Econometric Theory, **37**, 2021, 794–816.
doi:10.1017/S0266466620000304

FACTORISABLE MULTITASK QUANTILE REGRESSION*

SHIH-KANG CHAO
University of Missouri

WOLFGANG K. HÄRDLE
Humboldt-Universität zu Berlin
Xiamen University
Charles University

MING YUAN
Columbia University

A multivariate quantile regression model with a factor structure is proposed to study data with multivariate responses with covariates. The factor structure is allowed to vary with the quantile levels, which is more flexible than the classical factor models. Assuming the number of factors is small, and the number of responses and the input variables are growing with the sample size, the model is estimated with the nuclear norm regularization. The incurred optimization problem can only be efficiently solved in an approximate manner by off-the-shelf optimization methods. Such a scenario is often seen when the empirical loss is nonsmooth or the numerical procedure involves expensive subroutines, for example, singular value decomposition. To show that the approximate estimator is still statistically accurate, we establish a nonasymptotic bound on the Frobenius risk and prediction risk. For implementation, a numerical procedure that provably marginalizes the approximation error is proposed. The merits of our model and the proposed numerical procedures are demonstrated through the Monte Carlo simulation and an application to finance involving a large pool of asset returns.

1. INTRODUCTION

In a variety of applications in economics, interest often lies in the conditional quantiles of a response variable (Koenker and Hallock, 2001). Quantile regression (Koenker and Bassett, 1978) is arguably one of the most popular methods for estimating the quantile of a response variable. However, in a situation with multivariate responses having the same predictors, equation-by-equation quantile regression fails to capture the latent common structure. Factor models have been

*We would like to thank the Editor, Peter Phillips, the Co-Editor, Dennis Kristensen, and two anonymous referees for their very helpful and constructive comments. We also thank Scott Holan for providing many helpful comments. Address correspondence to Shih-Kang Chao, Department of Statistics, University of Missouri, Columbia, MO 65211, USA; e-mail: chaosh@missouri.edu.

© The Author(s), 2020. Published by Cambridge University Press.

https://doi.org/10.1017/S0266466620000304 Published online by Cambridge University Press

used for quantile regression with multivariate responses (Ando and Tsay, 2011; Chen, Dolado, and Gonzalo, 2019), but these models do not include the exogenous predictors and the factors are invariant to the quantile level.

We consider quantile level dependent factors $f_k^\tau(\boldsymbol{X})$ formed by covariates $\boldsymbol{X} \in \mathbb{R}^p$, where $\tau \in (0,1)$ is the quantile level. We assume that the conditional quantile $q_j(\tau|\boldsymbol{X})$ of Y_j, where Y_j is the jth component in the response vector $\boldsymbol{Y}$, satisfies

$$q_j(\tau|\boldsymbol{X}) = \sum_{k=1}^{r_\tau} \Psi_{\tau,kj} f_k^\tau(\boldsymbol{X}), \quad j = 1,\dots,m = \dim(\boldsymbol{Y}), \tag{1.1}$$

where $\Psi_{\tau,kj} \in \mathbb{R}$ is the factor loading, and $r_\tau \in \mathbb{N}$ is assumed small and is quantile level dependent. At first glance, (1.1) appears to be the factor-augmented regression model (FAR) (see, e.g., Stock and Watson, 2002; Bai and Ng, 2010), but in fact they are very different. The predictors and the response variables are both generated by factors in the FAR model, but (1.1) assumes that the factors are functions of predictors, where the functions are unknown but nonrandom.

From a practical perspective, the factors $f_k^\tau(\boldsymbol{X})$ in (1.1) are unobservable, so computing $\Psi_{\tau,kj}$ can be challenging. Instead of pre-estimating the factors, we adopt a *one-stage* approach that simultaneously estimates the factors and the loadings. In particular, we assume that $f_k^\tau(\boldsymbol{X}_i)$ is linear in $\boldsymbol{X} \in \mathbb{R}^p$, that is, $f_k^\tau(\boldsymbol{X}_i) \stackrel{\text{def}}{=} \Phi_{\tau,*k}^\top \boldsymbol{X}_i$, where $\Phi_{\tau,*k} \in \mathbb{R}^p$. Model (1.1) can be written as

$$q_j(\tau|\boldsymbol{X}_i) = \boldsymbol{\Gamma}_{\tau,*j}^\top \boldsymbol{X}_i, \quad i = 1,\dots,n, \quad j = 1,\dots,m = \dim(\boldsymbol{Y}), \tag{1.2}$$

where $\boldsymbol{\Gamma}_{\tau,*j} = \Phi_\tau \Psi_{\tau,*j}$ is the jth column of $\boldsymbol{\Gamma}_\tau$, which is the $p \times m$ coefficient matrix with p and m possibly diverging to infinity as $n \to \infty$. Note that this implies $\boldsymbol{\Gamma}_\tau = \Phi_\tau \Psi_\tau$, where $\Phi_\tau \in \mathbb{R}^{p \times r_\tau}$ linearly transforms $\boldsymbol{X}$ to factors by $\Phi_\tau^\top \boldsymbol{X}_i = (f_1^\tau(\boldsymbol{X}_i), f_2^\tau(\boldsymbol{X}_i), \dots, f_{r_\tau}^\tau(\boldsymbol{X}_i))^\top$, and $\Psi_\tau \in \mathbb{R}^{r_\tau \times m}$ are the loadings defined in (1.1), with the jth column $\Psi_{\tau,*j}$ corresponding to the output Y_j. Here, we normalize $\|\Psi_{\tau,k*}\|_2 = 1$ so that the model is identifiable. If the matrix $\boldsymbol{\Gamma}_\tau$ in (1.2) is known, the factors and loadings can be computed simultaneously by a factorization of $\boldsymbol{\Gamma}_\tau$; see Section 2.2.

The rank of $\boldsymbol{\Gamma}_\tau$ is the number of factors r_τ, so the rank of $\boldsymbol{\Gamma}_\tau$ is small whenever r_τ is. Since the rank corresponds to the number of nonzero singular values, assuming r_τ is small, we estimate $\boldsymbol{\Gamma}_\tau$ with the regularization on singular values:

$$\widehat{\boldsymbol{\Gamma}}_\tau \stackrel{\text{def}}{=} \arg\min_{\mathbf{S} \in \mathbb{R}^{p \times m}} \big\{ L_\tau(\mathbf{S}) \stackrel{\text{def}}{=} \widehat{Q}_\tau(\mathbf{S}) + \lambda \|\mathbf{S}\|_* \big\}, \tag{1.3}$$

where $\lambda > 0$ is a tuning parameter, $\|\mathbf{S}\|_*$ denotes the nuclear norm, which is the sum of singular values, and

$$\widehat{Q}_\tau(\mathbf{S}) \stackrel{\text{def}}{=} (mn)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \rho_\tau\big(Y_{ij} - \boldsymbol{X}_i^\top \mathbf{S}_{*j}\big). \tag{1.4}$$

Here, $\rho_\tau(u) = u(\tau - \mathbf{1}\{u \le 0\})$ is the "check function" (Koenker and Bassett, 1978). Note that $\widehat{Q}_\tau$ is similar to the loss function used in Koenker and Portnoy (1990) for the low dimensional case; that is, m and p do not diverge with n. For mean regression, the nuclear norm penalty has been considered; for example, see

https://doi.org/10.1017/S0266466620000304 Published online by Cambridge University Press

Yuan et al. (2007), Bunea, She, and Wegkamp (2011), Negahban and Wainwright (2011), Koltchinskii, Lounici, and Tsybakov (2011), Maurer and Pontil (2013), and Maurer, Pontil, and Romera-Paredes (2016).

In (1.3), $\widehat{\mathbf{\Gamma}}_\tau$ is the minimizer of a convex empirical loss and its convergence rate is known to be optimal among all the admissible estimators (Koltchinskii, 2011). Unfortunately, in practice, many off-the-shelf optimization methods only provide an approximate estimator $\widehat{\mathbf{\Gamma}}_{\tau,\delta}$ that solves the optimization problem (1.3) with an optimization error $\delta > 0$. For example, $\widehat{\mathbf{\Gamma}}_{\tau,\delta}$ may be computed with a surrogate loss function of L_τ that is easier to optimize, or $\widehat{\mathbf{\Gamma}}_{\tau,\delta}$ is the outcome of an iterative procedure, in which each iteration involves costly subroutines such as the singular value decomposition. A natural question that arises is: *Can* $\widehat{\mathbf{\Gamma}}_{\tau,\delta}$ *accurately estimate* $\mathbf{\Gamma}_\tau$ *in some statistical sense?*

The main goal of this paper is to provide an affirmative answer to the above question. In particular, we provide a nonasymptotic bound for the Frobenius risk and prediction risk of $\widehat{\mathbf{\Gamma}}_{\tau,\delta}$, which hold when δ is smaller than an explicit bound depending on n, m, and p. We also propose a simple algorithm that produces $\widehat{\mathbf{\Gamma}}_{\tau,\delta}$ with δ satisfying the bound. Theoretical tools developed in this paper may be potentially useful for other convex problems where finding an exact optimizer is expensive and/or unrealistic.

In addition to the theoretical guarantees, we evaluate our numerical procedure using Monte Carlo simulations with both i.i.d. and dependent designs. The response variables are generated from a two-piece normal distribution (Wallis, 2014). The two-piece normal distribution has been used for predicting the inflation rate by central banks (Wallis, 1999). The results show that our numerical procedure can correctly estimate the number of factors around 50%–60% of the time, and it either over or underestimates the true number of factors by at most two when the dimensionality is in several hundreds.

For an empirical analysis, we estimate the market systemic risk from a large pool of assets, and compute the exposure of each asset to the systemic risk. Our method is scalable to high dimensional data; therefore, we are able to overcome the computational barriers inherent in the existing methods (Adrian and Brunnermeier, 2016; White, Kim, and Manganelli, 2015).

We remark that multitask linear models have been considered in applications where the tails of distribution are of interest. A recent paper by Chao, Härdle, and Huang (2018) explores functional magnetic resonance imaging (fMRI) data with multitask *expectile* regression. They propose an iterative shrinkage algorithm, and show the finite-sample convergence rate of the estimator with nonzero optimization error. The task undertaken in the current paper is much more challenging than that in Chao et al. (2018) from both computational and theoretical aspects, because of the nonsmoothness of the quantile regression loss function. The model in this paper is more appropriate when the response variables are heavy-tailed as the moment conditions are not required.

The rest of this paper is organized as follows. Section 2 discusses the numerical procedure for estimating the coefficient matrix $\mathbf{\Gamma}_\tau$, as well as the factors and loadings. Selection of the tuning parameter λ is also discussed. Section 3 provides

https://doi.org/10.1017/S0266466620000304 Published online by Cambridge University Press

a nonasymptotic analysis for $\widehat{\mathbf{\Gamma}}_{\tau,\delta}$, based on which the estimation risk of the factors and loadings are analyzed. Simulation results are presented in Section 4. An application to the financial systemic risk is shown in Section 5. Section 6 concludes this paper. The Appendix contains the detailed development of the algorithm in Section 2 and the proof of key theoretical results. Technical lemmas and remaining proofs are shifted to the supplementary material.

