Electronic Journal of Statistics
Vol. 14 (2020) 143–173
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1664

Efficient estimation in expectile regression using envelope models

Tuo Chen

University of Florida
e-mail: chentuo@ufl.edu

Zhihua Su

University of Florida
e-mail: zhihuasu@stat.ufl.edu

Yi Yang

McGill University
e-mail: yi.yang6@mcgill.ca

and

Shanshan Ding

University of Delaware
e-mail: sding@udel.edu

Abstract: As a generalization of the classical linear regression, expectile regression (ER) explores the relationship between the conditional expectile of a response variable and a set of predictor variables. ER with respect to different expectile levels can provide a comprehensive picture of the conditional distribution of the response variable given the predictors. We adopt an efficient estimation method called the envelope model ([8]) in ER, and construct a novel envelope expectile regression (EER) model. Estimation of the EER parameters can be performed using the generalized method of moments (GMM). We establish the consistency and derive the asymptotic distribution of the EER estimators. In addition, we show that the EER estimators are asymptotically more efficient than the ER estimators. Numerical experiments and real data examples are provided to demonstrate the efficiency gains attained by EER compared to ER, and the efficiency gains can further lead to improvements in prediction.

Keywords and phrases: Sufficient dimension reduction, envelope model, expectile regression, generalized method of moments.

Received April 2018.

Contents

1 Introduction 144
2 Envelope expectile regression 147
2.1 Expectile regression 147
2.2 Envelope expectile regression 148

3 Estimation 150
4 Theoretical results 152
5 Simulations 154
6 Data analysis 161
6.1 state.x77 161
6.2 S&P 500 index 164
7 Extension to semiparametric settings 166
8 Discussion and future work 169
Supplementary Material 169
References 170

1. Introduction

The classical linear regression is a commonly used method when we want to study the relationship between a response variable Y and a p-dimensional predictor vector $\mathbf{X}$. It depicts the linear dependence between the conditional mean of Y and $\mathbf{X}$. However, it assumes that the conditional distribution of Y given $\mathbf{X}$ is homoscedastic, which is not always satisfied in real datasets. While the condition mean gives important information on the conditional distribution of Y and $\mathbf{X}$, it does not provide a complete picture of the distribution, especially for skewed distributions arising from price or income data. As a result, quantile regression (QR), which can overcome the limitations, gained considerable interest in recent years. QR studies the conditional quantile of Y given $\mathbf{X}$. It is a distribution free method and does not impose any assumption on the conditional distribution of the response Y. Investigation of multiple quantile levels gives us a comprehensive description of the distribution of Y conditional on $\mathbf{X}$.

An alternative to QR is expectile regression (ER). The idea of expectile was firstly studied in [32]. The πth expectile of Y, denoted by $f_\pi(Y)$, is defined as

$$f_\pi(Y) = \arg\min_{f\in\mathbb{R}} \mathrm{E}[\phi_\pi(Y-f)], \qquad \pi \in (0,1),$$

where $\phi_\pi(z) = |\pi - I(z < 0)|z^2$ is called the asymmetric least squares loss function. When $\pi = 0.5$, $\phi_\pi(z)$ is the least squares loss and $f_\pi(Y) = \mathrm{E}[Y]$. ER studies the conditional expectile of the response Y given the predictors $\mathbf{X}$. It has been explored in many statistics and econometric literature. [53] proposed a local linear polynomial estimator of the conditional expectiles with a one-dimensional predictor and established the corresponding asymptotic properties. [25] developed the conditional autoregressive expectile models (CARE), which investigated the impact of previous asset returns on the conditional expectile-based Value at Risk (EVaR) of current asset returns, and allowed different effects from positive and negative returns. They established the asymptotic normality of the CARE estimators, and extended the results in [32] to stationary and weakly dependent data. [50] proposed a varying-coefficient expectile model to estimate the conditional EVaR of asset returns. This approach

allows the coefficients to change with an effect modifying risk factor and provides a more flexible way to model the data. They showed that the varying-coefficient expectile model yields more stable estimators and more accurate prediction intervals compared to the CARE models. In recent years, many advances have been taken place on model selection in expecile regression. [19] studied the sparse expectile regression under high dimensional settings where the penalty functions include the Lasso and nonconvex penalties. [43] introduced a variety of model selection methods in semiparametric expectile regression. [28] provided asymptotic distributions of penalized expectile regression with SCAD and adaptive LASSO penalties for both *i.i.d.* and non-*i.i.d.* random errors. In addition, semiparametric ([37, 41, 40]) and nonparametric ([52, 51, 16, 22]) expectile estimation methods have been proposed in literature.

There are a lot of connections between QR and ER. Similar as QR, ER does not impose any distributional assumption on the response Y as well, and it can provide us with complete information on the conditional distribution of Y. Therefore, both QR and ER are able to overcome the previously mentioned limitation of the classical linear regression. Moreover, [53] pointed out there exists a one-to-one mapping between expectiles and quantiles. Hence ER can be interpreted as a flexible QR. In addition, expectiles can be used to estimate a quantile-based risk measure call expected shortfall [46]. However, at the same time, we should not ignore some different properties between QR and ER. Recall that the αth quantile of the response Y, denoted by $q_\alpha(Y)$, is defined as $q_\alpha(Y) = \arg\min_{q\in\mathbb{R}} \mathrm{E}\{\tau_\alpha(Y-q)\}$, where $\tau_\alpha(z) = |\alpha - I(z<0)||z|$ is the check loss function. Therefore quantiles minimize the expected check loss τ_α while expectiles minimize the expected asymmetric least squares loss ϕ_π. The check loss function is not differentiable while the asymmetric least squares loss function is differentiable because of the quadratic term. The difference in the loss functions give QR and ER their respective advantages. The main advantage of QR is that its results are easier to be interpreted and it is more robust to outliers than ER; while the main advantage of ER over QR is that ER is more computationally friendly, especially in a semiparametric model, as pointed out by [41]. Moreover, ER is more sensitive to the extreme values in datasets because ER takes the distance between an observation and the estimated expectile into account while QR only considers whether an observation is greater or less than the estimated quantile. This characteristic makes ER more desirable in many applications. One example is on the risk measures in the fields of econometrics and finance. Value at Risk (VaR) is a popular measure for evaluating portfolio risk based on quantiles. It provides crucial information about the potential loss, however, it is insensitive to the severity of more extreme realization since it does not depend on the tail shape of the distribution ([11, 15]). [25] proposed the risk measure EVaR and indicated EVaR can reflect the magnitude of the extreme losses for the underlying risk. Therefore, given a set of risk factors, studying conditional EVaR rather than conditional VaR may lead to a more proper respond to a catastrophic loss. At last, for the special case with $\alpha = 0.5$, QR degenerates

to the conditional median regression. For the special case with $\pi = 0.5$, ER degenerates to the standard linear regression.

In this paper we propose a new ER method, the *envelope expectile regression* (EER), for efficient estimation of the parameters in ER. It is motivated by the fact that some linear combinations of the predictors may be irrelevant to the distribution of the response. Our method takes this data structure into account, which results in more efficient estimation compared with the standard ER. We call those irrelevant combinations *the immaterial part* and the remaining part *the material part*. To identify the material part and the immaterial part, we employ a nascent technique called the *envelope* method. The immaterial part is then excluded from the subsequent analysis, leading to efficiency gains in estimation. To be noted, the immaterial part is different from the subset of inactive predictors in the popular penalized variable selection methods. We can see that both the penalized variable selection methods and the envelope method reduce the number of free parameters in the model but they are based on different assumptions. The penalized variable selection methods assume a subset of individual predictor variables are irrelevant to regression and thus have coefficients zero; while the envelope method assumes some linear combinations of the predictors are irrelevant to regression but all predictors can have nonzero coefficients. There are different application scenarios for the penalized variable selection methods and the envelope method. For example, suppose that a biologist wants to analyze the relationship between genes and a certain disease, given a dataset with measurements on the severity of the disease (response) and the expression levels of different genes (predictors). If the biologist believes only a few genes in the dataset have effects on the disease, then he would apply the penalized variable selection methods to identify those genes. On the other hand, if the biologist feels like each gene in the dataset is related to the disease in some way, then he would use dimension reduction methods such as principal component regression, partial least squares or the envelope model to find the linear combination(s) of the genes that affects the disease. The envelope method was first proposed in ([8]) under the context of multivariate linear regression for efficient estimation. It is then applied to many contexts in multivariate analysis including partial least squares ([6]), generalized linear models ([9]), elliptical linear regression ([17]), reduced rank regression ([7]), variable selection ([45]), bayesian analysis ([24]), matrix and tensor valued response ([12, 27]), and quantile regression ([13]). In this paper, we apply the idea of envelope method to expectile regression. The parameters are estimated by the generalized method of moments (GMM; [21]). And we demonstrate that the EER estimators have smaller asymptotic variance than the ER estimators both theoretically and numerically.

This paper is organized as follows. We derive the EER model in Section 2. Estimation of the EER model is discussed in Section 3. In Section 4, we investigate the theoretical properties of the EER estimators, and prove that the EER estimators are asymptotically more efficient than the ER estimators. We use simulations to demonstrate the performance of the EER model in estimation and prediction in Section 5. State murder rate data and S&P 500 index data are analyzed in Section 6 as examples. The extension to semiparametric settings

is discussed in Section 7. Section 8 concludes the paper with a discussion. All technical proofs are deferred to the supplementary material [5].

The following notations will be used in our discussion. Let $\mathbb{R}^{p\times u}$ be the set of all $p \times u$ matrices. For any matrix $\mathbf{A} \in \mathbb{R}^{p\times u}$, $\text{span}(\mathbf{A})$ represents the subspace spanned by the columns of $\mathbf{A}$. Let $\mathbf{P}_{\mathbf{A}}$ be the projection matrix onto $\text{span}(\mathbf{A})$ and $\mathbf{Q}_{\mathbf{A}} = \mathbf{I}_p - \mathbf{P}_{\mathbf{A}}$ be the projection matrix onto its orthogonal complement $\text{span}(\mathbf{A})^{\perp}$. For any subspace $\mathcal{S}$ in $\mathbb{R}^p$, $\mathbf{P}_{\mathcal{S}}$ represents the projection matrix onto $\mathcal{S}$ and $\mathbf{Q}_{\mathcal{S}}$ represents the projection matrix onto $\mathcal{S}^{\perp}$. Let "vec" represent the vectorization operator that vectorizes a matrix columnwise and "vech" represent the half-vectorization operator that vectorizes the lower triangle of a symmetric matrix. We use $\|\cdot\|$ to represent Frobenius norm and $\dagger$ to denote the Moore-Penrose inverse. We write $\xrightarrow{d}$ for convergence in distribution and $\xrightarrow{p}$ for convergence in probability. Moreover, a subspace $\mathcal{R}$ in $\mathbb{R}^p$ is a reducing subspace of a matrix $\mathbf{M} \in \mathbb{R}^{p\times p}$ if and only if $\mathbf{M}$ can be decomposed as $\mathbf{M} = \mathbf{P}_{\mathcal{R}}\mathbf{M}\mathbf{P}_{\mathcal{R}} + \mathbf{Q}_{\mathcal{R}}\mathbf{M}\mathbf{Q}_{\mathcal{R}}$.

2. Envelope expectile regression

2.1. Expectile regression

Consider a univariate response Y and a p-dimensional predictor vector $\mathbf{X}$. The standard ER considers the πth conditional expectile of Y as a linear function of $\mathbf{X}$

$$f_\pi(Y|\mathbf{X}) = \mu_\pi + \boldsymbol{\beta}_\pi^T\mathbf{X}, \tag{2.1}$$

where $f_\pi(Y|\mathbf{X})$ represents the πth conditional expectile of Y given $\mathbf{X}$, μ_π is the intercept and $\boldsymbol{\beta}_\pi \in \mathbb{R}^p$ contains the coefficients. When $\pi = 0.5$, $f_{0.5}(Y|\mathbf{X})$ degenerates to the conditional mean $\text{E}(Y|\mathbf{X})$ and ER degenerates to the standard linear regression.

To get the ER estimators, we use a property of the conditional expectile discussed in [32] that

$$(\mu_\pi, \boldsymbol{\beta}_\pi) = \underset{\mu\in\mathbb{R},\, \boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\ \text{E}\{\phi_\pi(Y - \mu - \boldsymbol{\beta}^T\mathbf{X})|\mathbf{X}\},$$

where $\phi_\pi(z) = |\pi - I(z < 0)|z^2$ is an asymmetric squared loss function. Given the random samples $\{(Y_i, \mathbf{X}_i) : i = 1, \ldots, n\}$ of $(Y, \mathbf{X})$, the ER estimators $\hat{\mu}_\pi$ and $\hat{\boldsymbol{\beta}}_\pi$ can be obtained by solving

$$(\hat{\mu}_\pi, \hat{\boldsymbol{\beta}}_\pi) = \underset{\mu_\pi\in\mathbb{R},\, \boldsymbol{\beta}_\pi\in\mathbb{R}^p}{\arg\min}\ \frac{1}{n}\sum_{i=1}^{n}\phi_\pi(Y_i - \mu_\pi - \boldsymbol{\beta}_\pi^T\mathbf{X}_i).$$

Taking the first derivative with respect to $(\mu_\pi,\ \boldsymbol{\beta}_\pi^T)^T$, the minimizer should satisfy the following estimating equations

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{W}_i(Y_i - \mu_\pi - \boldsymbol{\beta}_\pi^T\mathbf{X}_i)\left|I(Y_i < \mu_\pi + \mathbf{X}_i^T\boldsymbol{\beta}_\pi) - \pi\right| = 0, \tag{2.2}$$

where $\mathbf{W}_i = (1, \mathbf{X}_i^T)^T$.

2.2. Envelope expectile regression

EER is derived from the motivation that certain linear combinations of the predictors are irrelevant to some conditional expectiles, e.g. $\mathrm{E}(Y|\mathbf{X})$, of the response. For example, a stock index may be related to only a few combinations of the economic factors, while these combinations are uncorrelated with the other combinations that are not responsible for the variation in the index. Following this observation, we suppose that there exists a subspace $\mathcal{S}_\pi$ in the full predictor space $\mathbb{R}^p$ such that the πth conditional expectile of the response Y is related to the predictor vector $\mathbf{X}$ only through $\mathbf{P}_{\mathcal{S}_\pi}\mathbf{X}$. Specifically, we assume that

$$(a) \quad f_\pi(Y|\mathbf{X}) = f_\pi(Y|\mathbf{P}_{\mathcal{S}_\pi}\mathbf{X}) \quad \text{and} \quad (b) \quad \mathrm{Cov}(\mathbf{Q}_{\mathcal{S}_\pi}\mathbf{X}, \mathbf{P}_{\mathcal{S}_\pi}\mathbf{X}) = 0. \tag{2.3}$$

The conditions in (2.3) imply that $\mathbf{Q}_{\mathcal{S}_\pi}\mathbf{X}$ does not provide any information to the πth conditional expectile of Y neither from itself nor from its association with $\mathbf{P}_{\mathcal{S}_\pi}\mathbf{X}$. We call $\mathbf{P}_{\mathcal{S}_\pi}\mathbf{X}$ the material part of $\mathbf{X}$ and $\mathbf{Q}_{\mathcal{S}_\pi}\mathbf{X}$ the immaterial part of $\mathbf{X}$.

Remark 1. *Conditions in* (2.3) *incorporate the idea of sufficient dimension reduction in expectile regression. When* $\pi = 0.5$ *and the conditions in* (2.3) *are superimposed, model* (2.1) *is the envelope-based partial least squares (PLS) regression of* Y *on* $\mathbf{X}$ *([6]). It is well known that PLS regression improves prediction performance over ordinary least squares regression (e.g., [1, 4, 23]). For general* $\pi \in (0, 1)$*, conditions in* (2.3) *establish a general asymmetric envelope-based PLS regression framework, which improves estimation efficiency and enhances prediction performance over ER both theoretically and numerically as evidenced in Sections 4-6.*

Under the parameterization of ER (2.1), condition (2.3a) is equivalent to (i) $\boldsymbol{\beta}_\pi \in \mathcal{S}_\pi$, and condition (2.3$b$) holds if and only if (ii) $\mathcal{S}_\pi$ is a reducing subspace of $\boldsymbol{\Sigma}_\mathbf{X}$ ([8]), where $\boldsymbol{\Sigma}_\mathbf{X}$ is the covariance matrix of $\mathbf{X}$. If we find such a subspace $\mathcal{S}_\pi$, we can identify the immaterial part $\mathbf{Q}_{\mathcal{S}_\pi}\mathbf{X}$ when evaluating the relationship between $f_\pi(Y|\mathbf{X})$ and $\mathbf{X}$. Then estimation efficiency gains can be achieved by accounting for the immaterial variation in subsequent analysis. There may exist more than one subspace satisfying (i) and (ii). For instance, the full space $\mathbb{R}^p$ always satisfies (i) and (ii). To achieve the most efficiency gains, we focus on the smallest subspace (i.e., the subspace with smallest dimension) that satisfies (i) and (ii), such that we can identify all the immaterial information. We call this subspace the $\boldsymbol{\Sigma}_\mathbf{X}$-envelope of $\boldsymbol{\beta}_\pi$, and denote it by $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_\pi)$ or $\mathcal{E}_\pi$ if it appears in subscripts. The dimension of the envelope subspace $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_\pi)$ is denoted by u_π ($0 \le u_\pi \le p$). [8] discussed the existence and uniqueness of the envelope subspace $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_\pi)$.

Before deriving the EER model, we first discuss the parameterization of the envelope subspace $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_\pi)$. The envelope subspace can be determined by its basis. However, there can be many bases of $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_\pi)$. To make the parameters identifiable, we define one representative basis $\boldsymbol{\Gamma}_\pi$ for each envelope subspace: Take an arbitrary basis $\mathbf{G}_\pi$ of the envelope subspace $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_\pi)$. Since the dimension of the envelope subspace $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_\pi)$ is u_π, $\mathbf{G}_\pi$ has rank u_π and we can

find u_π rows in $\mathbf{G}_\pi$ that constitute a nonsingular matrix. If there are multiple combinations of such u_π rows, we take the rows with smallest indices. Without loss of generality, we assume that the first u_π rows of $\mathbf{G}_\pi$ constitute a nonsingular matrix, and we write $\mathbf{G}_\pi$ as $(\mathbf{G}_{1\pi}^T, \mathbf{G}_{2\pi}^T)^T$, where $\mathbf{G}_{1\pi} \in \mathbb{R}^{u_\pi \times u_\pi}$ and $\mathbf{G}_{2\pi} \in \mathbb{R}^{(p-u_\pi)\times u_\pi}$. Define the representative basis $\boldsymbol{\Gamma}_\pi = \mathbf{G}_\pi \mathbf{G}_{1\pi}^{-1} \equiv (\mathbf{I}_{u_\pi}, \mathbf{A}^T)^T$, where $\mathbf{A} \in \mathbb{R}^{(p-u_\pi)\times u_\pi}$. It can be shown that the representing basis defined as above is unique for $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_\pi)$. Therefore we have established a one-to-one correspondence between $\mathbf{A}$ and the envelope subspace $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_\pi)$. And if we know $\mathbf{A}$, we can completely decide $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_\pi)$. The following lemma shows that the basis of the orthogonal complement of $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_\pi)$ can also be expressed as a function of $\mathbf{A}$, which is useful for establishing the EER model. The proof is given in Section 1.1 of the supplementary material.

Lemma 1. *If* $\boldsymbol{\Gamma}_\pi = (\boldsymbol{I}_{u_\pi}, \mathbf{A}^T)^T$ *is a basis of* $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_\pi)$, *then* $\boldsymbol{\Gamma}_{0\pi} = (-\mathbf{A}, \boldsymbol{I}_{p-u_\pi})^T$ *is a basis of* $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_\pi)^\perp$.

Using the representative bases $\boldsymbol{\Gamma}_\pi$ and $\boldsymbol{\Gamma}_{0\pi}$, we reparametrize $\boldsymbol{\beta}_\pi$ and $\boldsymbol{\Sigma}_\mathbf{X}$ to derive the EER model. Since $\boldsymbol{\beta}_\pi \in \mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_\pi)$, there exists a u_π-dimensional coordinate vector $\boldsymbol{\eta}_\pi$ such that $\boldsymbol{\beta}_\pi = \boldsymbol{\Gamma}_\pi \boldsymbol{\eta}_\pi$. Because $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_\pi)$ is a reducing subspace of $\boldsymbol{\Sigma}_\mathbf{X}$, $\boldsymbol{\Sigma}_\mathbf{X}$ can be decomposed as $\boldsymbol{\Sigma}_\mathbf{X} = \mathbf{P}_{\mathcal{E}_\pi} \boldsymbol{\Sigma}_\mathbf{X} \mathbf{P}_{\mathcal{E}_\pi} + \mathbf{Q}_{\mathcal{E}_\pi} \boldsymbol{\Sigma}_\mathbf{X} \mathbf{Q}_{\mathcal{E}_\pi}$, where $\mathbf{P}_{\mathcal{E}_\pi} \boldsymbol{\Sigma}_\mathbf{X} \mathbf{P}_{\mathcal{E}_\pi}$ is the variance of the material part $\mathbf{P}_{\mathcal{E}_\pi} \mathbf{X}$ and $\mathbf{Q}_{\mathcal{E}_\pi} \boldsymbol{\Sigma}_\mathbf{X} \mathbf{Q}_{\mathcal{E}_\pi}$ is the variance of the immaterial part $\mathbf{Q}_{\mathcal{E}_\pi} \mathbf{X}$. With $\boldsymbol{\Gamma}_\pi$ and $\boldsymbol{\Gamma}_{0\pi}$ defined in Lemma 1, the coordinate form of the EER model is as follows

$$\begin{aligned} f_\pi(Y|\mathbf{X}) &= \mu_\pi + \boldsymbol{\eta}_\pi^T \boldsymbol{\Gamma}_\pi^T \mathbf{X} \\ \boldsymbol{\Sigma}_\mathbf{X} &= \boldsymbol{\Gamma}_\pi \boldsymbol{\Omega}_\pi \boldsymbol{\Gamma}_\pi^T + \boldsymbol{\Gamma}_{0\pi} \boldsymbol{\Omega}_{0\pi} \boldsymbol{\Gamma}_{0\pi}^T. \end{aligned} \tag{2.4}$$

In the EER model, we can see the predictor vector $\mathbf{X}$ affects the conditional expectile $f_\pi(Y|\mathbf{X})$ only through its linear combinations $\boldsymbol{\Gamma}_\pi^T \mathbf{X}$. The number of the linear combinations is u_π because $\boldsymbol{\Gamma}_\pi$ has u_π columns. Therefore, u_π represents the number of relevant linear combinations of the predictors, rather than the number of active predictors in penalized models. The u_π-dimensional vector $\boldsymbol{\eta}_\pi$ carries the coordinates of $\boldsymbol{\beta}_\pi$ relative to $\boldsymbol{\Gamma}_\pi$. The positive definite matrix $\boldsymbol{\Omega}_\pi \in \mathbb{R}^{u_\pi \times u_\pi}$ carries the coordinates of $\boldsymbol{\Sigma}_\mathbf{X}$ relative to $\boldsymbol{\Gamma}_\pi$ and the positive definite matrix $\boldsymbol{\Omega}_{0\pi} \in \mathbb{R}^{(p-u_\pi)\times(p-u_\pi)}$ carries the coordinates of $\boldsymbol{\Sigma}_\mathbf{X}$ relative to $\boldsymbol{\Gamma}_{0\pi}$. The parameter vector in the EER model is then $\boldsymbol{\zeta} = (\mu_\pi, \boldsymbol{\eta}_\pi^T, \mathrm{vec}(\mathbf{A})^T, \mathrm{vech}(\boldsymbol{\Omega}_\pi)^T, \mathrm{vech}(\boldsymbol{\Omega}_{0\pi})^T, \boldsymbol{\mu}_\mathbf{X}^T)^T$, where $\boldsymbol{\mu}_\mathbf{X}$ is the mean of $\mathbf{X}$. Here, we include $\boldsymbol{\mu}_\mathbf{X}$ in the parameter vector since it is involved in the estimating equations in Section 3. The total number of parameters is $1 + u_\pi + p + p(p+1)/2$. The parameter count is as follows: μ_π has 1 parameter, $\boldsymbol{\eta}_\pi$ has u_π parameters, $\mathbf{A}$ has $u_\pi(p-u_\pi)$ parameters, $\boldsymbol{\Omega}_\pi$ and $\boldsymbol{\Omega}_{0\pi}$ are both symmetric matrices, and they have $u_\pi(u_\pi+1)/2$ and $(p-u_\pi)(p-u_\pi+1)/2$ parameters respectively, and $\boldsymbol{\mu}_\mathbf{X}$ has p parameters. When $u_\pi = p$, the EER model degenerates to the ER model (2.1). The parameters in ER are μ_π and $\boldsymbol{\beta}_\pi$. In this paper, we also consider $\boldsymbol{\Sigma}_\mathbf{X}$ and $\boldsymbol{\mu}_\mathbf{X}$ as parameters in ER to make it comparable with the EER model when asymptotic covariance of the estimators is concerned. The parameter vector in ER is $\boldsymbol{\theta} = (\mu_\pi, \boldsymbol{\beta}_\pi^T, \mathrm{vech}(\boldsymbol{\Sigma}_\mathbf{X})^T, \boldsymbol{\mu}_\mathbf{X}^T)^T$, and the total number of parameters is

$1 + 2p + p(p+1)/2$. Comparing the number of parameters, we can see that the EER model reduces the number of parameters by $p - u_\pi$.

3. Estimation

In this section, we derive the EER estimators. Since there is no distributional assumption on the predictors or the response, the maximum likelihood estimation is not applicable for the EER model. Instead we adopt the generalized method of moments (GMM; [21]) to obtain the EER estimators.

We first present the estimating equations under ER, and then reparameterize the equations for the estimation of the EER model. The estimating equations under ER are constructed from (2.2) and the moment conditions of $\mathbf{X}$:

$$e_n(\boldsymbol{\theta}) = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n \mathbf{W}_i(Y_i - \mu_\pi - \mathbf{X}_i^T\boldsymbol{\beta}_\pi)\left|I(Y_i < \mu_\pi + \mathbf{X}_i^T\boldsymbol{\beta}_\pi) - \pi\right| \\ \operatorname{vech}(\boldsymbol{\Sigma}_\mathbf{X}) - \operatorname{vech}(\mathbf{S}_\mathbf{X}) \\ \boldsymbol{\mu}_\mathbf{X} - \bar{\mathbf{X}} \end{pmatrix} = 0, \tag{3.1}$$

where $\bar{\mathbf{X}} = \sum_{i=1}^n \mathbf{X}_i/n$ is the sample mean and $\mathbf{S}_\mathbf{X} = \sum_{i=1}^n (\mathbf{X}_i - \boldsymbol{\mu}_\mathbf{X})(\mathbf{X}_i - \boldsymbol{\mu}_\mathbf{X})^T/n$ is the sample covariance matrix given $\boldsymbol{\mu}_\mathbf{X}$. Let $\tilde{\boldsymbol{\theta}}$ denote the ER estimator by solving (3.1).

