Multivariate Factorizable Expectile Regression with Application to fMRI Data*

Shih-Kang Chao† Wolfgang K. Härdle‡ Chen Huang§

September 28, 2017

Abstract

A multivariate expectile regression model is proposed to analyze the tail events of large cross-sectional and spatial data, when the tail events are linked by a latent factor structure. The computational advantage of the method is demonstrated, and the estimation risk is analyzed for every *fixed number of iteration* and *fixed sample size*, when the latent factors are either exactly or approximately sparse. The proposed method is applied on the functional magnetic resonance imaging (fMRI) data taken during an experiment of investment decisions making. It is shown that the negative extreme blood oxygenation level dependent (BOLD) responses may be relevant to the risk preferences.

Keywords: multivariate regression, factor analysis, expectile regression, functional magnetic resonance imaging, risk preference

1 Introduction

Analyzing cross-sectional or spatial data is of critical interest in many scientific fields. Particularly, the interests in these fields are mostly in the *tail events*, which are the

*Financial support from the Deutsche Forschungsgemeinschaft via CRC 649 "Economic Risk" and IRTG 1792 "High Dimensional Non Stationary Time Series", Humboldt-Universität zu Berlin, is gratefully acknowledged. Shih-Kang Chao is partially supported by the Office of Naval Research of the U.S.A.

†Department of Statistics, Purdue University, 250 N University St., West Lafayette IN 47907-2066, U.S.A.

‡Ladislaus von Bortkiewicz Chair of Statistics, C.A.S.E. - Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany. Sim Kee Boon Institute for Financial Economics, Singapore Management University, 50 Stamford Road, Singapore 178899, Singapore.

§Corresponding author. Faculty of Mathematics and Statistics, University of St. Gallen, Bodanstrasse 6, 9000 St. Gallen, Switzerland. Email: chen.huang@unisg.ch.

extreme events that occur with very small (or very large) probability. For example, in finance, Value-at-Risk (VaR) defined by the 1% quantile of the distribution of investment portfolio is widely used for measuring the market risk. In climatology, one of the major interests is the prediction of extreme precipitation defined by the tail quantile with level very close to 1. The estimation or prediction of tail events is often complicated by high dimensionality, which is common in many modern applications. However, the latent factors that influence all the cross-sections or spatial points may be sparse.

Multivariate regression (Izenman, 1975; Reinsel and Velu, 1998) is a classical tool for analyzing the cross-sectional or spatial data, and the penalization methods with matrix nuclear norm (Yuan et al., 2007; Negahban and Wainwright, 2011; Negahban et al., 2012) is applied to handle high dimensionality. However, the literature in multivariate regression is mostly silent about the estimation and prediction of tail events. On the other hand, quantile regression proposed by Koenker and Bassett (1978) is a well-known method for estimating the conditional quantiles, which is done through optimizing a non-differentiable loss function. Koenker and Portnoy (1990) generalize the quantile regression to a multivariate regression framework, but it cannot be applied to modern high dimensional data. To deal with high dimensional data with certain sparsity structure, it seems necessary to use some penalization methods, but the non-differentiable loss function of quantile regression is less convenient when being optimized with a penalty that is also non-differentiable, such as the nuclear norm.

In this paper, we propose to estimate the tail events of a factorizable multivariate model using expectile regression (see (2.3) in Section 2 for the specific form of the model). Expectiles illustrate the tail events, and are closely related to quantiles (see, e.g. Section 2 of Rossi and Harvey (2009)). The expectile regression is proposed by Newey and Powell (1987) and is done through optimizing a smooth loss function. The smooth loss function of expectile regression yields computational advantages when being combined with a non-differentiable penalty, which will be shown in the algorithmic convergence analysis in Section 2.2. Furthermore, our method can be easily and efficiently implemented with the fast iterative shrinkage-thresholding algorithm of Beck and Teboulle (2009).

In addition to the convergence analysis, we *jointly* analyze the algorithmic and stochastic risk of our iterative estimator in Theorem 2.3, which characterizes the estimation error for each *fixed sample size* and *fixed number of iteration*. In particular, the theorem shows that our estimator is consistent as long as $\max\{p, m\} \ll n$ while $p, m \to \infty$, where p is the dimension of the covariates, m is the number of cross-sections or spatial points, and n is the sample size obtained in each cross-section or spatial point. The theorem is established under the weak assumption that the number of latent factors jointly influencing all the cross-sections or spatial points are approximately sparse.

Much interest has concentrated on using the functional magnetic resonance imaging (fMRI) data to understand the risk perception of humans (Heekeren et al., 2008). While the positive blood oxygenation level dependent (BOLD) signals are the focus in most studies, an increasing number of researchers are intrigued by the observed negative BOLD signals and their implications. Many hypotheses on the causes and implications of the negative BOLD are proposed, but they are still highly debatable (Mullinger et al., 2014).

We apply our method on the BOLD signals measured on the human subjects during an experiment on investment decisions making, and shed light on how the negative BOLD responses may be relevant in the decision making process. Using the same data, Majer et al. (2015) retrieve factor loadings from a dynamic factor model of the BOLD signals, and apply these loadings on explaining the subjects' risk attitude. However, their analysis only focus on the mean, and neglect the tail information of the BOLD signals. We apply our method on the BOLD responses obtained from 19 subjects, and estimate the factors and loadings at both high and low extreme expectile levels. We find that the factor loadings from the negative tail of the BOLD signals could not only well explain the revealed risk preference of the subjects in term of R^2, but also *predict* the revealed risk preference. The prediction performance of the negative extreme BOLD is generally similar to that of the positive extreme BOLD, but sometimes they can be more accurate. Nonetheless, we note that our results do not yield any conclusions on the source of the negative BOLD responses.

The rest of the paper is arranged as follows. Section 2 introduces the model setting, estimation method and theoretical properties of the estimator. Simulation studies of our method are shown in Section 3. Section 4 illustrates the empirical application with the fMRI data. Section 5 concludes this paper. Proofs and auxiliary results are provided in the appendices. The codes to implement the algorithms are publicly accessible via the website Q www.quantlet.de.

2 Method

2.1 Model

We start with defining some notations. Denote a matrix $\mathbf{S} = (s_{lj}) = [\mathbf{S}_{\cdot 1} ... \mathbf{S}_{\cdot m}] \in \mathbb{R}^{p \times m}$, where $\mathbf{S}_{\cdot j} \in \mathbb{R}^p$ are the column vectors. Let $\|\mathbf{S}\|_{\mathrm{F}}$, $\|\mathbf{S}\|_*$ and $\|\mathbf{S}\|$ be the matrix Frobenius, nuclear and spectral norm. Denote $\sigma_{\min}(\mathbf{S})$ and $\sigma_{\max}(\mathbf{S})$ the smallest and largest singular values. For a vector $\mathbf{v} \in \mathbb{R}^p$, $\|\mathbf{v}\|_2$ is the Euclidean norm. Define $\langle\langle \mathbf{A}, \mathbf{B} \rangle\rangle \stackrel{\text{def}}{=} \mathrm{tr}(\mathbf{A}^\top \mathbf{B})$. $\mathbf{I}_m$ is the identity matrix with dimension m.

Let $\{(\boldsymbol{X}_i, Y_{i1}, ..., Y_{im})\}_{1\le i\le n}$ be the samples with $Y_{ij} \in \mathbb{R}$ and $\boldsymbol{X}_i \in \mathbb{R}^p$. Specifically, Y_{ij} represents the value observed from the response j at the time point i, and $\{\boldsymbol{X}_i\}_{i=1}^n$ are the covariates. For simplicity, we assume that the samples are i.i.d. over i. For $\tau \in (0,1)$, the conditional expectile $e_j(\tau|\boldsymbol{X}_i)$ of Y_{ij} given $\boldsymbol{X}_i$ is defined by

$$e_j(\tau|\boldsymbol{X}_i) = \boldsymbol{X}_i^\top \boldsymbol{\gamma}_j(\tau), \tag{2.1}$$

where

$$\boldsymbol{\gamma}_j(\tau) \overset{\text{def}}{=} \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^p} \mathsf{E}\left[\rho_\tau(Y_{ij} - \boldsymbol{X}^\top\boldsymbol{\gamma})\right], \tag{2.2}$$

and $\rho_\tau(u) \overset{\text{def}}{=} |\tau - \mathbf{1}\{u<0\}||u|^2$. Define the coefficient matrix

$$\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_\tau \overset{\text{def}}{=} [\boldsymbol{\gamma}_1(\tau) ... \boldsymbol{\gamma}_m(\tau)].$$

We assume that the expectiles $e_1(\tau|\boldsymbol{X}_i), ..., e_m(\tau|\boldsymbol{X}_i)$ are related through a factor model:

$$e_j(\tau|\boldsymbol{X}_i) = \sum_{k=1}^r \psi_{j,k}(\tau) f_k^\tau(\boldsymbol{X}_i), \tag{2.3}$$

where $f_k^\tau(\boldsymbol{X}_i)$ is the kth factor, r is the number of factors, and $\psi_{j,k}(\tau)$ are the factor loadings. Furthermore, factors are constructed by linear combinations of covariates $\boldsymbol{X}_i$:

$$f_k^\tau(\boldsymbol{X}_i) = \boldsymbol{\varphi}_k(\tau)^\top \boldsymbol{X}_i. \tag{2.4}$$

where $\boldsymbol{\varphi}_k(\tau) = (\varphi_{k,1}(\tau), ..., \varphi_{k,p}(\tau))^\top$. By substituting (2.4) into (2.3), it can be seen that the factor structure yields the reparametrization $\boldsymbol{\Gamma}^\top = \boldsymbol{\Psi}_\tau\boldsymbol{\Phi}_\tau$, where the matrix $\boldsymbol{\Psi}_\tau = (\psi_{j,k}(\tau))_{j\le m, k\le r}$ and $\boldsymbol{\Phi} = (\varphi_{k,l}(\tau))_{k\le r, l\le p}$. Unfortunately, the matrix factorization is in general not unique, so the factors and loadings may not be identifiable from $\boldsymbol{\Gamma}$. We alleviate the identifiability issue by imposing the normalization restrictions as Eq. (2.14) on page 28 of Reinsel and Velu (1998):

$$\boldsymbol{\Psi}_\tau^\top\boldsymbol{\Psi}_\tau = \mathbf{I}_m, \quad \boldsymbol{\Phi}_\tau\boldsymbol{\Phi}_\tau^\top = \mathrm{diag}(\sigma_1(\boldsymbol{\Gamma}), ..., \sigma_{p\wedge m}(\boldsymbol{\Gamma})). \tag{2.5}$$

The restrictions (2.5) make the factors and loadings associated with the nonzero singular values of $\boldsymbol{\Gamma}$ identifiable up to sign, if the nonzero singular values are distinct. When there exist repeated singular values, $\boldsymbol{\Psi}_\tau$ and $\boldsymbol{\Phi}_\tau$ cannot be uniquely identified; see Remark 2.1. If $\boldsymbol{\Gamma}$ has the singular value decomposition $\mathbf{U}\mathbf{D}\mathbf{V}^\top$, we have $\boldsymbol{\Psi}_\tau = \mathbf{V}$ and $\boldsymbol{\Phi}_\tau = \mathbf{D}^\top\mathbf{U}^\top$. Suppose an estimator $\widehat{\boldsymbol{\Gamma}}$ is available, we can estimate the kth factor by $\widehat{f}_k^\tau(\boldsymbol{X}_i) = \boldsymbol{X}_i^\top\widehat{\boldsymbol{\varphi}}_k(\tau) = \sigma_k\boldsymbol{X}_i^\top\widehat{\mathbf{U}}_{\cdot k}$ and the factor loadings for the jth response by $\widehat{\boldsymbol{\psi}}_j(\tau) = \widehat{\mathbf{V}}_{j\cdot}$, where $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ are unitary matrices obtained from the singular value decomposition $\widehat{\boldsymbol{\Gamma}} = \widehat{\mathbf{U}}\widehat{\mathbf{D}}\widehat{\mathbf{V}}^\top$.

Remark 2.1 (Identifiability)**.** If there exist repeated singular values, then the singular vectors associated with these repeated singular values are not unique, and the factors and loadings are not uniquely identifiable. Since the sign in the matrix factorization cannot be determined, the sign of the loadings and factors are not identifiable. In our empirical analysis in Section 4, we only use the absolute value of the loadings.

The factor model (2.3) implies that $\mathbf{\Gamma}$ is of rank r, and the model (2.1) corresponds to a multivariate linear regression model. For the standard regression with square loss, Reinsel and Velu (1998) propose to estimate $\mathbf{\Gamma}$ with the reduced-rank regression under the knowledge of r. However, r is usually unknown in practice. Yuan et al. (2007) propose to perform the multivariate regression with the nuclear norm penalty, which does not require the knowledge of r. The latter inspired the use of the nuclear norm penalty in the next section. However, Yuan et al. (2007) do not provide an algorithm that can scale up to large dimensions.

2.2 Algorithm

To estimate our model under the factor model (2.3), we combine an asymmetric loss with the nuclear norm penalty. To be more specific, we estimate $\mathbf{\Gamma}$ (defined in Section 2.1) by solving:

$$\widehat{\mathbf{\Gamma}}_{\tau,\lambda} \stackrel{\text{def}}{=} \arg\min_{\mathbf{\Gamma}\in\mathbb{R}^{p\times m}} F(\mathbf{\Gamma}), \tag{2.6}$$

$$F(\mathbf{\Gamma}) \stackrel{\text{def}}{=} (mn)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\rho_\tau(Y_{ij} - \boldsymbol{X}_i^\top \mathbf{\Gamma}_{\cdot j}) + \lambda\|\mathbf{\Gamma}\|_*, \tag{2.7}$$

where λ is a tuning parameter, $\mathbf{\Gamma}_{\cdot j}$ is the jth column of $\mathbf{\Gamma}$. The second term $\|\mathbf{\Gamma}\|_* = \sum_{l=1}^{\min(p,m)}\sigma_l(\mathbf{\Gamma})$, where the singular values $\sigma_1(\mathbf{\Gamma}) \geq \sigma_2(\mathbf{\Gamma}) \geq \ldots \geq \sigma_{\min(p,m)}(\mathbf{\Gamma})$. We note that (2.7) is a convex optimization problem. The number of factors r in (2.3) does not need to be specified. To simplify the notation, we denote $\widehat{\mathbf{\Gamma}}$ for $\widehat{\mathbf{\Gamma}}_{\tau,\lambda}$ hereinafter.