NOTATION. *For a matrix* $\mathbf{A} = (A_{ij}) \in \mathbb{R}^{p\times m}$, *denote the singular values of* $\mathbf{A}$: $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \cdots \geq \sigma_{p\wedge m}(\mathbf{A})$. $\sigma_{\max}(\mathbf{A})$ *and* $\sigma_{\min}(\mathbf{A})$ *for the largest and smallest singular values of* $\mathbf{A}$. *Let* $\|\mathbf{A}\| = \sigma_{\max}(\mathbf{A})$ *be the spectral norm,* $\|\mathbf{A}\|_*$ *be the nuclear norm and* $\|\mathbf{A}\|_{\mathrm{F}}$ *be the Frobenius norm. Denote* $\mathbf{A}_{*j}$ *and* $\mathbf{A}_{i*}$ *as the jth column vector and the ith row vector of* $\mathbf{A}$. $\mathbf{I}_p$ *denotes the* $p \times p$ *identity matrix. For vectors* $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m$ *in* $\mathbb{R}^p$, *denote* $[\boldsymbol{a}_1\ \boldsymbol{a}_2\ \ldots\ \boldsymbol{a}_m] \in \mathbb{R}^{p\times m}$ *the matrix with* $\boldsymbol{a}_j$ *being its jth column.* $\mathbf{1}(S)$ *is the indicator function, which is unity when the statement S is true.*

2. APPROXIMATE ESTIMATOR AND ESTIMATION

In this paper, the approximate estimator $\widehat{\mathbf{\Gamma}}_{\tau,\delta}$ is assumed to satisfy

$$0 \leq L_\tau(\widehat{\mathbf{\Gamma}}_{\tau,\delta}) - L_\tau(\widehat{\mathbf{\Gamma}}_\tau) \leq \delta, \tag{2.1}$$

for some $\delta \geq 0$, where L_τ is the empirical loss in (1.3).

Section 2.1 presents an algorithm for computing $\widehat{\mathbf{\Gamma}}_{\tau,\delta}$ with a characterization of its optimization error δ. Section 2.2 describes how to estimate the factors and loadings from $\widehat{\mathbf{\Gamma}}_{\tau,\delta}$. Section 2.3 discusses the selection of the tuning parameter λ.

2.1. The Coefficient Matrix

The proposed estimation procedure combines the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) of Beck and Teboulle (2009) and the smoothing technique of Nesterov (2005). A similar approach has been applied in mean regression with penalties encouraging complex structural sparsity (Chen et al., 2012). This approach is more scalable to higher dimensions than the semidefinite programming (Fazel, Hindi, and Boyd, 2001; Srebro, Rennie, and Jaakkola, 2005), and is more stable than the nonconvex reformulation (Rennie and Srebro, 2005; Weimer et al., 2008a, 2008b). See Ciliberto, Stamos, and Pontil (2017) for a recent account.

Specifically, we first smooth the nondifferentiable $\hat{Q}_\tau$ in (1.4) and obtain a surrogate $\hat{Q}_{\tau,\kappa}$ parameterized by a smoothing parameter κ (Nesterov, 2005), whose gradient $\nabla\hat{Q}_{\tau,\kappa}$ is globally Lipschitz with constant $Lip = (\kappa m^2 n^2)^{-1}\|\mathbf{X}\|^2$, where $\mathbf{X} = [\boldsymbol{X}_1 \ldots \boldsymbol{X}_n]^\top \in \mathbb{R}^{n\times p}$ is the design matrix. As $\kappa \to 0$, $\hat{Q}_{\tau,\kappa}$ converges to $\hat{Q}_\tau$. Next, the FISTA (Beck and Teboulle, 2009) is applied on $\hat{Q}_{\tau,\kappa}$ with the step size Lip^{-1}. The procedure is summarized below, with comprehensive details provided in Section A.1.

https://doi.org/10.1017/S0266466620000304 Published online by Cambridge University Press

Step 1: Given $\kappa > 0$, we smooth $\hat{Q}_\tau$ to obtain $\hat{Q}_{\tau,\kappa}$;

Step 2: Given $\lambda > 0$ and an initial estimator $\mathbf{\Gamma}_{\tau,0}$, for each $t = 1, 2, \ldots, T$, apply the FISTA step on the minimization problem $\min_{\mathbf{S}\in\mathbb{R}^{p\times m}}\{\tilde{L}_\tau(\mathbf{S}) \stackrel{\text{def}}{=} \hat{Q}_{\tau,\kappa}(\mathbf{S}) + \lambda\|\mathbf{S}\|_*\}$ with the step size $\kappa m^2 n^2 \|\mathbf{X}\|^{-2}$. Return the last iterate $\mathbf{\Gamma}_{\tau,T}$.

A suitable choice of κ should result from the tradeoff between the smoothing error and the speed of convergence. In particular, a smaller κ leads to a smaller smoothing error, while slowing down the convergence. Corollary 3.4 will provide a theoretical result on the selection of κ and T. The next theorem shows that $\mathbf{\Gamma}_{\tau,T}$ is an approximate estimator in the sense of (2.1).

THEOREM 2.1. *Recall that $\widehat{\mathbf{\Gamma}}_\tau$ is the solution for minimizing (1.3) and let $\mathbf{\Gamma}_{\tau,\infty} \stackrel{\text{def}}{=} \lim_{T\to\infty} \mathbf{\Gamma}_{\tau,T} = \arg\min_{\mathbf{S}}\{\tilde{L}_\tau(\mathbf{S}) = \widehat{Q}_{\tau,\kappa}(\mathbf{S}) + \lambda\|\mathbf{S}\|_*\}$. Then for any T, $\widehat{\mathbf{\Gamma}}_{\tau,\delta} = \mathbf{\Gamma}_{\tau,T}$ satisfies (2.1) with*

$$\delta = \delta(T, n, \kappa, \tau, \mathbf{X}) = \frac{3\kappa mn(\tau \vee \{1-\tau\})^2}{2} + \frac{2\|\mathbf{\Gamma}_{\tau,0} - \mathbf{\Gamma}_{\tau,\infty}\|_{\mathrm{F}}^2}{(T+1)^2}\frac{\|\mathbf{X}\|^2}{\kappa m^2 n^2}. \tag{2.2}$$

See Section S.1.2 in the supplementary material for a proof for Theorem 2.1. The first term on the right-hand side of (2.2) is related to the smoothing error (Step 1), while the second term is related to the convergence rate of the FISTA algorithm (Step 2). The quantile level τ enters (2.2) by the term $\tau \vee \{1-\tau\}$, which increases when τ approaches the boundary of $[0, 1]$.

2.2. The Factors number, Factors, and Loadings

After factorizing the true coefficient matrix $\mathbf{\Gamma}_\tau = \Phi_\tau \Psi_\tau$, we can compute the factors $f_k^\tau(\boldsymbol{X}) = \Phi_{\tau,*k}^\top \boldsymbol{X}_i$ and loadings $\Psi_{\tau,kj}$ for $j = 1, \ldots, m$ and $k = 1, \ldots, r_\tau$. However, the decomposition $\mathbf{\Gamma}_\tau = \Phi_\tau \Psi_\tau$ is not unique. In particular, for any invertible matrix $\mathbf{P} \in \mathbb{R}^{r\times r}$, we have $\Phi_\tau \Psi_\tau = \Phi_\tau \mathbf{P}\mathbf{P}^{-1}\Psi_\tau$. Therefore, to fix such a matrix $\mathbf{P}$, we apply the constraint in eqn. (2.14) on p. 28 of Reinsel and Velu (1998): namely, the singular value decomposition $\mathbf{\Gamma}_\tau = \mathbf{U}_\tau \mathbf{D}_\tau \mathbf{V}_\tau^\top$, where $\mathbf{U}_\tau^\top \mathbf{U}_\tau = \mathbf{V}_\tau^\top \mathbf{V}_\tau = \mathbf{I}_{r_\tau}$. We set

$$\Psi_\tau = \mathbf{V}_\tau \text{ and } \Phi_\tau = \mathbf{D}_\tau^\top \mathbf{U}_\tau^\top. \tag{2.3}$$

In practice, using the singular value decomposition $\widehat{\mathbf{\Gamma}}_{\tau,\delta} = \tilde{\mathbf{U}}_\tau \tilde{\mathbf{D}}_\tau \tilde{\mathbf{V}}_\tau^\top$, the factors and loadings can be estimated similar to (2.3):

$$\begin{aligned} \hat{f}_k^\tau(\boldsymbol{X}_i) &= \tilde{\sigma}_k \tilde{\mathbf{U}}_{\tau,*k}^\top \boldsymbol{X}_i, \\ \hat{\Psi}_\tau &= \tilde{\mathbf{V}}_\tau, \end{aligned} \tag{2.4}$$

where $\tilde{\sigma}_k$ is the kth largest singular value of $\widehat{\mathbf{\Gamma}}_{\tau,\delta}$.

https://doi.org/10.1017/S0266466620000304 Published online by Cambridge University Press

The nuclear norm penalty in our loss function (1.3) shrinks most singular values to 0, so that $\tilde{r}_\tau = \max\{k : \tilde{\sigma}_k > 0\}$ is small. Note that when r_τ is not small, the proposed method can still work if the importance (measured by the magnitude of singular values) of the latter factors is small; see Remark 3.5 for details. In practice, we estimate the number of factors r_τ by $\hat{r}_\tau = \max\{k : \tilde{\sigma}_k > \epsilon\}$ with, for example, $\epsilon = 10^{-10}$.

2.3. Tuning

We propose two ways to select λ. The first method is based on simulation. In particular, define the random variable

$$\Lambda_\tau \stackrel{\text{def}}{=} (nm)^{-1}\|\mathbf{X}^\top \widetilde{\mathbf{W}}_\tau\|, \tag{2.5}$$

where $(\widetilde{W}_\tau)_{ij} = \mathbf{1}(U_{ij} \le \tau) - \tau$, and $\{U_{ij}\}$ are i.i.d. uniform (0,1) random variables for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, independent from $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. The random variable Λ_τ is pivotal conditioning on $\mathbf{X}$, as it does not depend on $\boldsymbol{\Gamma}_\tau$. The formula in (2.5) arises from the subgradient $\nabla \widehat{Q}_\tau(\boldsymbol{\Gamma}_\tau) = (nm)^{-1}\mathbf{X}^\top \widetilde{\mathbf{W}}_\tau$. Set

$$\lambda = \lambda_\tau = 2 \cdot \Lambda_\tau(1 - \eta|\mathbf{X}), \tag{2.6}$$

where $\Lambda_\tau(1-\eta|\mathbf{X})$ is the $(1-\eta)$-quantile of Λ_τ conditional on $\mathbf{X}$. The level η should be close to 0, and we take $\eta = 0.1$. The constant 2 in (2.6) is mainly for the convenience of theoretical development, and it can be replaced by any constant greater than one. In practice, when n is large enough, the constant has little effect on the estimated number of factors, which is shown in our empirical study; see the left panel of Figure 5.1.

The value of λ given by (2.6) can sometimes underestimate r_τ. Alternatively, given a validation dataset $\{(\check{\boldsymbol{X}}_i, \check{\boldsymbol{Y}}_i)\}_{i=1}^{n_v}$ that is independent from the data used to compute $\widehat{\boldsymbol{\Gamma}}_{\tau,\delta}$, we propose choosing λ by minimizing the penalized validation loss:

$$\min_\lambda \left\{ (mn_v)^{-1} \sum_{i=1}^{n_v}\sum_{j=1}^{m} \rho_\tau\big(\check{Y}_{ij} - \check{\boldsymbol{X}}_i^\top (\widehat{\boldsymbol{\Gamma}}^\lambda_{\tau,\delta})_{*j}\big) + \lambda \|\widehat{\boldsymbol{\Gamma}}^\lambda_{\tau,\delta}\|_* \right\}, \tag{2.7}$$

where we include the superscript λ in $\widehat{\boldsymbol{\Gamma}}_{\tau,\delta}$ to emphasize its dependence on λ. The minimizer of (2.7) can be found using a grid search and the value in (2.6) can be used as the upper bound of the grid points.