Under the EER model (2.4), we have $\boldsymbol{\beta}_\pi = \boldsymbol{\Gamma}_\pi\boldsymbol{\eta}_\pi$ and $\boldsymbol{\Sigma}_\mathbf{X} = \boldsymbol{\Gamma}_\pi\boldsymbol{\Omega}_\pi\boldsymbol{\Gamma}_\pi^T + \boldsymbol{\Gamma}_{0\pi}\boldsymbol{\Omega}_{0\pi}\boldsymbol{\Gamma}_{0\pi}^T$. Then we can build a map ψ between the EER parameter vector $\boldsymbol{\zeta}$ and the ER parameter vector $\boldsymbol{\theta}$, i.e., $\psi(\boldsymbol{\zeta}) = \boldsymbol{\theta}$. Then we can reparameterize (3.1) to get the estimating equations for the EER model:

$$e_n^*(\boldsymbol{\zeta}) = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n \mathbf{W}_i(Y_i - \mu_\pi - \mathbf{X}_i^T\boldsymbol{\Gamma}_\pi\boldsymbol{\eta}_\pi)\left|I(Y_i < \mu_\pi + \mathbf{X}_i^T\boldsymbol{\Gamma}_\pi\boldsymbol{\eta}_\pi) - \pi\right| \\ \operatorname{vech}(\boldsymbol{\Gamma}_\pi\boldsymbol{\Omega}_\pi{\boldsymbol{\Gamma}_\pi}^T + \boldsymbol{\Gamma}_{0\pi}\boldsymbol{\Omega}_{0\pi}{\boldsymbol{\Gamma}_{0\pi}}^T) - \operatorname{vech}(\mathbf{S}_\mathbf{X}) \\ \boldsymbol{\mu}_\mathbf{X} - \bar{\mathbf{X}} \end{pmatrix} = 0. \tag{3.2}$$

Note that in the EER model (2.4), other than $\boldsymbol{\zeta} = (\mu_\pi, \boldsymbol{\eta}_\pi^T, \operatorname{vec}(\mathbf{A})^T, \operatorname{vech}(\boldsymbol{\Omega}_\pi)^T, \operatorname{vech}(\boldsymbol{\Omega}_{0\pi})^T, \boldsymbol{\mu}_\mathbf{X}^T)^T$, the dimension of the envelope subspace u_π is also an important parameter. Here we first discuss the estimation of $\boldsymbol{\zeta}$ assuming u_π is known. In the estimating equation (3.2), there are $1 + 2p + p(p+1)/2$ equations and $1 + u_\pi + p + p(p+1)/2$ unknown parameters. As a result, it is possible that no solution exists. We then apply the GMM approach ([21]) to obtain the EER estimator $\hat{\boldsymbol{\zeta}}$. Let $\mathbf{Z} = (\mathbf{X}^T, Y)^T$ and $\mathbf{W} = (1, \mathbf{X}^T)^T$. Define $s(\mathbf{Z}; \boldsymbol{\theta})$ to be the population version of the moment conditions in (3.1):

$$s(\mathbf{Z}; \boldsymbol{\theta}) = \begin{pmatrix} s_1(\mathbf{Z}; \boldsymbol{\theta}) \\ s_2(\mathbf{Z}; \boldsymbol{\theta}) \\ s_3(\mathbf{Z}; \boldsymbol{\theta}) \end{pmatrix} = \begin{pmatrix} \mathbf{W}(Y - \mu_\pi - \mathbf{X}^T\boldsymbol{\beta}_\pi)\left|I(Y < \mu_\pi + \mathbf{X}^T\boldsymbol{\beta}_\pi) - \pi\right| \\ \operatorname{vech}(\boldsymbol{\Sigma}_\mathbf{X}) - \operatorname{vech}\{(\mathbf{X} - \boldsymbol{\mu}_\mathbf{X})(\mathbf{X} - \boldsymbol{\mu}_\mathbf{X})^T\} \\ \boldsymbol{\mu}_\mathbf{X} - \mathbf{X} \end{pmatrix}. \tag{3.3}$$

The estimation procedure can be summarized in the following steps.

Step 1: Get the intermediate estimator $\hat{\boldsymbol{\zeta}}^*$ by minimizing $e_n^*(\boldsymbol{\zeta})^T e_n^*(\boldsymbol{\zeta})$.

Step 2: Compute the scale matrix

$$\hat{\boldsymbol{\Delta}} = \left[\frac{1}{n}\sum_{i=1}^{n} s(\mathbf{Z}_i; \psi(\hat{\boldsymbol{\zeta}}^*))s(\mathbf{Z}_i; \psi(\hat{\boldsymbol{\zeta}}^*))^T\right]^{-1}.$$

Step 3: Obtain the GMM estimator $\hat{\boldsymbol{\zeta}}$ by minimizing $e_n^*(\boldsymbol{\zeta})^T \hat{\boldsymbol{\Delta}} e_n^*(\boldsymbol{\zeta})$.

Once we obtain $\hat{\boldsymbol{\zeta}}$, the EER estimators of $\boldsymbol{\beta}_\pi$ and $\boldsymbol{\Sigma}_{\mathbf{X}}$ are $\hat{\boldsymbol{\beta}}_\pi = \hat{\boldsymbol{\Gamma}}_\pi \hat{\boldsymbol{\eta}}_\pi$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{X}} = \hat{\boldsymbol{\Gamma}}_\pi \hat{\boldsymbol{\Omega}}_\pi \hat{\boldsymbol{\Gamma}}_\pi^T + \hat{\boldsymbol{\Gamma}}_{0\pi} \hat{\boldsymbol{\Omega}}_{0\pi} \hat{\boldsymbol{\Gamma}}_{0\pi}^T$. And we use $\hat{\boldsymbol{\theta}}$ to denote the EER estimator of $\boldsymbol{\theta}$: $\hat{\boldsymbol{\theta}} = (\hat{\mu}_\pi, \hat{\boldsymbol{\beta}}_\pi^T, \text{vech}(\hat{\boldsymbol{\Sigma}}_{\mathbf{X}})^T, \hat{\boldsymbol{\mu}}_{\mathbf{X}}^T)^T$. An analysis of the computational burden of the estimation procedure is given in Section 8 of the Supplement.

Remark 2. *In some envelope literature, e.g., [8], a different parameterization is adopted for* $\boldsymbol{\Gamma}_\pi$*:* $\boldsymbol{\Gamma}_\pi$ *is required to be a semi-orthogonal matrix, i.e.,* $\boldsymbol{\Gamma}_\pi^T \boldsymbol{\Gamma}_\pi = \mathbf{I}_{u_\pi}$*. In this case, span*$(\boldsymbol{\Gamma}_\pi)$ *is a point on a* $p \times u_\pi$ *Grassmann manifold, where a* $p \times u_\pi$ *Grassmann manifold is the set of all* u_π*-dimensional subspaces in a* p*-dimensional space. The above procedure can still be used to estimate the parameters, except that Step 1 and Step 3 involve Grassmann manifold optimization, which could be complicated and slow in sizable problems. For more information about Grassmann manifold optimization algorithms, please refer to [14, 29].*

Now we discuss the selection of u_π. Similar to other dimension reduction based methods (such as principal component analysis, partial least squares or reduced rank regression), the model selection for the EER model (2.4) is essentially the selection of the dimension u_π. In existing envelope models and methods, the dimension u_π is usually chosen by AIC, BIC or log-likelihood ratio testing. Since AIC, BIC as well as log-likelihood ratio testing all requires a likelihood function, they are not applicable in the context of the EER model. As a result, we adopt a nonparameteric method, robust cross validation (RCV; [33]), for the selection of u_π. Following [18, page 244], we use the "one-standard error" rule with RCV to choose the most parsimonious model which has about the same predictive accuracy as the best model. RCV is performed in the following steps:

Step 1: Randomly split the data into K folds. Usually K takes the value of 5 or 10. Successively use the kth fold for testing and the remaining folds for training, $k = 1, \ldots, K$.

Step 2: For each possible u_π $(0 \le u_\pi \le p)$, compute the mean expectile loss $\text{RCV}(u_\pi) = \frac{1}{n}\sum_{i=1}^{n} \phi_\pi(Y_i - \hat{\mu}_{\pi,-k(i)} - \hat{\boldsymbol{\beta}}_{\pi,-k(i)}^T \mathbf{X}_i)$, where $\hat{\mu}_{\pi,-k(i)}$ and $\hat{\boldsymbol{\beta}}_{\pi,-k(i)}$ are the EER estimators using the data excluding the kth fold that contains the ith observation.

Step 3: Instead of choosing the $\tilde{u}_\pi$ which achieves the smallest mean expectile loss $\text{RCV}(\tilde{u}_\pi)$, we select the smallest $\hat{u}_\pi$ whose mean expectile loss is less than one standard error above $\text{RCV}(\tilde{u}_\pi)$.

We provide an implementation of the EER model in the `R` package `expectEnv` which is available at `https://github.com/chentuo1993/expectEnv`. Using the

GMM approach, this package provides the EER estimator for parameters $\boldsymbol{\zeta}$ and $\boldsymbol{\theta}$. It also implements RCV for the selection of the envelope dimension u_π and computes bootstrap standard deviations for the EER estimators.

4. Theoretical results

In this section, we prove the EER estimator $\hat{\boldsymbol{\theta}}$ is asymptotically more efficient than or as efficient as the ER estimator $\tilde{\boldsymbol{\theta}}$. We first derive the asymptotic distribution of $\tilde{\boldsymbol{\theta}}$. [32] proved that the ER estimator $\tilde{\boldsymbol{\beta}}_\pi$ is consistent and asymptotically normal under some regularity conditions. We extend this result from $\tilde{\boldsymbol{\beta}}_\pi$ to $\tilde{\boldsymbol{\theta}}$ in Theorem 1. Next we establish the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ in Theorem 2, and show that the EER estimator is at least as efficient as the ER estimator by comparing the asymptotic covariance matrices. Because of the non-smoothness in the estimating equations and the EER model does not impose any assumptions on the distribution of Y, the traditional likelihood approach in envelope literature cannot be applied to derive the asymptotic distribution of $\hat{\boldsymbol{\theta}}$. Furthermore, there is over-parameterization in the estimating equation in the sense that the number of equations are greater than the number of the parameters. The theoretical tools we use to overcome these issues are Theorem 2.1 in [31] and Proposition 4.1 in [39]. To simplify the notations, we use $\text{avar}(\sqrt{n}\tilde{\boldsymbol{\theta}})$ to denote the asymptotic covariance matrix of $\tilde{\boldsymbol{\theta}}$ and $\text{avar}(\sqrt{n}\hat{\boldsymbol{\theta}})$ to denote the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ hereafter.

First we derive the asymptotic distribution of the ER estimator, and the following regularity conditions are assumed.

(C1) For each sample size n, $\mathbf{Z}_i = (\mathbf{X}_i^T, Y_i)^T$ is independent and identically distributed. Let f_1 be the conditional density of Y given $\mathbf{X}$ and f_2 be the marginal density function of $\mathbf{X}$. Then $\mathbf{Z}_i$ has a probability density function $f_1(Y_i|\mathbf{X}_i)f_2(\mathbf{X}_i)$ with respect to a measure $\mu_{\mathbf{Z}}$. Also, the conditional density $f_1(Y|\mathbf{X})$ is continuous in Y for almost all $\mathbf{X}$.

(C2) There is a constant $d \geq 0$ and a measurable function $\alpha(\mathbf{Z})$ that satisfy $f_1(Y|\mathbf{X}) \leq \alpha(\mathbf{Z})$, $\int |\mathbf{Z}|^{4+d}\alpha(\mathbf{Z})f_2(\mathbf{X})d\mu_{\mathbf{Z}} \leq +\infty$, and $\int \alpha(\mathbf{Z})f_2(\mathbf{X})d\mu_{\mathbf{Z}} \leq +\infty$.

(C3) $\mathrm{E}(\mathbf{X}\mathbf{X}^T)$ is nonsingular.

(C4) Let $\boldsymbol{\theta}_0$ be the true value of $\boldsymbol{\theta}$. The expectation $\mathrm{E}_{\boldsymbol{\theta}_0}[s(\mathbf{Z};\boldsymbol{\theta})]$ is twice differentiable at $\boldsymbol{\theta}_0$ with $\left.\frac{\partial \mathrm{E}_{\boldsymbol{\theta}_0}[s(\mathbf{Z};\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ having full rank and a finite Frobenius norm. The matrix $\mathrm{E}_{\boldsymbol{\theta}_0}[s(\mathbf{Z};\boldsymbol{\theta}_0)s(\mathbf{Z};\boldsymbol{\theta}_0)^T]$ is positive definite and has a finite Frobenius norm.

(C5) The support $\boldsymbol{\Theta}$ of $\boldsymbol{\theta}$ is a compact set and $\boldsymbol{\theta}_0$ is an interior point of $\boldsymbol{\Theta}$.

Conditions (C1)–(C3) are the conditions used in Theorem 3 of [32] to prove the consistency of the expectile estimator $\tilde{\boldsymbol{\beta}}_\pi$ in ER. Let $\boldsymbol{\theta}_1 = (\mu_\pi, \boldsymbol{\beta}_\pi^T)^T$, $\boldsymbol{\theta}_2 = (\text{vech}(\boldsymbol{\Sigma}_{\mathbf{X}})^T, \boldsymbol{\mu}_{\mathbf{X}}^T)^T$, and let $\boldsymbol{\theta}_{10}$ and $\boldsymbol{\theta}_{20}$ be the true value of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ respectively. Then the dependence of s_1, s_2 and s_3 in (3.3) on $\boldsymbol{\theta}$ can be specified as $s_1(\mathbf{Z};\boldsymbol{\theta}) = s_1(\mathbf{Z};\boldsymbol{\theta}_1)$, $s_2(\mathbf{Z};\boldsymbol{\theta}) = s_2(\mathbf{Z};\boldsymbol{\theta}_2)$ and $s_3(\mathbf{Z};\boldsymbol{\theta}) = s_3(\mathbf{Z};\boldsymbol{\theta}_2)$.

Theorem 1. *Assume the above regularity conditions (C1)–(C5) are satisfied, then we have*

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{C}^{-1}\mathbf{G}\mathbf{C}^{-1}),$$

where

$$\begin{aligned}\mathbf{C} &= \frac{\partial \mathrm{E}_{\boldsymbol{\theta}_0}[s(\mathbf{Z};\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\ &= \begin{pmatrix} -\mathrm{E}_{\boldsymbol{\theta}_0}\left[\mathbf{W}\mathbf{W}^T \left|I(Y < \mathbf{W}^T\boldsymbol{\theta}_{10}) - \pi\right|\right] & 0 & 0 \\ 0 & \boldsymbol{I}_{p(p+1)/2} & 0 \\ 0 & 0 & \boldsymbol{I}_p \end{pmatrix}\end{aligned}$$

and

$$\mathbf{G} = \mathrm{E}_{\boldsymbol{\theta}_0}[s(\mathbf{Z};\boldsymbol{\theta}_0)s(\mathbf{Z};\boldsymbol{\theta}_0)^T] = \begin{pmatrix} \mathbf{G}_{1,1} & 0 & 0 \\ 0 & \mathbf{G}_{2,2} & \mathbf{G}_{2,3} \\ 0 & \mathbf{G}_{2,3}^T & \mathbf{G}_{3,3} \end{pmatrix}$$

with

$\mathbf{G}_{1,1} = \mathrm{E}_{\boldsymbol{\theta}_0}\left[\mathbf{W}\mathbf{W}^T(Y - \mathbf{W}^T\boldsymbol{\theta}_{10})^2 \left|I(Y < \mathbf{W}^T\boldsymbol{\theta}_{10}) - \pi\right|^2\right]$,
$\mathbf{G}_{2,2} = \mathrm{Var}_{\boldsymbol{\theta}_0}\{\mathrm{vech}[(\mathbf{X} - \boldsymbol{\mu}_0)(\mathbf{X} - \boldsymbol{\mu}_0)^T]\}$,
$\mathbf{G}_{2,3} = \mathrm{E}_{\boldsymbol{\theta}_0}\{\mathrm{vech}[(\mathbf{X} - \boldsymbol{\mu}_0)(\mathbf{X} - \boldsymbol{\mu}_0)^T](\boldsymbol{\mu}_0 - \mathbf{X})^T\}$,
$\mathbf{G}_{3,3} = \mathrm{Var}_{\boldsymbol{\theta}_0}(\mathbf{X})$.

The proof of Theorem 1 is given in Section 1.2 of the supplementary material. Since both $\mathbf{C}$ and $\mathbf{G}$ are block diagonal, $\tilde{\boldsymbol{\theta}}_1 = (\tilde{\mu}_\pi, \tilde{\boldsymbol{\beta}}_\pi^T)^T$ is asymptotically independent of $\tilde{\boldsymbol{\theta}}_2 = (\mathrm{vech}(\tilde{\boldsymbol{\Sigma}}_{\mathbf{X}})^T, \tilde{\boldsymbol{\mu}}_{\mathbf{X}}^T)^T$. Theorem 1 provides the asymptotic distribution for all the parameters, and the asymptotic distribution of $\tilde{\boldsymbol{\beta}}_\pi$ agrees with the results in Theorem 3 of [32]. Now we established the asymptotic distribution of the EER estimator.

Theorem 2. *Suppose that the EER model (2.4) holds. Under the regularity conditions (C1)–(C5), if* $\frac{\partial \mathrm{E}_{\boldsymbol{\theta}_0}[s(\mathbf{Z};\boldsymbol{\theta})s(\mathbf{Z};\boldsymbol{\theta})^T]}{\partial \boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ *has a finite Frobenius norm and the support of* $\boldsymbol{\zeta}$ *is compact, then*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Psi}(\boldsymbol{\Psi}^T\mathbf{C}\mathbf{G}^{-1}\mathbf{C}\boldsymbol{\Psi})^\dagger\boldsymbol{\Psi}^{\mathrm{T}}),$$

where $\boldsymbol{\Psi} = \partial\psi(\boldsymbol{\zeta})/\partial\boldsymbol{\zeta}$.

The proof of Theorem 2 is given in Section 1.3 of the supplementary material. Theorem 2 indicates that the EER estimator is $\sqrt{n}$-consistent, and is asymptotically normal. Since we have the explicit form of the asymptotic covariance matrices of the ER estimator and the EER estimator, we can compare the efficiency of the two estimators.

Corollary 1. *Under the same conditions in Theorem 2,* $avar(\sqrt{n}\hat{\boldsymbol{\theta}}) \leq avar(\sqrt{n}\tilde{\boldsymbol{\theta}})$.

The proof of Corollary 1 is given in Section 1.4 of the supplementary material. Corollary 1 asserts that the EER estimator is asymptotically more efficient than or as efficient as the ER estimator.

5. Simulations

In this section, we demonstrate the estimation efficiency gains of the EER model via numerical experiments. In the simulations and examples in Section 6, we consider a set of expectile levels of the distribution, and estimation is performed for one level at a time instead of all levels simultaneously. Schnabel and Eilers [38] describes how to determine the complete distribution from a set of expectiles using penalized least squares.

We consider the following simulation settings:

$$Y_i = 3 + \boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{X}_i + (8 + \boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{X}_i)\epsilon_i, \qquad \text{for} \quad i = 1, \ldots, n.$$

We set $p = 12$ and $u_\pi = 2$. Both $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ were p-dimensional vectors. All elements in $\boldsymbol{\alpha}_1$ were 4. The first $p/2$ elements in $\boldsymbol{\alpha}_2$ were 0.1 and the rest $p/2$ elements were 0. Four types of error distribution were used to generate ϵ: standard normal distribution $\epsilon \sim \mathcal{N}(0, 1)$, student's t-distribution with 4 degrees of freedom $\epsilon \sim t_4$, mixed normal distribution $\epsilon \sim 0.9\mathcal{N}(0, 1) + 0.1\mathcal{N}(1, 5)$, and exponential distribution with mean 1, i.e., $\epsilon \sim \mathrm{Exp}(1)$.

Based upon the settings, the πth conditional expectile of Y had the following form

$$f_\pi(Y|\mathbf{X}) = 3 + \boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{X} + (8 + \boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{X})f_\pi(\epsilon) = 3 + 8f_\pi(\epsilon) + (\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 f_\pi(\epsilon))^T\mathbf{X},$$

where $f_\pi(\epsilon)$ represented the πth expectile of the error distribution. The slope coefficients are contained in $\boldsymbol{\beta}_\pi = \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 f_\pi(\epsilon)$ and the intercept is $\mu_\pi = 3 + 8f_\pi(\epsilon)$. The predictor vector $\mathbf{X}$ followed a normal distribution with mean 0 and covariance matrix $\boldsymbol{\Sigma}_\mathbf{X} = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^T + \boldsymbol{\Phi}_0\boldsymbol{\Lambda}_0\boldsymbol{\Phi}_0^T$, where $\boldsymbol{\Lambda}$ was a $u_\pi \times u_\pi$ diagonal matrix with diagonal elements 100 and 9, and $\boldsymbol{\Lambda}_0$ was a $(p - u_\pi) \times (p - u_\pi)$ identity matrix. The matrix $\boldsymbol{\Phi} \in \mathbb{R}^{p_1 \times u_\pi}$ was a semi-orthogonal matrix with the first $p/2$ rows being $(\sqrt{6}/6,\ 0)$ and the remaining $p/2$ rows being $(0, \sqrt{6}/6)$. And the matrix $\boldsymbol{\Phi}_0 \in \mathbb{R}^{p \times (p - u_\pi)}$ was a semi-orthogonal matrix that satisfied $\boldsymbol{\Phi}^T\boldsymbol{\Phi}_0 = 0$. Since $\boldsymbol{\alpha}_1 = \boldsymbol{\Phi} \cdot (4\sqrt{6}, 4\sqrt{6})^T$ and $\boldsymbol{\alpha}_2 = \boldsymbol{\Phi} \cdot (\sqrt{6}/10, 0)^T$, we have $\boldsymbol{\beta}_\pi \in \mathrm{span}(\boldsymbol{\Phi}) = \mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_\pi)$. Thus $f_\pi(Y|\mathbf{X})$ and $\mathbf{X}$ satisfied the EER model (2.4).

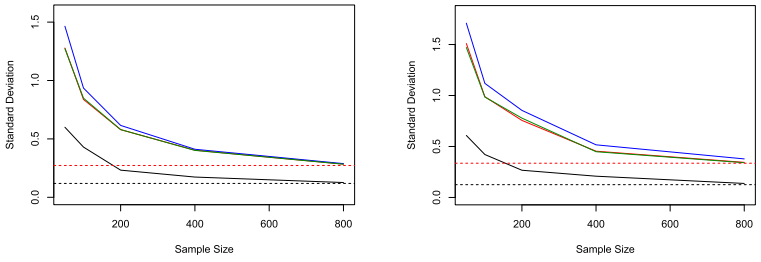
We varied the sample size n from 50 to 800. For each sample size, 100 replications were generated. For each replication, we computed the EER estimator, the ER estimator, the boosting estimator (componentwise gradient boosting expectile regression assuming each predictor has a linear effect) as well as the sparse ER estimator ([19]) of $\boldsymbol{\beta}_\pi$. The boosting estimator was computed by `R` package `expectreg` [42] with its default settings, i.e. the maximum number of boosting iterations is 4000, and cross validation is used to determine

the optimal amount of boosting iterations between 1 and 4000. The sparse ER estimator was computed by `R` package `SALES` [20]. We use the default choice of the tuning parameter, i.e. the largest value whose cross validation error is within one standard error of the minimum cross validation error. An alternative choice of the tuning parameter is the value that gives the minimum cross validation error. Results based on this alternative choice is included in Section 10 of the supplementary materials. For each element in $\boldsymbol{\beta}_\pi$, we computed the sample standard deviation for the 100 EER estimators, 100 ER estimators, 100 boosting estimators and 100 sparse ER estimators. We took expectile levels 0.10, 0.25, 0.50, 0.75 and 0.90 as examples. The results of a randomly chosen element in $\boldsymbol{\beta}_\pi$ with $\pi = 0.50$ and $\pi = 0.90$ are summarized in Figure 1. They reflect the mean and the upper tail properties of the response. The results of other expectile levels are given in Section 7 of the supplementary materials.

Figure 1 shows substantial efficiency gains achieved by the EER model in the estimation of $\boldsymbol{\beta}_\pi$. With all error distributions and expectile levels π, the sample standard deviations of the EER estimators are much smaller than the sample standard deviations of the ER estimators, the boosting estimators and the sparse ER estimators for all sample sizes. Take the first plot as an example (normal errors and $\pi = 0.5$), with sample size 200, the standard deviation of the EER estimator is already smaller than the asymptotic standard deviation of the ER estimator. The asymptotic standard deviation of the ER estimator in this case is 0.27 and the asymptotic standard deviation of the EER estimator is 0.12. We notice that the boosting estimator and the sparse ER estimator have about the same sample standard deviations as the ER estimator. This is because the variation from the immaterial part affects these two methods in a similar way as to ER. The sample standard deviations of the ER estimators and EER estimators both approach their corresponding asymptotic standard deviations as n increases, which confirms the asymptotic distributions established in Theorem 1 and Theorem 2.

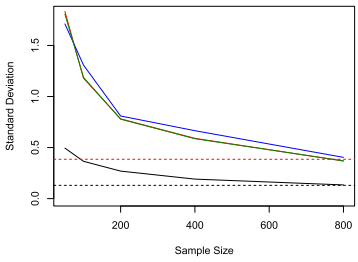
Moreover, we computed the bootstrap standard deviations of the four estimators for each component in $\boldsymbol{\beta}_\pi$ using the paired bootstrap method with 100 bootstrap repetitions. For demonstration purpose, we use the normal error as an example and include the results for $\pi = 0.5$ and $\pi = 0.9$ in Figure 2. The results indicate that the bootstrap standard deviation is a good approximation to the actual sample standard deviation. Therefore we use the bootstrap standard deviation as an estimator of the actual standard deviation in data analysis in Section 6.