To solve the optimization problem (2.7), we apply the fast iterative shrinkage-thresholding algorithm (FISTA) of Beck and Teboulle (2009). FISTA solves the optimization problems of the form:

$$\min_{\mathbf{\Gamma}}\{g(\mathbf{\Gamma}) + h(\mathbf{\Gamma})\}, \tag{2.8}$$

where g is a smooth convex function with Lipschitz continuous gradient ∇g,

$$\|\nabla g(\mathbf{\Gamma}_1) - \nabla g(\mathbf{\Gamma}_2)\|_{\text{F}} \leq L_{\nabla g}\|\mathbf{\Gamma}_1 - \mathbf{\Gamma}_2\|_{\text{F}}, \forall \mathbf{\Gamma}_1, \mathbf{\Gamma}_2 \in \mathbb{R}^{p\times m}, \tag{2.9}$$

where $L_{\nabla g}$ is the Lipschitz constant of ∇g and h is a continuous convex (possibly non-

smooth) function (Ji and Ye, 2009).

In view of (2.7), this corresponds to

$$g(\mathbf{\Gamma}) \stackrel{\text{def}}{=} (mn)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \rho_\tau(Y_{ij} - \boldsymbol{X}_i^\top \mathbf{\Gamma}_{\cdot j}), \tag{2.10}$$

$$h(\mathbf{\Gamma}) \stackrel{\text{def}}{=} \lambda \|\mathbf{\Gamma}\|_*. \tag{2.11}$$

The Lipschitz constant of ∇g is $L_{\nabla g} = 2(mn)^{-1} \max(\tau, 1-\tau) \|\mathbf{X}\|_{\text{F}}^2$; see Appendix A.1.

Algorithm 1 is an application of FISTA, with g and h chosen as (2.10) and (2.11).

Algorithm 1: FISTA for expectile regression with nuclear norm penalty.

Input: $\{\boldsymbol{Y}_i\}_{i=1}^n$, $\{\boldsymbol{X}_i\}_{i=1}^n$, λ

Output: $\widehat{\mathbf{\Gamma}} = \mathbf{\Gamma}_T$

Initialization: $\mathbf{\Gamma}_0 = 0$, $\mathbf{\Omega}_1 = 0$, step size $\delta_1 = 1$;

for $t = 1, 2, \ldots, T$ **do**

  $\mathbf{\Gamma}_t = \mathtt{SVT}_{\lambda, g}\big(\mathbf{\Omega}_t - L_{\nabla g}^{-1} \nabla g(\mathbf{\Omega}_t)\big)$;

  $\delta_{t+1} = \frac{1 + \sqrt{1 + 4\delta_t^2}}{2}$;

  $\mathbf{\Omega}_{t+1} = \mathbf{\Gamma}_t + \frac{\delta_t - 1}{\delta_{t+1}} (\mathbf{\Gamma}_t - \mathbf{\Gamma}_{t-1})$;

end

The subroutine $\mathtt{SVT}_{\lambda,g}$ in Algorithm 1 is the singular value thresholding operator given by $\mathtt{SVT}_{\lambda,g}\big(\mathbf{S}\big) \stackrel{\text{def}}{=} \mathbf{U}_{\mathbf{S}} \big(\mathbf{D}_{\mathbf{S}} - (\lambda / L_{\nabla g}) \mathbf{I}_{p \times m}\big)_+ \mathbf{V}_{\mathbf{S}}^\top$, where SVD implies $\mathbf{S} = \mathbf{U}_{\mathbf{S}} \mathbf{D}_{\mathbf{S}} \mathbf{V}_{\mathbf{S}}^\top$, $\mathbf{I}_{p \times m}$ is a rectangular identity matrix with main diagonal elements equal to 1, and $(\mathbf{S})_+ = (\max\{0, s_{ij}\})$.

Remark 2.2 (Initialization and the stopping rule)**.** We suggest to initialize the algorithm with $\mathbf{\Gamma}_0 = 0$ in Algorithm 1, but because the optimization problem is convex, this can be replaced by any matrix. Of course, the algorithm converges faster if we initialize it with a matrix that is close to the minimizer. We suggest to stop the algorithm at iteration T satisfying $|F(\widehat{\mathbf{\Gamma}}_{T+1}) - F(\widehat{\mathbf{\Gamma}}_T)| \leq \epsilon$, for some small $\epsilon > 0$. In the simulation and empirical analysis of this paper, $\epsilon = 10^{-6}$.

The convergence of Algorithm 1 in terms of the loss function is guaranteed by the following theorem.

Theorem 2.1 (Bounds for the loss difference and convergence rate in Algorithm 1)**.** *Let* $\{\mathbf{\Gamma}_t\}_{t=0}^T$ *be the sequence obtained by the iteration of Algorithm 1. Then*

$$|F(\mathbf{\Gamma}_t) - F(\widehat{\mathbf{\Gamma}})| \leq \frac{4(mn)^{-1} \max(\tau, 1-\tau) \|\mathbf{X}\|_{\text{F}}^2 \|\mathbf{\Gamma}_0 - \widehat{\mathbf{\Gamma}}\|_{\text{F}}^2}{(t+1)^2}. \tag{2.12}$$

In particular, if for $\epsilon > 0$,

$$t \geq \frac{2\sqrt{\max(\tau, 1-\tau)}\|\mathbf{X}\|_{\mathrm{F}}\|\boldsymbol{\Gamma}_0 - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}}{\sqrt{mn\epsilon}} - 1, \tag{2.13}$$

then $|F(\boldsymbol{\Gamma}_t) - F(\widehat{\boldsymbol{\Gamma}})| \leq \epsilon$.

The bound (2.12) comes from a careful calculation of the Lipschitz constant of the gradient of g. The proof of Theorem 2.1 can be found in Appendix A.1.

Theorem 2.1 shows that to get an ϵ-accurate solution, it requires $1/\sqrt{\epsilon}$ steps when holding other parameters fixed. This is smaller than $1/\epsilon$ steps given by quantile regression and $1/\epsilon^2$ by the general subgradient methods, see Theorem 2.3 and Remark 2.4 in Chao et al. (2016). In view of (2.13), when τ is approaching 0 or 1, the number of iterations that is required to achieve an ϵ-accurate solution would increase.

Furthermore, utilizing the strong convexity of g, we can obtain a bound for $\|\boldsymbol{\Gamma}_t - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2$. For this purpose, additional assumptions on the design $\boldsymbol{X}$ are required.

(A1) Suppose $\mathsf{E}\,\boldsymbol{X}_i = 0$, $\mathsf{E}\,\boldsymbol{X}_i\boldsymbol{X}_i^\top = \boldsymbol{\Sigma}$ with $\sigma_{\min}(\boldsymbol{\Sigma}) > C_1$ and $\sigma_{\max}(\boldsymbol{\Sigma}) < C_2$ for some constants $C_1, C_2 > 0$ uniformly in p. For some sequence $0 < a_n < 1$, constants $c_1, c_2 > 0$,

$$\mathrm{P}\left[\sigma_{\min}\left(\frac{\mathbf{X}^\top\mathbf{X}}{n}\right) \geq c_1\sigma_{\min}(\boldsymbol{\Sigma}), \sigma_{\max}\left(\frac{\mathbf{X}^\top\mathbf{X}}{n}\right) \leq c_2\sigma_{\max}(\boldsymbol{\Sigma})\right] \geq 1 - a_n. \tag{2.14}$$

Assumption (A1) holds for Gaussian design $\boldsymbol{X}$ with $c_1 = 1/9$, $c_2 = 9$ and $a_n = 4\exp(-n/2)$. See Wainwright (2009). It can be shown that (A1) holds for the sub-gaussian designs; see Vershynin (2012a) for details.

The following theorem characterizes the convergence in the Frobenius norm.

Theorem 2.2. *Given (A1), the sequence* $\boldsymbol{\Gamma}_t$ *obtained from Algorithm 1 satisfy*

$$\|\boldsymbol{\Gamma}_t - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2 \leq \frac{36}{n(t+1)^2}\frac{\max(\tau, 1-\tau)}{\min(\tau, 1-\tau)}\frac{\|\mathbf{X}\|_{\mathrm{F}}^2}{\sigma_{\min}(\boldsymbol{\Sigma})}\|\boldsymbol{\Gamma}_0 - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2, \tag{2.15}$$

with probability greater than $1 - a_n$. *In particular, if for* $\epsilon > 0$,

$$t \geq 6\sqrt{\frac{\max(\tau, 1-\tau)}{\min(\tau, 1-\tau)}}\frac{\|\mathbf{X}\|_{\mathrm{F}}\|\boldsymbol{\Gamma}_0 - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}}{\sqrt{n\sigma_{\min}(\boldsymbol{\Sigma})\epsilon}} - 1, \tag{2.16}$$

then $\|\boldsymbol{\Gamma}_t - \widehat{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2 \leq \epsilon$ *holds with probability greater than* $1 - a_n$.

The proof of Theorem 2.2 is in Section A.2. We discuss the estimation of the number of factors in the following remark.

Remark 2.3 (Estimation of the number of factors)**.** The number of factors r defined in Section 2.1 can be estimated by $\text{rank}(\mathbf{\Gamma}_T)$, which is the estimator generated by Algorithm 1. If the number of factors is exactly sparse, $\text{rank}(\mathbf{\Gamma}_T)$ is usually a good estimator; see the simulation study in Section 3.

2.3 Oracle Inequalities

In this section, we derive the bounds for the difference between the sequence $\mathbf{\Gamma}_t$ generated by Algorithm 1 and the true matrix $\mathbf{\Gamma}$. These results heavily rely on the strong convexity of ρ_τ.

We make the following assumptions.

(A2) There exists $C > 0$ such that for $u_{ij} \overset{\text{def}}{=} Y_{ij} - \boldsymbol{X}_i^\top \mathbf{\Gamma}_{\cdot j}$, $\mathrm{P}(|u_{ij}| > s) \leq \exp(1 - s^2/C^2)$, $\forall s \geq 0$) with sub-gaussian norm $\|u_{ij}\|_{\psi_2} \overset{\text{def}}{=} \sup_{p \geq 1} p^{-1/2} (\mathsf{E}\,|u_{ij}|^p)^{1/p}$, and let $K_u \overset{\text{def}}{=} \max_{1 \leq j \leq m} \|u_{ij}\|_{\psi_2}$.

(A3) Conditional on $\boldsymbol{X}_i$, Y_{ij} are independent over j.

(A2) regulates the tails of Y_{ij}. (A3) is required for obtaining the bounds on the tail probabilities of the estimation error.

In Theorem 2.3 below, we state a non-asymptotic bound for $\|\mathbf{\Gamma}_t - \mathbf{\Gamma}\|_{\mathrm{F}}$ in the general situation that the number of factors can be increasing with n.

Theorem 2.3 (Approximately sparse factors)**.** *Under (A1)-(A3),* $\lambda = 2cm^{-1} \max(\tau, 1-\tau) K_u \sqrt{\|\mathbf{\Sigma}\|} \sqrt{\frac{p+m}{n}}$ *for some absolute constant* $c > 0$*. Then for any* $q \in \{1, ..., p \wedge m\}$*, the sequence* $\mathbf{\Gamma}_t$ *obtained by Algorithm 1 satisfy*

$$\|\mathbf{\Gamma}_t - \mathbf{\Gamma}\|_{\mathrm{F}}^2 \leq c'' \Big(\frac{R_t}{n} + 1 \Big) \sqrt{\frac{p+m}{n}} \zeta_\tau \left\{ \sqrt{\frac{p+m}{n}} \zeta_\tau q + \sum_{j=q+1}^{p \wedge m} \sigma_j(\mathbf{\Gamma}) \right\} + \frac{c'' R_t}{n} \|\mathbf{\Gamma}_0 - \mathbf{\Gamma}\|_{\mathrm{F}}^2, \tag{2.17}$$

with probability greater than $1 - 3 \cdot 8^{-(p+m)} - a_n$*, where* $c'' > 0$ *is an absolute constant,* $R_t \overset{\text{def}}{=} \frac{1}{(t+1)^2} \frac{\max(\tau, 1-\tau)}{\min(\tau, 1-\tau)} \frac{\|\mathbf{X}\|_{\mathrm{F}}^2}{\sigma_{\min}(\mathbf{\Sigma})}$ *and* $\zeta_\tau \overset{\text{def}}{=} \frac{\max(\tau, 1-\tau)}{\min(\tau, 1-\tau)} \frac{\sqrt{\|\mathbf{\Sigma}\|}}{\sigma_{\min}(\mathbf{\Sigma})} K_u$.

Please see Appendix B for a proof of Theorem 2.3. Note that (2.17) holds for any $q \in \{1, ..., p \wedge m\}$. The optimal bound is obtained by selecting q that balances $\sqrt{\frac{p+m}{n}} \zeta_\tau q$

and $\sum_{j=q+1}^{p\wedge m}\sigma_j(\mathbf{\Gamma})$. For a fixed number of iterations t in Algorithm 1 and τ, a sufficient condition for (2.17) tending to zero is that the number of factors r is *approximately sparse* ($\mathbf{\Gamma}$ is *approximately low rank*): there exists an increasing sequence $q = q_n \in \mathbb{N}$ such that

$$\lim_{n\to\infty}\frac{p+m}{n}\zeta_\tau^2 q = 0 \quad \text{and} \quad \lim_{n\to\infty}\left\{\sum_{j=q+1}^{p\wedge m}\sqrt{\frac{p+m}{n}}\zeta_\tau\sigma_j(\mathbf{\Gamma})\right\} = 0, \tag{2.18}$$

where p and m can be growing sequences in n. The quantity R_t characterizes how the computational cost influences the error bound. We can increase the number of iterations in Algorithm 1 to shrink R_t, but this also increases the computational cost. Similar to Theorem 2.1 and B.1, when τ is approaching to the boundaries of $(0, 1)$, the bound in (2.17) will increase. Furthermore, heavier tails for Y_{ij} make higher K_u, and lead to higher error bounds.