3. THEORY

Section 3.1 develops a general nonasymptotic bound on the Frobenius risk and prediction risk of $\widehat{\boldsymbol{\Gamma}}_{\tau,\delta}$. Next, we apply this bound on the estimator $\boldsymbol{\Gamma}_{\tau,T}$ proposed in Section 2.1 and then on the factors and loadings estimator in Section 3.2.

https://doi.org/10.1017/S0266466620000304 Published online by Cambridge University Press

3.1. Risk of the Approximate Estimator $\widehat{\mathbf{\Gamma}}_{\tau,\delta}$

The following assumptions are needed.

(A1) (Sampling setting) Samples $(\boldsymbol{X}_1, \boldsymbol{Y}_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{Y}_n)$ are i.i.d. copies of $(\boldsymbol{X}, \boldsymbol{Y})$ random vectors in $\mathbb{R}^{p+m}$ with $p, m \geq 3$. $F^{-1}_{Y|X}(\tau|\boldsymbol{x}) = \boldsymbol{x}^\top \mathbf{\Gamma}_{\tau,*j}$, where Y_j is the jth variable in $\boldsymbol{Y}$. Moreover, for each i, $\boldsymbol{W}_{\tau,i*} \overset{\text{def}}{=} \{(\mathbf{1}(Y_{ij} - \boldsymbol{X}_i^\top \mathbf{\Gamma}_{\tau,*j} \leq 0) - \tau)\}_{1\leq j\leq m}$ are i.i.d.

(A2) (Covariates) $\boldsymbol{X}$ is centered with covariance matrix $\mathbf{\Sigma}_X$. Assume the density function of $\boldsymbol{X}$ exists. Suppose $0 < \sigma_{\min}(\mathbf{\Sigma}_X) < \sigma_{\max}(\mathbf{\Sigma}_X) < \infty$, and there exist constants $B_p \geq 1$, $c_1, c_2 > 0$ such that $\|\boldsymbol{X}_i\|$ and the sample covariance matrix $\widehat{\mathbf{\Sigma}}_X = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$ satisfies

$$\mathrm{P}\big\{\sigma_{\min}(\widehat{\mathbf{\Sigma}}_X) \geq c_1\sigma_{\min}(\mathbf{\Sigma}_X), \sigma_{\max}(\widehat{\mathbf{\Sigma}}_X) \leq c_2\sigma_{\max}(\mathbf{\Sigma}_X), \|\boldsymbol{X}_i\| \leq B_p\big\} \geq 1 - \gamma_n, \tag{3.1}$$

for a sequence $\gamma_n \to 0$.

(A3) (Conditional densities) There exist constants $\bar{f} > 0, \underline{f}^\tau > 0$ and $\bar{f}' < \infty$ such that

$$\max_{j\leq m} \sup_{\boldsymbol{x},y} \big|f_{Y_j|\boldsymbol{X}}(y|\boldsymbol{x})\big| \leq \bar{f}, \ \max_{j\leq m} \sup_{\boldsymbol{x},y} \left|\frac{\partial}{\partial y_j} f_{Y_j|\boldsymbol{X}}(y|\boldsymbol{x})\right| \leq \bar{f}',$$
$$\min_{j\leq m} \inf_{\boldsymbol{x}} f_{Y_j|\boldsymbol{X}}(\boldsymbol{x}^\top \mathbf{\Gamma}_{\tau,*j}|\boldsymbol{x}) \geq \underline{f}^\tau,$$

where $f_{Y_j|\boldsymbol{X}}$ is the conditional density function of Y_j on $\boldsymbol{X}$.

The i.i.d. condition in Assumption (A1) is mainly for convenience to apply random matrix theory (see Remark S.2.6 in the supplementary material). Relaxation to m-dependence or weak dependence (Nze and Doukhan, 2004) is left for future research. In Assumption (A2), B_p can be assumed bounded by a constant (e.g., p. 2 of Maurer and Pontil, 2013 and Thm. 1 of Yousefi et al., 2018), but generally $B_p \asymp \sqrt{p}$ if each component of $\boldsymbol{X}$ is bounded almost surely. Eigenvalue bounds in (3.1) hold when the components in $\boldsymbol{X}$ have light tails; see, for example, Vershynin (2012). (A3) is a standard condition in the quantile regression literature (Belloni and Chernozhukov, 2011). Note that the constant $\underline{f}^\tau$ decreases when τ approaches 0 or 1.

The next lemma gives a bound for $n^{-1}\|\mathbf{X}^\top\mathbf{W}_\tau\|$, where $\mathbf{W}_\tau = \{W_{\tau,ij}\}_{ij} = \{(\mathbf{1}(Y_{ij} - \boldsymbol{X}_i^\top \mathbf{\Gamma}_{\tau,*j} \leq 0) - \tau)\}_{1\leq i\leq n, 1\leq j\leq m}$ is an $n \times m$ random matrix. The proof is provided in the supplementary material.

LEMMA 3.1. *Assume (A1) and (A2) hold.*

1. *For arbitrary $u > 1$, with probability greater than $1 - 3e^{-(u-1)(p+m)\log 8} - \gamma_n$,*

$$\frac{1}{n}\|\mathbf{X}^\top\mathbf{W}_\tau\| \leq C^*\sqrt{u\sigma_{\max}(\mathbf{\Sigma}_X)K(\tau)}\sqrt{\frac{p+m}{n}}, \tag{3.2}$$

https://doi.org/10.1017/S0266466620000304 Published online by Cambridge University Press

where $C^* = 4\sqrt{\frac{c_2}{C'}\log 8} \vee 1$, C' *and* c_2 *are the absolute constants given by S.4.3 in the supplementary material and Assumption (A2), and*

$$K(\tau) \stackrel{\text{def}}{=} \begin{cases} 0, & \tau = 0, 1; \\ \frac{2\tau - 1}{2\{\log\tau - \log(1-\tau)\}}, & \tau \in (0,1)\backslash\{1/2\}; \\ 1/4, & \tau = 1/2. \end{cases} \tag{3.3}$$

2. *With probability 1, for any* $0 < \eta < 1$,

$$\Lambda_\tau(1-\eta|\mathbf{X}) \le \bar{\lambda} := \frac{C^*}{m}\sqrt{\Big(1 - \frac{\eta - \gamma_n}{3(p+m)\log 8}\Big)\sigma_{\max}(\mathbf{\Sigma}_X)K(\tau)}\sqrt{\frac{p+m}{n}}, \tag{3.4}$$

where $\gamma_n \to 0$ *is defined in Assumption (A2).*

See Section S.2.1 in the supplementary material for a proof of Lemma 3.1. The constant $K(\tau)$ is the sub-Gaussian norm of the binary random variable $W_{\tau,ij}$ (Buldygin and Moskvichova, 2013, Thm. 3.1). Particularly, $K(\tau)$ is symmetric about $\tau = 1/2$ and is concave in $\tau \in [0,1]$ with the unique maximum 1/4 at $\tau = 1/2$. In addition, $K(\tau) \ge \tau(1-\tau)$ (eqn. (9) on p. 36 of Buldygin and Moskvichova, 2013). See Lem. 2.1 of Buldygin and Moskvichova (2013) for more on $K(\tau)$.

Next we present nonasymptotic bounds on the Frobenius and prediction risk of $\widehat{\mathbf{\Gamma}}_{\tau,\delta}$. The key ingredients in the proof are the convexity argument and a new tail probability bound for the empirical process $\mathbb{G}_n\{\widehat{Q}_\tau(\mathbf{\Gamma}_\tau + \mathbf{\Delta}) - \widehat{Q}_\tau(\mathbf{\Gamma}_\tau)\}$, that is based on a sharp bound on the spectral norm of a partial sum of random matrices (Maurer and Pontil, 2013; Tropp, 2011). Define

$$\nu_\tau(\delta) \stackrel{\text{def}}{=} \frac{3}{8}\frac{\underline{f}^\tau}{\bar{f}'} \inf_{\substack{\mathbf{\Delta}\in\mathcal{K}(\mathbf{\Gamma}_\tau;\delta) \\ \mathbf{\Delta}\neq 0}} \frac{\big(\sum_{j=1}^m \mathsf{E}[|\boldsymbol{X}_i^\top \mathbf{\Delta}_{*j}|^2]\big)^{3/2}}{\sum_{j=1}^m \mathsf{E}[|\boldsymbol{X}_i^\top \mathbf{\Delta}_{*j}|^3]}, \tag{3.5}$$

where $\mathcal{K}(\mathbf{\Gamma}_\tau;\delta)$ is a star-shaped set of matrices defined in Section A.2 of the Appendix. We note the fact that $\mathcal{K}(\mathbf{\Gamma}_\tau;\delta_1) \subset \mathcal{K}(\mathbf{\Gamma}_\tau;\delta_2)$ for all $0 \le \delta_1 \le \delta_2$.

THEOREM 3.2. *Assume that (A1)–(A3) hold, and* λ *satisfies*

$$2\Lambda_\tau(1-\eta|\mathbf{X}) \le \lambda \le 2\bar{\lambda}, \tag{3.6}$$

where $\bar{\lambda}$ *is defined in (3.4). Let* $\delta \le C\lambda m^{1/2}n^{-1/2}$ *for some constant* $C > 0$*, where* δ *is in (2.1). For some* $u > 1$*, assume that* $r_\tau = \operatorname{rank}(\mathbf{\Gamma}_\tau)$ *satisfies*

$$u\epsilon_{n,\tau,r} \stackrel{\text{def}}{=} u\frac{(384\sqrt{2} + 96C^*)}{\underline{f}^\tau \wedge 1}\sqrt{\frac{\sigma_{\max}(\mathbf{\Sigma}_X)\vee 1}{\sigma_{\min}(\mathbf{\Sigma}_X)\wedge 1}}\sqrt{\frac{r_\tau(m+p\vee B_p)(\log p + \log m)}{mn}} < \nu_\tau(2Cm^{1/2}n^{-1/2}), \tag{3.7}$$

https://doi.org/10.1017/S0266466620000304 Published online by Cambridge University Press

where C^* is the constant in (3.2). Then,

$$\|\widehat{\mathbf{\Gamma}}_{\tau,\delta} - \mathbf{\Gamma}_\tau\|_{L_2(P_X)} \le u\epsilon_{n,\tau,r} \tag{3.8}$$

with probability at least $1 - \eta - \gamma_n - 16(pm)^{1-u^2} - 3\exp\{-(p+m)\log 8\}$, where $\|\cdot\|^2_{L_2(P_X)} \overset{\text{def}}{=} m^{-1}E_{P_X}\|\cdot^\top \boldsymbol{X}_i\|_2^2$ is the prediction risk; in addition, the Frobenius risk satisfies

$$\|\widehat{\mathbf{\Gamma}}_{\tau,\delta} - \mathbf{\Gamma}_\tau\|_{\mathrm{F}} \le u\left(\frac{m}{\sigma_{\min}(\mathbf{\Sigma}_X)}\right)^{1/2}\epsilon_{n,\tau,r}. \tag{3.9}$$

See Section S.2.2 in the supplementary material for a proof of Theorem 3.2. The bounds in Theorem 3.2 hold regardless of which algorithm is used to compute $\widehat{\mathbf{\Gamma}}_{\tau,\delta}$, as long as the optimization error $\delta \le C\lambda m^{1/2}n^{-1/2}$. Corollary 3.4 provides an application of Theorem 3.2 on the estimator $\mathbf{\Gamma}_{\tau,T}$ proposed in Section 2.1.

The $\epsilon_{n,\tau,r}$ defined in (3.7) is not related to the optimization error δ. If p and m are fixed, then $\epsilon_{n,\tau,r} = O(n^{-1/2})$. In $\epsilon_{n,\tau,r}$, the quantity $r_\tau(p+m)$ can be viewed as the actual number of unknown parameters, and it has to be much smaller than n. The estimation risk is higher when τ is close to 0 or 1, as $1/\underline{f}^\tau$ increases.