Now we compared the prediction performance of the EER model with the ER model, the boosting model and the sparse ER model. Under the same simulation settings, we fixed the sample size at $n = 800$ and generated 300 replications. For each replication, 400 samples were used for training and another 400 samples were used for testing. With each of the four models, we first fitted the model to the training samples. Then for each $(\mathbf{X}_i, Y_i)$ in the testing samples, we computed $\hat{f}_\pi(Y_i|\mathbf{X}_i)$, the predicted πth conditional expectile of Y_i given $\mathbf{X}_i$. The prediction

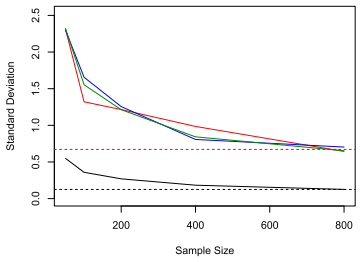


(a) Standard normal with $\pi = 0.5$

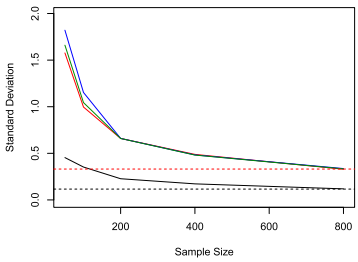
(b) Standard normal with $\pi = 0.9$



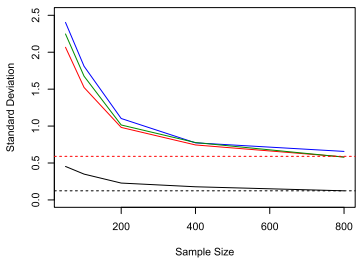
(c) t_4 with $\pi = 0.5$



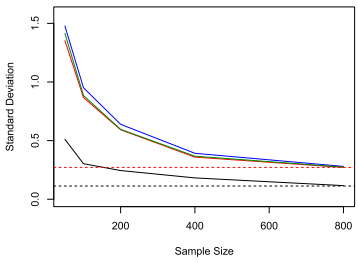
(d) t_4 with $\pi = 0.9$



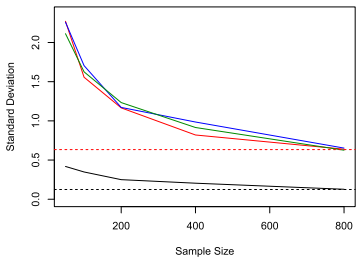
(e) Mixed normal with $\pi = 0.5$



(f) Mixed normal with $\pi = 0.9$

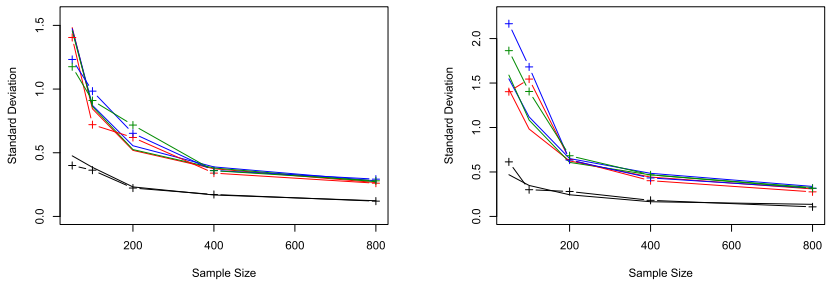


(g) Exp(1) with $\pi = 0.5$



(h) Exp(1) with $\pi = 0.9$

Fig 1: Comparison of the sample standard deviations. Red lines mark the ER estimator. Blue lines mark the sparse ER estimator. Green lines mark the boosting estimator. Black lines mark the EER estimator. The horizontal lines mark the asymptotic standard deviations of the ER estimator (the upper line in each panel) and the EER estimator (the lower line in each panel).



(a) Standard normal with $\pi = 0.5$ (b) Standard normal with $\pi = 0.9$

Fig 2: Sample standard deviations and bootstrap standard deviations. Red lines mark the ER estimator. Blue lines mark the sparse ER estimator. Green lines mark the boosting estimator. Black lines mark the EER estimator. Lines with "+" mark the bootstrap standard deviations for the corresponding estimators.

performance is measured by the root mean square error (RMSE) defined as

$$\text{RMSE} = \sqrt{\frac{1}{400}\sum_{i=1}^{400}\left[f_\pi(Y_i|\mathbf{X}_i) - \hat{f}_\pi(Y_i|\mathbf{X}_i)\right]^2}.$$

Table 1 and Figures 3–6 summarize the RMSEs under the four models with different error distributions. The EER estimator shows a superior prediction performance in all the cases. The boosting estimator and the ER estimator has comparable performance. Among the four models, the prediction performance of the sparse ER model is the worst because it tends to fit an overly sparse model, which introduces bias to its estimator. Take $\pi = 0.5$ as an example, the EER reduces the average RMSE by 37.7% to 46.3% compared to ER estimator, 37.7% to 46.5% compared to the boosting estimator, and 67.4% to 76.4% compared to the sparse ER estimator.

As shown in Section 8 of the supplementary material, the computation complexity of the EER estimator is a polynomial function of the number of the predictors p. Here we outline a comparison on the run time of the ER estimator, the EER estimator, the sparse ER estimator and the boosting estimator under the same setting that produced Table 1 with $\epsilon \sim \mathcal{N}(0, 1)$ and $\pi = 0.10$. Table 2 displays the average run time for each estimator in the 300 replications. We notice that the ER estimator takes the least time to compute. Compare to the ER estimator, the EER estimator takes about three times longer to compute (assuming u_π is known), the sparse ER estimator takes about 30 times longer to compute and the boosting estimator takes about 700 times longer to compute. The trend is similar for other error distributions and expectile levels.

In addition, we examined the performance of RCV in the selection of u_π. In the same settings that generated Figure 1, we applied RCV to choose the envelope dimension u_π. We performed 100 replications for each sample size. The

TABLE 1
Comparison of the RMSEs, averaged over 300 *replications.*

(a) $\epsilon \sim \mathcal{N}(0,1)$

	EER	ER	Boosting	Sparse ER
$\pi = 0.10$	1.04	1.85	1.86	2.99
$\pi = 0.25$	0.93	1.60	1.60	2.79
$\pi = 0.50$	0.90	1.52	1.52	2.76
$\pi = 0.75$	0.95	1.61	1.62	2.88
$\pi = 0.90$	1.10	1.87	1.91	3.17

(b) $\epsilon \sim t_4$

	EER	ER	Boosting	Sparse ER
$\pi = 0.10$	1.84	3.49	3.50	6.62
$\pi = 0.25$	1.28	2.46	2.47	5.27
$\pi = 0.50$	1.15	2.14	2.15	4.88
$\pi = 0.75$	1.31	2.44	2.45	5.39
$\pi = 0.90$	1.85	3.46	3.51	7.00

(c) $\epsilon \sim 0.9\mathcal{N}(0,1) + 0.1\mathcal{N}(1,5)$

	EER	ER	Boosting	Sparse ER
$\pi = 0.10$	1.20	2.21	2.22	3.95
$\pi = 0.25$	1.05	1.87	1.87	3.69
$\pi = 0.50$	1.05	1.86	1.87	3.88
$\pi = 0.75$	1.24	2.21	2.22	4.59
$\pi = 0.90$	1.75	3.14	3.18	6.05

(d) $\epsilon \sim \text{Exp}(1)$

	EER	ER	Boosting	Sparse ER
$\pi = 0.10$	0.70	0.76	0.79	1.96
$\pi = 0.25$	0.77	1.07	1.07	2.58
$\pi = 0.50$	0.96	1.54	1.54	3.40
$\pi = 0.75$	1.34	2.27	2.28	4.51
$\pi = 0.90$	1.99	3.37	3.40	6.03

TABLE 2
The run time (in seconds) of the EER estimator, the ER estimator, the boosting estimator and the sparse ER estimator given $\epsilon \sim \mathcal{N}(0,1)$ *and* $\pi = 0.10$ *in the prediction performance comparison simulation.*

ER	EER	Sparse ER	Boosting
0.042	0.13	1.32	29.44

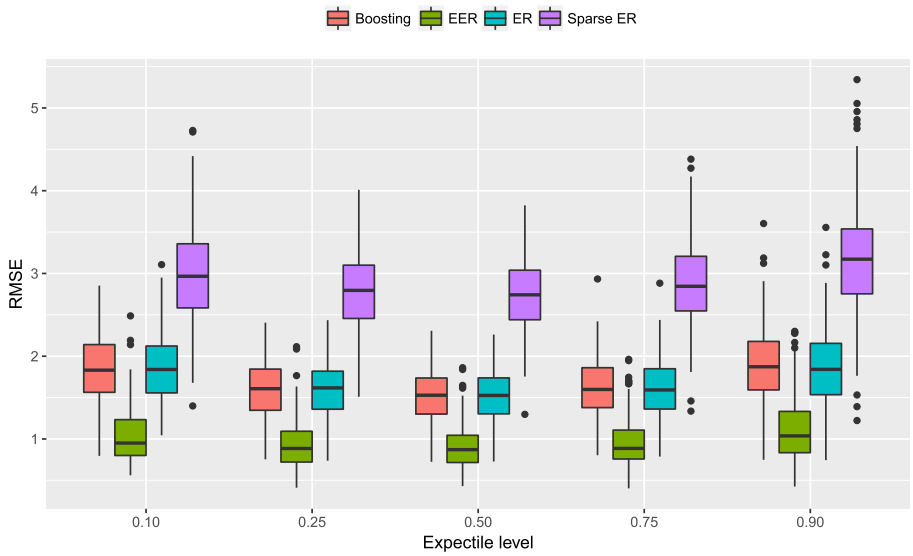


Fig 3: Boxplots of the RMSEs under the four models with $\epsilon \sim \mathcal{N}(0, 1)$.

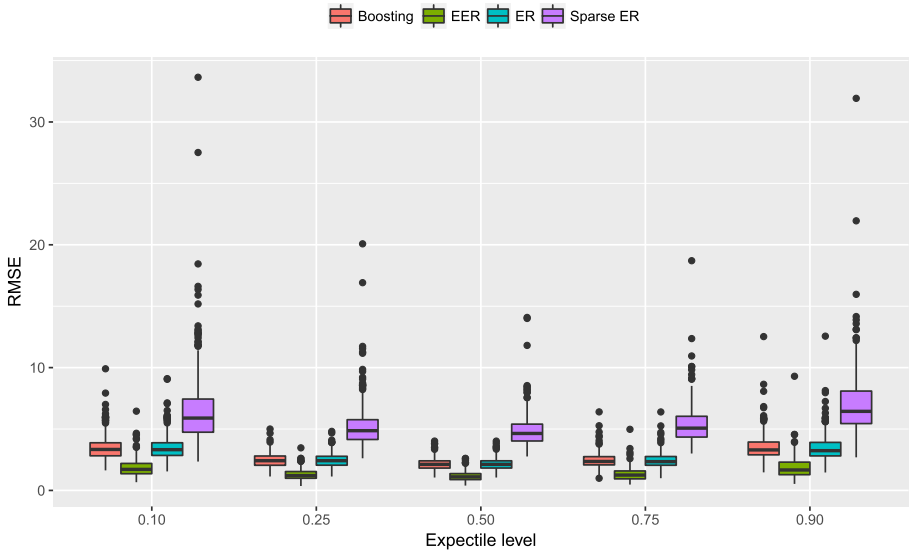


Fig 4: Boxplots of the RMSEs under the four models with $\epsilon \sim t_4$.

fraction that RCV selects the true dimension $u_\pi = 2$ is summarized in Table 3. RCV shows a stable performance in the selection of u_π. With a small sample size 25, RCV selects the true dimension more than 75% of the time. And its accuracy increases to 90% when sample size reaches 50. When RCV does not pick the

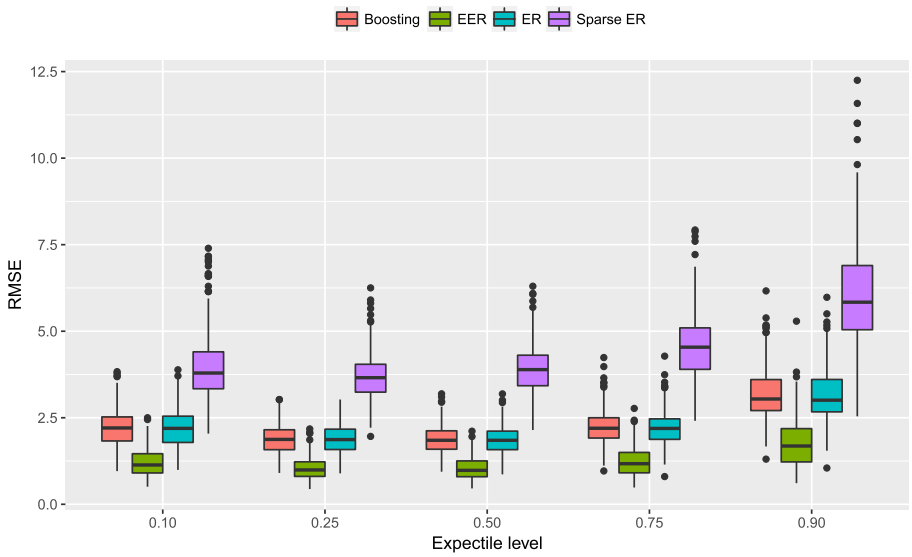


Fig 5: Boxplots of the RMSEs under the four models with $\epsilon \sim 0.9\mathcal{N}(0,1) + 0.1\mathcal{N}(1,5)$.

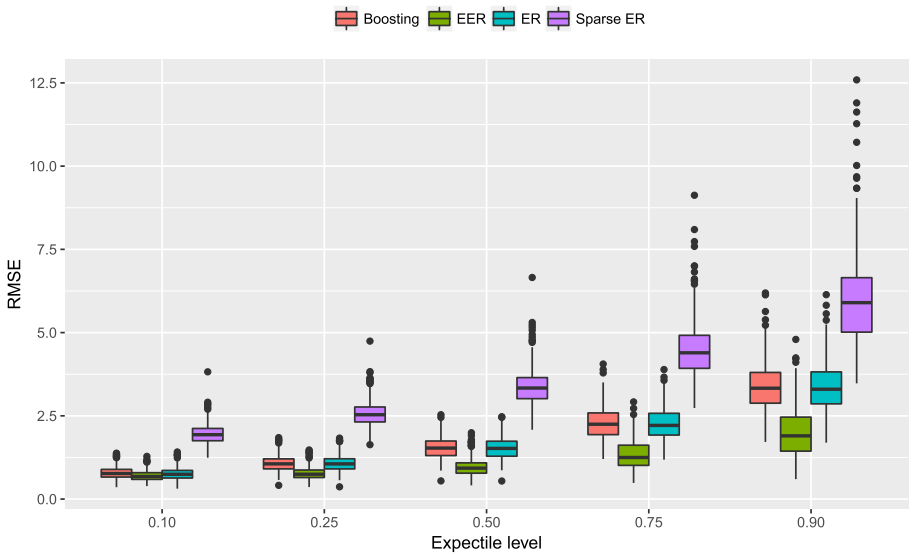


Fig 6: Boxplots of the RMSEs under the four models with $\epsilon \sim \text{Exp}(1)$.

correct dimension, it always overestimates the dimension. A bigger u_π yields a more conservative model, the resulting EER estimator loses some efficiency (compared to the EER model with the correct u_π), but it keeps all the material

information and does not introduce bias. Therefore we use RCV to choose u_π in data analysis examples.

In this section, the data is generated from an EER model, and the EER estimator can achieve efficiency gains in estimation and better prediction performance over the boosting model and the sparse ER model. However, the results can be different if the data were generated from a different underlying structure. Since the boosting model and the sparse ER model are variable selection methods, if the sparsity structure rather than the envelope structure is present, or in other words, some predictors are inactive and have coefficients zero, these two models can be more efficient than the EER model by making use of the sparsity structure. A simulation under such setting is included in Section 2 of the supplementary materials. Therefore we cannot conclude if the EER estimator is more efficient or less efficient than the boosting estimator and the sparse ER estimator in general. It depends on the underlying relationship between the response and the predictors, if the envelope structure holds or the sparsity structure holds. If a particular predictor has coefficient zero, then the sparsity structure holds. If $\boldsymbol{\beta}_\pi$ is contained in the subspace spanned by some eigenvectors of $\boldsymbol{\Sigma}_{\mathbf{X}}$, then the envelope structure holds. A potentially interesting scenario is that the data have both the envelope structure and the sparsity structure at the same time. A simulation under such setting is included in Section 3 of the supplementary materials. For completion, a simulation with no immaterial part is included in Section 4 of the supplement. In such case, any non-degenerate EER model ($u_\pi < p$) does not hold. However we can still expect to have a smaller mean squared error (MSE) from an approximate EER estimator in some cases due to the bias-variance tradeoff. Since quantiles and expectiles have a one-to-one mapping [53], we also computed the envelope quantile regression estimator [13] and compared it with the EER estimator using the same simulation setting in this section and the S&P 500 data. Details are included in Section 9 of the supplement.

6. Data analysis

6.1. state.x77

The dataset "state.x77" (contained in `datasets` package in `R`) contains eight measurements including population, average income, illiteracy, life expectancy, murder rate, high-school graduates percentage, land area and frost level for the 50 states in the United States of America. The dataset has been used in [36] as an example of multiple linear regression. Following [36], we took the murder rate as response, and population, average income, illiteracy rate and frost levels as the predictors. The density plot of murder rate indicated it was a bimodal distribution. In addition, we fitted a standard linear regression model on the dataset and checked for the homoscedasticity by Breush-Pagan test [2]. A p-value of 0.03 showed evidence of heteroskedasticity. Therefore, ER is more appropriate to fit the dataset compared to the standard linear regression. We fitted the data

TABLE 3
The fraction that RCV selects the true u_π with different error distributions.

(a) $\epsilon \sim \mathcal{N}(0,1)$

	$\pi = 0.10$	$\pi = 0.25$	$\pi = 0.50$	$\pi = 0.75$	$\pi = 0.90$
$n = 25$	76%	80%	81%	79%	86%
$n = 50$	90%	92%	90%	94%	91%
$n = 100$	98%	99%	100%	100%	98%
$n = 200$	100%	100%	100%	100%	100%
$n = 400$	100%	100%	100%	100%	100%
$n = 800$	100%	100%	100%	100%	100%

(b) $\epsilon \sim t_4$

	$\pi = 0.10$	$\pi = 0.25$	$\pi = 0.50$	$\pi = 0.75$	$\pi = 0.90$
$n = 25$	85%	87%	83%	84%	79%
$n = 50$	94%	92%	96%	94%	93%
$n = 100$	96%	98%	98%	100%	96%
$n = 200$	100%	100%	100%	100%	100%
$n = 400$	100%	100%	100%	100%	100%
$n = 800$	100%	100%	100%	100%	100%

(c) $\epsilon \sim 0.9\mathcal{N}(0,1) + 0.1\mathcal{N}(1,5)$

	$\pi = 0.10$	$\pi = 0.25$	$\pi = 0.50$	$\pi = 0.75$	$\pi = 0.90$
$n = 25$	85%	89%	89%	92%	90%
$n = 50$	98%	98%	95%	92%	92%
$n = 100$	95%	100%	98%	97%	97%
$n = 200$	100%	100%	100%	100%	100%
$n = 400$	100%	100%	100%	100%	99%
$n = 800$	100%	100%	100%	100%	100%

(d) $\epsilon \sim \text{Exp}(1)$

	$\pi = 0.10$	$\pi = 0.25$	$\pi = 0.50$	$\pi = 0.75$	$\pi = 0.90$
$n = 25$	79%	82%	78%	83%	81%
$n = 50$	91%	96%	96%	94%	91%
$n = 100$	100%	100%	100%	100%	99%
$n = 200$	100%	100%	100%	100%	98%
$n = 400$	100%	100%	100%	100%	100%
$n = 800$	100%	100%	100%	100%	100%

TABLE 4
The estimated regression coefficients for the standardized predictors given by the EER model, the ER model, the boosting model and the sparse ER model.

	EER				ER			
	Population	Income	Illiteracy	Frost	Population	Income	Illiteracy	Frost
$\pi = 0.10$	0.949	-0.349	1.382	-1.056	1.288	-0.698	2.010	-0.480
$\pi = 0.25$	0.755	-0.683	1.372	-1.345	1.152	-0.403	2.227	-0.280
$\pi = 0.50$	0.667	-0.648	1.327	-1.298	0.999	0.040	2.525	0.030
$\pi = 0.75$	0.954	-0.023	1.451	-1.074	0.921	0.338	2.729	0.271
$\pi = 0.90$	0.379	-0.404	0.820	-0.829	0.791	0.441	2.794	0.410
	Boosting				Sparse ER			
	Population	Income	Illiteracy	Frost	Population	Income	Illiteracy	Frost
$\pi = 0.10$	1.272	-0.689	1.965	-0.520	0.506	0.000	1.444	-0.490
$\pi = 0.25$	1.029	-0.253	2.187	-0.262	0.361	0.000	1.763	-0.153
$\pi = 0.50$	0.846	0.000	2.313	0.000	0.000	0.000	1.472	0.000
$\pi = 0.75$	0.529	0.000	2.045	0.000	0.000	0.000	1.293	0.000
$\pi = 0.90$	0.283	0.000	1.780	0.000	0.000	0.000	0.563	0.000

TABLE 5
Dimension selection results and efficiency comparison among the EER estimator, the ER estimator, the boosting estimator and the sparse ER estimator. Columns 3-4 contain the bootstrap standard deviation ratios of the ER estimator versus the EER estimator. Columns 5-6 contain the bootstrap standard deviation ratios of the sparse ER estimator versus the EER estimator. And columns 7-8 contain the bootstrap standard deviation ratios of the boosting estimator versus the EER estimator.

	$\hat{u}_\pi$	ER to EER		Sparse ER to EER		Boosting to EER	
		Range	Average	Range	Average	Range	Average
$\pi = 0.10$	2	1.21–1.75	1.50	1.32–1.66	1.49	1.04–1.46	1.26
$\pi = 0.25$	1	1.20–2.72	1.91	1.01–1.88	1.45	1.13–2.00	1.55
$\pi = 0.50$	1	1.35–2.77	2.08	1.56–1.56	1.56	1.21–1.50	1.36
$\pi = 0.75$	2	1.05–1.58	1.32	1.22–1.22	1.22	1.10–1.23	1.17
$\pi = 0.90$	1	1.90–3.00	2.32	2.26–2.26	2.26	1.70–2.01	1.86

using the ER with π varied in different levels at $0.10, 0.25, 0.50, 0.75$ and 0.90. Because the four predictors are in quite different scales, they were scaled to have unit standard deviation before the analysis. The analysis results of non-scaled predictors are given in Section 5 of the supplementary materials. RCV was used to select the dimensions of the envelope subspace for different expectile levels. Then the ER estimator, the boosting estimator, the sparse ER estimator and the EER estimator of $\boldsymbol{\beta}_\pi$ were computed. For each component in $\boldsymbol{\beta}_\pi$, we calculated the bootstrap standard deviations for the four estimators with 200 bootstrap repetitions. Since the boosting method and the sparse ER model are variable selection methods, some of the four predictors were selected as active and others were selected as inactive. The bootstrap standard deviation comparison of these methods with EER method is only performed on the selected active variables. The dimension selection results for the EER model, the estimated regression coefficients and the efficiency comparison are summarized in Table 4 and Table 5.

From Table 4 we can see the estimated regression coefficients given by the EER model and the ER model differ from each other. Take the predictor income (the per capita income) as an example, the ER model suggests that income has a negative effect on murder rate at lower quantiles ($\pi = 0.10$ and $\pi = 0.25$) and has a positive effect to murder rate at high quantile levels ($\pi = 0.50$, $\pi = 0.75$ and $\pi = 0.90$) while the EER models indicates that income has a negative effect on murder rate for all quantile levels. The results from the EER model seems to be more meaningful.

Table 5 shows the efficiency gains of the EER estimator over the other estimators at all investigated expectile levels. Taking $\pi = 0.90$ for example, RCV selected $u_\pi = 1$. The ratios of the bootstrap standard deviations of the ER estimator versus the EER estimator range from 1.92 to 2.99 with an average of 2.31. To achieve the same efficiency gains under the ER, we need to increase the sample size to $2.31^2 \approx 5.3$ times the original sample size. Similar efficiency gains are also noticed when comparing with the sparse ER estimator and the boosting estimator. The efficiency gains also lead to different interpretations of the data. For instance, with $\pi = 0.5$, population and illiteracy rate are significant predictors with positive coefficients under ER. Income and frost level (number of days with minimum temperature below freezing) are not significant. Sparse ER model selects illiteracy rate as the only active predictor with a positive coefficient. The boosting method selects population and illiteracy rate as active predictors with positive coefficients, and income and frost level as inactive predictors. However, under the EER model, all predictors are significant. While population and illiteracy rate have positive coefficients, income and frost level have negative association with murder rate. This indicates that the EER estimator can detect weaker signal from the data from improved estimation efficiency. The predictive performance of the EER estimator is slightly worse than the ER estimator because the RCV tends to select a parsimonious model, which may have a larger predicted expectile loss than the ER model. The details are included in Section 6 of the supplementary materials.

6.2. S&P 500 index

Now we provide another example using the S&P 500 Index data to show that the efficiency gains from the EER model can lead to a better prediction performance. The data [26] contains 351 quarterly economic observations from January, 1927 to December, 2014. The response was the equity premium, which is the return on the S&P 500 Index minus the return on treasury bill. We used 11 quarterly predictors following [3, 49, 10, 30, 34, 26]. The predictors included dividend yield (the difference between the log of dividends and the log of lagged prices), earnings–price ratio (the difference between the log of earnings and the log of prices), book-to-market ratio (the ratio of book value to market value for the Dow Jones Industrial Average), net equity expansion (ratio of 12-month moving sums of net issues by NYSE-listed stocks divided by the total market capitalization of NYSE stocks), stock variance (the sum of squared daily returns

TABLE 6

Mean and standard deviations of the predicted expectile losses under the ER and the EER model with different expectile levels. The 2nd and 3rd columns give the mean of the predicted expectile losses (in the unit of 10^{-3}). The 5th and 6th columns give the standard deviations of the predicted expectile losses (in the unit of 10^{-3}). The 4th and 7th columns represent percentage reduction of the EER model compared to the ER on the relative quantity.

	Mean			Standard Deviation		
	ER	EER	Reduction	ER	EER	Reduction
$\pi = 0.10$	3.34	2.26	32.26%	7.56	4.98	34.05%
$\pi = 0.25$	4.67	2.98	36.19%	10.15	5.88	42.05%
$\pi = 0.50$	5.32	3.10	41.74%	13.82	5.61	59.39%
$\pi = 0.75$	4.59	2.70	41.16%	15.19	4.19	72.43%
$\pi = 0.90$	3.58	2.29	36.06%	17.44	11.70	32.90%

on the S&P 500), treasury bill rate (the 3-month rate), term spread (difference between the long-term yield on government bonds and the treasury bill rate), long-term rate of return for government bonds, default yield spread (difference between BAA- and AAA-rated corporate bond yields), default return spread (the difference between the return on long-term corporate bonds and the return on long-term government bonds) and inflation. Some of the predictors were stock characteristics and the others reflected operation conditions of the selected companies. All of them had significant impacts on S&P 500 Index. Since an investigation on the dataset showed strong evidence of heteroskedasticity, the standard linear regression is not adequate to explore the relationship between the response and the predictors. Moreover, there were two extreme values in the response. Hence instead of a QR, we conducted an ER on the dataset which is more sensitive to the extreme values and could make more efficient use of the information in the dataset.

We fitted the data using the model:

$$f_\pi(Y_{t+1}|\mathbf{X}_t) = \mu_\pi + \boldsymbol{\beta}_\pi^T \mathbf{X}_t,$$

where Y_{t+1} was the equity premium at time $t+1$ and $\mathbf{X}_t$ was the predictor vector at time t. Both the ER and the EER model were applied to predict the conditional expectile of the response $f_\pi(Y_{t+1}|\mathbf{X}_t)$ in a moving window with a size of 80 quarters, i.e., use observations $\{(\mathbf{X}_t, Y_{t+1}), t = t_0 - 80, ..., t_0 - 1\}$ to predict $f_\pi(Y_{t_0+1}|\mathbf{X}_{t_0})$, where $t_0 = 81, ..., 351$. We used 80 as the size of the moving window because [26] showed that a 20-years estimation window delivers better results than alternative estimation windows. The predicted expectile loss is an important measure of the prediction performance in ER. Once we have $\hat{\mu}_\pi$ and $\hat{\boldsymbol{\beta}}_\pi$, the predicted expectile loss is computed as $\phi_\pi(Y_{t_0+1} - \hat{\mu}_\pi - \hat{\boldsymbol{\beta}}_\pi^T \mathbf{X}_{t_0})$.

