Highlights

Hierarchical Composite Quantile Regression with Adaptive Lasso

Hou Jian, Meng Tan, Maozai Tian

- This study introduces the HCQR model, which integrates adaptive Lasso regularization to effectively handle high-dimensional hierarchical data, demonstrating robustness and superior variable selection capabilities.
- Simulation results show that HCQR significantly outperforms traditional methods in terms of variable selection accuracy, error control, and robustness to outliers, making it well-suited for complex data scenarios.
- In analyzing Parkinson's disease data, HCQR accurately identifies key variables and provides interpretable results, showcasing its potential for addressing real-world challenges in hierarchical data analysis.

Hierarchical Composite Quantile Regression with Adaptive Lasso

Hou Jian[a], Meng Tan[a], Maozai Tian[a]

[a] *Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing, 100872, China*

Abstract

Hierarchical (multi-level) data commonly exhibit heavy-tailed noise and high-dimensional fixed-effect structures. To address these challenges, we propose a Hierarchical Composite Quantile Regression (HCQR) framework regularized with an adaptive Lasso penalty. Our approach integrates the statistical efficiency of composite quantile loss for robustness and the adaptive Lasso for sparse variable selection. We develop an efficient estimation procedure based on an Expectation-Maximization (EM) algorithm, with an embedded ADMM routine to solve the penalized M-step. This allows for simultaneous estimation of fixed effects, variance components, and prediction of random effects, while the quantile-based loss function ensures robustness against non-Gaussian error distributions. Simulation studies under various error distributions, including normal, t, Cauchy, and mixture-normal, confirm that HCQR achieves near-oracle performance in variable selection and prediction, even when the number of covariates exceeds the sample size. An application to a Parkinson's disease tele-monitoring dataset further illustrates how HCQR effectively uncovers clinically meaningful voice biomarkers while accounting for patient-specific disease progression.

Preprint submitted to JSPI *June 21, 2025*

Keywords: hierarchical model, composite quantile regression, adaptive lasso

PACS: 0000, 1111

2000 MSC: 0000, 1111

---

1. Introduction

In the realm of high-dimensional data analysis, researchers face not only the challenge of managing a vast number of variables but also addressing the structural complexity inherent in such data[1]. This complexity becomes particularly pronounced in longitudinal data, which is commonly observed in disciplines such as social sciences, medical research, and economics. These datasets often exhibit stratified or grouped characteristics, leading to variability among individuals. Ignoring these correlations can result in biased or inaccurate regression outcomes[2, 3].

To address this issue, hierarchical models, also known as multi-level models, have emerged as a powerful analytical tool[4, 5]. These models account for variability across different levels by incorporating random effects, thereby significantly improving both the accuracy and interpretability of regression analyses. Unlike mean-based mixed models, quantile regression reveals heterogeneous tail behaviour caused by group structure and heavy-tailed errors. When coupled with quantile regression techniques, they offer additional advantages, such as robust handling of outliers and the ability to provide a more nuanced understanding of data distribution.

Quantile regression[6], with its capacity to model the conditional quantiles of a response variable, has become increasingly popular across various

research domains. Composite quantile regression (CQR)[7] extends this approach by simultaneously modeling multiple quantiles, capturing the dynamics of the response variable across different sections of the data distribution. This technique mitigates the risk of insufficient data representation at extreme quantiles, thereby enabling a more comprehensive regression analysis. The integration of CQR with hierarchical models is particularly advantageous for datasets exhibiting hierarchical heterogeneity. It allows researchers to uncover intricate relationships among individuals at different quantiles, providing a robust framework for analyzing complex data structures.

The hierarchical linear model was first extended to hierarchical quantile regression by Tian and Chen (2006)[8] using the EQ algorithm, enabling hierarchical models to effectively handle outliers. Subsequent advances included the development by Luo (2012)[9] and Yu (2015)[10] of a Bayesian framework and Chen's (2014)[11] extension of quantile regression to composite quantile regression. However, estimation and variable selection in high-dimensional settings remain inadequately explored, presenting a critical gap that this study aims to address.