The rate in (3.9) achieves the same (up to a constant) convergence rate as the multivariate regression for the mean (Koltchinskii et al., 2011; Negahban and Wainwright, 2011),[1] which is unimprovable up to a logarithmic factor in the minimax sense (Koltchinskii et al., 2011). Hence, we conjecture that our bounds are unimprovable in their dependence on $p, m,$ and n, and $\widehat{\mathbf{\Gamma}}_{\tau,\delta}$ achieves the same rate as the exact estimator $\widehat{\mathbf{\Gamma}}_\tau$.

Remark 3.3 (Comment on the growth condition (3.7)). In Theorem 3.2, the growth condition (3.7) guarantees that $Q_\tau(\mathbf{S}+\mathbf{\Delta}) - Q_\tau(\mathbf{S})$ is minorized by a quadratic function of $\mathbf{\Delta}$, for $\mathbf{\Delta}$ inside a well-behaved set. Moreover, it can be easily seen that $\nu_\tau(\delta_1) < \nu_\tau(\delta_2)$ for all $\delta_1 > \delta_2$ from the definition of $\nu_\tau(\delta)$ in (3.5). Section S.3.1 in the supplementary material discusses (3.7) in detail.

Corollary 3.4 below proves that the estimator $\mathbf{\Gamma}_{\tau,T}$ in Section 2.1 achieves the bound in Theorem 3.2, provided that the smoothing parameter κ and the number of iterations T satisfy certain conditions.

COROLLARY 3.4. *Assume the conditions of Theorem 3.2 and $C = 1$ in (3.7). Suppose in addition that the true coefficient matrix $\mathbf{\Gamma}_\tau$ satisfies $\|\mathbf{\Gamma}_\tau\|_{\mathrm{F}}^2 \le C_\tau pm$ for some constant $C_\tau > 0$, and $B_p r(p+m)(\log p + \log m) = o(n)$. Let the initial estimator be $\mathbf{\Gamma}_{\tau,t=0} = 0$ in the algorithm in Section 2.1, and that*

[1] Their regression problem (in the mean) is analogous to our multivariate quantile regression setting by adjusting their n to mn. See Ex. 1 on p. 1075 of Negahban and Wainwright (2011). Note also that $\|\mathbf{\Delta}\|_{\mathrm{F}} \le (m/\sigma_{\min}(\mathbf{\Sigma}_X))^{1/2}\|\mathbf{\Delta}\|_{L_2(P_X)}$ for a well-behaved $\mathbf{\Delta}$ in our setting.

https://doi.org/10.1017/S0266466620000304 Published online by Cambridge University Press

$$\kappa \le \frac{\lambda}{3m^{1/2}n^{3/2}\{\tau \vee (1-\tau)\}^2}, \quad \textit{(smoothing parameter)} \tag{3.10}$$

$$T = T(\kappa) \ge \frac{4C_\tau^{1/2}p^{1/2}m^{1/2}\|\mathbf{X}\|}{\lambda^{1/2}\kappa^{1/2}m^{5/4}n^{3/2}} - 1, \quad \textit{(number of iterations)} \tag{3.11}$$

then for $u > 1$, (3.8) and (3.9) hold with $\widehat{\mathbf{\Gamma}}_{\tau,\delta} = \mathbf{\Gamma}_{\tau,T}$ with probability at least $1 - 2(\eta + \gamma_n + 16(pm)^{1-u^2} + 3\exp\{-(p+m)\log 8\})$.

See Section A.3 for a proof of Corollary 3.4. The main component of the proof is to verify that the optimization error of $\mathbf{\Gamma}_{\tau,T}$ is less than $\lambda m^{1/2}n^{-1/2}$ under (3.10) and (3.11). The conditions in both (3.10) and (3.11) can be improved, and we adopt the current form for a transparent exposition. Although it is theoretically appealing, selecting κ based on (3.10) may be too conservative, and a larger κ often yields a better performance.

Remark 3.5 (Many factors). When $\mathbf{\Gamma}_\tau$ is not exactly small in rank, that is, when r_τ is not a fixed number and possibly $r_\tau = \min\{p, m\}$. In this case, we can still bound $\widehat{\mathbf{\Gamma}}_{\tau,\delta} - \mathbf{\Gamma}_\tau$ using the method of Negahban et al. (2012). See Section S.3.2 of the supplementary material for more details.

3.2. Estimation Risk for Factors and Loadings

The estimation risk for the factors and loadings will be stated in terms of the Frobenius risk $\|\widehat{\mathbf{\Gamma}}_{\tau,\delta} - \mathbf{\Gamma}_\tau\|_{\mathrm{F}}$. Theorem 3.2 can be applied to find the explicit rate.

First by Mirsky's theorem (see, e.g., Thm. 4.11 on p. 204 of Stewart and Sun, 1990), the singular values can be consistently estimated if $\mathbf{\Gamma}_\tau$ can.

LEMMA 3.6. *If $\widehat{\mathbf{\Gamma}}_{\tau,\delta}$ satisfies (2.1), then*

$$\sum_{j=1}^{p\wedge m}\left\{\sigma_j(\widehat{\mathbf{\Gamma}}_{\tau,\delta}) - \sigma_j(\mathbf{\Gamma}_\tau)\right\}^2 \le \|\widehat{\mathbf{\Gamma}}_{\tau,\delta} - \mathbf{\Gamma}_\tau\|_{\mathrm{F}}^2. \tag{3.12}$$

Next, we consider the factors and loadings.

THEOREM 3.7. *If the nonzero singular values of matrix $\mathbf{\Gamma}_\tau$ are distinct, then with the choice of $\hat{\Psi}_\tau$ and $\hat{f}_k^\tau(\boldsymbol{X}_i)$ in (2.4),*

$$1 - |(\hat{\Psi}_\tau)_{*j}^\top(\Psi_\tau)_{*j}| \le \frac{2(2\|\mathbf{\Gamma}_\tau\| + \|\widehat{\mathbf{\Gamma}}_{\tau,\delta} - \mathbf{\Gamma}_\tau\|_{\mathrm{F}})\|\widehat{\mathbf{\Gamma}}_{\tau,\delta} - \mathbf{\Gamma}_\tau\|_{\mathrm{F}}}{\min\{\sigma_{j-1}^2(\mathbf{\Gamma}_\tau) - \sigma_j^2(\mathbf{\Gamma}_\tau), \sigma_j^2(\mathbf{\Gamma}_\tau) - \sigma_{j+1}^2(\mathbf{\Gamma}_\tau)\}}. \tag{3.13}$$

If, in addition, let the SVDs $\mathbf{\Gamma}_\tau = \mathbf{U}_\tau\mathbf{D}_\tau\mathbf{V}_\tau^\top$ and $\widehat{\mathbf{\Gamma}}_{\tau,\delta} = \tilde{\mathbf{U}}_\tau\tilde{\mathbf{D}}_\tau\tilde{\mathbf{V}}_\tau^\top$, suppose $(\tilde{\mathbf{U}}_\tau)_{*j}^\top(\mathbf{U}_\tau)_{*j} \ge 0$, then

https://doi.org/10.1017/S0266466620000304 Published online by Cambridge University Press

$$\left|\hat{f}_k^\tau(\boldsymbol{X}_i) - f_k^\tau(\boldsymbol{X}_i)\right| \leq \|\boldsymbol{X}_i\| \left(\|\widehat{\boldsymbol{\Gamma}}_{\tau,\delta} - \boldsymbol{\Gamma}_\tau\|_{\mathrm{F}} + 2\sigma_k(\boldsymbol{\Gamma}_\tau) \sqrt{\frac{(2\|\boldsymbol{\Gamma}_\tau\| + \|\widehat{\boldsymbol{\Gamma}}_{\tau,\delta} - \boldsymbol{\Gamma}_\tau\|_{\mathrm{F}}) \|\widehat{\boldsymbol{\Gamma}}_{\tau,\delta} - \boldsymbol{\Gamma}_\tau\|_{\mathrm{F}}}{\min\{\sigma_{k-1}^2(\boldsymbol{\Gamma}_\tau) - \sigma_k^2(\boldsymbol{\Gamma}_\tau), \sigma_k^2(\boldsymbol{\Gamma}_\tau) - \sigma_{k+1}^2(\boldsymbol{\Gamma}_\tau)\}}} \right). \tag{3.14}$$

See S.2.4 in the supplementary material for a proof of Theorem 3.7. The proof is based on a type of Davis-Kahan inequality in Yu, Wang, and Samworth (2015).

Remark 3.8 (Repeated singular values). Theorem 3.7 requires that the singular values of $\boldsymbol{\Gamma}_\tau$ are distinct. If there are repeated singular values, then we can only obtain a bound for the "canonical angle" (see, e.g., Yu et al., 2015) between subspaces, where each subspace is spanned by the singular vectors associated with the repeated singular values.

4. SIMULATION

The performance of the estimator of the quantile factor model in Section 2 will be evaluated via Monte Carlo simulations in this section. Section 4.1 presents the simulated Frobenius risk. The performance on estimating r_τ is first studied with i.i.d. data in Section 4.2 and then with time series data in Section 4.3.

4.1. Estimation Error

Given two distinct matrices $\mathbf{S}_1, \mathbf{S}_2$ with *nonnegative* entries, $\mathrm{rank}(\mathbf{S}_1) = r_1$ and $\mathrm{rank}(\mathbf{S}_2) = r_2$, the response variables are simulated from the two-piece normal model (Wallis, 2014):

$$Y_{ij} = G_\sigma^{-1}(U_{ij}) \boldsymbol{X}_i^\top \left((\mathbf{S}_1)_{*j} \mathbf{1}\{U_{ij} \leq 0.5\} + (\mathbf{S}_2)_{*j} \mathbf{1}\{U_{ij} > 0.5\} \right), \quad i = 1, \ldots, n = 500;\ j = 1, \ldots, m = 300, \tag{4.1}$$

where U_{ij} are i.i.d. uniform (0,1) independent of $\boldsymbol{X}_i$; G_σ is the cdf of $\mathcal{N}(0, \sigma^2)$. $\boldsymbol{X}_i \in \mathbb{R}^p$ follows a multivariate uniform (0,1) distribution with covariance matrix $\boldsymbol{\Sigma}$ in which $\boldsymbol{\Sigma}_{ij} = 0.1 * 0.8^{|i-j|}$ for $i, j \in \{1, \ldots, p\}$ with $p = 300$. Covariates $\boldsymbol{X}_i$ are simulated by the method of Falk (1999). The number of simulation repetitions is 500.