If the number of factors is fixed and is not increasing with n ($\text{rank}(\mathbf{\Gamma})$ is fixed), then (2.17) is minimized by selecting $q = \text{rank}(\mathbf{\Gamma})$ and $\sum_{j=q+1}^{p\wedge m}\sqrt{\frac{p+m}{n}}\zeta_\tau\sigma_j(\mathbf{\Gamma}) = 0$ in (2.17). Hence, we have the following corollary.

Corollary 2.1 (Exactly sparse factors)**.** *Under the conditions of Theorem 2.3,*

$$\|\mathbf{\Gamma}_t - \mathbf{\Gamma}\|_{\text{F}}^2 \leq c''\Big(\frac{R_t}{n}+1\Big)\frac{p+m}{n}\zeta_\tau^2\,\text{rank}(\mathbf{\Gamma}) + \frac{c''R_t}{n}\|\mathbf{\Gamma}_0 - \mathbf{\Gamma}\|_{\text{F}}^2, \tag{2.19}$$

with probability greater than $1 - 3\cdot 8^{-(p+m)} - a_n$*, where* $c'' > 0$ *is an absolute constant,* $R_t = \frac{1}{(t+1)^2}\frac{\max(\tau,1-\tau)}{\min(\tau,1-\tau)}\frac{\|\mathbf{X}\|_{\text{F}}^2}{\sigma_{\min}(\mathbf{\Sigma})}$ *and* $\zeta_\tau = \frac{\max(\tau,1-\tau)}{\min(\tau,1-\tau)}\frac{\sqrt{\|\mathbf{\Sigma}\|}}{\sigma_{\min}(\mathbf{\Sigma})}K_u$.

Remark 2.4. As explained in Section 2.1, we estimate $\boldsymbol{\psi}_{j,t}(\tau)$ by $\mathbf{V}_{j\cdot,t}$ in the SVD $\mathbf{\Gamma}_t = \mathbf{U}_t\mathbf{D}_t\mathbf{V}_t^\top$. By Theorem 3.10 of Chao et al. (2016), we have:

$$1 - |\boldsymbol{\psi}_j^\top(\tau)\boldsymbol{\psi}_{j,t}(\tau)| \leq \frac{2(2\|\mathbf{\Gamma}\| + \|\mathbf{\Gamma}_t - \mathbf{\Gamma}\|_{\text{F}})\|\mathbf{\Gamma}_t - \mathbf{\Gamma}\|_{\text{F}}}{\min\left\{\sigma_{j-1}^2(\mathbf{\Gamma}) - \sigma_j^2(\mathbf{\Gamma}), \sigma_j^2(\mathbf{\Gamma}) - \sigma_{j+1}^2(\mathbf{\Gamma})\right\}}, \tag{2.20}$$

where $\boldsymbol{\psi}_j(\tau)$ is the true loadings. Theorem 2.3 (or Corollary 2.1) can be used with (2.20) to get an explicit bound.

3 Simulation Study

In this section, we apply our method on the simulated data to evaluate the estimation performance on the factors and loadings, as the number of factors varies.

Set $n = m = p = 100$. For $i = 1, \ldots, n$, $j = 1, \ldots, m$, let $\boldsymbol{X}_i \sim \mathcal{N}(0, \Sigma_{p\times p})$ with

$\Sigma_{jk} = 0.5^{|j-k|}$ and $\boldsymbol{\varepsilon}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_{m\times m})$, the response variables are generated by

$$Y_{ij} = \boldsymbol{X}_i^\top \boldsymbol{\Gamma}_{\cdot j} + \varepsilon_{ij} = \sum_{k=1}^{r} \psi_{jk} f_k(\boldsymbol{X}_i) + \varepsilon_{ij} = \sum_{k=1}^{r} \mathbf{V}_{jk} \mathbf{D}_{kk} \boldsymbol{X}_i^\top \mathbf{U}_{\cdot k} + \varepsilon_{ij}, \tag{3.1}$$

where $r = \text{rank}(\boldsymbol{\Gamma})$. We will set $r = 2, 5, 10$, and the nonzero diagonal components of $\mathbf{D}$ are $(19.01, 18.74, 18.65, 18.22, 17.80, 17.50, 17.21, 17.02, 16.57, 16.49)$. The columns of $\mathbf{V}$ and $\mathbf{U}$ are the orthonormal singular vectors of a matrix with components chosen from $\mathcal{N}(0, 1)$. We repeat the data generation 500 times.

We apply Algorithm 1 with $\boldsymbol{Y}$ and $\widetilde{\boldsymbol{X}} = (\mathbf{l}_n, \widetilde{\boldsymbol{X}})$, where $\mathbf{l}_n = (1, ..., 1)$ is the intercept. The tuning parameter λ is selected according to Lemma B.1, i.e., $\lambda = 2cm^{-1} \max(\tau, 1-\tau) K_u \sqrt{\|\boldsymbol{\Sigma}\|} \sqrt{\frac{p+m}{n}}$. We stop the algorithm as described in Remark 2.2. Denote the resulting estimator $\widetilde{\boldsymbol{\Gamma}}_1$, and obtain $\widetilde{\boldsymbol{\Gamma}}$ by removing the first row (the intercept) of $\widetilde{\boldsymbol{\Gamma}}_1$.

Table 3.1 reports the results for the estimated number of factors $\widehat{r}$, which is the number of nonzero singular values of $\widetilde{\boldsymbol{\Gamma}}$ that are greater than 10^{-10}. That is, the singular values smaller than 10^{-10} are treated as zero. We try several values of c in the formula for λ because we do not know its exact value. The true number of factors are generally well recovered by our algorithm, except for the expectiles that deviate more from $\tau = 0.5$. Furthermore, the estimated number of factors is robust to the model randomness as the standard deviations are very small. The results are similar for different values of c, so we fix $c = 1.3$ for all the later analysis.

τ	0.05	0.3	0.5	0.7	0.95
			r=10		
c=1.3	10.95 (0.22)	11.00 (0.00)	10.00 (0.00)	11.00 (0.00)	10.94 (0.23)
c=1.5	10.70 (0.47)	11.00 (0.00)	10.00 (0.00)	11.00 (0.00)	10.71 (0.46)
c=1.7	10.19 (0.61)	11.00 (0.00)	10.00 (0.00)	11.00 (0.00)	10.20 (0.60)
			r=5		
c=1.3	6.00 (0.00)	6.00 (0.00)	5.00 (0.04)	6.00 (0.00)	6.00 (0.00)
c=1.5	6.00 (0.00)	6.00 (0.00)	5.00 (0.00)	6.00 (0.00)	6.00 (0.00)
c=1.7	6.00 (0.00)	6.00 (0.06)	5.00 (0.00)	6.00 (0.04)	6.00 (0.00)
			r=2		
c=1.3	3.00 (0.00)	3.00 (0.00)	2.03 (0.18)	3.00 (0.00)	3.00 (0.00)
c=1.5	3.00 (0.00)	2.99 (0.12)	2.00 (0.00)	2.99 (0.09)	3.00 (0.00)
c=1.7	3.00 (0.00)	2.72 (0.45)	2.00 (0.00)	2.78 (0.41)	3.00 (0.00)

Table 3.1: The averaged estimated number of factors $\widehat{r}$ over simulation repetitions with respect to τ and c. Values in the parentheses are the standard deviations over the simulation repetitions.

The Frobenius error $\|\widetilde{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_{\mathrm{F}}$ is shown in Figure 3.1. The results are symmetric in τ around $\tau = 0.5$, and the estimation errors tend to be larger for the tail τ. In the models where r is larger, the Frobenius error is also larger. Our findings in the simulation studies are consistent with the roles of τ and r in the error bound in Corollary 2.1.

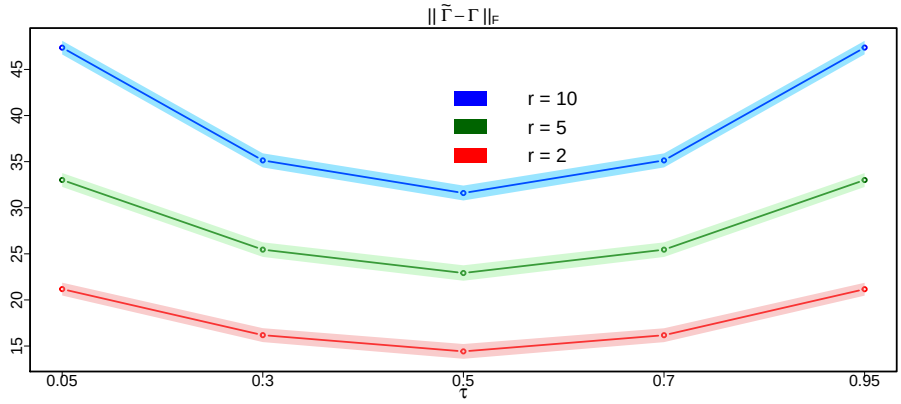

Figure 3.1: The averaged estimation error $\|\widetilde{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_{\mathrm{F}}$ ($c = 1.3$ in λ). The solid lines represent the averaged Frobenius errors over simulation repetitions, and the bands describe the standard deviations over the simulation repetitions.

We measure the estimation performance of the factors and loadings by

$$\begin{aligned} \|\mathbf{\Delta}^{\mathrm{fac}}_{k\cdot}\|_2/\mathbf{D}_{kk}, \text{ where } \mathbf{\Delta}^{\mathrm{fac}} \stackrel{\mathrm{def}}{=} |\widetilde{\mathbf{D}}\widetilde{\mathbf{U}}^\top| - |\mathbf{D}\mathbf{U}^\top|, \\ \|\mathbf{\Delta}^{\mathrm{load}}_{\cdot k}\|_2, \text{ where } \mathbf{\Delta}^{\mathrm{load}} \stackrel{\mathrm{def}}{=} |\widetilde{\mathbf{V}}| - |\mathbf{V}|, \end{aligned} \tag{3.2}$$

for $k = 1, ..., r$, where $\widetilde{\mathbf{V}}, \widetilde{\mathbf{D}}$ and $\widetilde{\mathbf{U}}$ are based on the SVD $\widetilde{\mathbf{\Gamma}} = \widetilde{\mathbf{U}}\widetilde{\mathbf{D}}\widetilde{\mathbf{V}}^\top$. We do not include the covariate $\boldsymbol{X}_i$ in the measure for the estimation error of the factors because all factors share the same $\boldsymbol{X}_i$. The results are presented in Figure 3.2. The results for $\tau = 0.05$ and 0.3 are omitted because they are similar to $\tau = 0.95$ and 0.7. Smaller r still gives smaller estimation error, but the associated standard deviation is larger. As τ deviating from 0.5, the error is larger. The patterns of the estimation errors for factors and loadings are similar.

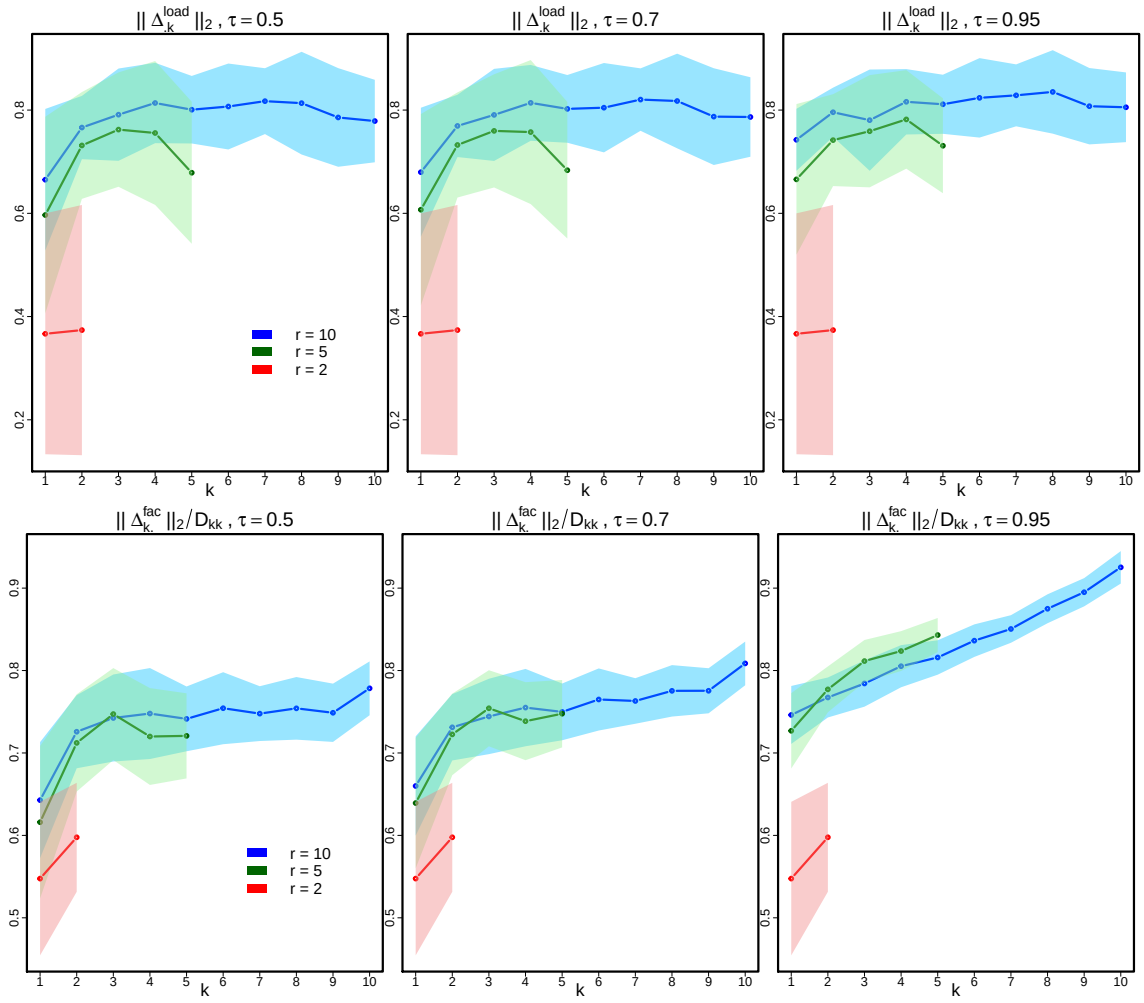

Figure 3.2: The estimation error $\|\mathbf{\Delta}^{\text{load}}_{\cdot k}\|_2$ for the loadings and $\|\mathbf{\Delta}^{\text{fac}}_{k\cdot}\|_2/\mathbf{D}_{kk}$ for the factors, defined in (3.2). The solid lines represent the averaged errors, and the bands describe the standard deviations over simulation repetitions; $c = 1.3$ in λ.

