} \left( \sqrt{n} + \sqrt{d_T} \right) \|X\theta(\tau) + \Pi(\tau) - \hat{\Pi}(\tau)\|_*
\]

for some positive constant \(c_1\),

\[\square\]

**Lemma 12.** Let

\[
\epsilon'(\eta) = \sup_{(\delta, \Delta) \in A_T \cap \mathcal{J}_T^T(\delta, \Delta) \leq \eta} \left| \hat{Q}_\tau(0, X\theta(\tau) + \Pi(\tau) + \Delta) - \hat{Q}_\tau(0, X\theta(\tau) + \Pi(\tau)) - Q_\tau(0, X\theta(\tau) + \Pi(\tau) + \Delta) + Q_\tau(0, X\theta(\tau) + \Pi(\tau)) \right|
\]

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

\[
\epsilon'(\eta) \leq \frac{\breve{C}_0 \eta c_T \phi_n \sqrt{T} (\sqrt{n} + \sqrt{d_T})}{\sqrt{nT}}
\]

for some constant \(\breve{C}_0 > 0\), with probability at least \(1 - \alpha_n\). Here, the sequence \(\{\alpha_n\}\) is independent of \(\eta\), and \(\alpha_n \to 0\).

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

\[\square\]
Lemma 13. Let

\[ A''_\tau = \{ \Delta \in A'_\tau : q(\Delta) \geq 2\eta_0, \ \Delta \neq 0 \} , \]

with

\[ \eta_0 = \tilde{C}_1 c_T \phi_n \sqrt{n + \sqrt{d_T}} \frac{\sqrt{nT}}{f}, \]

for an appropriate constant \( \tilde{C}_1 > 0 \), and

\[ q(\Delta) = \frac{3}{2} \frac{f^{3/2}}{F'} \left( \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T (\Delta_{i,t})^2 \right)^{3/2} \]

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

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

**Proof.** This follows as the proof of Lemma 8. \( \square \)

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

\[ \hat{\Delta} := \hat{\Pi}(\tau) - X\theta(\tau) - \Pi(\tau) \in A'_\tau, \]

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

\[
\frac{1}{\sqrt{nT}} \|\hat{\Delta}\|_F < \frac{4\eta \left( \sum_{i=1}^n \sum_{t=1}^T |\hat{\Delta}_{i,t}|^3 \right)^{1/2}}{3 \left( \sum_{i=1}^n \sum_{t=1}^T \hat{\Delta}_{i,t}^2 \right)^{1/2}} \leq \frac{4F' \|\hat{\Delta}\|_\infty \eta}{3 f^{3/2}}. \]

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, \( A''_\tau \) as \( A' \), and

\[ \kappa_0 = \frac{f_1^{1/2}}{3}. \]

in Assumption 3. This leads to

\[
\frac{1}{\sqrt{nT}} \|\hat{\Delta}\|_F \leq \eta, \]

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),

\[
v := \max \left\{ \min_{O \in \mathcal{O}_r} \| \hat{g}(\tau)O - g(\tau) \|_F, \min_{\tilde{O} \in \mathcal{O}_r} \| \tilde{\lambda}(\tau)O - \tilde{\lambda}(\tau) \|_F \right\} = O_p \left( \frac{(\sigma_1(\tau) + \sqrt{r} \text{Err})^2}{(\sigma_{r-1}(\tau))^2 - (\sigma_r(\tau))^2} \right).
\]

Furthermore,

\[
\frac{\| \tilde{\lambda}(\tau) - \lambda(\tau) \|^2}{nT} = \frac{1}{nT} \sum_{j=1}^r \| \lambda_{.,j}(\tau) - \hat{\lambda}_{.,j}(\tau) \|^2
\]

\[
\leq \frac{2}{nT} \sum_{j=1}^r (\sigma_j - \hat{\sigma}_j)^2 + \frac{2}{nT} \sum_{j=1}^r \sigma_j^2 \| \tilde{\lambda}_j(\tau) - \hat{\lambda}_j(\tau) \|^2
\]

\[
\leq \frac{2}{nT} \sum_{j=1}^r (\sigma_j - \hat{\sigma}_j)^2 + \frac{2}{nT} \sum_{j=1}^r \sigma_j^2 \| \tilde{\lambda}_j(\tau) - \hat{\lambda}_j(\tau) \|^2
\]

\[
\leq \frac{2 r^2}{nT} (\sigma_1 - \hat{\sigma}_1)^2 + \frac{2}{nT} \| \tilde{\lambda}(\tau) - \hat{\lambda}(\tau) \|^2
\]

\[
\leq \frac{2 r^2}{nT} \| \Pi(\tau) - \hat{\Pi}(\tau) \|^2 + \frac{2}{nT} \| \tilde{\lambda}(\tau) - \hat{\lambda}(\tau) \|^2
\]