We took expectile levels $\pi = 0.10, 0.25, 0.50, 0.75$ and 0.90 as examples. The results of the predicted expectile losses are summarized in Table 6. Table 6 shows that the EER model has smaller mean predicted expectile loss compared to the ER model with all the expectile levels. In addition, the EER model also

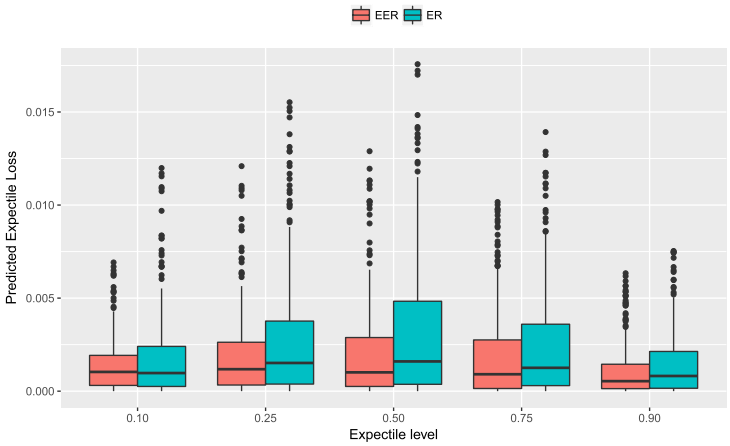


Fig 7: Boxplots of the trimmed predicted expectile losses.

has smaller standard deviations of the predicted expectile losses, which indicates that the EER model can give a more stable prediction. Figure 7 contains boxplots to graphically display the location and spread of the predicted expectile losses. Since there were some outliers, we trimmed the largest 5% of the predicted expectile losses under both the ER and the EER model for each expectile level to improve visibility. Both Table 6 and Figure 7 demonstrate substantial advantage of the EER model in prediction performance over the ER model.

We should note that the response in this example is actually weakly autocorrelated as revealed from its autocorrelation function (ACF) plot and the partial autocorrelation function (PACF) plot. Therefore an EER model that accommodates time dependent data is more suitable for the analysis of this dataset. [25] extended the asymptotic results of [32] to allow for stationary and weakly dependent data in the ER model. In some applications, the time effects can also be formulated by a mixed regression, and mixed expectile models are studied in [47]. The development of an EER model for time dependent data or for mixed expectile model is a potentially interesting future research direction and can have applications in financial, medical or meteorological datasets.

7. Extension to semiparametric settings

As an anonymous reviewer pointed out that one advantage to expectiles is the possibility to have very flexible semiparametric predictors. In this section, we consider a semiparametric ER model where the response is related to a combination of linear predictors $\mathbf{X} \in \mathbb{R}^{p_1}$ and nonlinear predictors $\mathbf{Z} \in \mathbb{R}^{p_2}$

$$f_\pi(Y|\mathbf{X}, \mathbf{Z}) = \mu_\pi + \boldsymbol{\beta}_\pi^T \mathbf{X} + g(\mathbf{Z}), \tag{7.1}$$

where g is a smooth function, μ_π is the intercept and $\boldsymbol{\beta}_\pi$ contains coefficients for the linear predictors. We assume that $\mathrm{E}(g(\mathbf{Z})) = 0$ for identification.

To impose the envelope structure on $\boldsymbol{\beta}_\pi$, we use the technique in partial envelope model [44] and consider the $\boldsymbol{\Sigma}_{\mathbf{X}}$-envelope of $\boldsymbol{\beta}_\pi$, denoted by $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}_\pi)$. Let u_π denote its dimension, and let $\boldsymbol{\Gamma}_\pi \in \mathbb{R}^{p\times u_\pi}$ and $\boldsymbol{\Gamma}_{0\pi} \in \mathbb{R}^{p\times(p-u_\pi)}$ be orthonormal bases of $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}_\pi)$ and $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}_\pi)^\perp$ respectively. Then $\boldsymbol{\beta}_\pi$ can be written as $\boldsymbol{\beta}_\pi = \boldsymbol{\Gamma}_\pi\boldsymbol{\eta}_\pi$, where $\boldsymbol{\eta}_\pi \in \mathbb{R}^{u_\pi}$ contains the coordinates of $\boldsymbol{\beta}_\pi$ with respect to $\boldsymbol{\Gamma}_\pi$. The covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}}$ can be written as $\boldsymbol{\Sigma}_{\mathbf{X}} = \boldsymbol{\Gamma}_\pi\boldsymbol{\Omega}_\pi\boldsymbol{\Gamma}_\pi^T + \boldsymbol{\Gamma}_{0\pi}\boldsymbol{\Omega}_{0\pi}\boldsymbol{\Gamma}_{0\pi}^T$, where $\boldsymbol{\Omega}_\pi$ contains the coordinates of $\boldsymbol{\Sigma}_{\mathbf{X}}$ with respect to $\boldsymbol{\Gamma}_\pi$ and $\boldsymbol{\Omega}_{0\pi}$ contains the coordinates of $\boldsymbol{\Sigma}_{\mathbf{X}}$ with respect to $\boldsymbol{\Gamma}_{0\pi}$. Then the semiparametric EER model is formulated as

$$f_\pi(Y|\mathbf{X},\mathbf{Z}) = \mu_\pi + \boldsymbol{\eta}_\pi^T\boldsymbol{\Gamma}_\pi^T\mathbf{X} + g(\mathbf{Z}), \quad \boldsymbol{\Sigma}_{\mathbf{X}} = \boldsymbol{\Gamma}_\pi\boldsymbol{\Omega}_\pi\boldsymbol{\Gamma}_\pi^T + \boldsymbol{\Gamma}_{0\pi}\boldsymbol{\Omega}_{0\pi}\boldsymbol{\Gamma}_{0\pi}^T. \quad (7.2)$$

Note that the envelope structure is only imposed on the linear predictor $\mathbf{X}$, not the entire predictor vector $(\mathbf{X}^T,\mathbf{Z}^T)^T$. This means that the conditional expectile depends on $\mathbf{X}$ only through $\boldsymbol{\Gamma}_\pi^T\mathbf{X}$, i.e. $f_\pi(Y|\mathbf{X},\mathbf{Z}) = f_\pi(Y|\boldsymbol{\Gamma}_\pi^T\mathbf{X},\mathbf{Z})$, and the variation of $\mathbf{X}$ can be decomposed into the variation of the material part and variation of the immaterial part, i.e., $\boldsymbol{\Sigma}_{\mathbf{X}} = \mathrm{var}(\mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{X}) + \mathrm{var}(\mathbf{Q}_{\boldsymbol{\Gamma}}\mathbf{X})$. By linking $\boldsymbol{\beta}_\pi$ to the material part, the semiparametric EER model (7.2) is expected to improve the efficiency in the estimation of $\boldsymbol{\beta}_\pi$. To impose envelope structure on $g(\mathbf{Z})$ involves completely different scopes and techniques, and we leave it as an important future research direction. To estimate from model (7.2), we use the following iterative algorithm. We denote the estimated linear part as $\hat{Y}_1$, and the estimated nonlinear part as $\hat{Y}_2$.

Step 1: Initialize $\hat{Y}_2 = 0$.

Step 2: Fit the EER model with $Y - \hat{Y}_2$ as the response and $\mathbf{X}$ as the predictors. Use the EER estimators to update $\hat{Y}_1$.

Step 3: Fit the ER additive model [41] with Y being the response, $\widehat{\boldsymbol{\Gamma}}_\pi^T\mathbf{X}$ being the linear predictors and $\mathbf{Z}$ being the nonlinear predictors, where $\widehat{\boldsymbol{\Gamma}}_\pi$ was calculated from Step 2. Update $\hat{Y}_2$.

Step 4: Repeat Step 2–Step 3 until the convergence of $\hat{Y}_1 + \hat{Y}_2$.

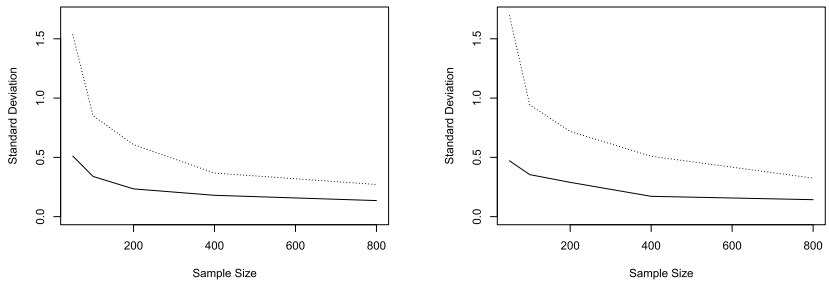
To illustrate the performance of semiparametric EER, we consider the following simulation settings:

$$Y_i = 3 + \boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{X}_i + h(\mathbf{Z}_i) + (8 + \boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{X}_i)\epsilon_i, \qquad \text{for} \quad i = 1,\ldots,n,$$

where $\mathbf{X} \in \mathbb{R}^{p_1}$ contains the linear predictors and $\mathbf{Z} \in \mathbb{R}^{p_2}$ contains the nonlinear predictors. We set $p_1 = 12$ and $p_2 = 4$. The nonlinear predictors $\mathbf{Z} = (Z_1, Z_2, Z_3, Z_4)$ followed the multivariate normal distribution $\mathcal{N}_4(0,\mathbf{I}_4)$ and the function $h(\mathbf{Z}) = \exp(\mathrm{Z}_1) + \sin(\mathrm{Z}_2) + \cos(\mathrm{Z}_3) + \mathrm{Z}_4^3$. The error ϵ was generated from the standard normal distribution.

Based upon the settings, the πth conditional expectile of Y has the following form

$$\begin{aligned} f_\pi(Y|\mathbf{X},\mathbf{Z}) &= 3 + \boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{X} + h(\mathbf{Z}) + (8 + \boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{X})f_\pi(\epsilon) \\ &= 3 + 8f_\pi(\epsilon) + \mathrm{E}(h(\mathbf{Z})) + (\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 f_\pi(\epsilon))^T\mathbf{X} + h(\mathbf{Z}) - \mathrm{E}(h(\mathbf{Z})), \end{aligned}$$



(a) $\pi = 0.5$ (b) $\pi = 0.9$

Fig 8: Sample standard deviations. Dashed lines mark the semiparametric ER estimator. Solid lines mark the semiparametric EER estimator.

where $f_\pi(\epsilon)$ represents the πth expectile of the standard normal distribution. For the linear part, the slope coefficients are contained in $\boldsymbol{\beta}_\pi = \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 f_\pi(\epsilon)$ and the intercept is $\mu_\pi = 3 + 8f_\pi(\epsilon) + \mathrm{E}(h(\mathbf{Z}))$. The parameters $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$ and the predictor vector $\mathbf{X}$ were generated in the same way as in Section 5. So the linear part follows an envelope structure. The nonlinear part was given by $g(\mathbf{Z}) = h(\mathbf{Z}) - \mathrm{E}(h(\mathbf{Z}))$.

We varied the sample size n from 50 to 800. For each sample size, 100 replications were generated. For each replication, we computed the semiparametric ER estimator of $\boldsymbol{\beta}_\pi$ by the ER additive model ([42]), and computed the semiparametric EER estimator of $\boldsymbol{\beta}_\pi$ using the preceding algorithm. Then for each component of $\boldsymbol{\beta}_\pi$, we computed the sample standard deviations for the semiparametric ER estimator and the semiparametric EER estimator based on the results from the 100 replications. We took expectile levels $\pi = 0.50$ and $\pi = 0.90$ as examples. The results of a randomly chosen component in $\boldsymbol{\beta}_\pi$ are summarized in Figure 8. The sample standard deviations of the semiparametric EER estimators are much smaller than the sample standard deviations of the semiparametric ER estimators. For example, when $n = 200$, the standard deviation of the semiparametric ER estimator is 0.61 while the standard deviation of the semiparametric EER estimator is 0.23 for $\pi = 0.5$. For $\pi = 0.9$, the standard deviations are 0.72 and 0.29 for the semiparametric ER estimator and the semiparametric EER estimator. This indicates that the envelope structure also yields substantial efficiency gains in estimation of $\boldsymbol{\beta}_\pi$ under the semiparametric ER context.

The baseball salary data was studied in [48] to determine whether the salary of a baseball player is affected by his offensive performance. The data contains the salary information from the 1992 season for 337 Major League Baseball (MLB) non-pitchers who played at least one game during both the 1991 and 1992 seasons. It also provides 12 offensive statistics for each player from the 1991 season including batting average, on-base percentage, number of runs, hits, doubles, triples, home runs, batted in, walks, strike-outs, stolen bases and errors.

We took the salary as response and the 12 offensive statistics as predictors. By exploring the scatter plots of the response versus each predictors, we identified five predictors for which the association with the response is not sufficiently explained by a linear relationship. The five predictors were batting average, on-base percentage, number of triples, strike-outs and errors. Therefore, they were used as nonlinear predictors and the remains were used as linear predictors. Before the analysis, each predictor was scaled to have unit standard deviation. We fitted the semiparametric ER model and the semiparametric EER model with $\pi = 0.5$ and 0.9 to the data. RCV suggested $u_\pi = 1$ for $\pi = 0.5$. For each element in $\boldsymbol{\beta}_\pi$, we calculated the bootstrap standard deviations for both the semiparametric ER estimator and the semiparametric EER estimator with 100 bootstrap samples. The ratios of bootstrap standard deviations of the semiparametric ER estimator versus the semiparametric EER estimator range from 3.99 to 13.65 with an average of 9.34. For $\pi = 0.9$, $u_\pi = 2$ was selected by RCV for the EER model. The ratios of the bootstrap standard deviations range from 1.32 to 10.87 with an average of 5.44. The results indicate the semiparametric EER model achieves substantial efficiency gains compared to the semiparametric ER model. To get the same average estimation efficiency under the semiparametric ER model, we need to increase the sample size to $9.34^2 \approx 87$ times the original sample size with $\pi = 0.5$ and $5.44^2 \approx 30$ times the original sample size with $\pi = 0.9$.

8. Discussion and future work

In this paper, we develop the EER model as an efficient estimation method for the ER. We estimate the parameters using GMM and established the asymptotic distribution of the estimators. Efficiency gains are demonstrated both theoretically and numerically. A potentially interesting extension is to perform variable selection to the predictors and develop a sparse EER model. In many applications such as the S&P 500 Index data, some predictors do not affect certain conditional expectile(s) of the response and have coefficients zero. It is of practical interest to identify those predictors, especially in high dimensional settings. Another direction is to derive the EER model with censored response. Censored data are often encountered in medical studies and econometrics when the variable of interest is only observed under certain conditions, such as top coding in income surveys, e.g. "$500,000 and above" ([35]). An EER model that can accommodate censored response will be applicable in such settings.

Supplementary Material

Supplementary material for "Efficient Estimation in Expectile Regression Using Envelope Model"
(doi: 10.1214/19-EJS1664SUPP; .pdf).

References

[1] Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (pls regression). *Wiley Interdisciplinary Reviews: Computational Statistics* **2**, 1, 97–106.

[2] Breusch, T. S. and Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, 1287–1294. MR0545960

[3] Campbell, J. Y. and Thompson, S. B. (2007). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies* **21**, 4, 1509–1531.

[4] Cengiz, C.-B. and Herwartz, H. (2011). Modeling stock index returns by means of partial least-squares methods: An out-of-sample analysis for three stock markets. *Applied Stochastic Models in Business and Industry* **27**, 3, 253–266. MR2858753

[5] Chen, T., Su, Z., Yang, Y., and Ding, S. (2020). Supplementary material for "Efficient estimation in expectile regression using envelope model". DOI: 10.1214/19-EJS1664SUPP

[6] Cook, R., Helland, I., and Su, Z. (2013). Envelopes and partial least squares regression. *Journal of the Royal Statistical Society: Series B* **75**, 5, 851–877. MR3124794

[7] Cook, R. D., Forzani, L., and Zhang, X. (2015). Envelopes and reduced-rank regression. *Biometrika* **102**, 2, 439–456. MR3371015

[8] Cook, R. D., Li, B., and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica* **3**, 927–960. MR2729839

[9] Cook, R. D. and Zhang, X. (2015). Foundations for envelope models and methods. *Journal of the American Statistical Association* **110**, 510, 599–611. MR3367250

[10] Dangl, T. and Halling, M. (2012). Predictive regressions with time-varying coefficients. *Journal of Financial Economics* **106**, 1, 157–181.

[11] Danielsson, J., Embrechts, P., Goodhart, C., Keating, C., Muennich, F., Renault, O., and Shin, H. S. (2001). An academic response to basel ii. *Special Paper-LSE Financial Markets Group*.

[12] Ding, S. and Cook, R. D. (2018). Matrix variate regressions and envelope models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**, 387–408. MR3763697

[13] Ding, S., Su, Z., Zhu, G., and Wang, L. (2019). Envelope quantile regression. *Statistica Sinica*, To appear. MR3821117

[14] Edelman, A., Arias, T. A., and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications* **20**, 2, 303–353. MR1646856

[15] Embrechts, P., Puccetti, G., Rüschendorf, L., Wang, R., and Beleraj, A. (2014). An academic response to basel 3.5. *Risks* **2**, 1, 25–48.

[16] Farooq, M. and Steinwart, I. (2017). An svm-like approach for expectile regression. *Computational Statistics & Data Analysis* **109**, 159–181.

MR3603647

[17] Forzani, L. and Su, Z. (2019). Envelopes for elliptical multivariate linear regression. *Statistica Sinica*, To appear. MR2729848

[18] Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*. Springer series in statistics. New York, NY, USA: Springer. MR2722294

[19] Gu, Y. and Zou, H. (2016a). High-dimensional generalizations of asymmetric least squares regression and their applications. *The Annals of Statistics* **44**, 6, 2661–2694. MR3576557

[20] Gu, Y. and Zou, H. (2016b). Sales: Elastic net and (adaptive) lasso penalized sparse asymmetric least squares (sales) and coupled sparse asymmetric least squares (cosales) using coordinate descent and proximal gradient algorithms. *R package version 1.0.0*.

[21] Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, 1029–1054. MR0666123

[22] Jiang, C., Jiang, M., Xu, Q., and Huang, X. (2017). Expectile regression neural network model with applications. *Neurocomputing* **247**, 73–86.

[23] Kelly, B. and Pruitt, S. (2013). Market expectations in the cross-section of present values. *The Journal of Finance* **68**, 5, 1721–1756.

[24] Khare, K., Pal, S., and Su, Z. (2017). A bayesian approach for envelope models. *The Annals of Statistics* **45**, 1, 196–222. MR3611490

[25] Kuan, C., Yeh, J., and Hsu, Y. (2009). Assessing value at risk with care, the conditional autoregressive expectile models. *Journal of Econometrics* **150**, 2, 261–270. MR2535521

[26] Li, J. and Tsiakas, I. (2017). Equity premium prediction: The role of economic and statistical constraints. *Journal of Financial Markets* **36**, 56–75.

[27] Li, L. and Zhang, X. (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association* **112**, 519, 1131–1146. MR3735365

[28] Liao, L., Park, C., and Choi, H. (2018). Penalized expectile regression: an alternative to penalized quantile regression. *Annals of the Institute of Statistical Mathematics*, 1–30. MR3915424

[29] Liu, X., Srivastava, A., and Gallivan, K. (2003). Optimal linear representations of images for object recognition. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings*, Vol. **1**. IEEE, I–I.

[30] Neely, C. J., Rapach, D. E., Tu, J., and Zhou, G. (2014). Forecasting the equity risk premium: the role of technical indicators. *Management Science* **60**, 7, 1772–1791.

[31] Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics* **4**, 2111–2245. MR1315971

[32] Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica* **55**, 4, 819–847. MR0906565

[33] Oh, H.-S., Nychka, D., Brown, T., and Charbonneau, P. (2004).

Period analysis of variable stars by robust smoothing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **53**, 1, 15–30. MR2037881

[34] Pettenuzzo, D., Timmermann, A., and Valkanov, R. (2014). Forecasting stock returns under economic constraints. *Journal of Financial Economics* **114**, 3, 517–553.

[35] Rigobon, R. and Stoker, T. M. (2007). Estimation with censored regressors: Basic issues. *International Economic Review* **48**, 4, 1441–1467. MR2375632

[36] Sarkar, D. (2008). *Lattice: multivariate data visualization with R*. Springer Science & Business Media.

[37] Schnabel, S. K. and Eilers, P. H. (2009). Optimal expectile smoothing. *Computational Statistics & Data Analysis* **53**, 12, 4168–4177. MR2744314

[38] Schnabel, S. K. and Eilers, P. H. (2013). A location-scale model for non-crossing expectile curves. *Stat* **2**, 1, 171–183.

[39] Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *Journal of the American Statistical Association* **81**, 393, 142–149. MR0830574

[40] Sobotka, F., Kauermann, G., Waltrup, L. S., and Kneib, T. (2013). On confidence intervals for semiparametric expectile regression. *Statistics and Computing* **23**, 2, 135–148. MR3016934

[41] Sobotka, F. and Kneib, T. (2012). Geoadditive expectile regression. *Computational Statistics & Data Analysis* **56**, 4, 755–767. MR2888723

[42] Sobotka, F., Schnabel, S., Schulze Waltrup, L., Eilers, P., Kneib, T., and Kauermann, G. (2014). expectreg: expectile and quantile regression. *R package version 0.39*. MR3403125

[43] Spiegel, E., Sobotka, F., and Kneib, T. (2017). Model selection in semiparametric expectile regression. *Electronic Journal of Statistics* **11**, 2, 3008–3038. MR3694575

[44] Su, Z. and Cook, R. D. (2011). Partial envelopes for efficient estimation in multivariate linear regression. *Biometrika* **98**, 1, 133–146. MR2804215

[45] Su, Z., Zhu, G., Chen, X., and Yang, Y. (2016). Sparse envelope model: efficient estimation and response variable selection in multivariate linear regression. *Biometrika* **103**, 3, 579–593. MR3551785

[46] Taylor, J. W. (2008). Estimating value at risk and expected shortfall using expectiles. *Journal of Financial Econometrics* **6**, 2, 231–252.

[47] Waltrup, L. S. (2015). Extensions of semiparametric expectile regression. Ph.D. thesis, Universitätsbibliothek der Ludwig-Maximilians-Universität.

[48] Watnik, M. R. (1998). Pay for play: Are baseball salaries based on performance? *Journal of Statistics Education* **6**, 2.

[49] Welch, I. and Goyal, A. (2007). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies* **21**, 4, 1455–1508.

[50] Xie, S., Zhou, Y., and Wan, A. T. (2014). A varying-coefficient expectile model for estimating value at risk. *Journal of Business & Economic Statistics* **32**, 4, 576–592. MR3272888

[51] Yang, Y., Zhang, T., and Zou, H. (2017). Flexible expectile regression in

reproducing kernel Hilbert spaces. *Technometrics* **60**, 1, 26–35. MR3768047

[52] Yang, Y. and Zou, H. (2015). Nonparametric multiple expectile regression via er-boost. *Journal of Statistical Computation and Simulation* **85**, 7, 1442–1458. MR3306805

[53] Yao, Q. and Tong, H. (1996). Asymmetric least squares regression estimation: A nonparametric approach. *Journal of Nonparametric Statistics* **6**, 2-3, 273–292. MR1383055

Electronic Journal of Statistics
ISSN: 1935-7524

Supplementary material for "Efficient Estimation in Expectile Regression Using Envelope Model"

Tuo Chen

University of Florida
e-mail: chentuo@ufl.edu

Zhihua Su

University of Florida
e-mail: zhihuasu@stat.ufl.edu

Yi Yang

McGill University
e-mail: yi.yang6@mcgill.ca

and

Shanshan Ding

University of Delaware
e-mail: sding@udel.edu

Contents

1 Technical Proof 2
1.1 Proof of Lemma 1 2
1.2 Proof of Theorem 1 2
1.3 Proof of Theorem 2 5
1.4 Corollary 8
2 Simulations Under Variable Selection Settings (Without Envelope Structure) 8
3 Simulations Under Variable Selection Settings (With Envelope Structure) 10
4 Simulations Under No Immaterial Part Settings 17
5 Analysis of "state.x77" with Predictors in Original Scale 18
6 Prediction Performance Comparison on "state.x77" 19
7 Simulation Results at More Expectile Levels 20
8 Analysis of Computational Complexity of the GMM Algorithm 20
9 Comparison Between EQR and EER 22
10 Simulation Results for Sparse Expectile Regression Estimator with an Alternative Tuning Parameter 27
References 30

1. Technical Proof

1.1. Proof of Lemma 1

If $\mathbf{\Gamma}_\pi$ takes the form of $(\mathbf{I}_{u_\pi}, \mathbf{A}^T)^T$, then any basis matrix of $\mathcal{E}_{\mathbf{\Sigma_X}}(\boldsymbol{\beta}_\pi)$ has its first u_π rows being a non-singular matrix. We denote $\mathbf{\Gamma}^*_\pi$ as an orthonormal basis matrix of $\mathcal{E}_{\mathbf{\Sigma_X}}(\boldsymbol{\beta}_\pi)$ and $\mathbf{\Gamma}^*_{0\pi}$ as an orthonormal basis matrix of $\mathcal{E}_{\mathbf{\Sigma_X}}(\boldsymbol{\beta}_\pi)^\perp$. Thus, $(\mathbf{\Gamma}^*_\pi,\ \mathbf{\Gamma}^*_{0\pi})$ is an orthogonal matrix and we rewrite it as a 2 by 2 block matrix:

$$(\mathbf{\Gamma}^*_\pi, \mathbf{\Gamma}^*_{0\pi}) = \begin{pmatrix} \mathbf{\Gamma}^*_{\pi 1} & \mathbf{\Gamma}^*_{0\pi 1} \\ \mathbf{\Gamma}^*_{\pi 2} & \mathbf{\Gamma}^*_{0\pi 2} \end{pmatrix},$$

where $\mathbf{\Gamma}^*_{\pi 1}$ is the matrix containing the first u_π rows of $\mathbf{\Gamma}^*_\pi$. Since both $(\mathbf{\Gamma}^*_\pi,$ $\mathbf{\Gamma}^*_{0\pi})$ and $\mathbf{\Gamma}^*_{\pi 1}$ are non-singular, the schur complement of $\mathbf{\Gamma}^*_{\pi 1}$, donated by $\boldsymbol{Q}$, is nonsingular. In this case, the inverse of $(\mathbf{\Gamma}^*_\pi,\ \mathbf{\Gamma}^*_{0\pi})$ is

$$\begin{aligned}(\mathbf{\Gamma}^*_\pi, \mathbf{\Gamma}^*_{0\pi})^{-1} &= \begin{pmatrix} \mathbf{\Gamma}^{*-1}_{\pi 1} + \mathbf{\Gamma}^{*-1}_{\pi 1}\mathbf{\Gamma}^*_{0\pi 1}\boldsymbol{Q}^{-1}\mathbf{\Gamma}^*_{\pi 2}\mathbf{\Gamma}^{*-1}_{\pi 1} & -\mathbf{\Gamma}^{*-1}_{\pi 1}\mathbf{\Gamma}^*_{0\pi 1}\boldsymbol{Q}^{-1} \\ -\boldsymbol{Q}^{-1}\mathbf{\Gamma}^*_{\pi 2}\mathbf{\Gamma}^{*-1}_{\pi 1} & \boldsymbol{Q}^{-1} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{\Gamma}^{*T}_{\pi 1} & \mathbf{\Gamma}^{*T}_{\pi 2} \\ \mathbf{\Gamma}^{*T}_{0\pi 1} & \mathbf{\Gamma}^{*T}_{0\pi 2} \end{pmatrix}.\end{aligned}$$

The second equality sign in the equation above comes from the fact that the inverse of an orthogonal matrix is the transpose of the orthogonal matrix. It turns out that $\mathbf{\Gamma}^{*T}_{0\pi 2} = \boldsymbol{Q}^{-1}$, which are nonsingular. Therefore, $\mathbf{\Gamma}_{0\pi 2}$ is nonsingular and invertible. It indicates that $\mathbf{\Gamma}^*_{0\pi}$ has its last $(p - u_\pi)$ rows being a nonsingular matrix. Then, we can decompose $\mathbf{\Gamma}^*_{0\pi}$ as

$$\mathbf{\Gamma}^*_{0\pi} = \begin{pmatrix} \mathbf{\Gamma}^*_{0\pi 1} \\ \mathbf{\Gamma}^*_{0\pi 2} \end{pmatrix} = \begin{pmatrix} \mathbf{\Gamma}^*_{0\pi 1}\mathbf{\Gamma}^{*-1}_{0\pi 2} \\ \mathbf{I}_{p-u_\pi} \end{pmatrix} \mathbf{\Gamma}^*_{0\pi 2} \equiv \begin{pmatrix} \mathbf{B} \\ \mathbf{I}_{p-u_\pi} \end{pmatrix} \mathbf{\Gamma}^*_{0\pi 2} \equiv \mathbf{\Gamma}_{0\pi}\mathbf{\Gamma}^*_{0\pi 2}. \quad (1.1)$$

Apparently, $\mathbf{\Gamma}_{0\pi}$ is a basis matrix of $\mathcal{E}_{\mathbf{\Sigma_X}}(\boldsymbol{\beta}_\pi)^\perp$ and we have $\mathbf{\Gamma}^T_\pi\mathbf{\Gamma}_{0\pi} = 0$, which means $\mathbf{B} + \mathbf{A}^T = 0$ and $\mathbf{B} = \text{-}\mathbf{A}^T$. Therefore, $\mathbf{\Gamma}_{0\pi}$ takes the form of $(-\mathbf{A}, \mathbf{I}_{p-u_\pi})^T$.