4 Empirical Analysis: Predicting Risk Attitude with fMRI Data

In this section, we apply our method on the fMRI data to predict the risk attitude on the investment decisions making. To understand how human brain responds to reward and risk is an important research topic in neuropsychology, financial economics and neuroeconomics (Heekeren et al., 2008; Camerer, 2007; Schultz, 2015). Previous research mainly focuses on the identification of the region of interest (ROI) with Blood Oxygenation Level Dependent (BOLD) signals (see Schultz (2015) and the references therein). However, only a few research uses fMRI on predicting the risk attitude of subjects. Helfinstein et al. (2014) train support vector machines with the BOLD signals recorded in a Ballon Analog Risk Task (BART) on several combinations of brain regions, and this classifier can predict

subjects' next choice with over 70% accuracy; van Bömmel et al. (2014) and Majer et al. (2015) retrieve factor loadings from a dynamic factor model on BOLD and apply these loadings on predicting subjects' risk attitude.

We focus on predicting the risk attitude of the subjects using the BOLD signals, but we differ from the previous studies in that we separately analyze the *positive* and *negative* BOLD signals observed in the cortical regions. The positive BOLD signals are known to be closely associated with increased neuronal activities, but the interpretation of large *negative* BOLD responses (NBR) is still controversial. Mullinger et al. (2014) argue that the best explanation for NBR at the cortical layer might be a decrease in cerebral blood flow (CBF) with a lesser reduction in the neuronal activities, which is measured by the cerebral metabolic rate of oxygen consumption ($CMRO_2$). This explanation is proven to be an important complement of the more classical "blood/vascular stealing" hypothesis (see p.263 of Mullinger et al. (2014)). However, Mullinger et al. (2014) also argue that there may exist deeper neuronal reasons for NBR than simply the inversion of the neurovascular coupling mechanism of the positive BOLD responses. Following the interpretation of NBR of Mullinger et al. (2014), we suspect that NBR also contain valuable information for predicting the risk attitude. Using our expectile based approach, we study whether the positive and negative extreme BOLD responses are relevant to the risk attitude.

4.1 Data

Our data come from a rapid event-related design experiment on investment decisions making, and this data set is firstly analyzed in Majer et al. (2015). The experiment was done as follows: 19 subjects were requested to make choices in 256 investment decision tasks and each task lasts 7 seconds. The fMRI was taken every two seconds (temporal resolution = 2 seconds), and this resulted in 1400 images for each subject. We have also acquired the answer for each task from each subject. Majer et al. (2015) identify three brain regions Anterior insula (left and right aINS) and dorsomedial prefrontal cortex (DMPFC) as the active regions related to investment decisions via spectral clustering method. We will only focus on the BOLD responses of the voxels in these three regions.

We integrate the information of each region (left and right aINS and DMPFC) spatially by taking the *quantiles* of the BOLD responses over all voxels in these regions. At each fMRI scan i of the sth subject, we take the quantiles with levels $\omega \in \{0.1, 0.5, 0.9\}$ of the BOLD responses over all voxels in the regions $b = 1$ (aINS_L), $b = 2$ (aINS_R) and $b = 3$ (DMPFC) to construct a single time series $\nu_i(s, b, \omega)$, where $i = 1, ..., N = 1400$. Figure 4.1 gives an illustration of the BOLD time series of each cluster. For each clus-

ter, the series of 19 subjects at ω are averaged (the solid lines) and the bands show the dispersion of the 19 time series. We observe that the series for $\omega = 0.9$ is positive, which summarizes the information of the positive BOLD responses, while the series for $\omega = 0.1$ is negative, which corresponds to the negative BOLD responses. The series for $\omega = 0.5$ is stationary and varying around the origin. From Figure 4.1, we observe that the series with each different ω shows different volatility, and this may imply that the series with different ω contains different information. We will show in Table 4.1 that the series with $\omega = 0.1$ and 0.9 tend to contain more information than $\omega = 0.5$.

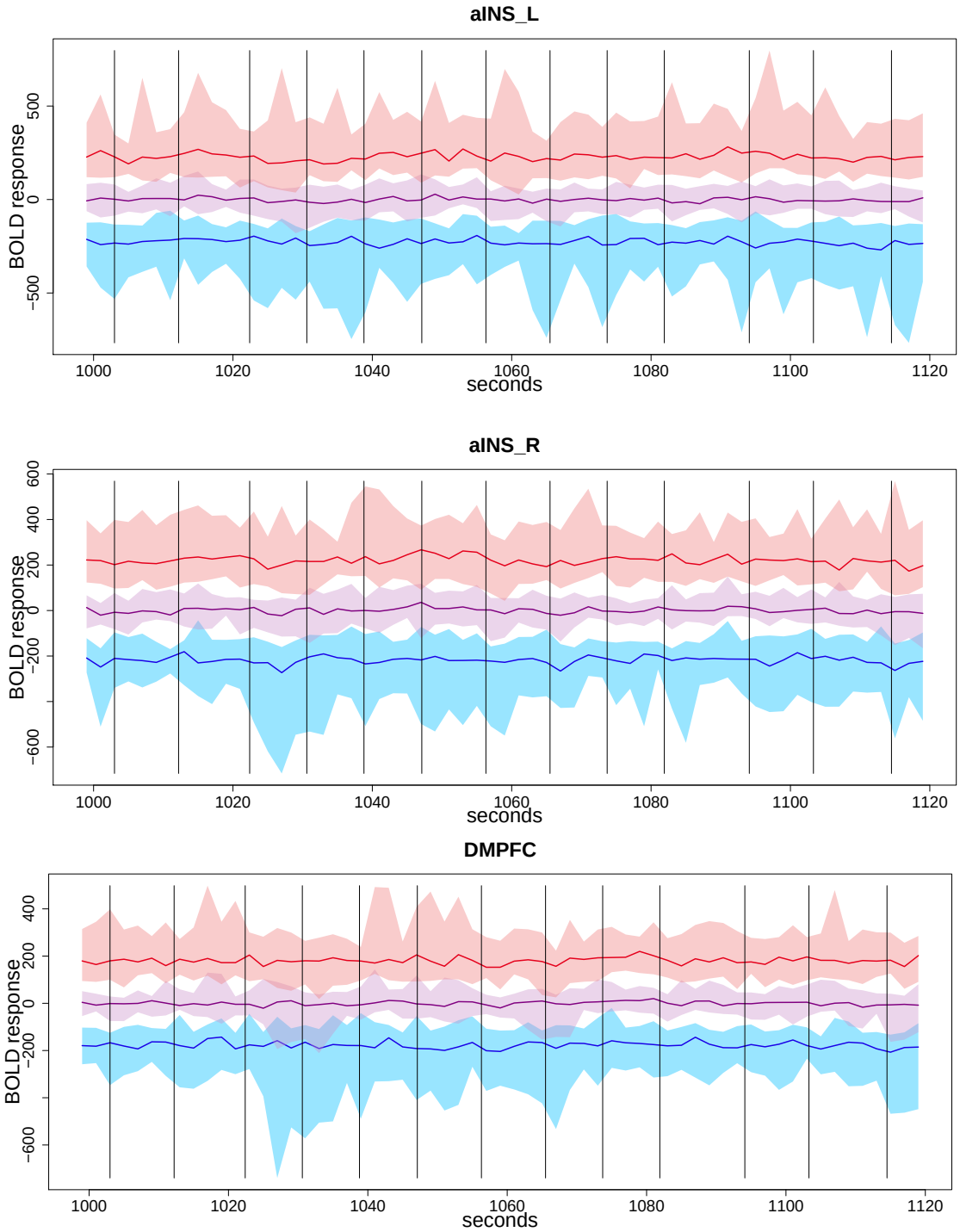

Figure 4.1: In each region, the ω quantiles of the BOLD responses over all the voxels between 1000-1120 seconds of the experiment are shown. In each subfigure (region), lowest (resp., middle, highest) solid lines represent the median of $\omega = 0.1$ (resp., $\omega = 0.5$, 0.9) quantiles of all 19 subjects, and the upper and lower boundaries of the bands present the maximum and the minimum of the ω quantiles of the 19 subjects. Vertical lines indicate the occurrences of the stimuli (the beginning of each task).

4.2 Method

4.2.1 Factor loadings at each region b and quantile level ω

For each ω and a single region b, we consider two approaches to construct the variable Y_{ij}:

(C1) Whole time series: set $Y_{ij}^{b,\omega} = \nu_i(j, b, \omega)$, where $i = 1, ..., n$ with $n = N = 1400$, $j = 1, ..., 19$ (subject). Thus, we have $m = 19$ curves in each region b and at each quantile level ω.

(C2) Analyzing each task separately (task-wise): we divide the whole time series in each region b and at each quantile level ω into subseries based on the beginning and the end of each task. Let $\mathcal{I}_q \subset \{1, ..., N\}$ be the set containing the indices of the images taken during the qth task. In our data, each $|\mathcal{I}_q| = 3$ or 4. We linearly interpolate the points $\{\nu_i(s, b, \omega)\}_{i \in \mathcal{I}_q}$ for each fixed s, b, and ω. Denote $\widetilde{\nu}_i(s, b, q, \omega)$ by the value on the interpolated curve at the ith point in n equally distant grid on the interval $(\min(\mathcal{I}_q), \max(\mathcal{I}_q))$, where $i = 1, ..., n = 50$. Let $Y_{ij}^{b,\omega} = \widetilde{\nu}_i(s, b, q, \omega)$ with $j = 256(s-1) + q$, where $s = 1, ..., 19$ (index for subject) and $q = 1, ..., 256$ (index for task) for each ω, b. Thus, there are $m = 19 * 256 = 4864$ curves in each b and ω.

The variable $\boldsymbol{X}_i$ is a vector of basis functions that need to be flexible enough to capture the various shapes of the fMRI BOLD sequences. For this purpose, we use the cubic B-spline basis $\{B_k\}_{k=1}^p$ with equally spaced knots on $[0, 1]$, and set $\boldsymbol{X}_i = \big(B_1(i/n), B_2(i/n), ..., B_p(i/n)\big)^\top$, where $i = 1, ..., n$. Note that $n = 1400$ in **(C1)** and $n = 50$ in **(C2)**. B-splines are suitable for estimating the hemodynamic response function, see Degras and Lindquist (2014) for more details. We select $p = \lceil n^{0.8} \rceil$ of basis functions in each approach above, where $\lceil \cdot \rceil$ takes the smallest integer that is greater than the argument. The power 0.8 is greater than the (asymptotic) optimal rate 0.4, because the nuclear norm penalty alleviates the issue of overfitting. As the result, there are 329 basis functions in the approach **(C1)** and 23 in **(C2)**.

We compute the matrix $\widehat{\boldsymbol{\Gamma}}^{b,\omega}$ with expectile level $\tau = 0.1, 0.5, 0.9$ using Y_{ij} and $\boldsymbol{X}_i$ by Algorithm 1, where Y_{ij} is chosen under either **(C1)** or **(C2)** with $\lambda^{b,\omega}$ selected by the standard 5-fold cross-validation for each region b and each quantile level ω. Please see Appendix D.1 for the exact value of λ for each pair (b, ω). Using SVD $\widehat{\boldsymbol{\Gamma}}^{b,\omega} = \widehat{\mathbf{U}}_\tau^{b,\omega} \widehat{\mathbf{D}}_\tau^{b,\omega} (\widehat{\mathbf{V}}_\tau^{b,\omega})^\top$, where $(\widehat{\mathbf{V}}_\tau^{b,\omega})^\top$ is regarded as the factor loadings. We note that the size of the matrix $\widehat{\mathbf{V}}_\tau^{b,\omega}$ is 19×19 if we define $Y_{ij}^{b,\omega}$ by following **(C1)**, and 4864×4864 by following **(C2)**. Note that the sign of the factor loadings cannot be determined exactly (see Remark 2.1).

Remark 4.1 (On the computation of λ)**.** The model error of the BOLD signals typically demonstrates autocorrelation following AR(k) or ARMA(1,1) (Lindquist, 2008, page 446) under the temporal resolution 2 seconds. A major consequence of the presence of temporal correlation is that the usual cross-validation could potentially underestimate λ, which leads to undersmoothing and overfitting (Opsomer et al., 2001, Section 2). This problem is especially important for the setting **(C2)**, where the dimensionality is high because we separate each task. However, we observe that the estimated number of factors for the setting **(C2)** is typically very sparse (less than five factors). Overall, the overfitting does not cause a big issue and the usual cross-validation works well in our model.