Because the elements in $\boldsymbol{X}_i$ are non-negative, the conditional quantile function $q_j(\tau|\boldsymbol{x})$ of Y_{ij} on $\boldsymbol{x}$ for the distribution of Y_{ij} is

$$q_j^l(\tau|\boldsymbol{x}) = \boldsymbol{x}^\top G_\sigma^{-1}(\tau) \left(\mathbf{S}_1 \mathbf{1}\{\tau \leq 0.5\} + \mathbf{S}_2 \mathbf{1}\{\tau > 0.5\} \right) \stackrel{\mathrm{def}}{=} \boldsymbol{x}^\top \boldsymbol{\Gamma}_{\tau,*j}. \tag{4.2}$$

It follows that $\boldsymbol{\Gamma}_\tau = G_\sigma^{-1}(\tau)\mathbf{S}_1$ for $\tau \leq 1/2$ and $\boldsymbol{\Gamma}_\tau = G_\sigma^{-1}(\tau)\mathbf{S}_2$ for $\tau > 1/2$. In particular, $\boldsymbol{\Gamma}_{1/2} = 0$ because $G_\sigma^{-1}(1/2) = 0$ for all $\sigma > 0$. This implies the 1/2-quantile of all responses are 0, so this is the least interesting case and is not investigated. To select $\mathbf{S}_1$ and $\mathbf{S}_2$, we first fix $r_1 = \mathrm{rank}(\mathbf{S}_1) = 2$. While for $\mathbf{S}_2$:

https://doi.org/10.1017/S0266466620000304 Published online by Cambridge University Press

TABLE 4.1. Averaged Frobenius risk with standard deviations. "Or." denotes the oracle estimator, which is estimated under the knowledge of true rank. The numbers in parentheses are standard deviation of the risks

τ	0.05	0.1	0.2	0.8	0.9	0.95
			$\sigma = 0.5$			
Symmetric	60.995	48.746	34.302	33.973	48.375	60.604
	(0.253)	*(0.227)*	*(0.209)*	*(0.202)*	*(0.217)*	*(0.247)*
Symmetric Or.	57.261	44.926	30.006	29.853	44.735	57.007
	(0.191)	*(0.152)*	*(0.116)*	*(0.118)*	*(0.152)*	*(0.184)*
Asymmetric	60.978	48.724	34.289	60.487	85.997	108.310
	(0.263)	*(0.220)*	*(0.207)*	*(0.539)*	*(0.567)*	*(0.820)*
Asymmetric Or.	57.239	44.911	30.002	54.922	80.583	102.663
	(0.202)	*(0.164)*	*(0.120)*	*(0.744)*	*(0.464)*	*(0.572)*
			$\sigma = 1$			
Symmetric	118.245	93.419	64.289	63.634	92.519	117.365
	(0.570)	*(0.420)*	*(0.387)*	*(0.382)*	*(0.372)*	*(0.438)*
Symmetric Or.	113.636	88.781	58.913	58.593	88.365	113.099
	(0.427)	*(0.338)*	*(0.238)*	*(0.221)*	*(0.301)*	*(0.378)*
Asymmetric	118.259	93.434	64.291	120.338	170.904	217.185
	(0.530)	*(0.412)*	*(0.380)*	*(1.151)*	*(1.273)*	*(1.547)*
Asymmetric Or.	113.647	88.788	58.911	108.754	161.303	205.371
	(0.387)	*(0.308)*	*(0.224)*	*(0.711)*	*(0.929)*	*(1.188)*

I. Symmetric model: $\mathbf{S}_2^{Sym}$ with $r_2^{Sym} = \text{rank}(\mathbf{S}_2^{Sym}) = 2$;
II. Asymmetric model: $\mathbf{S}_2^{Asym}$ with $r_2^{Asym} = \text{rank}(\mathbf{S}_2^{Asym}) = 6$.

The singular values of each of the matrices $\mathbf{S}_1$, $\mathbf{S}_2^{Sym}$, and $\mathbf{S}_2^{Asym}$ are distinct. The details on selecting these matrices are in Section S.3.3 of the supplementary material.

The algorithm in Section 2.1 (or Algorithm A.1) is applied with $\tau = 5\%$, 10%, 20%, 80%, 90%, and 95% to compute the estimator $\widehat{\mathbf{\Gamma}}^l_{\tau,\delta}$ for $\mathbf{\Gamma}^l_\tau$, where $l \in \{Sym, Asym\}$. We set $\kappa = 10^{-4}$ and stop the algorithm when the change in the loss function is less than 10^{-6}. The tuning parameter λ is selected by (2.6). We compare $\widehat{\mathbf{\Gamma}}^l_{\tau,\delta}$ with an oracle estimator, which is computed under the knowledge of the true r_τ.

The Frobenius risks are reported in Table 4.1. When the model variance is larger ($\sigma = 1$), the Frobenius risk is greater. The Frobenius risk is higher when τ is close to 0 or 1. Comparing the performance of the estimators under the symmetric model with the asymmetric model, we find similar performance when $\tau < 0.5$, because the r_τ of the two models are the same. However, when $\tau > 0.5$, the Frobenius risk

https://doi.org/10.1017/S0266466620000304 Published online by Cambridge University Press

under the asymmetric model is greater than that of the symmetric model under the same τ and σ, because $r_2^{Asym} = 6 > 2 = r_2^{Sym}$. These observations are consistent with Theorem 3.2. The Frobenius risk of the oracle estimator has similar patterns, but its risk and standard deviation are lower. The prediction risk has similar patterns as the Frobenius risk and is omitted for brevity.

4.2. Estimating the Number of Factors for i.i.d. Data

Using the same data generating process (4.1), this section shows the performance of estimating r_τ. We will focus on the estimation performance of two quantile levels $\tau = 0.2$ and 0.8 with $r_{\tau=0.2} = 2$ and $r_{\tau=0.8} = 6$.

The tuning parameter λ is selected by minimizing the *penalized validation loss* in (2.7) through the grid search, while using (2.6) as the maximal value of the grid. To compute the penalized validation loss (2.7) at a given λ, we independently generate data 150 times and for each time we compute $\mathbf{\Gamma}_{\tau,T}$. For each $\mathbf{\Gamma}_{\tau,T}$, a validation loss is computed with a validation data set of size 3,000 independent of the data used for estimation. Finally, we take an average of the 150 validation losses. The algorithm in Section 2.1 (Algorithm A.1) is applied with $T = 4{,}000$ iterations. We estimate r_τ using $\hat{r}_\tau = \max\{k : \hat{\sigma}_k > \epsilon\}$ with $\epsilon = 10^{-10}$, where $\hat{\sigma}_1 \geq \hat{\sigma}_2 \geq \cdots \geq \hat{\sigma}_{300}$ are the singular values of $\mathbf{\Gamma}_{\tau,T}$.

Figure 4.1 shows the relative frequency of $\hat{r}_\tau$ and the penalized validation loss as a function of λ. The penalized validation loss shows a quadratic shape and, thus, there exists a minimum for both $\tau = 0.2$ and 0.8. The minima of the two quantiles are essentially the same because the two quantile levels are symmetric about 0.5. For $\tau = 0.2$ with $r_{\tau=0.2} = 2$, the number of factors is correctly estimated over 50% of the time and, in the worst case scenario, the number of factors is misestimated by one. Ensuring accurate estimation for the case of $\tau = 0.8$ requires a smaller optimization error $\delta > 0$ than the case of $\tau = 0.2$, because $r_{\tau=0.8} = 6 > 2 = r_{\tau=0.2}$. Equation (2.2) in Theorem 2.1 implies that one should select a smaller κ and a greater T. Therefore, we take a slightly smaller $\kappa = 4 \times 10^{-7}$, but keep using $T = 4{,}000$. The bottom left panel of Figure 4.1 shows that the correct r_τ is identified almost 60% of the time and at worst it is misestimated by two. We note that with $\lambda = 0.0189$ selected by (2.6), only one factor is selected in unpublished simulations.

4.3. Estimating the Number of Factors for Dependent Data

In order to better study the empirical relevance of our method, we consider a time dependent design

$$\boldsymbol{X}_i = \mathbf{B}\boldsymbol{X}_{i-1} + \boldsymbol{\varepsilon}_i, \tag{4.3}$$

where $\boldsymbol{\varepsilon}_i$ is a multivariate Gaussian vector with mean zero and covariance $\Sigma_{\boldsymbol{\varepsilon}}$. The matrices $\mathbf{B}$ and $\Sigma_{\boldsymbol{\varepsilon}}$ are both 460×460, which are selected to imitate the temporal and cross-sectional dependence of the covariates in (5.1) of Section 5.

https://doi.org/10.1017/S0266466620000304 Published online by Cambridge University Press

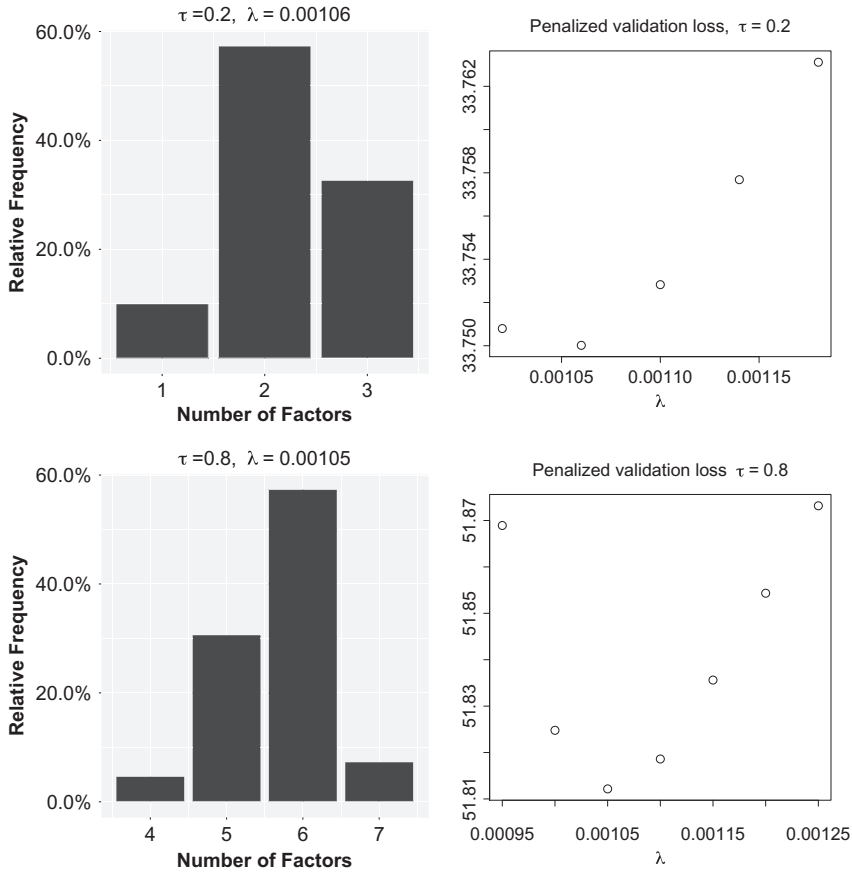


FIGURE 4.1. Relative frequency of the estimated number of factors with 150 Monte Carlo repetitions, and the plot for the penalized validation loss, $\tau = 0.2$ and 0.8. Data are generated as (4.1) where $\boldsymbol{X}_i$ is i.i.d. multivariate uniform (0,1) with positive correlations. $r_{\tau=0.2} = 2$ and $r_{\tau=0.8} = 6$. $\kappa = 6.66 \times 10^{-7}$ for $\tau = 0.2$ and $\kappa = 4 \times 10^{-7}$ for $\tau = 0.8$. $T = 4,000$ for both quantiles.

$\mathbf{B}$ is a sparse coefficient matrix, taken as the estimator obtained from fitting a vector autoregressive model on the covariates in (5.1) with the ℓ_1 norm penalty (Davis, Zang, and Zheng, 2016; Nicholson, Matteson, and Bien, 2017). Note that the four leading singular values of $\mathbf{B}$ are greater than one, so (4.3) is unstable. The details for selecting $\mathbf{B}$ and $\Sigma_{\boldsymbol{\epsilon}}$ are in Section S.5 in the supplementary material.