1.2. Proof of Theorem 1

We apply Theorem 3.3 of [5] to derive the asymptotic distribution of $\tilde{\boldsymbol{\theta}}$. There are five conditions (i)–(v) in their Theorem and we need to check them. We denote $e(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}_0}[s(\mathbf{Z}; \boldsymbol{\theta})]$.

Based on the conditions (C1)–(C3) and Theorem 3 of [4], we have $\tilde{\boldsymbol{\theta}}_1 \xrightarrow{p} \boldsymbol{\theta}_{10}$ and $\boldsymbol{\theta}_{10}$ is the unique point satisfying $\mathrm{E}_{\boldsymbol{\theta}_0}[s_1(\mathbf{Z}; \boldsymbol{\theta}_{10})] = 0$. Therefore, it is obvious that $\tilde{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_0$ is the unique point in $\mathbf{\Theta}$ satisfying $e(\boldsymbol{\theta}) = 0$.

Because $\tilde{\boldsymbol{\theta}}$ is the minimizer of $||e_n(\boldsymbol{\theta})||$, the condition (i) holds. The conditions (ii) and (v) automatically hold given (C4) and (C5). By Central Limit Theorem, $\sqrt{n}(e_n(\boldsymbol{\theta}_0) - \mathrm{E}_{\boldsymbol{\theta}_0}[s(\mathbf{Z}_i; \boldsymbol{\theta_0})]) \xrightarrow{d} \mathcal{N}(0, \mathbf{G})$, where $\mathrm{E}_{\boldsymbol{\theta}_0}[s(\mathbf{Z}_i; \boldsymbol{\theta}_0)] = e(\boldsymbol{\theta}_0) = 0$ and $\mathbf{G} = \mathrm{E}_{\boldsymbol{\theta}_0}[s(\mathbf{Z}; \boldsymbol{\theta}_0)s(\mathbf{Z}; \boldsymbol{\theta}_0)^T]$. The condition (iv) holds.

To prove the condition (*iii*) holds, we need to prove the following Lemma as first.

Lemma 1. $\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le\delta_n} \|e_n(\boldsymbol{\theta}) - e(\boldsymbol{\theta}) - e_n(\boldsymbol{\theta}_0)\| = o_p(n^{-1/2})$, *where* δ_n *is any sequence of positive numbers with limitation 0.*

Proof. Let w_j, μ_j, σ_j, $s_{1,j}$, $s_{2,j}$ and $s_{3,j}$ represent the jth component of $\mathbf{W}$, $\boldsymbol{\mu}_{\mathbf{X}}$, $\text{vech}(\boldsymbol{\Sigma}_{\mathbf{X}})$, $s_1(\mathbf{Z};\boldsymbol{\theta}_1)$, $s_2(\mathbf{Z};\boldsymbol{\theta}_2)$ and $s_3(\mathbf{Z};\boldsymbol{\theta}_2)$ respectively. For any $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ and $j = 1, \ldots, p+1$,

$$\begin{aligned}|s_{1,j}(\mathbf{Z};\boldsymbol{\theta}_1) - s_{1,j}(\mathbf{Z};\boldsymbol{\theta}_1^*)|^2 &= w_j^2\big((Y-\mathbf{W}^T\boldsymbol{\theta}_1)\left|I(Y<\mathbf{W}^T\boldsymbol{\theta}_1)-\pi\right| \\ &\quad -(Y-\mathbf{W}^T\boldsymbol{\theta}_1^*)\left|I(Y<\mathbf{W}^T\boldsymbol{\theta}_1^*)-\pi\right|\big)^2.\end{aligned}$$

If $I(Y<\mathbf{W}^T\boldsymbol{\theta}_1) = I(Y<\mathbf{W}^T\boldsymbol{\theta}_1^*)$, then

$$\begin{aligned}&\big((Y-\mathbf{W}^T\boldsymbol{\theta}_1)\left|I(Y<\mathbf{W}^T\boldsymbol{\theta}_1)-\pi\right| - (Y-\mathbf{W}^T\boldsymbol{\theta}_1^*)\left|I(Y<\mathbf{W}^T\boldsymbol{\theta}_1^*)-\pi\right|\big)^2 \\ &= (I(Y<\mathbf{W}^T\boldsymbol{\theta}_1)-\pi)(Y-\mathbf{W}^T\boldsymbol{\theta}_1 - Y + \mathbf{W}^T\boldsymbol{\theta}_1^*))^2 \\ &= (I(Y<\mathbf{W}^T\boldsymbol{\theta}_1)-\pi)\big(\mathbf{W}^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1)\big)^2 \\ &\le \big(\mathbf{W}^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1)\big)^2 \le \|\mathbf{W}\|^2\|\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1\|^2.\end{aligned}$$

If $I(Y<\mathbf{W}^T\boldsymbol{\theta}_1) \ne I(Y<\mathbf{W}^T\boldsymbol{\theta}_1^*)$, then

$$\begin{aligned}&\big((Y-\mathbf{W}^T\boldsymbol{\theta}_1)\left|I(Y<\mathbf{W}^T\boldsymbol{\theta}_1)-\pi\right| - (Y-\mathbf{W}^T\boldsymbol{\theta}_1^*)\left|I(Y<\mathbf{W}^T\boldsymbol{\theta}_1^*)-\pi\right|\big)^2 \\ &\le (Y-\mathbf{W}^T\boldsymbol{\theta}_1 - Y + \mathbf{W}^T\boldsymbol{\theta}_1^*))^2 \\ &= \big(\mathbf{W}^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1)\big)^2 \le \|\mathbf{W}\|^2\|\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1\|^2.\end{aligned}$$

Therefore, by condition (C2), there exists a positive constant c_1 such that

$$\begin{aligned}&\mathrm{E}_{\boldsymbol{\theta}_0}\big[\sup_{\boldsymbol{\theta}^*:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\le\delta_n}|s_{1,j}(\mathbf{Z};\boldsymbol{\theta}_1)-s_{1,j}(\mathbf{Z};\boldsymbol{\theta}_1^*)|^2\big] \\ \le&\mathrm{E}_{\boldsymbol{\theta}_0}\big[w_j^2\|\mathbf{W}\|^2\|\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1\|^2\big] \le \delta_n^2\mathrm{E}_{\boldsymbol{\theta}_0}\big[\|\mathbf{W}\|^4\big] \le c_1\delta_n^2.\end{aligned} \tag{1.2}$$

Let μ_j^*, σ_j^* and $\text{vech}[\cdot]_j$ represent the jth component of $\boldsymbol{\mu}_{\mathbf{X}}^*$, $\text{vech}(\boldsymbol{\Sigma}_{\mathbf{X}}^*)$ and $\text{vech}[\cdot]$. Then for $j = 1, \ldots, (p+1)p/2$,
$|s_{2,j}(\mathbf{Z};\boldsymbol{\theta}_2) - s_{2,j}(\mathbf{Z};\boldsymbol{\theta}_2^*)|^2 = \big(\sigma_j - \text{vech}[(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}})^T]_j - \sigma_j^* + \text{vech}[(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}}^*)(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}}^*)^T]_j\big)^2$. By (C2), it is easy to verify there exists a positive constant c_2 such that

$$\mathrm{E}_{\boldsymbol{\theta}_0}\big[\sup_{\boldsymbol{\theta}^*:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\le\delta_n}|s_{2,j}(\mathbf{Z};\boldsymbol{\theta}_2)-s_{2,j}(\mathbf{Z};\boldsymbol{\theta}_2^*)|^2\big] \le c_2\delta_n^2. \tag{1.3}$$

Similarly, for $j = 1, \ldots, p$, there exists a positive constant c_3 such that

$$\mathrm{E}_{\boldsymbol{\theta}_0}\big[\sup_{\boldsymbol{\theta}^*:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\le\delta_n}|s_{3,j}(\mathbf{Z};\boldsymbol{\theta}_2)-s_{3,j}(\mathbf{Z};\boldsymbol{\theta}_2^*)|^2\big] = \mathrm{E}_{\boldsymbol{\theta}_0}\big[\sup_{\boldsymbol{\theta}^*:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\le\delta_n}(\mu_j-\mu_j^*)^2\big] \le c_3\delta_n^2. \tag{1.4}$$

Combining the results in (1.2), (1.3) and (1.4), we know $s(\mathbf{Z};\boldsymbol{\theta})$ is $L^2(P)$ continuous at $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. By applying Lemma 2.17 in [5], we have

$$\begin{aligned} & n^{-1/2} \sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\delta_n} \left\| \sum_{i=1}^{n} \{s(\mathbf{Z};\boldsymbol{\theta}) - \mathrm{E}_{\boldsymbol{\theta}_0}[s(\mathbf{Z};\boldsymbol{\theta})] - s(\mathbf{Z};\boldsymbol{\theta}_0)\} \right\| \\ =& n^{-1/2} \sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\delta_n} \|ne_n(\boldsymbol{\theta}) - ne(\boldsymbol{\theta}) - ne_n(\boldsymbol{\theta}_0)\| = o_p(1). \end{aligned}$$

Thus, $\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\delta_n} \|e_n(\boldsymbol{\theta}) - e(\boldsymbol{\theta}) - e_n(\boldsymbol{\theta}_0)\| = o_p(n^{-1/2})$. □

With the result of Lemma 2,

$$\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\delta_n} \frac{\|e_n(\boldsymbol{\theta}) - e(\boldsymbol{\theta}) - e_n(\boldsymbol{\theta}_0)\|}{n^{-1/2} + \|e_n(\boldsymbol{\theta})\| + \|e(\boldsymbol{\theta})\|} \leq o_p(1).$$

The condition (iii) holds. We have already verified all the conditions of Theorem 3.3 in [5]. With the result of Theorem 3.3, we have

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{G}\mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1}),$$

where $\mathbf{C}=\frac{\partial \mathrm{E}_{\boldsymbol{\theta}_0}[s(\mathbf{Z};\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ and $\mathbf{G} = \mathrm{E}_{\boldsymbol{\theta}_0}[s(\mathbf{Z};\boldsymbol{\theta}_0)s(\mathbf{Z};\boldsymbol{\theta}_0)^T]$.

According to [4], we know that

$$\frac{\partial \mathrm{E}_{\boldsymbol{\theta}_0}[s_1(\mathbf{Z};\boldsymbol{\theta}_1)]}{\partial \boldsymbol{\theta}_1^T}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = -\mathrm{E}_{\boldsymbol{\theta}_0}\left[\mathbf{W}\mathbf{W}^T \left|I(Y < \mathbf{W}^T\boldsymbol{\theta}_{10}) - \pi\right|\right].$$

As a result, it is easy to give the expression of $\mathbf{C}$ as

$$\mathbf{C} = \begin{pmatrix} -\mathrm{E}_{\boldsymbol{\theta}_0}\left[\mathbf{W}\mathbf{W}^T \left|I(Y < \mathbf{W}^T\boldsymbol{\theta}_{10}) - \pi\right|\right] & 0 & 0 \\ 0 & \mathbf{I}_{p(p+1)/2} & 0 \\ 0 & 0 & \mathbf{I}_p \end{pmatrix}.$$

Next, we give the expression of $\mathbf{G}$ in the form of $(\mathbf{G}_{ij})_{i,j=1,2,3}$. It is easy to check

$\mathbf{G}_{11} = \mathrm{E}_{\boldsymbol{\theta}_0}[s_1(\mathbf{Z};\boldsymbol{\theta}_{10})s_1(\mathbf{Z};\boldsymbol{\theta}_{10})^T] = \mathrm{E}_{\boldsymbol{\theta}_0}\left[\mathbf{W}\mathbf{W}^T(Y-\mathbf{W}^T\boldsymbol{\theta}_{10})^2 \left|I(Y < \mathbf{W}^T\boldsymbol{\theta}_{10}) - \pi\right|^2\right]$;
$\mathbf{G}_{22} = \mathrm{E}_{\boldsymbol{\theta}_0}[s_2(\mathbf{Z};\boldsymbol{\theta}_{20})s_2(\mathbf{Z};\boldsymbol{\theta}_{20})^T] = \mathrm{Var}_{\boldsymbol{\theta}_0}\{\mathrm{vech}[(\mathbf{X}-\boldsymbol{\mu}_0)(\mathbf{X}-\boldsymbol{\mu}_0)^T]\}$;
$\mathbf{G}_{33} = \mathrm{E}_{\boldsymbol{\theta}_0}[s_3(\mathbf{Z};\boldsymbol{\theta}_{20})s_3(\mathbf{Z};\boldsymbol{\theta}_{20})^T] = \mathrm{Var}_{\boldsymbol{\theta}_0}[\mathbf{X}]$;
$\mathbf{G}_{23} = \mathrm{E}_{\boldsymbol{\theta}_0}[s_2(\mathbf{Z};\boldsymbol{\theta}_{20})s_3(\mathbf{Z};\boldsymbol{\theta}_{20})^T] = \mathrm{E}_{\boldsymbol{\theta}_0}\{\mathrm{vech}[(\mathbf{X}-\boldsymbol{\mu}_0)(\mathbf{X}-\boldsymbol{\mu}_0)^T](\boldsymbol{\mu}_0 - \mathbf{X})^T\}$
and

$$\begin{aligned} \mathbf{G}_{12} &= \mathrm{E}_{\boldsymbol{\theta}_0}[s_1(\mathbf{Z};\boldsymbol{\theta}_{10})s_2(\mathbf{Z};\boldsymbol{\theta}_{20})^T] \\ &= \mathrm{E}_{\boldsymbol{\theta}_0}\left\{\mathbf{W}s_2(\mathbf{Z};\boldsymbol{\theta}_{20})^T\mathrm{E}_{\boldsymbol{\theta}_0}\left[(Y - \mathbf{W}^T\boldsymbol{\theta}_{10}) \left|I(Y < \mathbf{W}^T\boldsymbol{\theta}_{10}) - \pi\right| \,\middle|\, \mathbf{W}\right]\right\} = 0. \end{aligned}$$

Similarly, $\mathbf{G}_{13} = 0$. Since $\mathbf{C}$ is full rank and symmetric, we have

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{C}^{-1}\mathbf{G}\ \mathbf{C}^{-1}).$$

We complete the proof of Theorem 1.

1.3. Proof of Theorem 2

For notation simplicity, let $Q_n(\boldsymbol{\theta}) = e_n^T(\boldsymbol{\theta})\hat{\boldsymbol{\Delta}}e_n(\boldsymbol{\theta})$ and $Q(\boldsymbol{\theta}) = e^T(\boldsymbol{\theta})\boldsymbol{\Delta}e(\boldsymbol{\theta})$, where $\boldsymbol{\Delta} = \mathbf{G}^{-1} = \{\mathrm{E}_{\boldsymbol{\theta}_0}[s(\mathbf{Z};\boldsymbol{\theta}_0)s(\mathbf{Z};\boldsymbol{\theta}_0)^T]\}^{-1}$ and $\hat{\boldsymbol{\Delta}} = \left[\frac{1}{n}\sum_{i=1}^n s(\mathbf{Z}_i;\psi(\hat{\boldsymbol{\zeta}}^*))s(\mathbf{Z}_i;\psi(\hat{\boldsymbol{\zeta}}^*))^T\right]^{-1}$. Let $l_n(\boldsymbol{\gamma}) = e_n(\boldsymbol{\gamma}/\sqrt{n}+\boldsymbol{\theta}_0)$ and $l(\boldsymbol{\gamma}) = e(\boldsymbol{\gamma}/\sqrt{n}+\boldsymbol{\theta}_0)$. Let $T_n(\boldsymbol{\gamma}) = l_n^T(\boldsymbol{\gamma})\hat{\boldsymbol{\Delta}}l_n(\boldsymbol{\gamma})$ and $T(\boldsymbol{\gamma}) = l^T(\boldsymbol{\gamma})\boldsymbol{\Delta}l(\boldsymbol{\gamma})$. In addition, let $\epsilon_n(\boldsymbol{\gamma}) = [l_n(\boldsymbol{\gamma})-l_n(0)-l(\boldsymbol{\gamma})]/[1+\|\boldsymbol{\gamma}\|]$, $\kappa_n(\boldsymbol{\gamma}) = \epsilon_n^T(\boldsymbol{\gamma})\hat{\boldsymbol{\Delta}}\epsilon_n(\boldsymbol{\gamma})+2l_n(0)^T\hat{\boldsymbol{\Delta}}\epsilon_n(\boldsymbol{\gamma})$ and $\rho_n(\boldsymbol{\gamma}) = n[T_n(\boldsymbol{\gamma})-\kappa_n(\boldsymbol{\gamma})-T_n(0)-\hat{\mathbf{D}}^T\boldsymbol{\gamma}/\sqrt{n}-T(\boldsymbol{\gamma})]$, where $\hat{\mathbf{D}} = 2\mathbf{C}\hat{\boldsymbol{\Delta}}l_n(0)$. We firstly prove three Lemmas.

Lemma 2. *Under the same conditions in Theorem 2,* $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$.

Proof. Let $\mathcal{F} = \{s(\mathbf{Z};\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. Based on the fact that $s(\mathbf{Z};\boldsymbol{\theta})$ is a continuous function of $\boldsymbol{\theta}$ and conditions (C2) and (C5), it is easy to verify $\mathcal{F}$ satisfies all the conditions of Uniform Law of Large Numbers. Therefore, we have $\sup_{\boldsymbol{\theta}:\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|e_n(\boldsymbol{\theta}) - e(\boldsymbol{\theta})\| \xrightarrow{a.s.} 0$. As a result, $Q_n(\boldsymbol{\theta})$ uniformly converges to $Q(\boldsymbol{\theta})$ in probability in the domain $\boldsymbol{\Theta}_e = \{\boldsymbol{\theta} : \boldsymbol{\theta} \in \boldsymbol{\Theta} \text{ and } \boldsymbol{\theta} = \psi(\boldsymbol{\zeta})\}$. Since $\boldsymbol{\Theta}_e$ is compact, $Q(\boldsymbol{\theta})$ is continuous and $\boldsymbol{\theta}_0$ is the unique minimizer of $Q(\boldsymbol{\theta})$, all the conditions of Theorem 2.1 in [3] are satisfied. By the result of Theorem 2.1, we have $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$. □

Lemma 3. *Under the same conditions in Theorem 2,* $\sup_{\boldsymbol{\gamma}:\|\boldsymbol{\gamma}\|/\sqrt{n}\le\delta_n} \frac{|\rho_n(\boldsymbol{\gamma})|}{\|\boldsymbol{\gamma}\|(1+\|\boldsymbol{\gamma}\|)} = o_p(1)$*, where* δ_n *is any sequence of positive numbers with limitation 0.*

Proof. From the definition of $\epsilon_n(\boldsymbol{\gamma})$, we can decompose $T_n(\boldsymbol{\gamma})$ as

$$\begin{aligned} T_n(\boldsymbol{\gamma}) &= (1+\|\boldsymbol{\gamma}\|)^2\epsilon_n^T(\boldsymbol{\gamma})\hat{\boldsymbol{\Delta}}\epsilon_n(\boldsymbol{\gamma}) + l_n^T(0)\hat{\boldsymbol{\Delta}}l_n(0) + l^T(\boldsymbol{\gamma})\hat{\boldsymbol{\Delta}}l(\boldsymbol{\gamma}) \\ &\quad + 2(1+\|\boldsymbol{\gamma}\|)\epsilon_n^T(\boldsymbol{\gamma})\hat{\boldsymbol{\Delta}}l_n(0) + 2(1+\|\boldsymbol{\gamma}\|)\epsilon_n^T(\boldsymbol{\gamma})\hat{\boldsymbol{\Delta}}l(\boldsymbol{\gamma}) + 2l_n^T(0)\hat{\boldsymbol{\Delta}}l(\boldsymbol{\gamma}). \end{aligned}$$

It can be shown that $|\rho_n(\boldsymbol{\gamma})|/(\|\boldsymbol{\gamma}\|(1+\|\boldsymbol{\gamma}\|)) \le \sum_{i=1}^n B_j(\boldsymbol{\gamma})$, where

$B_1(\boldsymbol{\gamma}) = n(2+\|\boldsymbol{\gamma}\|)\epsilon_n^T(\boldsymbol{\gamma})\hat{\boldsymbol{\Delta}}\epsilon_n(\boldsymbol{\gamma})/(1+\|\boldsymbol{\gamma}\|)$, $B_2(\boldsymbol{\gamma}) = 2n|\epsilon_n^T(\boldsymbol{\gamma})\hat{\boldsymbol{\Delta}}l_n(0)|/(1+\|\boldsymbol{\gamma}\|)$,
$B_3(\boldsymbol{\gamma}) = 2n|\epsilon_n^T(\boldsymbol{\gamma})\hat{\boldsymbol{\Delta}}l(\boldsymbol{\gamma})|/\|\boldsymbol{\gamma}\|$, $B_4(\boldsymbol{\gamma}) = n|2l_n^T(0)\hat{\boldsymbol{\Delta}}l(\boldsymbol{\gamma})-\hat{\mathbf{D}}^T\boldsymbol{\gamma}/\sqrt{n}|/(\|\boldsymbol{\gamma}\|(1+\|\boldsymbol{\gamma}\|))$
$B_5(\boldsymbol{\gamma}) = n|l^T(\boldsymbol{\gamma})(\hat{\boldsymbol{\Delta}}-\boldsymbol{\Delta})l(\boldsymbol{\gamma})|/(\|\boldsymbol{\gamma}\|(1+\|\boldsymbol{\gamma}\|))$.

From Lemma 2, we know $\sup_{\boldsymbol{\gamma}:\|\boldsymbol{\gamma}\|/\sqrt{n}\le\delta_n} \|\epsilon_n(\boldsymbol{\gamma})\|)^2 = o_p(n^{-1/2})$. We define $\nu = \{\boldsymbol{\gamma} : \|\boldsymbol{\gamma}\|/\sqrt{n} \le \delta_n\}$ and consider B_1–B_5 separately. We have

$$\begin{aligned} &\sup_\nu B_1(\boldsymbol{\gamma}) = n\sup_\nu \frac{2+\|\boldsymbol{\gamma}\|}{1+\|\boldsymbol{\gamma}\|}\epsilon_n^T(\boldsymbol{\gamma})\hat{\boldsymbol{\Delta}}\epsilon_n(\boldsymbol{\gamma}) \le n\left\|\hat{\boldsymbol{\Delta}}\right\|\sup_\nu \frac{2+\|\boldsymbol{\gamma}\|}{1+\|\boldsymbol{\gamma}\|}(\sup_\nu\|\epsilon_n(\boldsymbol{\gamma})\|)^2 = o_p(1) \text{ and,} \\ &\sup_\nu B_2(\boldsymbol{\gamma}) \le \sup_\nu 2n|\epsilon_n^T(\boldsymbol{\gamma})\hat{\boldsymbol{\Delta}}l_n(0)| \le 2n\sup_\nu\|\epsilon_n(\boldsymbol{\gamma})\|\left\|\hat{\boldsymbol{\Delta}}\right\|\|l_n(0)\| \\ &\qquad = 2\left\|\hat{\boldsymbol{\Delta}}\right\|\left\|\sqrt{n}l_n(0)\right\|\sqrt{n}\sup_\nu\|\epsilon_n(\boldsymbol{\gamma})\| = o_p(1). \end{aligned}$$

By Taylor expansion, $l(\boldsymbol{\gamma}) = e(\boldsymbol{\gamma}/\sqrt{n} + \boldsymbol{\theta}_0) = \mathbf{C}\boldsymbol{\gamma}/\sqrt{n} + o(\boldsymbol{\gamma}/\sqrt{n})$. Thus,

$$\begin{aligned}
\sup_{\nu} B_3(\boldsymbol{\gamma}) &= \sup_{\nu} 2n|\epsilon_n^T(\boldsymbol{\gamma})\hat{\boldsymbol{\Delta}}(e(\boldsymbol{\gamma}/\sqrt{n} + \boldsymbol{\theta}_0))|/\left\|\boldsymbol{\gamma}\right\| \\
&\leq \sup_{\nu} 2n\left\|\epsilon_n(\boldsymbol{\gamma})\right\| \left\|\hat{\boldsymbol{\Delta}}\right\| (\left\|\mathbf{C}\right\| \left\|\boldsymbol{\gamma}\right\| /\sqrt{n} + o(\left\|\boldsymbol{\gamma}\right\| /\sqrt{n}))/\left\|\boldsymbol{\gamma}\right\| \\
&= 2\left\|\hat{\boldsymbol{\Delta}}\right\| (\left\|\mathbf{C}\right\| + o(1))\sqrt{n}\sup_{\nu}\left\|\epsilon_n(\boldsymbol{\gamma})\right\| \\
&= o_p(1)
\end{aligned}$$

$$\begin{aligned}
\sup_{\nu} B_4(\boldsymbol{\gamma}) &= \sup_{\nu} n|2l_n^T(0)\hat{\boldsymbol{\Delta}}(l(\boldsymbol{\gamma}) - \mathbf{C}\boldsymbol{\gamma}/\sqrt{n})|/(\left\|\boldsymbol{\gamma}\right\| (1 + \left\|\boldsymbol{\gamma}\right\|)) \\
&\leq 2n\sup_{\nu}\left\|l_n(0)\right\| \left\|\hat{\boldsymbol{\Delta}}\right\| o(1/\sqrt{n}) \\
&= 2o(1)\left\|\hat{\boldsymbol{\Delta}}\right\| \sqrt{n}\left\|l_n(0)\right\| \\
&= o_p(1).
\end{aligned}$$

Finally,

$$\begin{aligned}
\sup_{\nu} B_5(\boldsymbol{\gamma}) &\leq \sup_{\nu} n\left\|l(\boldsymbol{\gamma})\right\|^2 \left\|\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\right\| /(\left\|\boldsymbol{\gamma}\right\| (1 + \left\|\boldsymbol{\gamma}\right\|)) \\
&\leq \sup_{\nu}\left\|\boldsymbol{\gamma}\right\|^2 (\left\|\mathbf{C}\right\| + o(1))^2 \left\|\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\right\| /\left\|\boldsymbol{\gamma}\right\|^2 \\
&= \left\|\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\right\| (\left\|\mathbf{C}\right\| + o(1))^2 = o_p(1).
\end{aligned}$$

Therefore, $\sup_{\nu} \frac{|\rho_n(\boldsymbol{\gamma})|}{\left\|\boldsymbol{\gamma}\right\|(1+\left\|\boldsymbol{\gamma}\right\|)} = o_p(1)$. □

Before stating the next Lemma, we define $\hat{\boldsymbol{\gamma}} = \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. Note that $T_n(\boldsymbol{\gamma})$ is minimized at $\hat{\boldsymbol{\gamma}}$.