4.2.2 Predicting risk attitude

To evaluate the prediction performance, we need to obtain the subjects' risk attitude β_s, where $s = 1, ..., 19$ denotes the subject. We follow the approach of Majer et al. (2015) and estimate β_s using the investment decisions made by the subjects to each task with logistic regression; see Appendix D.2 for more details. In essence, higher β_s means the subject s is *less* risk-averse.

In order to use the loadings $\widehat{\mathbf{V}}_\tau^{b,\omega}$ to predict β_s, we apply the standard linear regression models. In particular, in the case **(C1)**, we construct a model for β_s using the first two factor loadings

$$\begin{aligned}\beta_s = \alpha_0^{\omega,\tau} &+ \alpha_{11}^{\omega,\tau}\big|(\widehat{\mathbf{V}}_\tau^{1,\omega})_{s1}\big| + \alpha_{12}^{\omega,\tau}\big|(\widehat{\mathbf{V}}_\tau^{2,\omega})_{s1}\big| + \alpha_{13}^{\omega,\tau}\big|(\widehat{\mathbf{V}}_\tau^{3,\omega})_{s1}\big| \\ &+ \alpha_{21}^{\omega,\tau}\big|(\widehat{\mathbf{V}}_\tau^{1,\omega})_{s2}\big| + \alpha_{22}^{\omega,\tau}\big|(\widehat{\mathbf{V}}_\tau^{2,\omega})_{s2}\big| + \alpha_{23}^{\omega,\tau}\big|(\widehat{\mathbf{V}}_\tau^{3,\omega})_{s2}\big| + \varepsilon_s, \quad s = 1, ..., 19, \quad \textbf{(M1)}\end{aligned}$$

where $\{\alpha_0^{\omega,\tau}, \alpha_{11}^{\omega,\tau}, \alpha_{12}^{\omega,\tau}, \alpha_{13}^{\omega,\tau}, \alpha_{21}^{\omega,\tau}, \alpha_{22}^{\omega,\tau}, \alpha_{23}^{\omega,\tau}\} \in \mathbb{R}^7$ are the intercept and the coefficients associated with the regions left and right Anterior insula, and dorsomedial prefrontal cortex.

In the case **(C2)**, define the averaged loadings of all tasks for each s

$$\mu_s^{b,\omega,\tau} \stackrel{\text{def}}{=} \frac{1}{256}\sum_{q=1}^{256}\big|(\widehat{\mathbf{V}}_\tau^{b,\omega})_{256(s-1)+q,1}\big|.$$

We construct another model for β_s using $\mu_s^{b,\omega,\tau}$:

$$\begin{aligned}\beta_s = \bar{\alpha}_0^{\omega,\tau} &+ \bar{\alpha}_{11}^{\omega,\tau}\mu_s^{1,\omega,\tau} + \bar{\alpha}_{12}^{\omega,\tau}\mu_s^{2,\omega,\tau} + \bar{\alpha}_{13}^{\omega,\tau}\mu_s^{3,\omega,\tau} \\ &+ \bar{\alpha}_{21}^{\omega,\tau}\mu_s^{1,\omega,\tau} + \bar{\alpha}_{22}^{\omega,\tau}\mu_s^{2,\omega,\tau} + \bar{\alpha}_{23}^{\omega,\tau}\mu_s^{3,\omega,\tau} + \varepsilon_s, \quad s = 1, ..., 19, \qquad \textbf{(M2)}\end{aligned}$$

where $\{\bar{\alpha}_0^{\omega,\tau}, \bar{\alpha}_{11}^{\omega,\tau}, \bar{\alpha}_{12}^{\omega,\tau}, \bar{\alpha}_{13}^{\omega,\tau}, \bar{\alpha}_{21}^{\omega,\tau}, \bar{\alpha}_{22}^{\omega,\tau}, \bar{\alpha}_{23}^{\omega,\tau}\} \in \mathbb{R}^7$. We take the absolute value of the

loadings $\widehat{\mathbf{V}}_{\tau}^{b,\omega}$ because we are only interested in the magnitude of the loadings, which describes the importance of the factors.

Remark 4.2. If sufficiently many subjects are available, then ideally we could use all the estimated factors as suggested by one of our referees. However, because we have only 19 subjects, the number of factor loadings that can be included is very limited. For example, according to the results of an extensive simulation study shown in Table 1 on page 438 in Knofczynski and Mundfrom (2008), the maximum number of predictors that guarantees the best prediction performance is perhaps only around 9 to 12, given the sample size 19. In an unreported analysis, we checked the out-of-sample performance of the models that include up to 3 and 4 factors loadings. We are not able to find strong evidences that more factor loadings improve the prediction performance.

4.2.3 In-sample and out-of-sample performance

We compare the in-sample and out-of-sample performance of the models (**M1**) and (**M2**). For the in-sample performance, R^2 of both regressions (**M1**) and (**M2**) are computed. In addition, in order to determine whether (**M1**) and (**M2**) correctly predict the *order* of risk-aversion of the subjects (rather than the exact value of β_s), we calculate the Spearman's and Kendall's rank correlations between the fitted $\widehat{\beta}_s$ (in-sample) and β_s.

To measure the out-of-sample performance, we calculate $\{\widetilde{\beta}_s\}_{s=1}^{19}$ by a leave-one-out procedure. The steps are as below:

(1) Fix s, where $s = 1, ..., 19$. Use the values of the remaining 18 subjects to compute the coefficients $\{\alpha_0^{\omega,\tau}, \alpha_{11}^{\omega,\tau}, \alpha_{12}^{\omega,\tau}, \alpha_{13}^{\omega,\tau}, \alpha_{21}^{\omega,\tau}, \alpha_{22}^{\omega,\tau}, \alpha_{23}^{\omega,\tau}\}$ in model (**M1**) or $\{\bar{\alpha}_0^{\omega,\tau}, \bar{\alpha}_{11}^{\omega,\tau}, \bar{\alpha}_{12}^{\omega,\tau}, \bar{\alpha}_{13}^{\omega,\tau}, \bar{\alpha}_{21}^{\omega,\tau}, \bar{\alpha}_{22}^{\omega,\tau}, \bar{\alpha}_{23}^{\omega,\tau}\}$ in model (**M2**) by the standard linear regression.

(2) Compute $\widetilde{\beta}_s$ by plugging in the coefficients computed in the last step in models (**M1**) and (**M2**), and input the loadings of the sth subject.

(3) Repeat steps (1) and (2) for each $s = 1, ..., 19$.

(4) Calculate the Spearman's and Kendall's rank correlations between $\{\widetilde{\beta}_s\}_{s=1}^{19}$ and $\{\beta_s\}_{s=1}^{19}$.

4.3 Empirical Results

In Table 4.1, we present the in-sample fitting and out-of-sample performance for models (**M1**) and (**M2**) with the constrained model that uses only the 1st factor ($\alpha_{21}^{\omega,\tau} = \alpha_{22}^{\omega,\tau} =$

$\alpha_{23}^{\omega,\tau} = 0$ in (**M1**) and $\bar{\alpha}_{21}^{\omega,\tau} = \bar{\alpha}_{22}^{\omega,\tau} = \bar{\alpha}_{23}^{\omega,\tau} = 0$ in (**M2**)) and the unconstrained model, under various (τ, ω) pairs.

For the in-sample fitting, cases $\omega = 0.1$ and $\omega = 0.9$ outperform the case $\omega = 0.5$. This shows that both extreme negative or positive BOLD can lead to good fitting for models (**M1**) and (**M2**). In particular, the fitting performance is the best when $\tau = 0.9$ for $\omega = 0.9$ and $\tau = 0.1$ for $\omega = 0.1$, which correspond to the upper boundary of the red area and the lower boundary of the blue area in each of the three panels in Figure 4.1.

For the out-of-sample performance, the constrained (**M2**) using only the first factor with the negative BOLD ($\omega = 0.1$, $\tau = 0.1$) nearly always outperforms all the other cases. In contrast, positive BOLD ($\omega = 0.9$) under the same model performs poorly. Moreover, the unconstrained model improves the prediction performance in most cases, particularly for (**M2**) under $\omega = 0.9$ and $\tau = 0.9$.

Majer et al. (2015) estimate a dynamic semiparametric factor model and extract the resulting factor loadings to predict the subjects' risk attitude. They evaluate the in-sample fitting (with all 19 subjects) by $R^2 = 0.47$ for a special case of our (**M1**) ($\tau = 0.5$ and $\alpha_{21}^{\omega,\tau} = \alpha_{22}^{\omega,\tau} = \alpha_{23}^{\omega,\tau} = 0$). Their fitting performance beats all the R^2 in our results, but we are able to describe the *predictive* abilities at several levels of τ, instead of only looking at $\tau = 0.5$. Our findings successfully confirm that the tails of the BOLD signals are more informative than their means in predicting the risk attitude.

		Constrained model (only 1st factor)						Unconstrained model (2 factors)					
		Whole series (**M1**)			Task-wise (**M2**)			Whole series (**M1**)			Task-wise (**M2**)		
	τ	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
		In-sample fitting											
ω=0.1	R^2	0.084	0.158	0.101	0.412	0.412	**0.413**	0.312	0.263	0.226	**0.455**	0.454	0.454
	Spearman's rank corr	0.149	0.377	0.328	0.595	0.595	**0.604**	0.532	0.526	0.396	**0.618**	**0.618**	**0.618**
	Kendall's rank corr	0.076	0.263	0.228	0.462	0.462	**0.474**	0.333	0.357	0.275	**0.474**	**0.474**	**0.474**
ω=0.5	R^2	0.070	0.043	0.030	0.134	0.136	0.135	0.307	0.260	0.352	**0.445**	0.440	0.441
	Spearman's rank corr	0.177	0.140	0.226	0.335	0.316	0.326	0.547	0.528	0.596	0.533	0.544	0.544
	Kendall's rank corr	0.135	0.088	0.135	0.205	0.193	0.205	0.427	0.333	**0.415**	0.368	0.380	0.380
ω=0.9	R^2	0.199	0.238	0.148	0.206	0.205	0.205	0.393	0.367	0.229	**0.487**	**0.496**	**0.500**
	Spearman's rank corr	0.435	0.540	0.181	0.412	0.412	0.412	0.588	**0.628**	0.582	0.596	**0.637**	**0.637**
	Kendall's rank corr	0.333	0.391	0.135	0.298	0.298	0.298	**0.439**	**0.439**	**0.439**	0.462	**0.497**	**0.497**
		Out-of-sample predicting											
ω=0.1	Spearman's rank corr	-0.453	-0.181	-0.321	**0.454**	0.451	0.440	-0.079	-0.133	0.072	0.298	0.298	0.298
	Kendall's rank corr	-0.322	-0.111	-0.240	**0.357**	0.345	0.345	-0.076	-0.088	0.041	0.216	0.216	0.216
ω=0.5	Spearman's rank corr	-0.444	-0.700	-0.658	-0.119	-0.119	-0.119	-0.035	-0.196	0.247	0.205	0.204	0.212
	Kendall's rank corr	-0.275	-0.509	-0.450	-0.064	-0.064	-0.064	-0.006	-0.146	0.135	0.123	0.111	0.123
ω=0.9	Spearman's rank corr	-0.207	0.204	-0.493	0.023	0.023	0.023	0.161	0.072	-0.447	0.293	0.307	0.307
	Kendall's rank corr	-0.170	0.135	-0.345	0.006	0.006	0.006	0.076	0.041	-0.298	0.205	0.216	0.216

Table 4.1: The goodness of fit R^2, Spearman's and Kendall's rank correlations for the in-sample fitting and out-of-sample prediction by (**M1**) or (**M2**) with/without constrains, under different τ, ω levels.

5 Conclusions

In this paper, we propose a factorizable multivariate expectile regression method for the high-dimensional cross-sectional or spatial data with sparse latent factors. Fast iterative shrinkage-thresholding algorithm is applied to estimate the model. The convergence of the algorithm and the non-asymptotic theoretical guarantee of the estimator are established. We apply our method on the fMRI data obtained from an investment decisions making experiment, and study the ranking accuracy of the subjects' risk preference using the factor loadings of the extreme BOLD responses. The results show that the negative BOLD signals could provide comparable prediction performance as the positive BOLD signals. This provides insights into the on-going debate on the meaning of the negative BOLD responses.

There are several possibilities for the future research. As many data in practice are time series, there is a need to relax the i.i.d. assumption and make our method compatible with richer temporal structure. Statistical inference is also an important issue for many applications.

References

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

Camerer, C. F. (2007). Neuroeconomics: Using neuroscience to make economic predictions, *The Economic Journal* **117**(519): C26–C42.

Chao, S.-K., Härdle, W. K. and Yuan, M. (2016). Factorisable multi-task quantile regression, *SFB 649 Discussion Paper 2016-057*, Sonderforschungsbereich 649, Humboldt Universität zu Berlin, Germany. Available at `http://sfb649.wiwi.hu-berlin.de/papers/pdf/SFB649DP2016-057.pdf`.

Chao, S.-K., Proksch, K., Dette, H. and Härdle, W. K. (2017). Confidence corridors for multivariate generalized quantile regression, *Journal of Business & Economic Statistics* **35**(1): 70–85.

Degras, D. and Lindquist, M. A. (2014). A hierarchical model for simultaneous detection and estimation in multi-subject fMRI studies, *NeuroImage* **98**: 61–72.

Fadili, J. M. and Peyré, G. (2011). Total variation projection with first order schemes, *Image Processing, IEEE Transactions* **20**(3): 657–669.

Heekeren, H. R., Marrett, S. and Ungerleider, L. G. (2008). The neural systems that mediate human perceptual decision making, *Nature Reviews Neuroscience* **9**(6): 467–479.