The response variable $\boldsymbol{Y}$ is generated as (4.1) with $\boldsymbol{X}_i$ in (4.3). Here, we fix $\text{rank}(\mathbf{S}_2) = 2$, while $\text{rank}(\mathbf{S}_1) = 2$ or 3. Under this design, the number of quantile factors $r_\tau = \text{rank}(\mathbf{S}_1)$ for all $\tau \in (0, 1)$ because the distribution of $\boldsymbol{X}$ in (4.3) is symmetric about the origin. To imitate the data in Section 5, we set $n = 2,765$ and $m = 230$. We select λ using (2.6), which yields $\lambda^* = 4.38 \times 10^{-3}$. Note that

https://doi.org/10.1017/S0266466620000304 Published online by Cambridge University Press

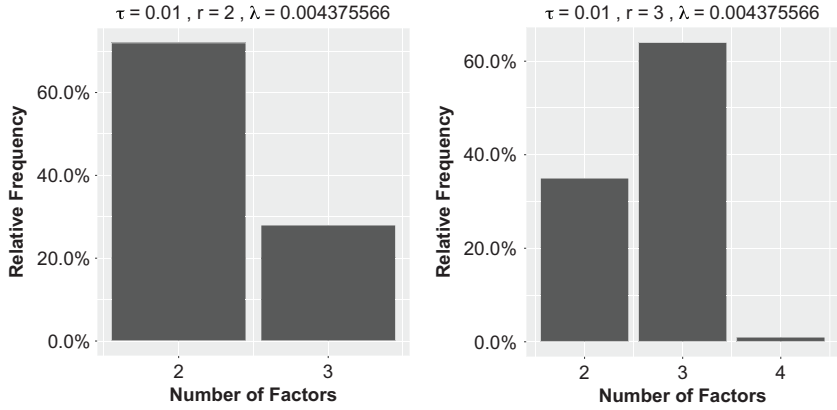


FIGURE 4.2. The distribution of the estimated number of factors for $r = 2$ (left) and $r = 3$ (right) with $\tau = 0.01$, with covariates simulated from (4.3). The distribution is obtained from 150 Monte Carlo repetitions. $(\kappa, T) = (1.5 \times 10^{-4}, 3,500)$. The λ is computed with the simulation method in (2.6).

(2.7) is feasible for selecting λ, but it is less computationally efficient. We set $\kappa = 1.5 \times 10^{-4}$ and $T = 3,500$ for the algorithm. We compute $\hat{r}_\tau = \max\{k : \hat{\sigma}_k > \epsilon\}$ with $\epsilon = 10^{-10}$ as in the last section.

Figure 4.2 shows the performance of estimating $r_{\tau=0.01}$. In the 150 Monte Carlo simulations, our method can correctly estimate r_τ in over 60% of the simulations. In the worst case scenario, r_τ is misestimated by one. Note that the quantile level $\tau = 0.01$ is more extreme than that in Section 4.2. This shows that our method performs well for extreme quantile levels and time series data, even though our theory is not designed for time series data. Additional simulations with a simpler AR(1) model can be found in Section S.6 of the supplementary material.

5. EMPIRICAL ANALYSIS: ESTIMATING THE SYSTEMIC RISK

In the aftermath of the financial crisis that began in 2007, governments and supervisory authorities have come to realize the need to quantify the impact of systemic risk on financial institutions, and this has motivated numerous studies; see Bisias et al. (2012) or Brunnermeier and Oehmke (2013) for a survey. Quantifying the impact of systemic risk is inherently a high dimensional problem, as a realistic model often includes hundreds or thousands of financial institutions. Unfortunately, due to the excessive computational cost, the models found in existing studies are low dimensional in nature. For example, Adrian and Brunnermeier (2016) estimated pairwise spillover effect between two institutions by conditioning on a set of variables; White et al. (2015) first obtained a proxy of the market shock and then estimated the impact of the market shock by performing bivariate vector autoregression between an institution and the market shock proxy. Our multitask quantile regression can fill this gap, as our method can easily scale up to

https://doi.org/10.1017/S0266466620000304 Published online by Cambridge University Press

hundreds of response variables. In addition, no proxy of the market shock needs to be precalculated, because our quantile factors naturally summarize the market information that is most relevant to the downside risk.

We analyze the same set of daily stock closing prices as White et al. (2015), with the same time frame from January 1, 2000 to August 6, 2010. The dataset is downloaded from Dr. Manganelli's personal website (http://www.simonemanganelli.org/Simone/Research.html). See Tab. 1 of White et al. (2015) for a detailed breakdown of the stocks by sector and country, as well as their averaged market value and leverage (the ratio of short and long term debt over common equity). There are $m = 230$ financial institutions. The daily log-returns of the closing prices are used and this results in $n = 2,765$.

Let $Y_{i,j}$ be the return of institution j at time i, where $j = 1, \ldots, m$ and $i = 1, \ldots, n$. For $\tau \in (0, 1)$, consider the quantile $q_j(\tau|\boldsymbol{X}_i) = \boldsymbol{X}_i^\top \boldsymbol{\Gamma}_{\tau,*j}$ for Y_j, where

$$\boldsymbol{X}_i = (|Y_{i-1,1}|, \ldots, |Y_{i-1,m}|, Y^-_{i-1,1}, \ldots, Y^-_{i-1,m})^\top \in \mathbb{R}^{2m} \tag{5.1}$$

and $Y^- \stackrel{\text{def}}{=} \max\{-Y, 0\}$. The choice of covariates allows for positive or negative lag returns to have unequal influence on the return today, as proposed by Engle and Manganelli (2004).

We estimate $\boldsymbol{\Gamma}_\tau$ using the algorithm in Section 2.1 (or Algorithm A.1) with two quantile levels $\tau = 0.01$ and 0.99. The algorithm is performed with $\kappa = 1.5 \times 10^{-4}$ and we stop the algorithm when the change in the loss function is less than 10^{-6}. The factors and loadings are estimated as (2.4) in Section 2.2. By using (2.6), $\lambda^* = 8.53 \times 10^{-3}$ for $\tau = 0.01$. The left panel of Figure 5.1 shows the estimated singular values for $\lambda \le \lambda^*$. The estimated number of factors for $\tau = 0.01$ is one even when we decrease λ^* by ten folds, so we set $\hat{r}_{\tau=0.01} = 1$. We use λ^* for both $\tau = 0.01$ and $\tau = 0.99$ by symmetry, and $\hat{r}_{\tau=0.99} = 1$.

The right panel of Figure 5.1 presents the estimated first factors at $\tau = 0.01$ and 0.99. Both first factors $f_1^{0.01}(X_i)$ and $f_1^{0.99}(X_i)$ are volatile and moving away from 0 at the end of 2008 and in the first quarter of 2009, which corresponds to the periods of the financial crisis.

The left panel of Figure 5.2 is the "tail to tail" plot, in which each point is a pair of factor loadings $((\hat{\Psi}_{0.01})_{1j}, (\hat{\Psi}_{0.99})_{1j})$ for the jth financial institution, $j = 1, \ldots, 230$, corresponding to the factors on the right panel of Figure 5.1. The values $((\hat{\Psi}_{0.01})_{1j}, (\hat{\Psi}_{0.99})_{1j})$ are all positive. The fact that they scatter around the 45 degree line suggests that the log-returns of these stocks are roughly equally associated to the two tail quantile factors, but the degree of association varies by institution.

The right panel of Figure 5.2 plots the institutions based on their averaged market value (x-axis) and leverage (y-axis), and the color represents the absolute magnitude of the $\tau = 0.01$ factor loading. A financial institution with a larger market value and a higher leverage tends to have a higher loading, as most red and yellow color points concentrate in the northeastern part of the figure. Hence, they are more vulnerable to the market shock, which is in line with the conclusion

https://doi.org/10.1017/S0266466620000304 Published online by Cambridge University Press

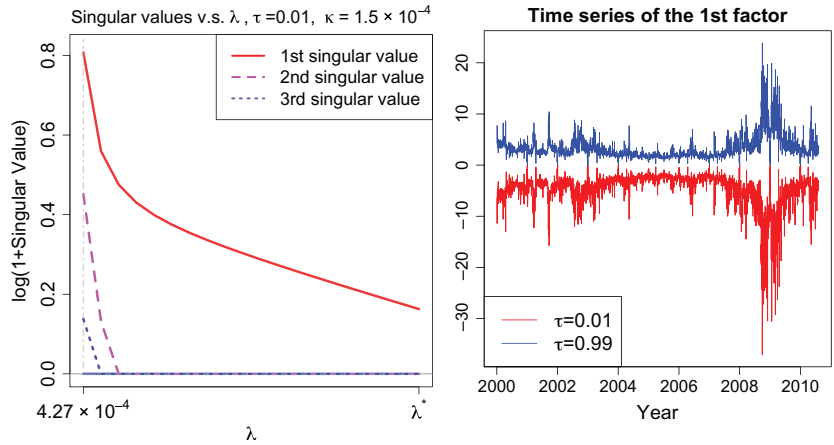


FIGURE 5.1. Left panel plots singular values versus λ. The right end point of x-axis is $\lambda^* = 8.53 \times 10^{-3}$ selected by the simulation method (2.6). Right panel: time series plot of the first factor of $\tau = 0.01$ (red) and $\tau = 0.99$ (blue).

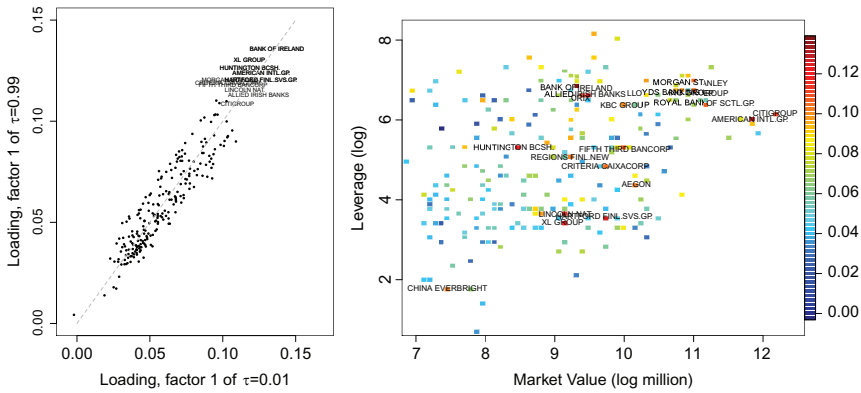


FIGURE 5.2. Left panel: tail to tail plot. Each point is a pair $((\hat{\Psi}_{0.01})_{1j}, (\hat{\Psi}_{0.99})_{1j})$ for stocks $j = 1, \ldots, 230$; Right panel: the plot of firms based on their averaged market value and leverage over the data period. The color scale corresponds to the magnitude of their $\tau = 0.01$ factor loading, and the firms with loading greater than 0.1 are labeled by name.

of White et al. (2015). The institutions that are more vulnerable to the market shock seem to form clusters and the geographical and financial properties of each cluster is left for future research.

Remark 5.1 (Extreme quantiles). As τ is close to zero or one, the $1/f^{\tau}$ increases and the bounds (3.8) and (3.9) in Theorem 3.2 become loose, suggesting the estimator may be inaccurate. Extreme quantile is often characterized through the low *extremal order* or *extremal rank* τn (Chernozhukov, 1999, 2005; Chernozhukov,

https://doi.org/10.1017/S0266466620000304 Published online by Cambridge University Press

Fernández-Val, and Kaji, 2017). In particular, if $n\tau \to c$ for some $c \geq 0$ as $\tau \to 0$ as $n \to \infty$ (respectively, $n(1-\tau) \to c$ as $\tau \to 1$ for the right extreme quantile, by symmetry we only discuss the left extreme quantile), classical asymptotic analysis breaks down if c is small or equal to 0, which is called the "extreme" quantile. Simulation study of asymptotic distribution in Chernozhukov (1999) suggests that $n\tau \geq 15$ might be large enough to be "nonextreme." In our application, extreme quantile issue may be mild as $n\tau = 27.65$ with $\tau = 1\%$. Complete analysis for the extreme quantiles under the multitask regression is left for future research.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a factor based multitask quantile regression model that is more flexible than the classical factor models, because our model allows different conditional quantiles of the response variables to be modeled with different factors. In addition, we propose an algorithm that is scalable to higher dimensions, making our model implementable in many high dimensional economic problems.