Lemma 4. *Under the same conditions in Theorem 2,* $\left\|\hat{\boldsymbol{\gamma}}\right\| = O_p(1)$.

Proof. Let ν be the same defined in Lemma 4. Firstly,

$$\begin{aligned}
\sup_{\nu}|\kappa_n(\boldsymbol{\gamma})| &= \sup_{\mu}|\epsilon_n^T(\boldsymbol{\gamma})\hat{\boldsymbol{\Delta}}\epsilon_n(\boldsymbol{\gamma}) + 2l_n(0)^T\hat{\boldsymbol{\Delta}}\epsilon_n(\boldsymbol{\gamma})| \\
&\leq \left\|\hat{\boldsymbol{\Delta}}\right\| (\sup_{\mu}\left\|\epsilon_n(\boldsymbol{\gamma})\right\|)^2 + \left\|\hat{\boldsymbol{\Delta}}\right\| \sup_{\mu}\left\|l_n(0)\right\| \left\|\epsilon_n(\boldsymbol{\gamma})\right\| \\
&= o_p(n^{-1}).
\end{aligned}$$

Since $T_n(\hat{\boldsymbol{\gamma}}) \leq T_n(0)$ and $\hat{\boldsymbol{\gamma}} \in \nu$, $T_n(\hat{\boldsymbol{\gamma}}) - \kappa_n(\hat{\boldsymbol{\gamma}}) = T_n(\hat{\boldsymbol{\gamma}}) + o_p(n^{-1}) \leq T_n(0) + o_p(n^{-1})$. We define

$$M = -n[T_n(\hat{\boldsymbol{\gamma}}) - \kappa_n(\hat{\boldsymbol{\gamma}}) - T_n(0) - o_p(n^{-1})] = -\rho_n(\hat{\boldsymbol{\gamma}}) - \sqrt{n}\hat{\mathbf{D}}^T\hat{\boldsymbol{\gamma}} - nT(\hat{\boldsymbol{\gamma}}) + o_p(1) \geq 0.$$

By Taylor expansion, we have $T(\hat{\boldsymbol{\gamma}}) = \hat{\boldsymbol{\gamma}}^T\mathbf{H}\hat{\boldsymbol{\gamma}}/2n + o(\left\|\hat{\boldsymbol{\gamma}}\right\|^2/n)$, where $\mathbf{H} = n\frac{\partial^2 T(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}\boldsymbol{\gamma}^T}\Big|_{\boldsymbol{\gamma}=0} = \frac{\partial^2 Q(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = 2\mathbf{C}\mathbf{G}^{-1}\mathbf{C}$. Because $\mathbf{H}$ is a positive definite

matrix by (C4), there exists a positive constant c such that with probability one $T(\hat{\boldsymbol{\gamma}}) \geq c \left\|\boldsymbol{\gamma}\right\|^2 / n$. Therefore, by applying Lemma 4, we have

$$\begin{aligned} M &\leq \left\|\hat{\boldsymbol{\gamma}}\right\| (1 + \|\hat{\hat{\boldsymbol{\gamma}}}\|) o_p(1) + \sqrt{n} \left\|\hat{\mathbf{D}}\right\|^T \left\|\hat{\boldsymbol{\gamma}}\right\| - c \left\|\hat{\boldsymbol{\gamma}}\right\|^2 + o_p(1) \\ &\leq \left\|\hat{\boldsymbol{\gamma}}\right\| (1 + \|\hat{\hat{\boldsymbol{\gamma}}}\|) o_p(1) + 2\sqrt{n} \left\|\mathbf{C}\right\| \left\|\hat{\boldsymbol{\Delta}}\right\| \left\|l_n(0)\right\| \left\|\hat{\boldsymbol{\gamma}}\right\| - c \left\|\hat{\boldsymbol{\gamma}}\right\|^2 + o_p(1) \\ &= \left\|\hat{\boldsymbol{\gamma}}\right\| (1 + \|\hat{\hat{\boldsymbol{\gamma}}}\|) o_p(1) + O_p(1) \left\|\hat{\boldsymbol{\gamma}}\right\| - c \left\|\hat{\boldsymbol{\gamma}}\right\|^2 + o_p(1) \\ &= [-c + o_p(1)] \left\|\hat{\boldsymbol{\gamma}}\right\|^2 + \left\|\hat{\boldsymbol{\gamma}}\right\| O_p(1) + o_p(1). \end{aligned}$$

Since $M \geq 0$,

$$(c - o_p(1)) \left\|\hat{\boldsymbol{\gamma}}\right\|^2 - O_p(1) \left\|\hat{\boldsymbol{\gamma}}\right\| \leq o_p(1) \Longrightarrow \left\|\hat{\boldsymbol{\gamma}}\right\|^2 - O_p(1) \left\|\hat{\boldsymbol{\gamma}}\right\| \leq o_p(1) \Longrightarrow \hat{\boldsymbol{\gamma}} = O_p(1).$$

□

To prove Theorem 2, we define $Z_n(\boldsymbol{\gamma}) = n[T_n(\boldsymbol{\gamma}) - T_n(0)]$. Obviously, $Z_n(\boldsymbol{\gamma})$ is minimized at $\hat{\boldsymbol{\gamma}}$. Based on Lemma 4, Lemma 5 and Taylor expansion, we have

$$Z_n(\boldsymbol{\gamma}) = \sqrt{n}\hat{\mathbf{D}}^T\boldsymbol{\gamma} + \frac{1}{2}\boldsymbol{\gamma}^T\mathbf{H}\boldsymbol{\gamma} + o(\|\boldsymbol{\gamma}\|^2) + \rho_n(\boldsymbol{\gamma}) + n\kappa_n(\boldsymbol{\gamma}) \xrightarrow{d} \mathbf{N}^T\boldsymbol{\gamma} + \frac{1}{2}\boldsymbol{\gamma}^T\mathbf{H}\boldsymbol{\gamma},$$

where $\mathbf{N}$ is a random vector distributed as $\mathcal{N}(0, 4\mathbf{C}\mathbf{G}^{-1}\mathbf{C})$. We define $Z(\boldsymbol{\gamma}) = \mathbf{N}^T\boldsymbol{\gamma} + \frac{1}{2}\boldsymbol{\gamma}^T\mathbf{H}\boldsymbol{\gamma}$. By Corollary 5.58 in [8], we have $\hat{\boldsymbol{\gamma}} \xrightarrow{d} \tilde{\boldsymbol{\gamma}}$, where

$$\tilde{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}/\sqrt{n}+\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_e}{\operatorname{argmin}} Z(\boldsymbol{\gamma}) = \underset{\boldsymbol{\gamma}/\sqrt{n}+\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_e}{\operatorname{argmin}} \frac{1}{2}(\boldsymbol{\gamma} + \mathbf{H}^{-1}\mathbf{N})^T\mathbf{H}(\boldsymbol{\gamma} + \mathbf{H}^{-1}\mathbf{N}).$$

The parameter vector $\boldsymbol{\gamma}$ is overparameterized. We apply Proposition 4.1 in [6] to solve this problem. The discrepancy function can be formed as

$$F(x, \xi) = \frac{1}{2}(\frac{\boldsymbol{\gamma}}{\sqrt{n}} + \frac{\mathbf{H}^{-1}\mathbf{N}}{\sqrt{n}})^T\mathbf{H}(\frac{\boldsymbol{\gamma}}{\sqrt{n}} + \frac{\mathbf{H}^{-1}\mathbf{N}}{\sqrt{n}}).$$

It is easy to check this discrepancy function satisfies Shapiro's assumptions and $\frac{\partial^2 F}{\partial \xi \xi^T} = \mathbf{H}$. In addition, $-\mathbf{H}^{-1}\mathbf{N} \xrightarrow{d} \mathcal{N}(0, \mathbf{C}^{-1}\mathbf{G}\ \mathbf{C}^{-1})$. Therefore, by Proposition 4.1 in [6], we have $\tilde{\boldsymbol{\gamma}} \xrightarrow{d} \mathcal{N}(0, \Lambda_g)$, where $\Lambda_g = \boldsymbol{\Psi}(\boldsymbol{\Psi}^T\mathbf{C}\mathbf{G}^{-1}\mathbf{C}\boldsymbol{\Psi})^\dagger\boldsymbol{\Psi}^{\mathrm{T}}$. Hence,

$$\hat{\boldsymbol{\gamma}} = \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Psi}(\boldsymbol{\Psi}^T\mathbf{C}\mathbf{G}^{-1}\mathbf{C}\boldsymbol{\Psi})^\dagger\boldsymbol{\Psi}^{\mathrm{T}}).$$

We complete the proof of Theorem 2.

1.4. Corollary

Proof. Let $\boldsymbol{\Upsilon} = \mathbf{C}^{-1}\mathbf{G}\mathbf{C}^{-1}$. According to the results in Theorem 1 and Theorem 2,

$$\begin{aligned} avar(\sqrt{n}\tilde{\boldsymbol{\theta}}) - avar(\sqrt{n}\hat{\boldsymbol{\theta}}) &= \mathbf{C}^{-1}\mathbf{G}\ \mathbf{C}^{-1} - \boldsymbol{\Psi}(\boldsymbol{\Psi}^T\mathbf{C}\mathbf{G}^{-1}\mathbf{C}\boldsymbol{\Psi})^{\dagger}\boldsymbol{\Psi}^{\mathrm{T}} \\ &= \boldsymbol{\Upsilon} - \boldsymbol{\Psi}(\boldsymbol{\Psi}^T\boldsymbol{\Upsilon}^{-1}\boldsymbol{\Psi})^{\dagger}\boldsymbol{\Psi}^T \\ &= \boldsymbol{\Upsilon}^{1/2}(\mathbf{I} - \mathbf{P}_{\boldsymbol{\Upsilon}^{-1/2}\boldsymbol{\Psi}})\boldsymbol{\Upsilon}^{1/2} \\ &= \boldsymbol{\Upsilon}^{1/2}\mathbf{Q}_{\boldsymbol{\Upsilon}^{-1/2}\boldsymbol{\Psi}}\boldsymbol{\Upsilon}^{1/2} \\ &\geq 0. \end{aligned}$$

□

2. Simulations Under Variable Selection Settings (Without Envelope Structure)

In this section, we investigate the performance of the ER model, the EER model, the boosting model and the sparse ER model under the settings in which sparsity structure exists but no (nontrivial) envelope structure exists. In this case, $u_\pi = p$, and the EER model degenerates to the ER model. We consider the following settings:

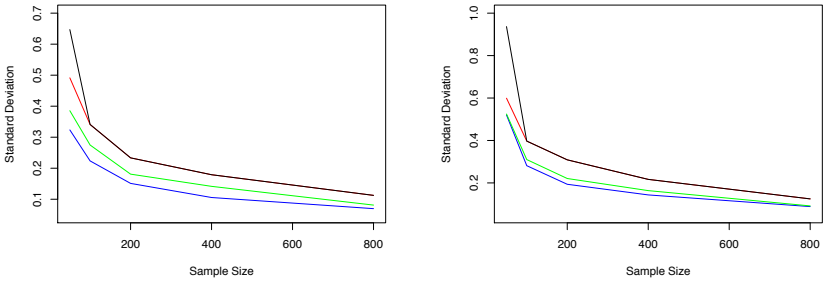
$$Y_i = 3 + \boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{X}_i + (2 + \boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{X}_i)\epsilon_i, \qquad \text{for} \quad i = 1, \ldots, n.$$

We set $p = 6$ and $p_A = 3$, where p_A denotes the number of active predictors. Both $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ were p-dimensional vectors. The first p_A elements in $\boldsymbol{\alpha}_1$ were 4 and the rest $p - p_A$ elements were 0. The first p_A elements in $\boldsymbol{\alpha}_2$ were 0.1 and the rest $p - p_A$ elements were 0. The error term ϵ was generated from standard normal distribution $\epsilon \sim \mathcal{N}(0, 1)$.

Based upon the settings, the πth conditional expectile of Y had the following form

$$f_\pi(Y|\mathbf{X}) = 3 + \boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{X} + (2 + \boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{X})f_\pi(\epsilon) = 3 + 2f_\pi(\epsilon) + (\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 f_\pi(\epsilon))^T\mathbf{X},$$

where $f_\pi(\epsilon)$ represented the πth expectile of the error distribution. Thus the coefficients were contained in $\boldsymbol{\beta}_\pi = \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 f_\pi(\epsilon)$ and the last $p - p_A$ elements of $\boldsymbol{\beta}_\pi$ were 0. This means that the first p_A predictors were active predictors, and the rest were inactive. The predictor vector $\mathbf{X}$ followed a normal distribution with mean 0 and covariance matrix $\boldsymbol{\Sigma}_\mathbf{X}$. The upper left $p_A \times p_A$ block of $\boldsymbol{\Sigma}_\mathbf{X}$ was a diagonal matrix with diagonal elements being 1, 2 and 4. The bottom right block was a $(p - p_A) \times (p - p_A)$ diagonal matrix with diagonal elements being 8, 16 and 32. The off-diagonal blocks of $\boldsymbol{\Sigma}_\mathbf{X}$ were $3\mathbf{M}$ and $3\mathbf{M}^T$, where $\mathbf{M}$ was a randomly generated $p_A \times (p - p_A)$ orthogonal matrix (generated using `randortho` function in `R` package `pracma`). In this case, the envelope subspace $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_\pi) = \mathbb{R}^p$ and the EER model reduces to the ER model since no immaterial



(a) $\pi = 0.5$ (b) $\pi = 0.9$

Fig 1: Comparison of the sample standard deviations. Red lines mark the ER estimator. Blue lines mark the sparse ER estimator. Green lines mark the boosting estimator. Black lines mark the EER estimator.

information is present. Therefore, for this scenario, the sparsity structure exists but no (nontrivial) envelope structure exists.

We varied the sample size n from 50 to 800. For each sample size, 100 replications were generated. For each replication, we computed the EER estimator (u_π chosen by RCV), the ER estimator, the boosting estimator as well as the sparse ER estimator of $\boldsymbol{\beta}_\pi$. For each element in $\boldsymbol{\beta}_\pi$, we computed the sample standard deviation from the 100 EER estimators, 100 ER estimators, 100 boosting estimators and 100 sparse ER estimators. We took expectile levels 0.50 and 0.90 as examples. The results of a randomly chosen nonzero element in $\boldsymbol{\beta}_\pi$ with $\pi = 0.50$ and $\pi = 0.90$ are summarized in Figure 1.

In each panel of Figure 1, the line for the EER estimator almost overlaps with the line for the ER estimator when sample size exceeds 100. This is expected as when RCV selected $u_\pi = p$ and the EER estimator degenerates to the ER estimator. With small sample size, there was a little variation in the model selection for the EER model, so the EER estimator was more variable than the ER estimator. The efficiency gains from the sparse ER model and the boosting model is obvious under this setting. Take $n = 200$ as an example, the standard deviation is 0.23 for the ER or EER estimator, 0.18 for the boosting estimator and 0.15 for the sparse ER estimator for $\pi = 0.5$. The efficient gains is because that the boosting estimator and the sparse ER estimator correctly identified the underlying sparsity structure. Therefore, in the case where there is sparsity structure but no (nontrivial) envelope structure, the boosting estimator and the sparse ER estimator achieves more efficiency gains than the EER estimator. However, if the sparsity structure and the envelope structure both exist, the EER estimator may be more efficient than the boosting and sparse ER estimator as shown in Section 3.

3. Simulations Under Variable Selection Settings (With Envelope Structure)

In this section, we investigate the performance of the EER model when the underlying model has both the sparsity structure and the envelope structure. We consider the following simulation settings:

$$Y_i = 3 + \boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{X}_i + (8 + \boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{X}_i)\epsilon_i, \qquad \text{for} \quad i = 1, \ldots, n.$$

We set $p = 12$, $u_\pi = 2$ and $p_A = 6$, where p_A denotes the number of active predictors. Both $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ were p-dimensional vectors. The first p_A elements in $\boldsymbol{\alpha}_1$ were 4 and the rest $p - p_A$ elements were 0. The first p_A elements in $\boldsymbol{\alpha}_2$ were 0.1 and the rest $p - p_A$ elements were 0. Four types of error distribution were used to generate ϵ: standard normal distribution $\epsilon \sim \mathcal{N}(0, 1)$, student's t–distribution with 4 degrees of freedom $\epsilon \sim t_4$, mixed normal distribution $\epsilon \sim 0.9\mathcal{N}(0, 1) + 0.1\mathcal{N}(1, 5)$, and exponential distribution $\epsilon \sim \mathrm{Exp}(1)$.

Based upon the settings, the πth conditional expectile of Y had the following form

$$f_\pi(Y|\mathbf{X}) = 3 + \boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{X} + (8 + \boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{X})f_\pi(\epsilon) = 3 + 8f_\pi(\epsilon) + (\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 f_\pi(\epsilon))^T\mathbf{X},$$

where $f_\pi(\epsilon)$ represented the πth expectile of the error distribution. Thus $\boldsymbol{\beta}_\pi = \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 f_\pi(\epsilon)$ and the last $p - p_A$ elements of $\boldsymbol{\beta}_\pi$ were 0, which means only the first p_A components in $\mathbf{X}$ were active predictors. The predictor vector $\mathbf{X}$ followed a normal distribution with mean 0 and covariance matrix $\boldsymbol{\Sigma}_\mathbf{X} = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^T + \boldsymbol{\Phi}_0\boldsymbol{\Lambda}_0\boldsymbol{\Phi}_0^T$, where $\boldsymbol{\Lambda}$ was a $u_\pi \times u_\pi$ diagonal matrix with diagonal elements 100 and 9, and $\boldsymbol{\Lambda}_0$ was a 2×2 block matrix. The upper left block of $\boldsymbol{\Lambda}_0$ was a $(p_A - u_\pi) \times (p_A - u_\pi)$ identity matrix and the bottom right block was a $(p - p_A) \times (p - p_A)$ identity matrix. The off-diagonal blocks of $\boldsymbol{\Lambda}_0$ were $0.8\boldsymbol{\Lambda}_{0*}$ and $0.8\boldsymbol{\Lambda}_{0*}^T$ where $\boldsymbol{\Lambda}_{0*}$ was a randomly generated $(p - p_A) \times (p_A - u_\pi)$ semi-orthogonal matrix. The matrix $\boldsymbol{\Phi} \in \mathbb{R}^{p \times u_\pi}$ was a semi-orthogonal matrix with the first $p_A/2$ rows being $(\sqrt{3}/3,\ 0)$, the following $p_A/2$ rows being $(0,\ \sqrt{3}/3)$ and the remaining $p - p_A$ rows being $(0,\ 0)$. The matrix $\boldsymbol{\Phi}_0 \in \mathbb{R}^{p \times (p - u_\pi)}$ was a semi-orthogonal matrix that satisfied $\boldsymbol{\Phi}^T\boldsymbol{\Phi}_0 = 0$. Since $\boldsymbol{\alpha}_1 = \boldsymbol{\Phi} \cdot (4\sqrt{3}, 4\sqrt{3})^T$ and $\boldsymbol{\alpha}_2 = \boldsymbol{\Phi} \cdot (\sqrt{3}/10, \sqrt{3}/10)^T$, $f_\pi(Y|\mathbf{X})$ and $\mathbf{X}$ satisfied the EER model with $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\boldsymbol{\beta}_\pi) = \mathrm{span}(\boldsymbol{\Phi})$.

Under this setting, we repeated the sample standard deviations comparison, the prediction performance comparison and the RCV performance examination as described in Section 5 of the paper. To be noted, here for the sample standard deviations comparison, we randomly choose an active component of $\boldsymbol{\beta}_\pi$ to display the outcomes. All results are given in Figures 2 – 7 and Tables 1 – 2.

Figure 2 shows substantial efficiency gains from the EER model in the estimation of $\boldsymbol{\beta}_\pi$. In all the plots with different error distributions and expectile levels π, the sample standard deviations of the EER estimators are much smaller than the sample standard deviations of the ER estimators, the boosting estimators and the sparse ER estimators under all sample sizes. As variable selection methods, the boosting model and the sparse ER model are more efficient than the

ER model since they correctly identify the underlying sparse structure. However, they do not account for the immaterial information in $\mathbf{X}$ in the estimation. The EER model can still be more efficient than the boosting model and the sparse ER model if the variation of the immaterial part has a large effect on estimation, such as in this example.

Figure 3 indicates that the bootstrap standard deviation is a good approximation to the actual sample standard deviation. Table 1 and Figures 4 – 7 summarize the RMSEs under the EER model, the ER model, the boosting model and the sparse ER model with different error distributions. We can see a notable improvement of the prediction performance for the EER model. Take Table 1 (a) as an example, the EER model reduces the average RMSE by about 40% comparing with the ER model, by about 30% comparing with the boosting model and by about 60% comparing with the sparse ER model. Both the sparse ER model and the boosting model identifies the active predictors. But the sparse ER model tends to put more shrinkage on the nonzero coefficients, while the boosting estimator does not over shrink the nonzero coefficients. Therefore, we notice that the boosting estimator has a better prediction performance than the ER estimator, but the sparse ER estimator has the largest prediction error.

Table 2 summaries the fraction that RCV selects the true dimension $u_\pi = 2$. RCV selects the true dimension more than 90% of the time when sample size reaches 100. And it still gives an accuracy over 75% with a small sample size 25.

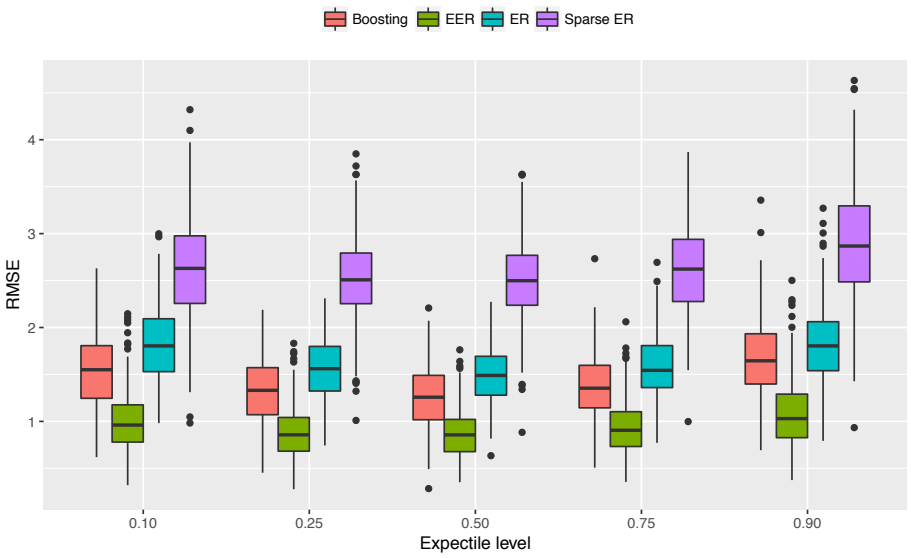
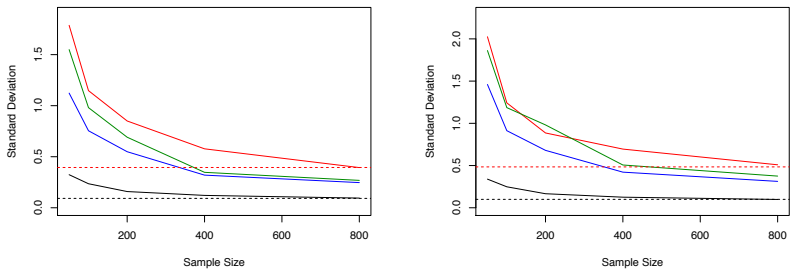
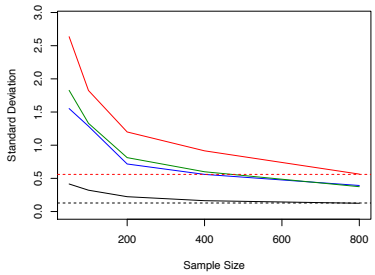


Fig 4: Boxplots of RMSEs under the four models with $\epsilon \sim \mathcal{N}(0, 1)$.

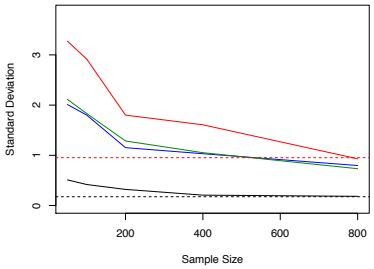


(a) Standard normal with $\pi = 0.5$

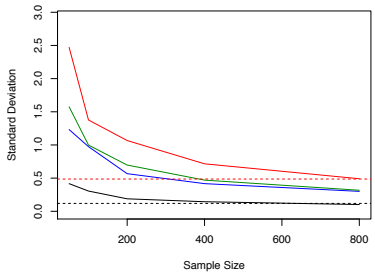
(b) Standard normal with $\pi = 0.9$



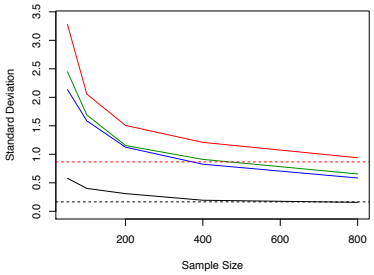
(c) t_4 with $\pi = 0.5$



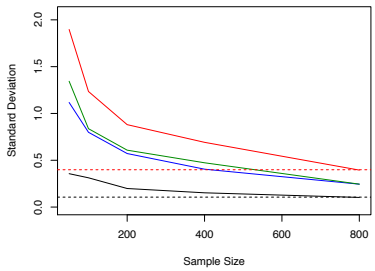
(d) t_4 with $\pi = 0.9$



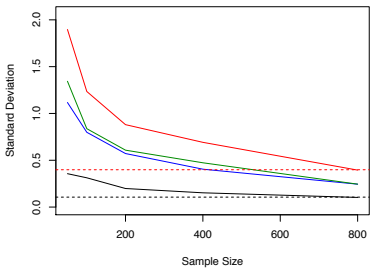
(e) Mixed normal with $\pi = 0.5$



(f) Mixed normal with $\pi = 0.9$



(g) Exp(1) with $\pi = 0.5$



(h) Exp(1) with $\pi = 0.9$

Fig 2: Sample standard deviations. Red lines mark the ER estimator. Blue lines mark the sparse ER estimator. Green lines mark the boosting estimator. Black lines mark the EER estimator. The horizontal lines mark the asymptotic standard deviations of the ER estimator (the upper line in each panel) and the EER estimator (the lower line in each panel).