Helfinstein, S. M., Schonberg, T., Congdon, E., Karlsgodt, K. H., Mumford, J. A., Sabb, F. W., Cannon, T. D., London, E. D., Bilder, R. M. and Poldrack, R. A. (2014). Predicting risky choices from brain activity patterns, *Proceedings of the National Academy of Sciences* **111**(7): 2470–2475.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model, *Journal of Multivariate Analysis* **5**(2): 248–264.

Ji, S. and Ye, J. (2009). An accelerated gradient method for trace norm minimization, *Proceedings of the 26th International Conference on Machine Learning* .

Knofczynski, G. T. and Mundfrom, D. (2008). Sample sizes when using multiple linear regression for prediction, *Educational and Psychological Measurement* **68**(3): 431–442.

Koenker, R. and Bassett, G. W. (1978). Regression quantiles, *Econometrica* **46**(1): 33–50.

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

Lindquist, M. A. (2008). The statistical analysis of fMRI data, *Statistical Science* **23**(4): 439–464.
URL: *http://dx.doi.org/10.1214/09-STS282*

Majer, P., Mohr, P. N. C., Heekeren, H. and Härdle, W. K. (2015). Portfolio decisions and brain reactions via the CEAD method, *Psychometrika* **81**(3): 881–903.

Mullinger, K., Mayhew, S., Bagshaw, A., Bowtell, R. and Francis, S. (2014). Evidence that the negative BOLD response is neuronal in origin: A simultaneous EEG–BOLD–CBF study in humans, *NeuroImage* **94**: 263 – 274.

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

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

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing, *Econometrica* **55**(4): 819–847.

Opsomer, J., Wang, Y. and Yang, Y. (2001). Nonparametric regression with correlated errors, *Statist. Sci.* **16**(2): 134–153.
URL: *http://dx.doi.org/10.1214/ss/1009213287*

Reinsel, G. C. and Velu, R. P. (1998). *Multivariate Reduced-Rank Regression*, Springer, New York.

Rossi, G. D. and Harvey, A. (2009). Quantiles, expectiles and splines, *Journal of Econometrics* **152**(2): 179 – 185. Nonparametric and Robust Methods in Econometrics.
URL: *http://www.sciencedirect.com/science/article/pii/S0304407609000153*

Schultz, W. (2015). Neuronal reward and decision signals: From theories to data, *Physiological Reviews* **95**(3): 853–951.

van Bömmel, A., Song, S., Majer, P., Mohr, P. N. C., Heekeren, H. R. and Härdle, W. K. (2014). Risk patterns and correlated brain activities. multidimensional statistical analysis of fmri data in economic decision making study, *Psychometrika* **79**(3): 489–514.

Vershynin, R. (2012a). How close is the sample covariance matrix to the actual covariance matrix?, *Journal of Theoretical Probability* **25**(3): 655–686.
URL: *http://dx.doi.org/10.1007/s10959-010-0338-z*

Vershynin, R. (2012b). Introduction to the non-asymptotic analysis of random matrices, *in* Y. Eldar and G. Kutyniok (eds), *Compressed Sensing, Theory and Applications*, Cambridge University Press, chapter 5, pp. 210–268.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso), *IEEE Transactions on Information Theory* **55**: 2183–2202.

Yuan, M., Ekici, A., Lu, Z. and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression, *Journal of the Royal Statistical Society: Series B* **69**(3): 329–346.

APPENDIX

APPENDIX A: Proofs for Section 2.2

A.1 Proof for Theorem 2.1

Theorem 4.4 in Beck and Teboulle (2009) gives the upper bound of the loss difference at iteration t by

$$|F(\mathbf{\Gamma}_t) - F(\widehat{\mathbf{\Gamma}})| \leq \frac{2L_{\nabla g}\|\mathbf{\Gamma}_0 - \widehat{\mathbf{\Gamma}}\|_{\mathrm{F}}^2}{(t+1)^2}, \tag{A.1}$$

where $L_{\nabla g}$ is the Lipschitz constant of $\nabla g(\mathbf{\Gamma})$ defined in (2.9).

We note that

$$\rho'_\tau(u) = \begin{cases} 2\tau u & \text{for } u \geq 0; \\ 2(1-\tau)u & \text{for } u < 0. \end{cases} \tag{A.2}$$

Hence, the gradient is

$$\nabla g(\mathbf{\Gamma}) = -(mn)^{-1}\mathbf{X}^\top\big\{\mathbf{W} \circ (\mathbf{Y} - \mathbf{X}\mathbf{\Gamma})\big\}, \tag{A.3}$$

where $\mathbf{W}(\mathbf{\Gamma}) = (w_{ij}) \in \mathbb{R}^{n\times m}$, $w_{ij} \overset{\text{def}}{=} 2\left\{\tau + \mathbf{1}(Y_{ij} \leq \boldsymbol{X}_i^\top \mathbf{\Gamma}_{\cdot j})(1-2\tau)\right\}$, "$\circ$" represents the Hadamard product.

To simplify the notations, define $\mathbf{U}(\mathbf{\Gamma}) = (Y_{ij} - \boldsymbol{X}_i^\top \mathbf{\Gamma}_{\cdot j}) \in \mathbb{R}^{n\times m}$. For all $\mathbf{\Gamma}_1, \mathbf{\Gamma}_2 \in \mathbb{R}^{p\times m}$, let $\mathbf{U}_1 = \mathbf{U}(\mathbf{\Gamma}_1)$, $\mathbf{U}_2 = \mathbf{U}(\mathbf{\Gamma}_2)$, $\mathbf{W}_1 = \mathbf{W}(\mathbf{\Gamma}_1)$ and $\mathbf{W}_2 = \mathbf{W}(\mathbf{\Gamma}_2)$.

$$\begin{aligned}
\|\nabla g(\mathbf{\Gamma}_1) - \nabla g(\mathbf{\Gamma}_2)\|_{\mathrm{F}} &= (mn)^{-1}\|\mathbf{X}^\top(\mathbf{W}_1 \circ \mathbf{U}_1) - \mathbf{X}^\top(\mathbf{W}_2 \circ \mathbf{U}_2)\|_{\mathrm{F}} \\
&\leq (mn)^{-1}\|\mathbf{X}\|_{\mathrm{F}}\|\mathbf{W}_1 \circ \mathbf{U}_1 - \mathbf{W}_2 \circ \mathbf{U}_2\|_{\mathrm{F}} \qquad \text{(by submultiplicity)} \\
&= (mn)^{-1}\|\mathbf{X}\|_{\mathrm{F}}\Big[\sum_{i=1}^{n}\sum_{j=1}^{m}\big\{\rho'_\tau(u_{1,ij}) - \rho'_\tau(u_{2,ij})\big\}^2\Big]^{1/2} \\
&\leq (mn)^{-1}\|\mathbf{X}\|_{\mathrm{F}}\Big[\sum_{i=1}^{n}\sum_{j=1}^{m}\big\{2\max(\tau, 1-\tau)\big\}^2(u_{1,ij} - u_{2,ij})^2\Big]^{1/2} \\
&= 2(mn)^{-1}\max(\tau, 1-\tau)\|\mathbf{X}\|_{\mathrm{F}}\|\mathbf{Y} - \mathbf{X}\mathbf{\Gamma}_1 - (\mathbf{Y} - \mathbf{X}\mathbf{\Gamma}_2)\|_{\mathrm{F}} \\
&\leq 2(mn)^{-1}\max(\tau, 1-\tau)\|\mathbf{X}\|_{\mathrm{F}}^2\|\mathbf{\Gamma}_1 - \mathbf{\Gamma}_2\|_{\mathrm{F}} \quad \text{(by submultiplicity)},
\end{aligned} \tag{A.4}$$

where the fourth line makes use of the fact that $\rho'_\tau(u)$ is Lipschitz continuous with Lipschitz constant $2\max(\tau, 1-\tau)$, see Chao et al. (2017).

Plug $L_{\nabla g} = 2(mn)^{-1}\max(\tau, 1-\tau)\|\mathbf{X}\|_{\mathrm{F}}^2$ into (A.1) yields

$$|F(\mathbf{\Gamma}_t) - F(\widehat{\mathbf{\Gamma}})| \leq \frac{4(mn)^{-1}\max(\tau, 1-\tau)\|\mathbf{X}\|_{\mathrm{F}}^2\|\mathbf{\Gamma}_0 - \widehat{\mathbf{\Gamma}}\|_{\mathrm{F}}^2}{(t+1)^2}. \tag{A.5}$$

Moreover, setting the right hand side of (A.5) to be ϵ ($\forall \epsilon > 0$) and solving for t gives

$$t \geq \frac{2\sqrt{\max(\tau, 1-\tau)}\|\mathbf{X}\|_{\mathrm{F}}\|\mathbf{\Gamma}_0 - \widehat{\mathbf{\Gamma}}\|_{\mathrm{F}}}{\sqrt{mn\epsilon}} - 1. \tag{A.6}$$

□

A.2 Proof for Theorem 2.2

Following the proof of Theorem 1 in Fadili and Peyré (2011), define

$$I(\mathbf{\Gamma}_t) \stackrel{\text{def}}{=} g(\mathbf{\Gamma}_t) - g(\widehat{\mathbf{\Gamma}}) - \langle\!\langle \nabla g(\mathbf{\Gamma}_t), \mathbf{\Gamma}_t - \widehat{\mathbf{\Gamma}} \rangle\!\rangle, \tag{A.7}$$

$$J(\mathbf{\Gamma}_t) \stackrel{\text{def}}{=} h(\mathbf{\Gamma}_t) - h(\widehat{\mathbf{\Gamma}}) + \langle\!\langle \nabla g(\mathbf{\Gamma}_t), \mathbf{\Gamma}_t - \widehat{\mathbf{\Gamma}} \rangle\!\rangle, \tag{A.8}$$

the sum of them gives

$$I(\mathbf{\Gamma}_t) + J(\mathbf{\Gamma}_t) = F(\mathbf{\Gamma}_t) - F(\widehat{\mathbf{\Gamma}}). \tag{A.9}$$

According to Lemma C.2, we have

$$\begin{aligned} I(\mathbf{\Gamma}_t) &\geq \kappa\|\mathbf{\Gamma}_t - \widehat{\mathbf{\Gamma}}\|_{\mathrm{F}}^2 \\ &= \frac{1}{9}m^{-1}\min(\tau, 1-\tau)\sigma_{\min}(\mathbf{\Sigma})\|\mathbf{\Gamma}_t - \widehat{\mathbf{\Gamma}}\|_{\mathrm{F}}^2, \end{aligned} \tag{A.10}$$

where the second line holds with probability greater than $1 - a_n$ under (A1).

Since $\widehat{\mathbf{\Gamma}}$ is the optimizer of (2.6), therefore,

$$\mathbf{0} \in \nabla g(\widehat{\mathbf{\Gamma}}) + \nabla h(\widehat{\mathbf{\Gamma}}), \tag{A.11}$$

which implies

$$-\nabla g(\widehat{\mathbf{\Gamma}}) \in \nabla h(\widehat{\mathbf{\Gamma}}). \tag{A.12}$$

As a result, we have

$$h(\mathbf{\Gamma}_t) - h(\widehat{\mathbf{\Gamma}}) \geq \langle\!\langle -\nabla g(\mathbf{\Gamma}_t), \mathbf{\Gamma}_t - \widehat{\mathbf{\Gamma}} \rangle\!\rangle, \tag{A.13}$$

i.e., $J(\mathbf{\Gamma}_t) \geq 0$.

Plugging (A.10) and (A.13) into (A.9) yields,

$$\begin{aligned}\|\mathbf{\Gamma}_t - \widehat{\mathbf{\Gamma}}\|_{\mathrm{F}}^2 &\leq \frac{9m}{\min(\tau, 1-\tau)\sigma_{\min}(\mathbf{\Sigma})}\big\{F(\mathbf{\Gamma}_t) - F(\widehat{\mathbf{\Gamma}})\big\} \\ &\leq \frac{36}{n(t+1)^2}\frac{\max(\tau, 1-\tau)}{\min(\tau, 1-\tau)}\frac{\|\mathbf{X}\|_{\mathrm{F}}^2}{\sigma_{\min}(\mathbf{\Sigma})}\|\mathbf{\Gamma}_0 - \widehat{\mathbf{\Gamma}}\|_{\mathrm{F}}^2, \end{aligned} \tag{A.14}$$

with probability greater than $1 - a_n$. The second line comes from the result of Theorem 2.1. □

APPENDIX B: Proof for Theorem 2.3

By triangle inequality, we have

$$\|\mathbf{\Gamma}_t - \mathbf{\Gamma}\|_{\mathrm{F}}^2 = \|\mathbf{\Gamma}_t - \widehat{\mathbf{\Gamma}} + \widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_{\mathrm{F}}^2 \leq 2\|\mathbf{\Gamma}_t - \widehat{\mathbf{\Gamma}}\|_{\mathrm{F}}^2 + 2\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_{\mathrm{F}}^2. \tag{B.1}$$

Combining the results of Lemma B.2 and Theorem 2.2, it follows that

$$\begin{aligned}\|\mathbf{\Gamma}_t - \mathbf{\Gamma}\|_{\mathrm{F}}^2 \leq & 18^3 c^2 \frac{p+m}{n}\frac{\max(\tau, 1-\tau)^2}{\min(\tau, 1-\tau)^2}\frac{\|\mathbf{\Sigma}\|}{\sigma_{\min}(\mathbf{\Sigma})^2}K_u^2 \dim(\overline{\mathcal{M}}) \\ & + 144c\sqrt{\frac{p+m}{n}}\frac{\max(\tau, 1-\tau)}{\min(\tau, 1-\tau)}\frac{\sqrt{\|\mathbf{\Sigma}\|}}{\sigma_{\min}(\mathbf{\Sigma})}K_u\|\mathbf{\Gamma}_{\mathcal{M}^\perp}\|_* \\ & + \frac{72}{n(t+1)^2}\frac{\max(\tau, 1-\tau)}{\min(\tau, 1-\tau)}\frac{\|\mathbf{X}\|_{\mathrm{F}}^2}{\sigma_{\min}(\mathbf{\Sigma})}\|\mathbf{\Gamma}_0 - \widehat{\mathbf{\Gamma}}\|_{\mathrm{F}}^2, \end{aligned} \tag{B.2}$$

holds with probability greater than $1 - 3 \times 8^{-(p+m)} - a_n$.
Furthermore, given

$$\|\mathbf{\Gamma}_0 - \widehat{\mathbf{\Gamma}}\|_{\mathrm{F}}^2 = \|\mathbf{\Gamma}_0 - \mathbf{\Gamma} + \mathbf{\Gamma} - \widehat{\mathbf{\Gamma}}\|_{\mathrm{F}}^2 \leq 2\|\mathbf{\Gamma}_0 - \mathbf{\Gamma}\|_{\mathrm{F}}^2 + 2\|\mathbf{\Gamma} - \widehat{\mathbf{\Gamma}}\|_{\mathrm{F}}^2, \tag{B.3}$$

and applying Lemma B.2 again we complete the proof of Theorem 2.3. □

Now we show auxiliary results used in the proof of Theorem 2.3. The next theorem is an application of Theorem 1 of Negahban et al. (2012).