A future research direction may be to generalize the linear model in (1.2) to a more flexible model. Particularly, nonparametric models may be applied to model $f_k^\tau(\boldsymbol{x})$ by regarding it as an element of a sieve space, so $f_k^\tau(\boldsymbol{x})$ can be expressed by basis functions. Estimating this model can be done by using our algorithm in Section 2. Illustrations of this idea using the Chinese temperature data are presented in an earlier version of this paper (Chao, Härdle, and Yuan, 2015, Sect. 7). The theoretical analysis does not follow straightforwardly from the analysis in this paper, so it is left for future research.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit: https://doi.org/10.1017/S0266466620000304

REFERENCES

Adrian, T. & M.K. Brunnermeier (2016) CoVaR. *American Economic Review* 106(7), 1705–1741.

Ando, T. & R.S. Tsay (2011) Quantile regression models with factor-augmented predictors and information criterion. *Econometrics Journal* 14, 1–24.

Bai, J. & S. Ng (2010) Instrumental variable estimation in a data rich environment. *Econometric Theory* 26(6), 1577–1606.

Beck, A. & M. Teboulle (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences* 2(1), 183–202.

Belloni, A. & V. Chernozhukov (2011) ℓ_1-penalized quantile regression in high-dimensional sparse models. *Annals of Statistics* 39(1), 82–130.

Bisias, D., M. Flood, A. Lo, & S. Valavanis (2012) A Survey of Systemic Risk Analytics, Office of Financial Research Working Paper.

https://doi.org/10.1017/S0266466620000304 Published online by Cambridge University Press

Brunnermeier, M.K. & M. Oehmke (2013) Chapter 18—Bubbles, financial crises, and systemic risk. *Handbook of the Economics of Finance*, vol. 2, pp. 1221–1288. Elsevier. http://www.sciencedirect.com/science/article/pii/B9780444594068000184.

Buldygin, V.V. & K.K. Moskvichova (2013 The sub-Gaussian norm of a binary random variable. *Theory of Probability and Mathematical Statistics* 86, 33–49.

Bunea, F., Y. She, & M.H. Wegkamp (2011) Optimal selection of reduced rank estimators of high-dimensional matrices. *Annals of Statistics* 39(2), 1282–1309.

Chao, S.-K., W.K. Härdle, & C. Huang (2018) Multivariate factorizable expectile regression with application to fMRI data. *Computational Statistics & Data Analysis* 121, 1 – 19.

Chao, S.-K., W.K. Härdle, & M. Yuan (2015) Factorisable sparse tail event curves. Preprint, arXiv:1507.03833

Chen, L., J.J. Dolado, & J. Gonzalo (2019) Quantile factor models. Preprint arXiv:1911.02173.

Chen, X., Q. Lin, S. Kim, J.G. Carbonell, & E.P. Xing (2012) Smoothing proximal gradient method for general structured sparse regression. *The Annals of Applied Statistics* 6(2), 719–752.

Chernozhukov, V. (1999) Conditional extremes and near-extremes: Estimation, inference, and economic applications, PhD thesis, Department of Economics, Stanford University.

Chernozhukov, V. (2005) Extremal quantile regression. *Annals of Statistics* 33(2), 806–839.

Chernozhukov, V., I. Fernández-Val, & T. Kaji (2017) Extremal quantile regression: An overview. Preprint, arXiv:1612.06850.

Ciliberto, C., D. Stamos, & M. Pontil (2017) Reexamining low rank matrix factorization for trace norm regularization. Preprint, arXiv:1706.08934

Davis, R.A., P. Zang, & T. Zheng (2016) Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics* 25(4), 1077–1096.

Engle, R. & S. Manganelli (2004) CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics* 22, 367–381.

Falk, M. (1999) A simple approach to the generation of uniformly distributed random variables with prescribed correlation. *Communications in Statistics—Simulation and Computation* 28(3), 785–791.

Fazel, M., H. Hindi, & S.P. Boyd (2001) A rank minimization heuristic with application to minimum order system approximation. *Proceedings of the American Control Conference*, pp. 4734–4739.

Koenker, R. & G.S. Bassett (1978) Regression quantiles. *Econometrica* 46(1), 33–50.

Koenker, R. & K.F. Hallock (2001) Quantile regression. *Journal of Economic Perspectives* 15(4), 143–156.

Koenker, R. & S. Portnoy (1990) *M* estimation of multivariate regressions. *Journal of American Statistical Association* 85(412), 1060–1068.

Koltchinskii, V. (2011) *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer.

Koltchinskii, V., K. Lounici, & A.B. Tsybakov (2011) Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Annals of Statistics* 39(5), 2243–2794.

Maurer, A. & M. Pontil (2013) Excess risk bounds for multitask learning with trace norm regularization. *JMLR: Workshop and Conference Proceedings* 30, 1–22.

Maurer, A., M. Pontil, & B. Romera-Paredes (2016) The benefit of multitask representation learning. *Journal of Machine Learning Research* 17, 1–32.

Negahban, S.N., P. Ravikumar, M.J. Wainwright, & B. Yu (2012) A unified framework for high-dimensional analysis of *M*-estimators with decomposable regularizers. *Statistical Science* 27(4), 538–557.

Negahban, S.N. & M.J. Wainwright (2011) Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics* 39(2), 1069–1097.

Nesterov, Y. (2005) Smooth minimization of non-smooth functions. *Mathematical Programming* 103(1), 127–152.

Nicholson, W.B., D.S. Matteson, & J. Bien (2017) VARX-L: Structured regularization for large vector autoregressions with exogenous variables. *International Journal of Forecasting* 33(3), 627–651.

https://doi.org/10.1017/S0266466620000304 Published online by Cambridge University Press

Nze, P.A. & P. Doukhan (2004) Weak dependence: Models and applications to econometrics. *Econometric Theory* 20(6), 995–1045.
Reinsel, G.C. & R.P. Velu (1998) *Multivariate Reduced-Rank Regression*. Springer.
Rennie, J.D.M. & N. Srebro (2005) Fast maximum margin matrix factorization for collaborative prediction. *Proceedings of the 22nd International Conference on Machine Learning, ICML '05*. ACM, pp. 713–719.
Srebro, N., J.D.M. Rennie, & T.S. Jaakkola (2005) Maximum-margin matrix factorization. *Proceedings of the Conference on Neural Information Processing Systems*, pp. 1329–1336.
Stewart, G.W. & J.-G. Sun (1990) *Matrix Perturbation Theory*. Academic Press.
Stock, J.H. & M.W. Watson (2002) Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics* 20(2), 147–162.
Tropp, J.A. (2011) User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics* 12(4), 389–434.
Vershynin, R. (2012) How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability* 25(3), 655–686.
Wallis, K.F. (1999) Asymmetric density forecasts of inflation and the Bank of England's fan chart. *National Institute Economic Review* 167, 106–112.
Wallis, K.F. (2014) The two-piece normal, binormal, or double Gaussian distribution: Its origin and rediscoveries, *Statistical Science* 29(1), 106–112.
Weimer, M., A. Karatzoglou, Q.V. Le, & A.J. Smola (2008) COFI RANK—Maximum margin matrix factorization for collaborative ranking. In J.C. Platt, D. Koller, Y. Singer, & S.T. Roweis (eds.), *Advances in Neural Information Processing Systems 20*. Curran Associates, Inc., pp. 1593–1600. http://papers.nips.cc/paper/3359-cofi-rank-maximum-margin-matrix-factorization-for-collaborative-ranking.pdf
Weimer, M., A. Karatzoglou, & A. Smola (2008) Improving maximum margin matrix factorization. *Machine Learning* 72(3), 263–276.
White, H., T.-H. Kim, & S. Manganelli (2015) VAR for VaR: Measuring systemic risk using multivariate regression quantiles. *Journal of Econometrics* 187, 169–188.
Yousefi, N., Y. Lei, M. Kloft, M. Mollaghasemi, & G. Anagnastapolous (2018) Local Rademacher complexity-based learning guarantees for multi-task learning. *Journal of Machine Learning Research* 19(1), 1385–1431.
Yu, Y., T. Wang, & R.J. Samworth (2015) A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika* 102(2), 315–323.
Yuan, M., A. Ekici, Z. Lu, & R. Monteiro (2007) Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B* 69(3), 329–346.

APPENDIX

Details on the algorithm in Section 2.1 are provided in Section A.1. Section A.3 provides a proof of Corollary 3.4.

A.1. Details for the Numerical Procedure in Section 2.1

The FISTA of Beck and Teboulle (2009) is a popular method for optimizing the loss function that is a sum of two convex functions. However, a challenge here is that the subgradient of $\widehat{Q}_\tau(\mathbf{S})$ is not Lipschitz, so the FISTA algorithm may not be stable. To resolve this problem, we apply the method of Nesterov (2005) to find a "nice" surrogate for $\widehat{Q}_\tau(\mathbf{S})$, as will be shown below.

https://doi.org/10.1017/S0266466620000304 Published online by Cambridge University Press

Recall from (1.3) that the objective function to be minimized is

$$L_\tau(\mathbf{S}) = (mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\rho_\tau\big(Y_{ij} - \boldsymbol{X}_i^\top \mathbf{S}_{*j}\big) + \lambda\|\mathbf{S}\|_* = \widehat{Q}_\tau(\mathbf{S}) + \lambda\|\mathbf{S}\|_*. \tag{A.1}$$

First, we introduce the dual variables Θ_{ij}:

$$\widehat{Q}_\tau(\mathbf{S}) = \max_{\Theta_{ij}\in[\tau-1,\,\tau]} (mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\Theta_{ij}\big(Y_{ij} - \boldsymbol{X}_i^\top \mathbf{S}_{*j}\big). \tag{A.2}$$

See Section S.1.1 in the supplementary material for a proof of (A.2). Denote the matrix $\boldsymbol{\Theta} = (\Theta_{ij})$ for $i = 1,\ldots,n, j = 1,\ldots,m$. To smooth $\widehat{Q}_\tau(\mathbf{S})$, we consider a smooth approximation to $\widehat{Q}_\tau(\mathbf{S})$ as in eqn. (2.5) of Nesterov (2005):

$$\widehat{Q}_{\tau,\kappa}(\mathbf{S}) \stackrel{\text{def}}{=} \max_{\Theta_{ij}\in[\tau-1,\,\tau]} \Big\{(mn)^{-1}\tilde{Q}_\tau(\mathbf{S},\boldsymbol{\Theta}) - \frac{\kappa}{2}\|\boldsymbol{\Theta}\|_{\mathrm{F}}^2\Big\}, \tag{A.3}$$

where $\tilde{Q}_\tau(\mathbf{S},\boldsymbol{\Theta}) \stackrel{\text{def}}{=} \sum_{i=1}^{n}\sum_{j=1}^{m}\Theta_{ij}\big(Y_{ij} - \boldsymbol{X}_i^\top \mathbf{S}_{*j}\big)$, and $\kappa > 0$ is a smoothing regularization constant depending on m, n and the desired accuracy. When $\kappa \to 0$, $\widehat{Q}_{\tau,\kappa}(\mathbf{S})$ converges to $\widehat{Q}_\tau(\mathbf{S})$. $\widehat{Q}_{\tau,\kappa}(\mathbf{S})$ defined in (A.3) has a gradient that is Lipschitz:

$$\nabla\widehat{Q}_{\tau,\kappa}(\mathbf{S}) \stackrel{\text{def}}{=} -(mn)^{-1}\mathbf{X}^\top[[(\kappa mn)^{-1}(\mathbf{Y} - \mathbf{XS})]]_\tau, \tag{A.4}$$

where $\mathbf{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2\ \ldots\ \boldsymbol{X}_n]^\top$, $[[\mathbf{A}]]_\tau = ([[A_{ij}]]_\tau)$ performs component-wise truncation on a real matrix $\mathbf{A}$ to the interval $[\tau - 1, \tau]$; in particular,

$$[[A_{ij}]]_\tau = \begin{cases} \tau, & \text{if } A_{ij} \geq \tau; \\ A_{ij}, & \text{if } \tau - 1 < A_{ij} < \tau; \\ \tau - 1, & \text{if } A_{ij} \leq \tau - 1. \end{cases}$$

Observe that (A.4) is similar to the subgradient $-\mathbf{X}\{\tau - \mathbf{1}(\mathbf{Y} - \mathbf{XS} \leq 0)\}$ of $\widehat{Q}_\tau(\mathbf{S})$, where the operator $\tau - \mathbf{1}(\cdot \leq 0)$ applies component-wise to the matrix $\mathbf{Y} - \mathbf{XS}$. The major difference lies in the fact that (A.4) replaces the discrete non-Lipschitz $\tau - \mathbf{1}(\mathbf{Y} - \mathbf{XS} \leq 0)$ with a Lipschitz function $[[\kappa^{-1}(\mathbf{Y} - \mathbf{XS})]]_\tau$.