\[
= O_p \left( \frac{r \phi^2_T (1 + s_T) \max \{ \log(p \mathcal{C}_T \vee n, r_T) \} \left( \frac{1}{n} + \frac{1}{d_T} \right) + \kappa_0 \phi f}{nT} \right) + O_p \left( \frac{(\sigma_1(\tau) + \sqrt{r} \text{Err})^2 \text{Err}^2}{nT ((\sigma_{r-1}(\tau))^2 - (\sigma_r(\tau))^2)} \right),
\]

where the third inequality follows from Weyl’s inequality, and the last one from (68).
Table 7: Firm Characteristics Construction.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Name</th>
<th>Construction</th>
</tr>
</thead>
<tbody>
<tr>
<td>acc</td>
<td>Working capital accruals</td>
<td>Annual income before extraordinary items (ib) minus operating cash flows (oancf) divided by average total assets (at)</td>
</tr>
<tr>
<td>agr</td>
<td>Asset growth</td>
<td>Annual percent change in total assets (at)</td>
</tr>
<tr>
<td>beta</td>
<td>Beta</td>
<td>Estimated market beta from weekly returns and equal weighted market returns for 3 years</td>
</tr>
<tr>
<td>bm</td>
<td>Book-to-market</td>
<td>Book value of equity (ceq) divided by end of fiscal year-end market capitalization</td>
</tr>
<tr>
<td>chinv</td>
<td>Change in inventory</td>
<td>Change in inventory (inv) scaled by average total assets (at)</td>
</tr>
<tr>
<td>chmom</td>
<td>Change in 6-month momentum</td>
<td>Cumulative returns from months t-6 to t-1 minus months t-12 to t-7</td>
</tr>
<tr>
<td>dolvol</td>
<td>Dollar trading volume</td>
<td>Natural log of trading volume times price per share from month t-2</td>
</tr>
<tr>
<td>dy</td>
<td>Dividend to price</td>
<td>Total dividends (dvt) divided by market capitalization at fiscal year-end</td>
</tr>
<tr>
<td>egr</td>
<td>Earnings announcement return</td>
<td>Annual percent change in book value of equity (ceq)</td>
</tr>
<tr>
<td>ep</td>
<td>Earnings to price</td>
<td>Annual income before extraordinary items (ib) divided by end of fiscal year market cap</td>
</tr>
<tr>
<td>gma</td>
<td>Gross profitability</td>
<td>Revenues (revt) minus cost of goods sold (cogs) divided by lagged total assets (at)</td>
</tr>
<tr>
<td>idiovol</td>
<td>Idiosyncratic return volatility</td>
<td>Standard deviation of residuals of weekly returns on weekly equal weighted market returns for 3 years prior to month end</td>
</tr>
<tr>
<td>ill</td>
<td>Illiquidity (Amihud)</td>
<td>Average of daily (absolute return / dollar volume).</td>
</tr>
<tr>
<td>invest</td>
<td>Capital expenditures and inventory</td>
<td>Annual change in gross property, plant, and equipment (ppegt) + annual change in inventories (invt) all scaled by lagged total assets (at)</td>
</tr>
<tr>
<td>lev</td>
<td>Leverage</td>
<td>Annual change in gross property, plant, and equipment (ppegt) + annual change in inventories (invt) all scaled by lagged total assets (at)</td>
</tr>
<tr>
<td>lgr</td>
<td>Growth in long-term debt</td>
<td>Annual percent change in total liabilities (lt)</td>
</tr>
<tr>
<td>mom1m</td>
<td>1-month momentum</td>
<td>1-month cumulative return</td>
</tr>
<tr>
<td>Variable</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>mom6m</td>
<td>6-month momentum</td>
<td>5-month cumulative returns ending one month before month end</td>
</tr>
<tr>
<td>mve</td>
<td>Size</td>
<td>Natural log of market capitalization at end of month t-1</td>
</tr>
<tr>
<td>operprof</td>
<td>Operating profitability</td>
<td>Revenue minus cost of goods sold - SG&amp;A expense - interest expense divided by lagged common shareholders’ equity</td>
</tr>
<tr>
<td>range</td>
<td>Range of stock price</td>
<td>Monthly average of daily price range: (high-low)/((high+low)/2) (alternative measure of volatility)</td>
</tr>
<tr>
<td>retvol</td>
<td>Return volatility</td>
<td>Standard deviation of daily returns from month t-1</td>
</tr>
<tr>
<td>roaq</td>
<td>Return on assets</td>
<td>Income before extraordinary items (ibq) divided by one quarter lagged total assets (atq)</td>
</tr>
<tr>
<td>roeq</td>
<td>Return on equity</td>
<td>Earnings before extraordinary items divided by lagged common shareholders’ equity</td>
</tr>
<tr>
<td>sue</td>
<td>Unexpected quarterly earnings</td>
<td>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</td>
</tr>
<tr>
<td>turn</td>
<td>Share turnover</td>
<td>Average monthly trading volume for most recent 3 months scaled by number of shares outstanding in current month</td>
</tr>
</tbody>
</table>

Note: Estimated under different values of turning parameter $\nu_2$, when $\nu_1 = 10^{-3}$ is fixed. The results are reported for quantiles 10%, 50% and 90%.