Theorem B.1 (Error bounds for the estimator)**.** *Under (A1), for any $q \in \{1, ..., p \wedge m\}$, any optimal solution $\widehat{\mathbf{\Gamma}}$ in the problem* (2.6) *with $\lambda \geq 2\|\nabla g(\mathbf{\Gamma})\|$ satisfies the bound*

$$\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_{\mathrm{F}}^2 \leq \frac{9m^2\lambda^2}{\big\{c_1 \min(\tau, 1-\tau)\sigma_{\min}(\mathbf{\Sigma})\big\}^2}q + \frac{36m\lambda}{\min(\tau, 1-\tau)\sigma_{\min}(\mathbf{\Sigma})}\sum_{j=q+1}^{p\wedge m}\sigma_j(\mathbf{\Gamma}), \tag{B.4}$$

with probability greater than $1 - a_n$, where $\sigma_j(\mathbf{\Gamma})$ is the jth singular value of $\mathbf{\Gamma}$.

Proof for Theorem B.1. The proof is an application of Theorem 1 of Negahban et al.

(2012). First, we observe that the nuclear norm is *decomposable* in the sense that

$$\|\mathbf{\Gamma} + \mathbf{\Delta}\|_* = \|\mathbf{\Gamma}\|_* + \|\mathbf{\Delta}\|_*, \forall \mathbf{\Gamma} \in \mathcal{M}_q, \mathbf{\Delta} \in \overline{\mathcal{M}}_q^{\perp}, \tag{B.5}$$

where

$$\begin{aligned} \mathcal{M}_q &= \mathcal{M}(U_q, V_q) \stackrel{\text{def}}{=} \{\mathbf{\Theta} \in \mathbb{R}^{p\times m} | \operatorname{col}(\mathbf{\Theta}) \subseteq U_q, \operatorname{row}(\mathbf{\Theta}) \subseteq V_q\}, \\ \overline{\mathcal{M}}_q^{\perp} &= \overline{\mathcal{M}}^{\perp}(U_q, V_q) \stackrel{\text{def}}{=} \{\mathbf{\Theta} \in \mathbb{R}^{p\times m} | \operatorname{col}(\mathbf{\Theta}) \subseteq U_q^{\perp}, \operatorname{row}(\mathbf{\Theta}) \subseteq V_q^{\perp}\}, \end{aligned} \tag{B.6}$$

where $\operatorname{row}(\mathbf{\Theta})$ and $\operatorname{col}(\mathbf{\Theta})$ denote the row and column spaces of $\mathbf{\Theta}$. It can be seen that $\mathcal{M}_q \subset \overline{\mathcal{M}}_q$ where $\overline{\mathcal{M}}_q \stackrel{\text{def}}{=} \{\mathbf{\Theta} \in \mathbb{R}^{p\times m} | \operatorname{tr}(\mathbf{\Theta}^\top \mathbf{S}) = 0, \ \forall \mathbf{S} \in \overline{\mathcal{M}}_q^{\perp}\}$. Similarly, $\mathcal{M}_q^{\perp} \stackrel{\text{def}}{=} \{\mathbf{\Theta} \in \mathbb{R}^{p\times m} | \operatorname{tr}(\mathbf{\Theta}^\top \mathbf{S}) = 0, \ \forall \mathbf{S} \in \mathcal{M}_q\}$.

We will verify its conditions (G1) and (G2). For condition (G1), it is already mentioned above that the nuclear norm $\|\cdot\|_*$ is decomposable with respect to $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$ defined in (B.6). For condition (G2), note that on the event

$$\Omega_1 \stackrel{\text{def}}{=} \left\{ \sigma_{\min}\left(\frac{\mathbf{X}^\top \mathbf{X}}{n}\right) \geq c_1 \sigma_{\min}(\mathbf{\Sigma}), \sigma_{\max}\left(\frac{\mathbf{X}^\top \mathbf{X}}{n}\right) \leq c_2 \sigma_{\max}(\mathbf{\Sigma}) \right\}, \tag{B.7}$$

the loss function g is restrictive strongly convex with coefficients κ and $\xi = 0$ (we replace $\tau_{\mathcal{L}}$ in Negahban et al. (2012) by ξ) shown in Lemma C.2. Since we measure the error in the Frobenius norm $\|\cdot\|_{\mathrm{F}}$, the subspace compatibility constant (Definition 3 of Negahban et al. (2012)) is

$$\Psi(\overline{\mathcal{M}}_q) \stackrel{\text{def}}{=} \sup_{\mathbf{S} \in \overline{\mathcal{M}}_q} \frac{\|\mathbf{S}\|_*}{\|\mathbf{S}\|_{\mathrm{F}}} \leq \sqrt{q}.$$

The conclusion of this Theorem follows from Theorem 1 of Negahban et al. (2012). □

Lemma B.1. *Under (A1)-(A3),*

$$\mathrm{P}\left(\|\nabla g(\mathbf{\Gamma})\| \leq c m^{-1} \max(\tau, 1-\tau) K_u \sqrt{\|\mathbf{\Sigma}\|} \sqrt{\frac{p+m}{n}} \right) \geq 1 - 3 \times 8^{-(p+m)} - a_n, \tag{B.8}$$

where $c > 0$ is an absolute constant.

Proof for Lemma B.1. Throughout the proof, we restrict on the event Ω_1 in (B.7). Recall the expression from (A.3) that

$$\nabla g(\mathbf{\Gamma}) = -(mn)^{-1} \mathbf{X}^\top \big\{ \mathbf{W} \circ (\mathbf{Y} - \mathbf{X}\mathbf{\Gamma}) \big\}.$$

and the matrix $\mathbf{U}(\mathbf{\Gamma}) = (u_{ij}) = (Y_{ij} - \boldsymbol{X}_i^\top \mathbf{\Gamma}_{\cdot j}) \in \mathbb{R}^{n\times m}$. Following the proof of Lemma 3

in Negahban and Wainwright (2011), we have

$$\begin{aligned}
\mathrm{P}\left(n^{-1}\|\mathbf{X}^\top(\mathbf{W}\circ\mathbf{U})\| \geq 4s\right) &= \mathrm{P}\left(\sup_{\substack{\boldsymbol{\beta}\in\mathcal{S}^{p-1},\\ \boldsymbol{\alpha}\in\mathcal{S}^{m-1}}} n^{-1}|\boldsymbol{\beta}^\top\mathbf{X}^\top(\mathbf{W}\circ\mathbf{U})\boldsymbol{\alpha}| \geq 4s\right) \\
&\leq 8^{p+m}\sup_{\substack{\boldsymbol{\beta}\in\mathcal{S}^{p-1},\\ \boldsymbol{\alpha}\in\mathcal{S}^{m-1}}} \mathrm{P}\left(n^{-1}|\langle\mathbf{X}\boldsymbol{\beta},(\mathbf{W}\circ\mathbf{U})\boldsymbol{\alpha}\rangle| \geq s\right) \\
&\leq 8^{p+m}\sup_{\substack{\boldsymbol{\beta}\in\mathcal{S}^{p-1},\\ \boldsymbol{\alpha}\in\mathcal{S}^{m-1}}} \mathrm{P}\left(n^{-1}\sum_{i=1}^{n}\langle\boldsymbol{\beta},\boldsymbol{X}_i\rangle\langle\boldsymbol{\alpha},(\mathbf{W}\circ\mathbf{U})_i\rangle \geq s\right),
\end{aligned} \tag{B.9}$$

where $\mathcal{S}^{m-1} \stackrel{\text{def}}{=} \{\boldsymbol{\alpha}\in\mathbb{R}^m : \|\boldsymbol{\alpha}\|_2 = 1\}$ is the Euclidean sphere in m-dimensions. $\forall s \geq 0$, there exists $C > 0$ such that $\mathrm{P}\left(|u_{ij}| > s\right) \leq \exp\left(1 - s^2/C^2\right)$. Since $|w_{ij}| \leq \max(\tau, 1-\tau)$, we have

$$\begin{aligned}
\mathrm{P}\left(|w_{ij}u_{ij}| > s\right) &\leq \mathrm{P}\left(\max(\tau,1-\tau)|u_{ij}| > s\right) \\
&= \mathrm{P}\left(|u_{ij}| > \frac{s}{\max(\tau,1-\tau)}\right) \\
&\leq \exp\left(1 - \frac{s^2}{\max(\tau,1-\tau)^2C^2}\right).
\end{aligned} \tag{B.10}$$

It means for each $j \in \{1,\dots m\}$, $w_{ij}u_{ij}$ are sub-gaussian. Moreover, the maximal sub-gaussian norm is bounded by

$$\begin{aligned}
\max_{1\leq j\leq m}\|w_{ij}u_{ij}\|_{\psi_2} &= \max_{1\leq j\leq m}\sup_{p\geq 1} p^{-1/2}\big(\mathsf{E}\,|w_{ij}u_{ij}|^p\big)^{1/p} \\
&\leq \max(\tau,1-\tau)\max_{1\leq j\leq m}\sup_{p\geq 1} p^{-1/2}\big(\mathsf{E}\,|u_{ij}|^p\big)^{1/p} \\
&= \max(\tau,1-\tau)K_u.
\end{aligned} \tag{B.11}$$

Then by Hoeffding's inequality (Proposition 5.10 of Vershynin, 2012b), we can conclude that $\left\langle\boldsymbol{\alpha},(\mathbf{W}\circ\mathbf{U})_i\right\rangle$ is also sub-guassian,

$$\begin{aligned}
\mathrm{P}\left(\left\langle\boldsymbol{\alpha},(\mathbf{W}\circ\mathbf{U})_i\right\rangle \geq s\right) &= \mathrm{P}\left(\Big|\sum_{j=1}^{m}\alpha_j w_{ij}u_{ij}\Big| \geq s\right) \\
&\leq \exp\left(1 - \frac{C's^2}{\max(\tau,1-\tau)^2K_u^2\|\boldsymbol{\alpha}\|_2^2}\right) \\
&= \exp\left(1 - \frac{C's^2}{\max(\tau,1-\tau)^2K_u^2}\right),
\end{aligned} \tag{B.12}$$

where $C' > 0$ is an absolute constant. Furthermore, its sub-gaussian norm is bounded by

$$\begin{aligned}
\big\|\langle \boldsymbol{\alpha}, (\mathbf{W}\circ\mathbf{U})_i\rangle\big\|_{\psi_2} &= \sup_{p\geq 1} p^{-1/2}\Big\{ \mathsf{E}\,\big|\langle \boldsymbol{\alpha}, (\mathbf{W}\circ\mathbf{U})_i\rangle\big|^p\Big\}^{1/p} \\
&= \sup_{p\geq 1} p^{-1/2}\bigg(\mathsf{E}\,\Big|\sum_{j=1}^m \alpha_j w_{ij}u_{ij}\Big|^p\bigg)^{1/p} \\
&\leq \max(\tau,1-\tau)\sup_{p\geq 1} p^{-1/2}\bigg(\mathsf{E}\,\Big|\sum_{j=1}^m \alpha_j u_{ij}\Big|^p\bigg)^{1/p} \\
&\leq \max(\tau,1-\tau) M K_u, \qquad \text{(B.13)}
\end{aligned}$$

where $M > 0$ is an absolute constant. The last line comes from Khintchine inequality (Corollary 5.12 of Vershynin, 2012b) and recall that $\|\boldsymbol{\alpha}\|_2 = 1$. Applying Hoeffding's inequality again we can obtain

$$\begin{aligned}
\mathrm{P}\left(n^{-1}\sum_{i=1}^n \langle \boldsymbol{\beta}, \boldsymbol{X}_i\rangle\langle \boldsymbol{\alpha}, (\mathbf{W}\circ\mathbf{U})_i\rangle \geq s\right) &\leq \exp\left(1 - \frac{C'' s^2 n}{\max(\tau,1-\tau)^2 M^2 K_u^2 n^{-1}\sum_{i=1}^n \langle \boldsymbol{\beta}, \boldsymbol{X}_i\rangle^2}\right) \\
&\leq \exp\left(1 - \frac{C'' s^2 n}{\max(\tau,1-\tau)^2 M^2 K_u^2 n^{-1}\|\mathbf{X}\boldsymbol{\beta}\|_2^2}\right), \\
&\leq \exp\left(1 - \frac{C'' s^2 n}{c_2 \max(\tau,1-\tau)^2 M^2 K_u^2 \|\boldsymbol{\Sigma}\|}\right), \qquad \text{(B.14)}
\end{aligned}$$

where C'' is an absolute constant. Combining (B.9) and (B.14) gives

$$\mathrm{P}\left(n^{-1}\big\|\mathbf{X}^\top(\mathbf{W}\circ\mathbf{U})\big\| \geq 4s\right) \leq \exp\left(1 - \frac{C'' s^2 n}{9\max(\tau,1-\tau)^2 M^2 K_u^2\|\boldsymbol{\Sigma}\|} + (p+m)\log 8\right). \qquad \text{(B.15)}$$