Now, we replace the optimization problem involving $L_\tau(\mathbf{S})$ in (A.1) by the one involving

$$\widetilde{L}_\tau(\mathbf{S}) \stackrel{\text{def}}{=} \widehat{Q}_{\tau,\kappa}(\mathbf{S}) + \lambda\|\mathbf{S}\|_*, \tag{A.5}$$

where we recall the definition of $\widehat{Q}_{\tau,\kappa}(\mathbf{S})$ in (A.3). We may apply FISTA of Beck and Teboulle (2009) for minimizing (A.5). Let $S_\lambda(\cdot)$ be the proximity operator on $\mathbb{R}^{p\times m}$:

$$S_\lambda(\mathbf{S}) \stackrel{\text{def}}{=} \mathbf{U}(\mathbf{D} - \lambda\mathbf{I}_{p\times m})_+\mathbf{V}^\top, \tag{A.6}$$

where $\mathbf{I}_{p\times m}$ is the $p \times m$ rectangular identity matrix with the main diagonal elements equal to 1, and the SVD $\mathbf{S} = \mathbf{UDV}^\top$. See Theorem S.4.2 in the supplementary material for more details for the proximity operator.

Specific steps are summarized in Algorithm A.1.

https://doi.org/10.1017/S0266466620000304 Published online by Cambridge University Press

ALGORITHM A.1.

Input: $\mathbf{Y}$, $\mathbf{X}$, $0 < \tau < 1$, $\lambda > 0$, $\kappa > 0$;
Initialization: $\Gamma_{\tau,0} = 0, \Omega_{\tau,1} = 0$, step size $\delta_1 = 1$, $M = \frac{1}{\kappa m^2 n^2}\|\mathbf{X}\|^2$;
for $t = 1, 2, \ldots, T$ **do**
  $\Gamma_{\tau,t} = S_{\lambda/M}\big(\Omega_{\tau,t} - \frac{1}{M}\nabla \widehat{Q}_{\tau,\kappa}(\Omega_{\tau,t})\big)$;
  $\delta_{t+1} = \frac{1+\sqrt{1+4\delta_t^2}}{2}$;
  $\Omega_{\tau,t+1} = \Gamma_{\tau,t} + \frac{\delta_t - 1}{\delta_{t+1}}(\Gamma_{\tau,t} - \Gamma_{\tau,t-1})$;
end
Output: $\Gamma_{\tau,T}$

A.2. On the Set $\mathcal{K}$

Definition A.1. For $\mathbf{A} \in \mathbb{R}^{p\times m}$ with rank r, the singular value decomposition of $\mathbf{A}$ is $\mathbf{A} = \sum_{j=1}^{r} \sigma(\mathbf{A})\mathbf{u}_j\mathbf{v}_j^\top$. The *support* of $\mathbf{A}$ is defined by (S_1, S_2) in which $S_1 = \text{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ and $S_2 = \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$. Define the projection matrix on S_1: $\mathbf{P}_1 \stackrel{\text{def}}{=} \mathbf{U}_{[1:r]}\mathbf{U}_{[1:r]}^\top$, in which $\mathbf{U}_{[1:r]} = [\mathbf{u}_1 \ldots \mathbf{u}_r] \in \mathbb{R}^{p\times r}$; $\mathbf{P}_2 \stackrel{\text{def}}{=} \mathbf{V}_{[1:r]}\mathbf{V}_{[1:r]}^\top$, where $\mathbf{V}_{[1:r]} = [\mathbf{v}_1 \ldots \mathbf{v}_r] \in \mathbb{R}^{m\times r}$. Denote $\mathbf{P}_1^\perp = \mathbf{I}_{p\times r} - \mathbf{P}_1$ and $\mathbf{P}_2^\perp = \mathbf{I}_{m\times r} - \mathbf{P}_2$. For any matrix $\mathbf{S} \in \mathbb{R}^{p\times m}$, define

$$\mathcal{P}_{\mathbf{A}}(\mathbf{S}) \stackrel{\text{def}}{=} \mathbf{P}_1\mathbf{S}\mathbf{P}_2; \quad \mathcal{P}_{\mathbf{A}}^\perp(\mathbf{S}) \stackrel{\text{def}}{=} \mathbf{P}_1^\perp\mathbf{S}\mathbf{P}_2^\perp.$$

Define for any $a \geq 0$,

$$\mathcal{K}(\mathbf{\Gamma}_\tau; a) \stackrel{\text{def}}{=} \left\{\mathbf{S} \in \mathbb{R}^{p\times m} : \|\mathcal{P}_{\mathbf{\Gamma}_\tau}^\perp(\mathbf{S})\|_* \leq 3\|\mathcal{P}_{\mathbf{\Gamma}_\tau}(\mathbf{S})\|_* + a\right\}. \tag{A.7}$$

We note that the nuclear norm is *decomposable* under the projection; namely, for any $\mathbf{S}, \mathbf{A} \in \mathbb{R}^{p\times m}$, $\|\mathbf{S}\|_* = \|\mathcal{P}_{\mathbf{A}}(\mathbf{S})\|_* + \|\mathcal{P}_{\mathbf{A}}^\perp(\mathbf{S})\|_*$. This is analogous to the ℓ_1 norm for vectors: for any vector $\boldsymbol{\nu}$ and support S, $\|\boldsymbol{\nu}\|_1 = \|\boldsymbol{\nu}_S\|_1 + \|\boldsymbol{\nu}_{S^c}\|_1$; see Definition 1 on page 541 of Negahban et al. (2012). Moreover, the rank of $\mathcal{P}_{\mathbf{A}}(\mathbf{S})$ is at most rank($\mathbf{A}$).

The shape of $\mathcal{K}(\mathbf{\Gamma}_\tau; a)$ is not a cone when $a > 0$, but is still a star-shaped set. This set has a similar shape as the set defined in eqn. (17) on p. 544 in Negahban et al. (2012). See also their Fig. 1 on p. 544.

A.3. Proof of Corollary 3.4

To apply Theorem 3.2, it is enough to find a bound for the optimization error δ of $\mathbf{\Gamma}_{\tau,T}$ that holds with high probability. Suppose the initial estimator is $\mathbf{\Gamma}_{\tau,t=0} = 0$. The bound in (2.2) suggests that

$$L_\tau(\mathbf{\Gamma}_{\tau,T}) - L_\tau(\widehat{\mathbf{\Gamma}}_\tau) \leq \underbrace{\frac{3\kappa mn(\tau \vee \{1-\tau\})^2}{2}}_{\text{(I)}} + \underbrace{\frac{4(\|\mathbf{\Gamma}_\tau\|_{\text{F}}^2 + \|\mathbf{\Gamma}_{\tau,\infty} - \mathbf{\Gamma}_\tau\|_{\text{F}}^2)}{(T+1)^2}\frac{\|\mathbf{X}\|^2}{\kappa m^2 n^2}}_{\text{(II)}}, \tag{A.8}$$

where we apply the bound $\|\mathbf{\Gamma}_{\tau,\infty}\|_{\text{F}}^2 \leq 2\|\mathbf{\Gamma}_{\tau,\infty} - \mathbf{\Gamma}_\tau\|_{\text{F}}^2 + 2\|\mathbf{\Gamma}_\tau\|_{\text{F}}^2$. It is sufficient to show that (I)+(II)$\leq \lambda(m/n)^{1/2}$, and the desired conclusion follows from Theorem 3.2.

https://doi.org/10.1017/S0266466620000304 Published online by Cambridge University Press

Using the bound on κ in (3.10), elementary calculation verifies that (I) in (A.8) is less than $\lambda(m/n)^{1/2}/2$. Under the event that $\|\mathbf{\Gamma}_{\tau,\infty} - \mathbf{\Gamma}_\tau\|_{\mathrm{F}}^2 = o(1)$, it follows by elementary calculation that (II)$\leq \lambda(m/n)^{1/2}/2$, and the corollary is proved.

It is left to show that $\|\mathbf{\Gamma}_{\tau,\infty} - \mathbf{\Gamma}_\tau\|_{\mathrm{F}}^2 = o(1)$ with high probability. Recall that $\mathbf{\Gamma}_{\tau,\infty} = \lim_{T\to\infty} \mathbf{\Gamma}_{\tau,T} = \arg\min_{\mathbf{S}}\{\tilde{L}_\tau(\mathbf{S}) = \widehat{Q}_{\tau,\kappa}(\mathbf{S}) + \lambda\|\mathbf{S}\|_*\}$. The $L_\tau(\mathbf{\Gamma}_{\tau,\infty}) - L_\tau(\widehat{\mathbf{\Gamma}}_\tau)$ of $\mathbf{\Gamma}_{\tau,\infty}$ can be estimated by

$$
\begin{aligned}
&L_\tau(\mathbf{\Gamma}_{\tau,\infty}) - L_\tau(\widehat{\mathbf{\Gamma}}_\tau) \qquad \textbf{(A.9)}\\
&\quad \leq (L_\tau(\mathbf{\Gamma}_{\tau,\infty}) - \widetilde{L}_\tau(\mathbf{\Gamma}_{\tau,\infty})) + \underbrace{(\widetilde{L}_\tau(\mathbf{\Gamma}_{\tau,\infty}) - \widetilde{L}_\tau(\widehat{\mathbf{\Gamma}}_\tau))}_{\leq 0} + \underbrace{(\widetilde{L}_\tau(\widehat{\mathbf{\Gamma}}_\tau) - L_\tau(\widehat{\mathbf{\Gamma}}_\tau))}_{\leq 0 \text{ by (S.1.1)}}\\
&\quad \leq L_\tau(\mathbf{\Gamma}_{\tau,\infty}) - \widetilde{L}_\tau(\mathbf{\Gamma}_{\tau,\infty})\\
&\quad \overset{(S.1.1)}{\leq} \kappa(\tau \vee \{1-\tau\})^2 \frac{nm}{2}\\
&\quad \leq \lambda(m/n)^{1/2}/6. \qquad \textbf{(A.9)}
\end{aligned}
$$

Therefore, $\mathbf{\Gamma}_{\tau,\infty}$ is an approximate optimizer. The growth condition (3.7) with $C = 1$ in the hypothesis of this corollary ensures (3.7) holds with $C = 1/6$. Hence, (3.9) in Theorem 3.2 yields

$$
\|\mathbf{\Gamma}_{\tau,\infty} - \mathbf{\Gamma}_\tau\|_{\mathrm{F}} \leq u\left(\frac{m}{\sigma_{\min}(\mathbf{\Sigma}_X)}\right)^{1/2} \epsilon_{n,\tau,r}, \qquad \textbf{(A.10)}
$$

with probability at least $1 - \eta - \gamma_n - 16(pm)^{1-u^2} - 3\exp\{-(p+m)\log 8\}$.

https://doi.org/10.1017/S0266466620000304 Published online by Cambridge University Press