TABLE 1
The average RMSEs of the 300 replications under the four models with different error distributions.

(a) $\epsilon \sim \mathcal{N}(0,1)$

	EER	ER	Boosting	Sparse ER
$\pi = 0.10$	1.05	1.82	1.54	2.65
$\pi = 0.25$	0.88	1.57	1.32	2.51
$\pi = 0.50$	0.87	1.49	1.26	2.50
$\pi = 0.75$	0.94	1.58	1.38	2.62
$\pi = 0.90$	1.09	1.83	1.67	2.89

(b) $\epsilon \sim t_4$

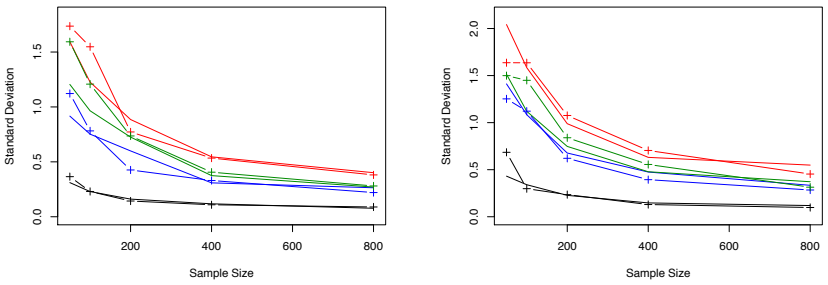
	EER	ER	Boosting	Sparse ER
$\pi = 0.10$	1.77	3.42	2.89	6.00
$\pi = 0.25$	1.20	2.40	2.02	4.84
$\pi = 0.50$	1.08	2.09	1.77	4.52
$\pi = 0.75$	1.24	2.39	2.06	5.04
$\pi = 0.90$	1.79	3.41	3.06	6.54

(c) $\epsilon \sim 0.9\mathcal{N}(0,1) + 0.1\mathcal{N}(1,5)$

	EER	ER	Boosting	Sparse ER
$\pi = 0.10$	1.15	2.16	1.86	3.53
$\pi = 0.25$	1.01	1.83	1.55	3.35
$\pi = 0.50$	1.01	1.81	1.54	3.55
$\pi = 0.75$	1.20	2.15	1.85	4.24
$\pi = 0.90$	1.72	3.06	2.72	5.59

(d) $\epsilon \sim \text{Exp}(1)$

	EER	ER	Boosting	Sparse ER
$\pi = 0.10$	0.64	0.74	0.67	1.85
$\pi = 0.25$	0.73	1.04	0.90	2.42
$\pi = 0.50$	0.93	1.51	1.29	3.17
$\pi = 0.75$	1.33	2.22	1.95	4.19
$\pi = 0.90$	1.96	3.31	2.99	5.53



(a) Standard normal with $\pi = 0.5$ (b) Standard normal with $\pi = 0.9$

Fig 3: Sample standard deviations and bootstrap standard deviations. Red lines mark the ER estimator. Blue lines mark the sparse ER estimator. Green lines mark the boosting estimator. Black lines mark the EER estimator. Lines with "+" mark the bootstrap standard deviations for the corresponding estimators.

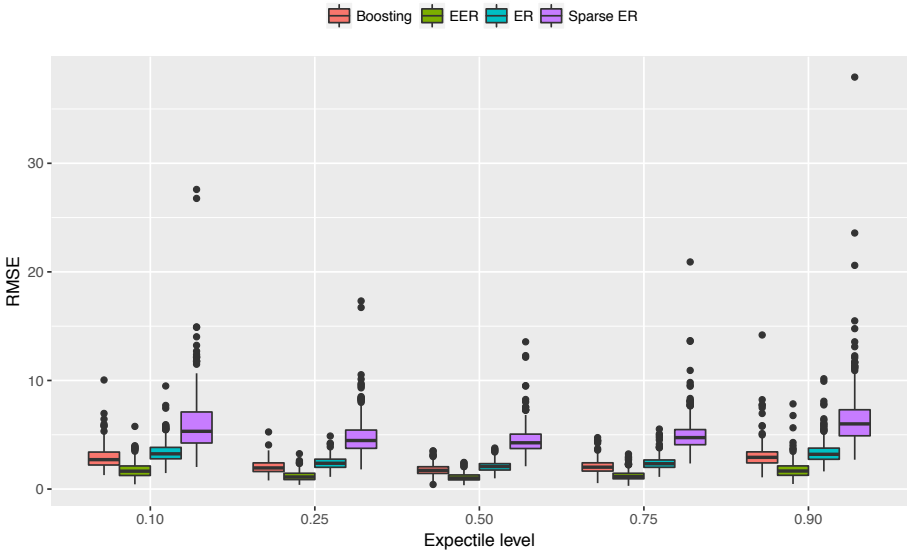


Fig 5: Boxplots of RMSEs under the four models with $\epsilon \sim t_4$.

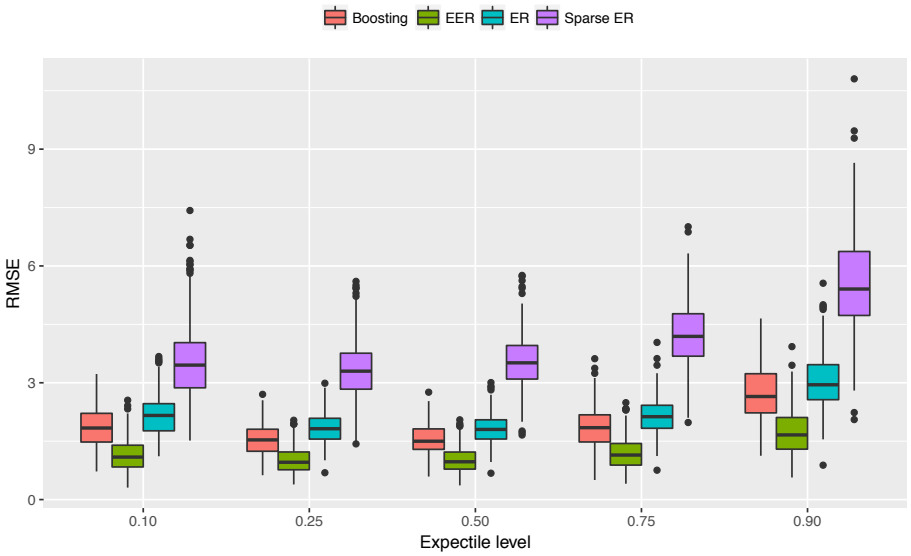


Fig 6: Boxplots of RMSEs under the four models with $\epsilon \sim 0.9\mathcal{N}(0,1) + 0.1\mathcal{N}(1,5)$.

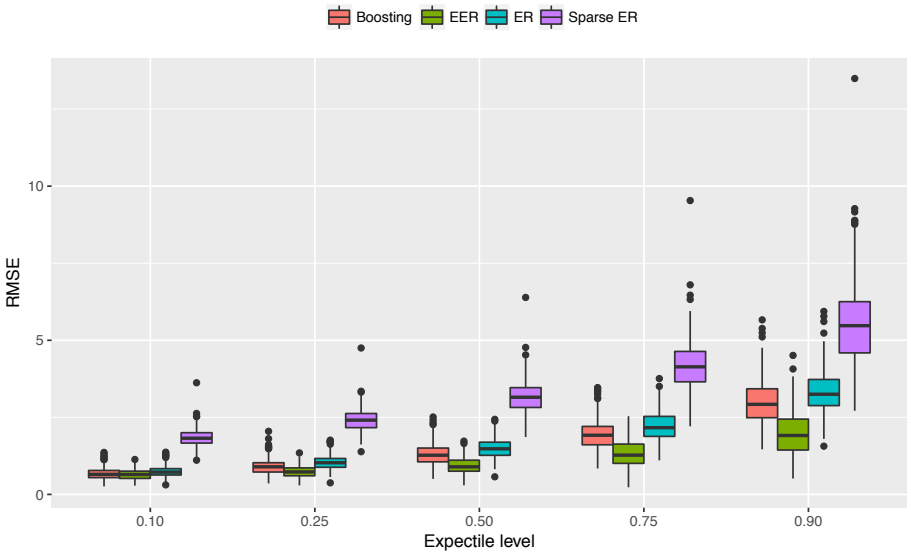


Fig 7: Boxplots of RMSEs under the four models with $\epsilon \sim \text{Exp}(1)$.

TABLE 2
The fraction that RCV selects the true u_π with different error distributions.

(a) $\epsilon \sim \mathcal{N}(0,1)$

	$\pi = 0.10$	$\pi = 0.25$	$\pi = 0.50$	$\pi = 0.75$	$\pi = 0.90$
$n = 25$	79%	83%	80%	79%	82%
$n = 50$	95%	94%	94%	92%	88%
$n = 100$	97%	97%	99%	98%	96%
$n = 200$	99%	100%	99%	100%	99%
$n = 400$	99%	100%	100%	100%	100%
$n = 800$	100%	100%	100%	100%	100%

(b) $\epsilon \sim t_4$

	$\pi = 0.10$	$\pi = 0.25$	$\pi = 0.50$	$\pi = 0.75$	$\pi = 0.90$
$n = 25$	84%	88%	89%	88%	80%
$n = 50$	90%	94%	95%	91%	93%
$n = 100$	98%	98%	99%	99%	96%
$n = 200$	100%	100%	100%	100%	100%
$n = 400$	100%	100%	100%	100%	100%
$n = 800$	100%	100%	100%	100%	100%

(c) $\epsilon \sim 0.9\mathcal{N}(0,1) + 0.1\mathcal{N}(1,5)$

	$\pi = 0.10$	$\pi = 0.25$	$\pi = 0.50$	$\pi = 0.75$	$\pi = 0.90$
$n = 25$	80%	85%	88%	84%	84%
$n = 50$	90%	96%	96%	91%	91%
$n = 100$	93%	99%	99%	98%	98%
$n = 200$	98%	100%	100%	100%	100%
$n = 400$	100%	100%	100%	100%	100%
$n = 800$	100%	100%	100%	100%	100%

(d) $\epsilon \sim \text{Exp}(1)$

	$\pi = 0.10$	$\pi = 0.25$	$\pi = 0.50$	$\pi = 0.75$	$\pi = 0.90$
$n = 25$	78%	78%	76%	79%	78%
$n = 50$	95%	95%	98%	94%	84%
$n = 100$	99%	99%	100%	99%	96%
$n = 200$	100%	100%	100%	99%	99%
$n = 400$	100%	100%	100%	100%	100%
$n = 800$	100%	100%	100%	100%	100%

4. Simulations Under No Immaterial Part Settings

In this section, we conduct a simulation study to investigate the performance of the EER model if no immaterial part exists. The data was generated from the following model

$$Y_i = 3 + \boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{X}_i + (8 + \boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{X}_i)\epsilon_i, \qquad \text{for} \quad i = 1, \ldots, 800.$$

We set $p = 6$ and each element in $\boldsymbol{\alpha}_1$ was drawn from independent standard normal distribution. Each elements in $\boldsymbol{\alpha}_2$ was 0.1. The predictor vector $\mathbf{X}$ followed a normal distribution with mean 0 and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}} = \mathbf{P}^T\mathbf{D}\mathbf{P}$, where $\mathbf{P}$ was a randomly generated orthogonal matrix (generated using `randortho` function in `R` package `pracma`), and $\mathbf{D}$ was a diagonal matrix with diagonal elements being 1, 2, 4, 8, 16 and 32. The error ϵ was generated from the normal distribution $\epsilon \sim \mathcal{N}(0, 5)$.

Based upon the settings, the πth conditional expectile of Y has the following form

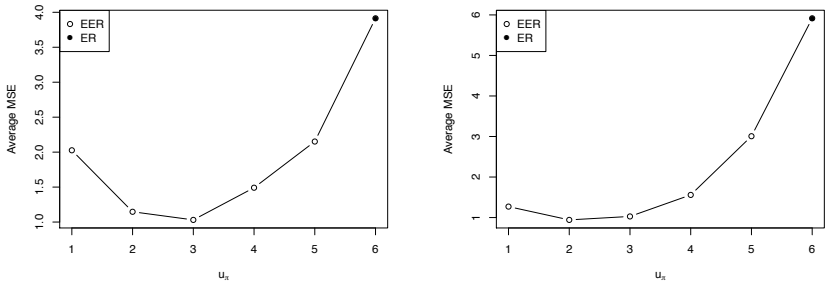
$$f_\pi(Y|\mathbf{X}) = 3 + \boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{X} + (8 + \boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{X})f_\pi(\epsilon) = 3 + 8f_\pi(\epsilon) + (\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 f_\pi(\epsilon))^T\mathbf{X},$$

where $f_\pi(\epsilon)$ represents the πth expectile of the error distribution, the intercept is $3+8f_\pi(\epsilon)$ and the coefficients are $\boldsymbol{\beta}_\pi = \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 f_\pi(\epsilon)$. In this case, $\boldsymbol{\Sigma}_{\mathbf{X}}$ does not have the decomposition as a sum of the variation of the material part (related to $\boldsymbol{\beta}_\pi$) and the variation of the immaterial part. So the envelope subspace is the full space $\mathbb{R}^p$, and there is no immaterial part.

Although no envelope structure is present, we will compute an approximate "EER" estimator and compare it with the ER estimator. We know that under an EER model, the envelope subspace $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}_\pi) = \mathrm{span}(\boldsymbol{\Gamma}_\pi)$ is spanned by the eigenvectors of $\boldsymbol{\Sigma}_{\mathbf{X}}$. Therefore, for each $1 \le u_\pi \le p$, we approximate $\boldsymbol{\Gamma}_\pi$ by $\hat{\boldsymbol{\Gamma}}_\pi$, which is a $p \times u_\pi$ matrix whose columns were the u_π eigenvectors of $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}$ corresponding to the u_π largest eigenvalues. We note that under the exact EER model, $\boldsymbol{\Gamma}_\pi$ is chosen to be the eigenvectors of $\boldsymbol{\Sigma}_{\mathbf{X}}$ that contains $\boldsymbol{\beta}_\pi$. They may not necessarily be the eigenvectors corresponding to the largest eigenvalues. Since the exact (nontrivial) EER model does not exist here, we are proposing a way to approximate the $\boldsymbol{\Gamma}_\pi$ such that its estimator is least variable. Then we defined the "EER" estimator as $\hat{\boldsymbol{\beta}}_\pi = \hat{\boldsymbol{\Gamma}}_\pi\hat{\boldsymbol{\eta}}_\pi$, where $\hat{\boldsymbol{\eta}}_\pi$ was the ER estimator with Y being the response and $\hat{\boldsymbol{\Gamma}}_\pi^T\mathbf{X}$ being the predictors.

We generated 100 replications and computed the mean squared error (MSE) $\|\hat{\boldsymbol{\beta}}_\pi - \boldsymbol{\beta}_\pi\|^2$ at expectile levels $\pi = 0.5$ and 0.9 for each replication. The average MSE are summarized in Figure 8. Because the true u_π equals p, a smaller u_π leads to larger bias, but its estimator is less variable. As u_π increases, the bias of $\hat{\boldsymbol{\beta}}_\pi$ becomes smaller but its variance becomes larger. When $\pi = 0.5$, the bias-variance tradeoff makes the average MSE reach its minimum 1.03 at $u_\pi = 3$. When $\pi = 0.9$, the average MSE reaches its minimum 0.94 at $u_\pi = 2$. Note that when $u_\pi = p$, the EER estimator $\hat{\boldsymbol{\beta}}_\pi$ reduces to the ER estimator. The MSE of the ER estimator is 3.91 for $\pi = 0.5$ and 5.92 for $\pi = 0.9$. The results shows that

when there is no immaterial part, we can still expect to have a smaller MSE from an approximate EER estimator in some cases due to the bias-variance tradeoff.



(a) $\pi = 0.5$ (b) $\pi = 0.9$

Fig 8: Average MSE with respect to u_π. Note that the MSE corresponding to $u_\pi = p = 6$ is the MSE of the ER estimator.

5. Analysis of "state.x77" with Predictors in Original Scale

We perform the same data analysis on the "state.x77" data again, but with the predictors in the original scale instead of the standardized predictors. We first select the dimension of the envelope subspace with RCV. For all quantile levels, RCV selects $u_\pi = 4(= p)$. This indicated that there is no immaterial part in the data and the EER estimator reduces to the ER estimator. In this case, the EER estimator and the ER estimator have the same efficiency. Detailed information about the estimated regression coefficients are provided in the following Table 3. Because $u_\pi = p$, the estimated regression coefficients given by the EER model and the ER model are exactly the same.

TABLE 3

The estimated regression coefficients for the predictors in the original scale given by the EER model, the ER model, the boosting model and the sparse ER model.

	EER				ER			
	Population	Income	Illiteracy	Frost	Population	Income	Illiteracy	Frost
$\pi = 0.10$	0.29×10^{-3}	-1.14×10^{-3}	3.30	-9.24×10^{-3}	0.29×10^{-3}	-1.14×10^{-3}	3.30	-9.24×10^{-3}
$\pi = 0.25$	0.26×10^{-3}	-0.66×10^{-3}	3.65	-5.39×10^{-3}	0.26×10^{-3}	-0.66×10^{-3}	3.65	-5.39×10^{-3}
$\pi = 0.50$	0.22×10^{-3}	0.06×10^{-3}	4.14	0.58×10^{-3}	0.22×10^{-3}	0.06×10^{-3}	4.14	0.58×10^{-3}
$\pi = 0.75$	0.21×10^{-3}	0.55×10^{-3}	4.48	5.22×10^{-3}	0.21×10^{-3}	0.55×10^{-3}	4.48	5.22×10^{-3}
$\pi = 0.90$	0.18×10^{-3}	0.72×10^{-3}	4.58	7.89×10^{-3}	0.18×10^{-3}	0.72×10^{-3}	4.58	7.89×10^{-3}
	Boosting				Sparse ER			
	Population	Income	Illiteracy	Frost	Population	Income	Illiteracy	Frost
$\pi = 0.10$	0.29×10^{-3}	-1.12×10^{-3}	3.22	-10.00×10^{-3}	0.10×10^{-3}	0.00×10^{-3}	2.28	-9.34×10^{-3}
$\pi = 0.25$	0.23×10^{-3}	-0.41×10^{-3}	3.59	-5.05×10^{-3}	0.08×10^{-3}	0.00×10^{-3}	2.89	-2.95×10^{-3}
$\pi = 0.50$	0.19×10^{-3}	0.00×10^{-3}	3.79	0.00×10^{-3}	0.04×10^{-3}	0.00×10^{-3}	2.70	0.00×10^{-3}
$\pi = 0.75$	0.12×10^{-3}	0.00×10^{-3}	3.36	0.00×10^{-3}	0.00×10^{-3}	0.00×10^{-3}	2.12	0.00×10^{-3}
$\pi = 0.90$	0.06×10^{-3}	0.00×10^{-3}	2.92	0.00×10^{-3}	0.00×10^{-3}	0.00×10^{-3}	0.73	0.00×10^{-3}

We took a close look at the data and found that the scales of the four predictors are quite different. For example, population varies from 365 to 21198 (thousand) while illiteracy level varies from 0.5 to 2.8 (percent). This makes the eigenvalues of $\mathbf{\Sigma_X}$ range from 0.17 to 1.99×10^7 and the eigenvectors are very close to the standard basis vectors, i.e, $(1,0,0,0)^T$, $(0,1,0,0)^T$, $(0,0,1,0)^T$ and $(0,0,0,1)^T$. In this case, if $\boldsymbol{\beta}_\pi$ belongs to an envelope subspace that is a proper subset of $\mathbb{R}^p$, then we are essentially performing variable selection. For example, if the dimension of envelope subspace is $u = 3$, then the envelope subspace is spanned by 3 out of the 4 eigenvectors of $\mathbf{\Sigma_X}$. Since $\boldsymbol{\beta}_\pi$ lies in the envelope subspace, one component of $\boldsymbol{\beta}_\pi$ has to be 0, which means the corresponding predictor is immaterial to the conditional expectile of the response. The dimension selection results from RCV indicates the EER model finds that all four predictors are material to the conditional expectile of the response at all investigated expectile levels.

This situation is also shared by other dimension reduction based methods such as principal component analysis (PCA). If one component is selected, which corresponds to direction $(0,1,0,0)^T$, then only one variable (income) is included in subsequent analysis. Thus when variables have drastically different scales, PCA normally standardize the variables. We followed this practice and presented the results with standardized predictor variables in Section 6 of the paper.

6. Prediction Performance Comparison on "state.x77"

We compared the prediction performance between the ER model and the EER model on "state.x77" using five fold cross-validation repeated with 50 random splits to compute the mean predicted expectile losses. The results are summarized in the following table. The predicted expectile losses from the EER model

TABLE 4
Mean of the predicted expectile losses under the ER and the EER model with different expectile levels.

	$\pi = 0.10$	$\pi = 0.25$	$\pi = 0.50$	$\pi = 0.75$	$\pi = 0.90$
ER	1.48	2.79	3.75	3.29	2.45
EER	1.58	2.82	3.86	3.85	2.72

are slightly larger than those from the ER model. This may due to the criterion we use to select the dimension of the envelope subspace u_π. We selected u_π by RCV with one standard deviation rule. In other words, instead of choosing the dimension that has the minimum RCV, we choose the smallest dimension having RCV less than one standard deviation above the minimum value of RCV. Therefore this criterion tends to select a more parsimonious model by sacrificing some predictive accuracy comparing to the best model. In this case, it is possible that the full model, i.e., the ER model, has an RCV that is closer to the minimum value of RCV compared to the selected model.

7. Simulation Results at More Expectile Levels

In Section 5 of the paper, we give the results at expectile levels 0.50 and 0.90 in Figure 1. Here we provide the results at expectile levels 0.10, 0.25 and 0.75 in Figure 9.

Figure 9 shows a similar pattern as Figure 1 of the paper. For every error distribution and expectile level, the sample standard deviations of the EER estimators are much smaller than the sample standard deviations of the ER estimators, the boosting estimators and the sparse ER estimators under all sample sizes.

8. Analysis of Computational Complexity of the GMM Algorithm

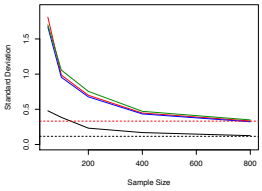
We give an analysis about the computational burden on the parameter estimation approach – generalized method of moments (GMM). There are three steps in GMM and we will count the number of flops for each step.

Step 1 : Get the intermediate estimator $\hat{\boldsymbol{\zeta}}^*$ by minimizing $e_n^*(\boldsymbol{\zeta})^T e_n^*(\boldsymbol{\zeta})$, where

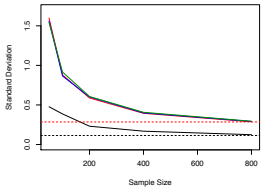
$$e_n^*(\boldsymbol{\zeta}) = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n \mathbf{W}_i(Y_i - \mu_\pi - \mathbf{X}_i^T\boldsymbol{\Gamma}_\pi\boldsymbol{\eta}_\pi)\left|I(Y_i < \mu_\pi + \mathbf{X}_i^T\boldsymbol{\Gamma}_\pi\boldsymbol{\eta}_\pi) - \pi\right| \\ \mathrm{vech}(\boldsymbol{\Gamma}_\pi\boldsymbol{\Omega}_\pi{\boldsymbol{\Gamma}_\pi}^T + \boldsymbol{\Gamma}_{0\pi}\boldsymbol{\Omega}_{0\pi}{\boldsymbol{\Gamma}_{0\pi}}^T) - \mathrm{vech}(\mathbf{S}_\mathbf{X}) \\ \boldsymbol{\mu}_\mathbf{X} - \bar{\mathbf{X}} \end{pmatrix}.$$

In this step, we apply Nelder-Mead method to find the minimum of the objective function. It is an iterative method and the number of flops in each iteration is $\mathcal{O}(T_f)$, where T_f represents the number of flops to compute the value of the objective function $e_n^*(\boldsymbol{\zeta})^T e_n^*(\boldsymbol{\zeta})$ for a given $\boldsymbol{\zeta}$ ([7]).

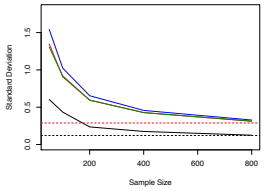
- Because $\mathbf{X}_i$ is a p-dimensional vector, $\boldsymbol{\Gamma}_\pi$ is a p by u_π matrix and $\boldsymbol{\eta}_\pi$ is a u_π-dimensional vector, it takes $\mathcal{O}(pu_\pi)$ flops to compute $\mathbf{X}_i^T\boldsymbol{\Gamma}_\pi\boldsymbol{\eta}_\pi$. Afterwards, because Y_i, μ_π and $\mathbf{X}_i^T\boldsymbol{\Gamma}_\pi\boldsymbol{\eta}_\pi$ are scalars, it takes $\mathcal{O}(1)$ flops to compute $(Y_i - \mu_\pi - \mathbf{X}_i^T\boldsymbol{\Gamma}_\pi\boldsymbol{\eta}_\pi)$ and $\left|I(Y_i < \mu_\pi + \mathbf{X}_i^T\boldsymbol{\Gamma}_\pi\boldsymbol{\eta}_\pi) - \pi\right|$. Next, because $\mathbf{W}_i$ is a $(p+1)$-dimensional vector, it takes $\mathcal{O}(p)$ flops to compute the product $\mathbf{W}_i(Y_i - \mu_\pi - \mathbf{X}_i^T\boldsymbol{\Gamma}_\pi\boldsymbol{\eta}_\pi)\left|I(Y_i < \mu_\pi + \mathbf{X}_i^T\boldsymbol{\Gamma}_\pi\boldsymbol{\eta}_\pi) - \pi\right|$. Finally, we need to perform the above multiplication for each sample and then take the average. Hence, the total number of flops to compute the first line in $e_n^*(\boldsymbol{\zeta})$ is $\mathcal{O}(npu_\pi)$.
- Because $\boldsymbol{\Gamma}_\pi$ is a p by u_π matrix and $\boldsymbol{\Omega}_\pi$ is a u_π by u_π matrix, it takes $\mathcal{O}(p^2u_\pi)$ flops to compute $\boldsymbol{\Gamma}_\pi\boldsymbol{\Omega}_\pi{\boldsymbol{\Gamma}_\pi}^T$. In addition, because $\boldsymbol{\Gamma}_{0\pi}$ is a p by $(p-u_\pi)$ matrix and $\boldsymbol{\Omega}_{0\pi}$ is a $(p-u_\pi)$ by $(p-u_\pi)$ matrix, it takes $\mathcal{O}(p(p-u_\pi)^2) = \mathcal{O}(p^3)$ flops to compute $\boldsymbol{\Gamma}_{0\pi}\boldsymbol{\Omega}_{0\pi}{\boldsymbol{\Gamma}_{0\pi}}^T$. Afterwards, it takes $\mathcal{O}(np^2)$ flops to compute $\mathbf{S}_\mathbf{X} = \sum_{i=1}^n(\mathbf{X}_i - \boldsymbol{\mu}_\mathbf{X})(\mathbf{X}_i - \boldsymbol{\mu}_\mathbf{X})^T/n$. Finally, because $\mathrm{vech}(\boldsymbol{\Gamma}_\pi\boldsymbol{\Omega}_\pi{\boldsymbol{\Gamma}_\pi}^T + \boldsymbol{\Gamma}_{0\pi}\boldsymbol{\Omega}_{0\pi}{\boldsymbol{\Gamma}_{0\pi}}^T)$ and $\mathrm{vech}(\mathbf{S}_\mathbf{X})$ are $\mathcal{O}(p^2)$-dimensional vectors, it takes $\mathcal{O}(p^2)$ flops to compute the difference between them. Hence, the total number of flops to compute the second line in $e_n^*(\boldsymbol{\zeta})$ is $\mathcal{O}(np^2 + p^3)$.