Set $s = \frac{1}{4} c \max(\tau,1-\tau) K_u \sqrt{\|\boldsymbol{\Sigma}\|}\sqrt{\frac{p+m}{n}}$, where $c \overset{\text{def}}{=} 4\cdot\sqrt{2\log 8 \frac{9M^2}{C''}}$, then we can conclude from the fact $P(\Omega_1) \geq 1 - a_n$,

$$\begin{aligned}
&\mathrm{P}\left(n^{-1}\big\|\mathbf{X}^\top(\mathbf{W}\circ\mathbf{U})\big\| \leq c\max(\tau,1-\tau)K_u\sqrt{\|\boldsymbol{\Sigma}\|}\sqrt{\frac{p+m}{n}}\right) \\
&\geq \big[1 - \exp\big(1 - (p+m)\log 8\big)\big] \times (1 - a_n) \\
&\geq \big[1 - 3\times 8^{-(p+m)}\big] \times (1 - a_n) \\
&\geq 1 - 3\times 8^{-(p+m)} - a_n \qquad (\text{as } p+m > 1). \qquad \text{(B.16)}
\end{aligned}$$

This finishes the proof. □

Lemma B.2. *Under (A1)-(A3), selecting* $\lambda = 2cm^{-1}\max(\tau,1-\tau)K_u\sqrt{\|\boldsymbol{\Sigma}\|}\sqrt{\frac{p+m}{n}}$, *for*

$n \geq 2\min(m,p)$, *any optimal solution* $\widehat{\mathbf{\Gamma}}$ *in the problem* (2.6) *satisfies the bound*

$$\begin{aligned}\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_{\mathrm{F}}^2 \leq & c' \frac{p+m}{n} \frac{\max(\tau, 1-\tau)^2}{\min(\tau, 1-\tau)^2} \frac{\|\mathbf{\Sigma}\|}{\sigma_{\min}(\mathbf{\Sigma})^2} K_u^2 \dim(\overline{\mathcal{M}}) \\ & + c' \sqrt{\frac{p+m}{n}} \frac{\max(\tau, 1-\tau)}{\min(\tau, 1-\tau)} \frac{\sqrt{\|\mathbf{\Sigma}\|}}{\sigma_{\min}(\mathbf{\Sigma})} K_u \|\mathbf{\Gamma}_{\mathcal{M}^\perp}\|_*, \end{aligned} \tag{B.17}$$

with probability greater than $1 - 3 \times 8^{-(p+m)} - a_n$, *where* $c, c' > 0$ *are absolute constants.*

Proof of Lemma B.2. Recall that Ω_1 is defined as (B.7), and let the event that (B.8) holds as Ω_2. On event $\Omega_1 \cap \Omega_2$, (B.17) can be achieved by simply plugging $\lambda = 2cm^{-1}\max(\tau, 1-\tau) K_u \sqrt{\|\mathbf{\Sigma}\|} \sqrt{\frac{p+m}{n}}$ into (B.4). We note that

$$\begin{aligned} \mathrm{P}(\Omega_2 \cap \Omega_1) = \mathrm{P}(\Omega_2 | \Omega_1)\, \mathrm{P}(\Omega_1) &\geq \left[1 - 3 \times 8^{-(p+m)}\right] \times (1 - a_n) \\ &\geq 1 - 3 \times 8^{-(p+m)} - a_n \qquad (\text{as } p + m > 1). \end{aligned} \tag{B.18}$$

□

APPENDIX C: Auxiliary Results

Lemma C.1. *For any* $u, \delta \in \mathbb{R}$ *and* $\tau \in (0,1)$,

$$\rho_\tau(u+\delta) - \rho_\tau(u) - \rho'_\tau(u)\delta \geq \min(\tau, 1-\tau)\delta^2. \tag{C.1}$$

Proof of Lemma C.1. When $u = 0$, we have $\rho_\tau(u) = \rho'_\tau(u) = 0$, therefore

$$\rho_\tau(\delta) = |\tau - \mathbf{1}\{\delta < 0\}|\delta^2 \geq \min(\tau, 1-\tau)\delta^2.$$

If $u > 0$, $u + \delta < 0$ ($\delta < 0$), we have

$$\rho_\tau(u+\delta) - \rho_\tau(u) - \rho'_\tau(u)\delta - \min(\tau, 1-\tau)\delta^2 = \begin{cases} (1-2\tau)(\delta+u)^2 \geq 0 & \text{for } \tau \leq 1-\tau; \\ (1-2\tau)(u+2\delta)u > 0 & \text{for } \tau > 1-\tau. \end{cases}$$

If $u > 0$, $u + \delta > 0$ ($\delta > 0$), we have

$$\rho_\tau(u+\delta) - \rho_\tau(u) - \rho'_\tau(u)\delta - \min(\tau, 1-\tau)\delta^2 = \begin{cases} (2\tau-1)(u+2\delta)u \geq 0 & \text{for } \tau \leq 1-\tau; \\ (2\tau-1)(u+\delta)^2 u > 0 & \text{for } \tau > 1-\tau. \end{cases}$$

In the other two cases,

$$\rho_\tau(u+\delta)-\rho_\tau(u)-\rho'_\tau(u)\delta=\begin{cases}\tau\delta^2\geq\min(\tau,1-\tau)\delta^2 & \text{for } u>0, u+\delta\geq 0;\\ (1-\tau)\delta^2\geq\min(\tau,1-\tau)\delta^2 & \text{for } u<0, u+\delta\leq 0.\end{cases}$$

Therefore, we can conclude that

$$\rho_\tau(u+\delta)-\rho_\tau(u)-\rho'_\tau(u)\delta\geq\min(\tau,1-\tau)\delta^2.$$

□

Lemma C.2. *$g(\mathbf{\Gamma})$ defined in (2.10) is κ-strongly convex and differentiable with $\kappa = m^{-1}\min(\tau,1-\tau)\sigma_{\min}(\frac{\mathbf{X}^\top\mathbf{X}}{n})$.*

Proof of Lemma C.2. Denote $\widetilde{u}_{ij}\overset{\text{def}}{=}Y_{ij}-\boldsymbol{X}_i^\top(\mathbf{\Gamma}_{\cdot j}+\mathbf{\Delta}_{\cdot j})$ and $u_{ij}\overset{\text{def}}{=}Y_{ij}-\boldsymbol{X}_i^\top\mathbf{\Gamma}_{\cdot j}$, for $i=1,\ldots,n$, $j=1,\ldots,m$, we have

$$\begin{aligned}\langle\!\langle\nabla g(\mathbf{\Gamma}),\mathbf{\Delta}\rangle\!\rangle &= \operatorname{tr}\left(\nabla g(\mathbf{\Gamma})^\top\mathbf{\Delta}\right)\\ &= -(mn)^{-1}\sum_{j=1}^m\sum_{l=1}^p\mathbf{\Delta}_{lj}\sum_{i=1}^n\rho'(u_{ij})X_{il}\\ &= -(mn)^{-1}\sum_{i=1}^n\sum_{j=1}^m\Big\{\sum_{l=1}^p\mathbf{\Delta}_{lj}\rho'(u_{ij})X_{il}\Big\}\\ &= -(mn)^{-1}\sum_{i=1}^n\sum_{j=1}^m\big\{\rho'(u_{ij})\boldsymbol{X}_i^\top\mathbf{\Delta}_{\cdot j}\big\}. \qquad \text{(C.2)}\end{aligned}$$

Therefore,

$$\begin{aligned}g(\mathbf{\Gamma}+\mathbf{\Delta})-g(\mathbf{\Gamma})-\langle\!\langle\nabla g(\mathbf{\Gamma}),\mathbf{\Delta}\rangle\!\rangle &= (mn)^{-1}\sum_{i=1}^n\sum_{j=1}^m\big\{\rho(\widetilde{u}_{ij})-\rho(u_{ij})+\rho'(u_{ij})\boldsymbol{X}_i^\top\mathbf{\Delta}_{\cdot j}\big\}\\ &\geq (mn)^{-1}\min(\tau,1-\tau)\sum_{i=1}^n\sum_{j=1}^m(\boldsymbol{X}_i^\top\mathbf{\Delta}_{\cdot j})^2 \text{ (by Lemma C.1)}\\ &= (mn)^{-1}\min(\tau,1-\tau)\|\mathbf{X}\mathbf{\Delta}\|_{\mathrm{F}}^2\\ &= (mn)^{-1}\min(\tau,1-\tau)\operatorname{tr}(\mathbf{\Delta}^\top\mathbf{X}^\top\mathbf{X}\mathbf{\Delta})\\ &\geq m^{-1}\min(\tau,1-\tau)\sigma_{\min}\big(\frac{\mathbf{X}^\top\mathbf{X}}{n}\big)\|\mathbf{\Delta}\|_{\mathrm{F}}^2. \qquad \text{(C.3)}\end{aligned}$$

□

APPENDIX D: Additional Details for Section 4

D.1 Tuning Parameters by Cross-Validation

Choosing $\omega = 0.1$, $b = 1$ (aINS_L cluster) in **(C1)** case as an example, Figure 5.1 illustrates the cross-validation error function in terms of λ under different τ levels. The optimal tuning parameters determined by 5-fold cross-validation under all cases are reported in Table 5.1.

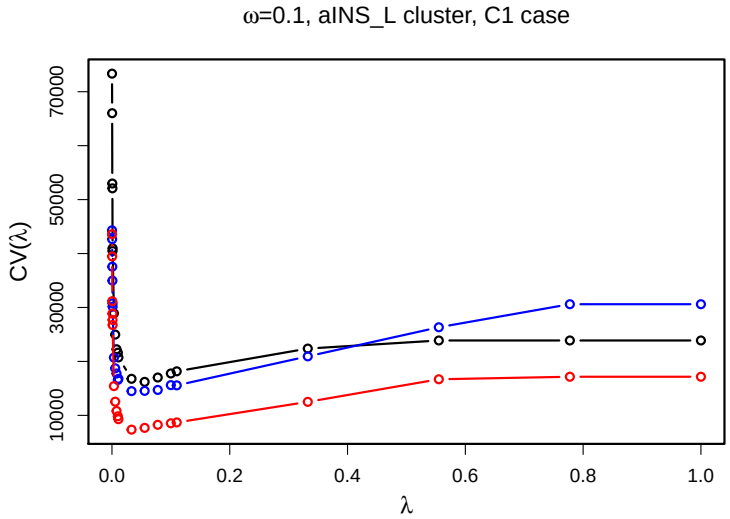

Figure 5.1: The cross-validation error function in terms of tuning parameter λ, with τ =0.1, 0.5, and 0.9, respectively.

		Whole series **(C1)**			Task-wise **(C2)**		
	τ	0.1	0.5	0.9	0.1	0.5	0.9
ω=0.1	$aINS_L$	0.0442	0.0552	0.0383	0.0008	0.0006	0.0008
	$aINS_R$	0.0303	0.0421	0.0293	0.0004	0.0008	0.0004
	DMPFC	0.0348	0.0504	0.0198	0.0004	0.0007	0.0006
ω=0.5	$aINS_L$	0.0181	0.0403	0.0153	0.0004	0.0006	0.0003
	$aINS_R$	0.0137	0.0393	0.0157	0.0006	0.0004	0.0005
	DMPFC	0.0195	0.0391	0.0143	0.0006	0.0002	0.0007
ω=0.9	$aINS_L$	0.0253	0.0408	0.0275	0.0006	0.0004	0.0004
	$aINS_R$	0.0243	0.0442	0.0200	0.0008	0.0002	0.0006
	DMPFC	0.0193	0.0474	0.0206	0.0005	0.0008	0.0008

Table 5.1: Tuning parameters by 5-fold cross validation.

D.2 Risk Attitude Parameter

The risk attitude parameter β is estimated by logistic model via maximum likelihood estimation (MLE)

$$\begin{aligned} \mathrm{P}\{\text{risky choice}|x\} &= \frac{1}{1+\exp[-\sigma\{\bar{x}-\beta S(x)-5\}]}, \\ \mathrm{P}\{\text{sure choice}|x\} &= 1-\frac{1}{1+\exp[-\sigma\{\bar{x}-\beta S(x)-5\}]}, \end{aligned} \tag{D.1}$$

where x is the return stream displayed to the individual, its mean and standard deviation are $\bar{x}$ and $S(x)$.

The estimated risk attitude parameters for 19 subjects in order are plotted in Figure 5.2, also see Majer et al. (2015). Negative parameters imply risk-seeking behaviors; while positive parameters indicate averse risk patterns. We can see most of the individuals are risk-averse and the two extremes #1 and #19 are the most risk-averse and most risk-seeking persons respectively.

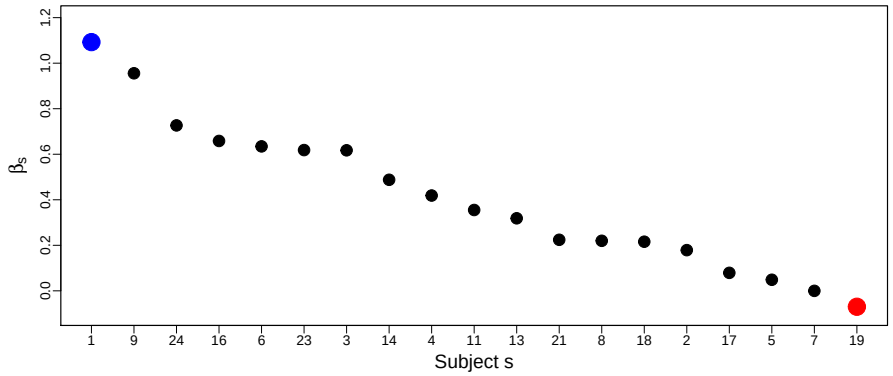

Figure 5.2: Estimated risk attitude for 19 subjects.