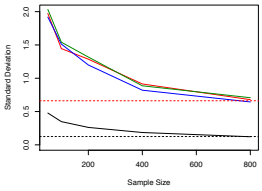
(a) Standard normal with $\pi = 0.1$



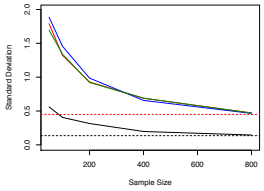
(b) Standard normal with $\pi = 0.25$



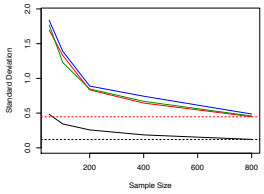
(c) Standard normal with $\pi = 0.75$



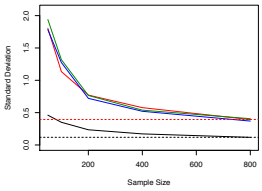
(d) t_4 with $\pi = 0.1$



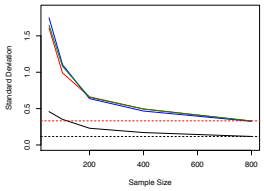
(e) t_4 with $\pi = 0.25$



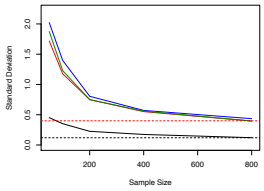
(f) t_4 with $\pi = 0.75$



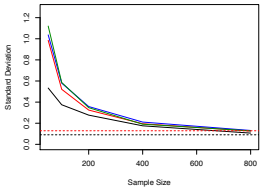
(g) Mixed normal with $\pi = 0.1$



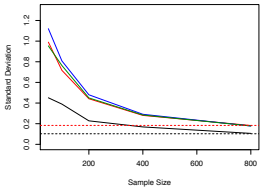
(h) Mixed normal with $\pi = 0.25$



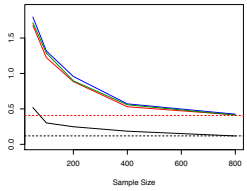
(i) Mixed normal with $\pi = 0.75$



(j) Exp(1) with $\pi = 0.1$



(k) Exp(1) with $\pi = 0.25$



(l) Exp(1) with $\pi = 0.75$

Fig 9: Comparison of the sample standard deviations. Red lines mark the ER estimator. Blue lines mark the sparse ER estimator. Green lines mark the boosting estimator. Black lines mark the EER estimator. The horizontal lines mark the asymptotic standard deviations of the ER estimator (the upper line in each panel) and the EER estimator (the lower line in each panel).

– Because $\boldsymbol{\mu}_{\mathbf{X}}$ is a p-dimensional vector and $\bar{\mathbf{X}}$ is also a p-dimensional vector, it takes $\mathcal{O}(p)$ flops to compute the difference between them. Hence, the total number of flops to compute the third line in $e_n^*(\boldsymbol{\zeta})$ is $\mathcal{O}(p)$.

To sum up, it takes $\mathcal{O}(np^2 + p^3)$ flops to compute $e_n^*(\boldsymbol{\zeta})$. Once we have $e_n^*(\boldsymbol{\zeta})$, it takes another $\mathcal{O}(p^2)$ flops to compute the objective function $e_n^*(\boldsymbol{\zeta})^T e_n^*(\boldsymbol{\zeta})$. So the total number of flops to compute the value of the objective function is T_f=$\mathcal{O}(np^2 + p^3)$. Therefore, the number of flops in each iteration of the Nelder-Mead algorithm is $\mathcal{O}(np^2 + p^3)$.

Step 2 : Compute the scale matrix

$$\hat{\boldsymbol{\Delta}} = \left[\frac{1}{n}\sum_{i=1}^{n} s(\mathbf{Z}_i; \psi(\hat{\boldsymbol{\zeta}}^*))s(\mathbf{Z}_i; \psi(\hat{\boldsymbol{\zeta}}^*))^T\right]^{-1},$$

where

$$s(\mathbf{Z}; \psi(\boldsymbol{\zeta})) = \begin{pmatrix} \mathbf{W}(Y - \mu_\tau - \mathbf{X}^T\boldsymbol{\Gamma}_\tau\boldsymbol{\eta}_\tau)\left|I(Y < \mu_\tau + \mathbf{X}^T\boldsymbol{\Gamma}_\tau\boldsymbol{\eta}_\tau) - \tau\right| \\ \text{vech}(\boldsymbol{\Gamma}_\tau\boldsymbol{\Omega}_\tau{\boldsymbol{\Gamma}_\tau}^T + \boldsymbol{\Gamma}_{0\tau}\boldsymbol{\Omega}_{0\tau}{\boldsymbol{\Gamma}_{0\tau}}^T) - \text{vech}\{(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})^T\} \\ \boldsymbol{\mu}_{\mathbf{X}} - \mathbf{X}. \end{pmatrix}$$

In this step, we firstly compute the matrix $\frac{1}{n}\sum_{i=1}^{n} s(\mathbf{Z}_i; \psi(\hat{\boldsymbol{\zeta}}^*))s(\mathbf{Z}_i; \psi(\hat{\boldsymbol{\zeta}}^*))^T$. Following similar calculations in Step 1, it takes $\mathcal{O}(p^3)$ flops to compute $s(\mathbf{Z}_i; \psi(\hat{\boldsymbol{\zeta}}^*))$. Upon we get $s(\mathbf{Z}_i; \psi(\hat{\boldsymbol{\zeta}}^*))$, it takes another $\mathcal{O}(p^4)$ flops to compute the multiplication $s(\mathbf{Z}_i; \psi(\hat{\boldsymbol{\zeta}}^*))s(\mathbf{Z}_i; \psi(\hat{\boldsymbol{\zeta}}^*))^T$. We need to do this multiplication for each sample and then take the average, then the number of flops to get the matrix $\frac{1}{n}\sum_{i=1}^{n} s(\mathbf{Z}_i; \psi(\hat{\boldsymbol{\zeta}}^*))s(\mathbf{Z}_i; \psi(\hat{\boldsymbol{\zeta}}^*))^T$ is $\mathcal{O}(np^4)$. Afterwards, we need to solve for the inversion of the matrix. Matrix inversion takes $\mathcal{O}(m^3)$ flops for an m by m matrix. In our case, it takes $\mathcal{O}(p^6)$ flops for the matrix inversion. So the number of flops in this step is $\mathcal{O}(np^4+p^6)$.

Step 3 : Obtain the GMM estimator $\hat{\boldsymbol{\zeta}}$ by minimizing $e_n^*(\boldsymbol{\zeta})^T\hat{\boldsymbol{\Delta}}e_n^*(\boldsymbol{\zeta})$.
Similar as Step 1, we apply Nelder-Mead method to find the minimum of the objective function $e_n^*(\boldsymbol{\zeta})^T\hat{\boldsymbol{\Delta}}e_n^*(\boldsymbol{\zeta})$. It takes $\mathcal{O}(np^2 + p^3)$ to compute $e_n^*(\boldsymbol{\zeta})$. Once we get $e_n^*(\boldsymbol{\zeta})$, it takes another $\mathcal{O}(p^4)$ flops to compute the objective function $e_n^*(\boldsymbol{\zeta})^T\hat{\boldsymbol{\Delta}}e_n^*(\boldsymbol{\zeta})$. So the total number of flops to compute the value of the objective function is $T_f = \mathcal{O}(np^2 + p^4)$. Therefore, the number of flops in each iteration of the Nelder-Mead algorithm is $\mathcal{O}(np^2 + p^4)$.

9. Comparison Between EQR and EER

In this section, we compare the performance between the envelope quantile regression (EQR; [1]) model and the EER model with simulated data and S&P 500 data.

For the simulated data, we use same settings in Section 5 of the paper. Since the true underlying distributions are known, we are able to map expectiles to

quantiles under each distribution. For example, 0.19 quantile is identical to 0.10 expectile under the standard normal distribution $\epsilon \sim \mathcal{N}(0, 1)$. The mappings for the four error distributions considered in the simulation are shown in the following Table 5.

TABLE 5
The mappings between expectiles and quantiles under the four types of distributions.

(a) $\epsilon \sim \mathcal{N}(0, 1)$

Expectile levels π	0.10	0.25	0.50	0.75	0.90
Quantile levels α	0.19	0.33	0.50	0.67	0.81

(b) $\epsilon \sim \text{Exp}(1)$

Expectile levels π	0.10	0.25	0.50	0.75	0.90
Quantile levels α	0.34	0.48	0.63	0.77	0.87

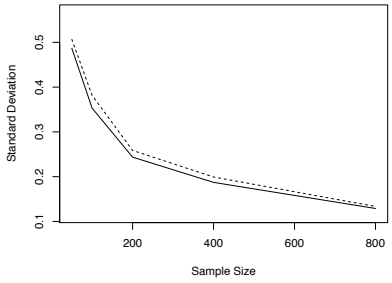
(c) $\epsilon \sim 0.9\mathcal{N}(0, 1) + 0.1\mathcal{N}(1, 5)$

Expectile levels π	0.10	0.25	0.50	0.75	0.90
Quantile levels α	0.19	0.34	0.52	0.70	0.84

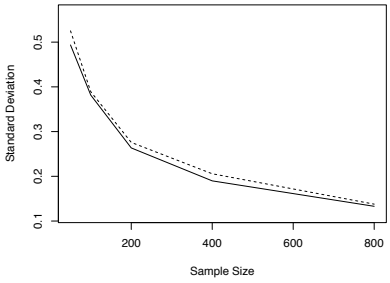
(d) $\epsilon \sim t_4$

Expectile levels π	0.10	0.25	0.50	0.75	0.90
Quantile levels α	0.16	0.30	0.50	0.70	0.84

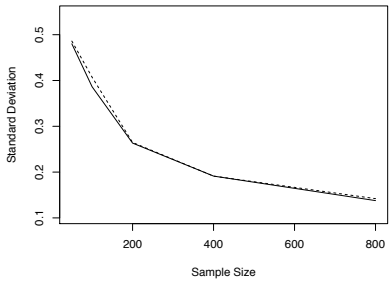
We repeated the simulation in Section 5 of the manuscript for the EQR model. Then we compared the sample standard deviations of the EER estimator and the EQR estimator, as well as the prediction performance. Note that the expectile levels investigated for the EER model were still 0.10, 0.25, 0.50, 0.75 and 0.90, while their corresponding quantile level mappings given in Table 5 were used for the EQR model. The results are summarized in Figure 10 and Tables 6.



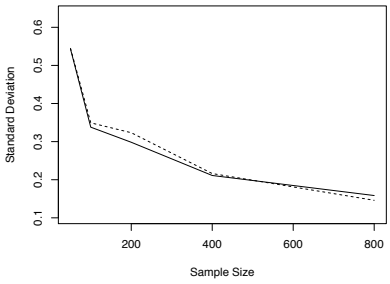
(a) Standard normal with $\pi = 0.5$



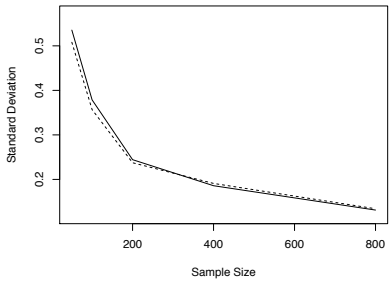
(b) Standard normal with $\pi = 0.9$



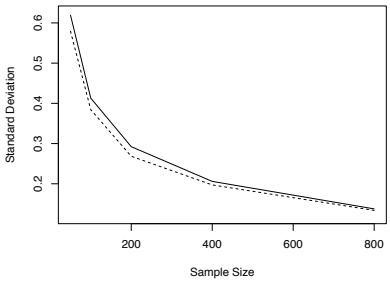
(c) Exp(1) with $\pi = 0.5$



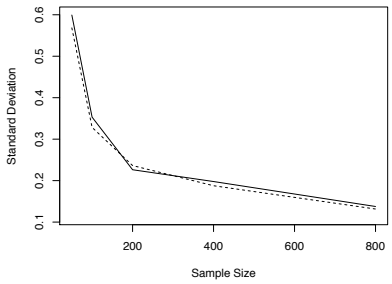
(d) Exp(1) with $\pi = 0.9$



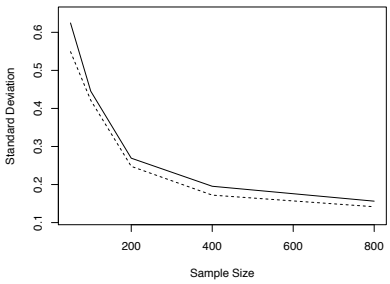
(e) Mixed normal with $\pi = 0.5$



(f) Mixed normal with $\pi = 0.9$



(g) t_4 with $\pi = 0.5$



(h) t_4 with $\pi = 0.9$

Fig 10: Sample standard deviations. Solid lines mark the EER estimator. Dashed lines mark the EQR estimator.

TABLE 6
The average RMSEs of the 300 replications under the EER model and EQR model with different error distributions.

(a) $\epsilon \sim \mathcal{N}(0,1)$

$\pi(\alpha)$	0.10 (0.19)	0.25 (0.33)	0.50 (0.50)	0.75 (0.67)	0.90 (0.81)
EER	1.05	0.94	0.93	0.98	1.12
EQR	1.06	1.00	0.98	1.01	1.09

(b) $\epsilon \sim \text{Exp}(1)$

$\pi(\alpha)$	0.10 (0.34)	0.25 (0.48)	0.50 (0.63)	0.75 (0.77)	0.90 (0.87)
EER	0.70	0.77	0.95	1.32	2.01
EQR	0.76	0.86	1.06	1.29	1.70

(c) $\epsilon \sim 0.9\mathcal{N}(0,1) + 0.1\mathcal{N}(1,5)$

$\pi(\alpha)$	0.10 (0.19)	0.25 (0.34)	0.50 (0.52)	0.75 (0.70)	0.90 (0.84)
EER	1.19	1.02	1.02	1.21	1.71
EQR	1.11	0.99	0.99	1.07	1.29

(d) $\epsilon \sim t_4$

$\pi(\alpha)$	0.10 (0.16)	0.25 (0.30)	0.50 (0.50)	0.75 (0.70)	0.90 (0.84)
EER	1.71	1.22	1.11	1.27	1.82
EQR	1.40	1.09	1.01	1.09	1.36

The estimation efficiency is similar for the EER estimator and the EQR estimator. The sample standard deviations of the EER estimators are very close to those of the EQR estimators, as indicated in Figure 10. Additionally, we observed that under the distributions with relatively smaller variance (standard normal and Exp(1)), the sample standard deviations of the EER estimators are slightly smaller than those of the EQR estimators. While under the distributions with relatively larger variance (mixed normal and t_4), the sample standard deviations of the EER estimators become slightly larger than those of the EQR estimators. The observation is consistent with the property that expectiles are more sensitive to extreme values. Under distributions with relatively larger variance, it is more likely to have extreme values in the data, which results in more variant EER estimators than the EQR estimators. Similar observation is shown in Table 6 as well. Under standard normal and Exp(1), the average RMSEs from the EER model are slightly smaller than those from the EQR model for most expectile levels. While under mixed normal and t_4, the average RMSEs from the EER model become larger than those from the EQR model.

For S&P 500 data, since we do not have direct knowledge on the underlying distribution, a grid of 23 levels (0.01, 0.02, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30,

0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 0.98 and 0.99) were investigated as quantiles under the EQR model and expectiles under the EER model. To compare prediction performance, we notice that the measure of prediction performance are different for the EQR model and the EER model: EQR model uses the quantile loss and EER model uses the expectile loss. Therefore we firstly computed the mean of the predicted quantile loss for each level under both the EQR model and the EER model. Among the 23 levels, the mean of predicted quantile loss under the EQR model ranges from 3.0×10^{-3} to 3.1×10^{-2} with an average of 2.0×10^{-2}. The mean of predicted quantile loss under the EER model ranges from 3.0×10^{-3} to 3.1×10^{-2} with an average of 2.1×10^{-2}. Boxplots of the predicted quantile loss under the two models are included in the left panel of Figure 11. Secondly we computed the mean of the predicted expectile loss under both models for each level, and the results are included in the boxplots in the right panel of Figure 11. Among the 23 levels, the mean of the predicted expectile loss under the EQR model ranges from 8.8×10^{-4} to 3.3×10^{-3} with an average of 2.5×10^{-3}. The mean of predicted expectile loss under the EER model ranges from 4.7×10^{-4} to 3.6×10^{-3} with an average of 2.3×10^{-3}. Based on the ranges and averages, we can not identify a statistically significant difference of the prediction performance between the EQR model and the EER model in this case.

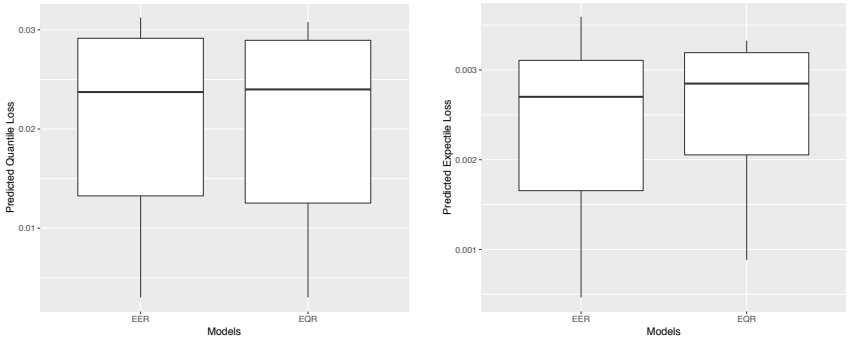


Fig 11: Boxplots of predicted quantile loss and expectile loss for the two models.

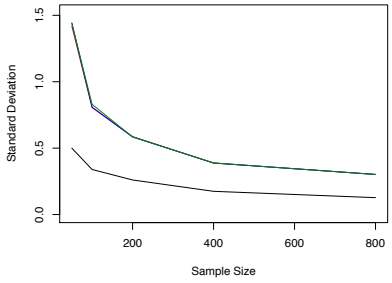
Similar to the relationship between the QR and the ER, the EQR model and the EER model have their unique advantages over each other, and neither approach is uniformly superior. We need to choose the appropriate model based on the goal and context of the problems. For example, if we want to evaluate the potential loss from a portfolio and we are strongly risk averse, then we may use the EER model because it is more sensitive to the extreme losses. If we hope to have a model that is easier to interpret, then we may want to use the EQR model.

10. Simulation Results for Sparse Expectile Regression Estimator with an Alternative Tuning Parameter

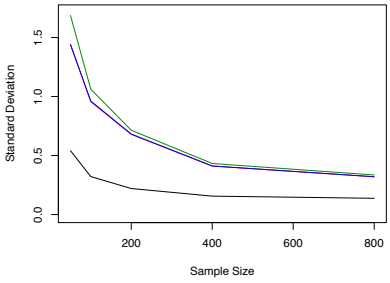
In the same setting as in Section 5 of the manuscript, we update the results of the sparse ER estimator with a different tuning parameter. The sparse ER estimator was computed by R package `SALES` [2]. It gives two choices of the tuning parameter λ: λ_{min}, which is the value of λ that minimizes the cross validation error, and λ_{1se}, which is the largest value of λ having its cross validation error within one standard error of the minimum cross validation error. The package takes λ_{1se} as the default value for the subsequent variable selection and parameter estimation, and the corresponding results are included in the manuscript. In this section, we update the results using λ_{min} as the tuning parameter. We included the ER estimator, EER estimator and the boosting estimator in all figures and tables for completeness. Note that the results for these estimators are unchanged.

The sample standard deviations are included in Figure 12. We do not observe a big difference in sample standard deviation between the sparse ER estimators using λ_{1se} (page 14 of the manuscript) and using λ_{min}. But it seems that the sparse ER estimator has a slightly smaller sample standard deviation with λ_{min}.

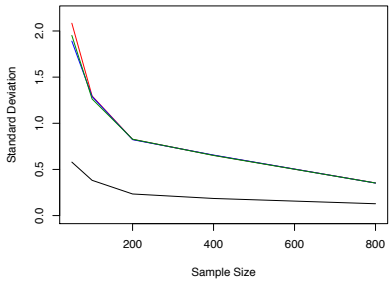
We also calculated the root mean squares errors (RMSE) of the sparse ER estimator using λ_{min} as the tuning parameter. The results are in Table 7. Compared to Table 1 of the manuscript (page 16), we notice that the performance of the sparse ER estimator gets better, and its RMSE is very close to the ER estimator and the boosting estimator. It seems that the default value λ_{1se} gives a model that is too parsimonious for this settings.



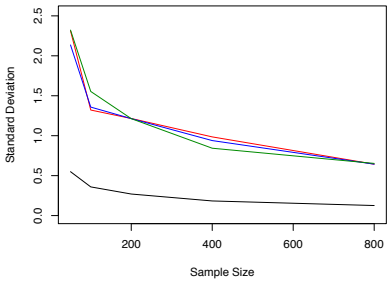
(a) Standard normal with $\pi = 0.5$



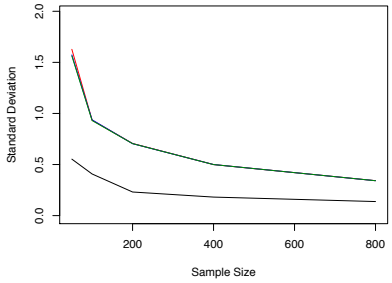
(b) Standard normal with $\pi = 0.9$



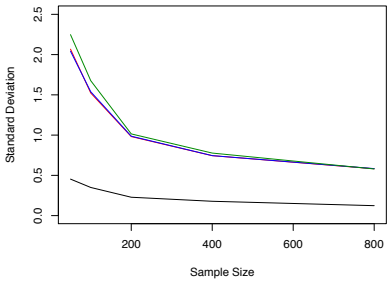
(c) t_4 with $\pi = 0.5$



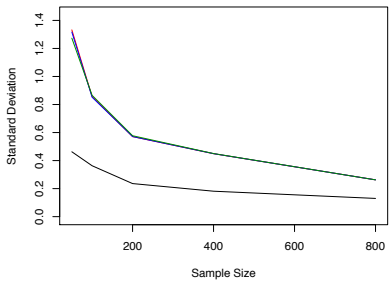
(d) t_4 with $\pi = 0.9$



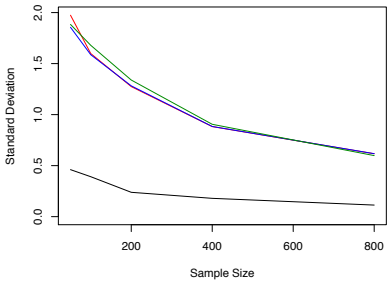
(e) Mixed normal with $\pi = 0.5$



(f) Mixed normal with $\pi = 0.9$



(g) Exp(1) with $\pi = 0.5$



(h) Exp(1) with $\pi = 0.9$

Fig 12: Comparison of the sample standard deviations. Red lines mark the ER estimator. Blue lines mark the sparse ER (with λ_{min} as the tuning parameter) estimator. Green lines mark the boosting estimator. Black lines mark the EER estimator.

TABLE 7
Comparison of the RMSEs, averaged over 300 *replications. Using* λ_{min} *as the selected value of* λ *for the sparse ER estimator.*

(a) $\epsilon \sim \mathcal{N}(0,1)$

	EER	ER	Boosting	Sparse ER
$\pi = 0.10$	1.04	1.85	1.86	1.86
$\pi = 0.25$	0.93	1.60	1.60	1.60
$\pi = 0.50$	0.90	1.52	1.52	1.52
$\pi = 0.75$	0.95	1.61	1.62	1.61
$\pi = 0.90$	1.10	1.87	1.91	1.88

(b) $\epsilon \sim t_4$

	EER	ER	Boosting	Sparse ER
$\pi = 0.10$	1.84	3.49	3.50	3.54
$\pi = 0.25$	1.28	2.46	2.47	2.47
$\pi = 0.50$	1.15	2.14	2.15	2.14
$\pi = 0.75$	1.31	2.44	2.45	2.44
$\pi = 0.90$	1.85	3.46	3.51	3.47

(c) $\epsilon \sim 0.9\mathcal{N}(0,1) + 0.1\mathcal{N}(1,5)$

	EER	ER	Boosting	Sparse ER
$\pi = 0.10$	1.20	2.21	2.22	2.23
$\pi = 0.25$	1.05	1.87	1.87	1.88
$\pi = 0.50$	1.05	1.86	1.87	1.86
$\pi = 0.75$	1.24	2.21	2.22	2.22
$\pi = 0.90$	1.75	3.14	3.18	3.16

(d) $\epsilon \sim \text{Exp}(1)$

	EER	ER	Boosting	Sparse ER
$\pi = 0.10$	0.70	0.76	0.79	0.75
$\pi = 0.25$	0.77	1.07	1.07	1.06
$\pi = 0.50$	0.96	1.54	1.54	1.54
$\pi = 0.75$	1.34	2.27	2.28	2.27
$\pi = 0.90$	1.99	3.37	3.40	3.39

References

[1] Ding, S., Su, Z., Zhu, G., and Wang, L. (2019). Envelope quantile regression. *Statistica Sinica*, To appear.

[2] Gu, Y. and Zou, H. (2016). Sales: Elastic net and (adaptive) lasso penalized sparse asymmetric least squares (sales) and coupled sparse asymmetric least squares (cosales) using coordinate descent and proximal gradient algorithms. *R package version 1.0.0*.

[3] Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics 4*, 2111–2245.

[4] Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica* **55**, 4, 819–47.

[5] Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica: Journal of the Econometric Society* **57**, 5, 1027–1057.

[6] Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *Journal of the American Statistical Association* **81**, 393, 142–149.

[7] Singer, S. and Singer, S. (1999). Complexity analysis of nelder-mead search iterations. In *Proceedings of the 1. Conference on Applied Mathematics and Computation, Dubrovnik, Croatia*. PMF–Matematički odjel, Zagreb, 185–196.

[8] Van der Vaart, A. W. (1998). *Asymptotic statistics*. Vol. **3**. Cambridge university press.