High-dimensional hierarchical data introduce unique challenges. The number of fixed effects can grow exponentially with the inclusion of additional levels and covariates. For instance, in a two-level hierarchical model with 10 fixed-effect covariates at each level, the total number of covariates reaches 20, and with three levels, it can escalate to 30 or more. To manage this complexity, we employ adaptive Lasso penalization[12], which is well-suited for variable selection in high-dimensional stratified composite quantile regression models.

Recognizing the inefficiencies of the EM algorithm in high-dimensional contexts, we propose an estimation framework inspired by the algorithm developed by Gu et al. (2018)[13]. To validate our approach, we conduct simulation studies to assess the model's robustness using heavy-tailed error distributions. Metrics such as the true positive rate (TP) and the false discovery rate (FDR) are used to evaluate variable selection performance, with the results compared against the Oracle estimator and existing methods[14]. Our findings demonstrate the superior performance of the proposed method in the handling of complex high-dimensional data structures.

The remainder of this manuscript is structured as follows. The Section 2 describes the stratified high-dimensional composite quantile regression model, including its key notations and assumptions. The Section 3 elaborates on the implementation of the proposed algorithm. The Section 4 provides a validation of the approach through simulation experiments, highlighting its efficacy in handling complex datasets. Finally, the Section 5 applies the proposed method to real datasets, providing a detailed analysis of its practical utility. The details of the proof are given in the Appendix.

2. Model and Methodology

2.1. Hierarchical Linear Model

We begin by describing a two-level hierarchical linear model. Consider a dataset with J groups, where each group j contains n_j elements. Let $Y_{ij} \in \mathbb{R}$ denote a real-valued response variable for the i-th element of the j-th group. For each element, we define $\mathbf{X}_{ij}$ as a row vector of dimension $1 \times L$ (including

the intercept). The first-level model is specified as:

$$Y_{ij} = \mathbf{X}_{ij}\,\boldsymbol{\beta}_j + \varepsilon_{ij}, \tag{1}$$

where $\boldsymbol{\beta}_j \in \mathbb{R}^{L\times 1}$ is a group-specific coefficient vector. The term ε_{ij} represents an i.i.d. random error, and we assume that ε_{ij} and $\mathbf{X}_{ij}$ are independent of each other.

The second-level model captures the between-group variation in the coefficients:

$$\boldsymbol{\beta}_j = \mathbf{Z}_j\,\boldsymbol{\gamma} + \mathbf{v}_j, \quad \mathbf{v}_j \sim N(\mathbf{0}, \mathbf{T}), \tag{2}$$

where $\boldsymbol{\gamma}$ is the vector of fixed effect parameters. For each group j, the matrix $\mathbf{Z}_j$ belongs to $\mathbb{R}^{L\times F}$, and $F = \sum_{l=1}^{L}\big(f_l + 1\big)$. Here, f_l is the number of group-level covariates (excluding intercept) for the l-th coefficient, so each coefficient at the first level may include its own intercept and additional group-level covariates. The vector $\mathbf{v}_j$ represents the random effects for group j, with $\mathbf{T}$ as its covariance matrix.

To provide a clearer specification of the group-level covariates, we can break down $\mathbf{Z}_j$ into block-diagonal form. For each coefficient l (from 1 to L) in the first-level model, we define the corresponding row vector of group-level covariates as $\mathbf{Z}_{jl} = [1, Z_{jl1}, Z_{jl2}, \ldots, Z_{jlf_l}]$, which has dimension $1 \times (f_l + 1)$. Then $\mathbf{Z}_j$ can be structured as a block-diagonal matrix:

$$\mathbf{Z}_j = \begin{bmatrix} \mathbf{Z}_{j1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_{j2} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Z}_{j3} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Z}_{jL} \end{bmatrix}, \tag{3}$$

where each $\mathbf{0}$ is a zero matrix of appropriate dimensions. This block-diagonal structure ensures that the group-level covariates for each first-level coefficient are properly aligned.

Correspondingly, the fixed effect parameter vector $\boldsymbol{\gamma}$ is organized as:

$$\boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \\ \vdots \\ \boldsymbol{\gamma}_L \end{bmatrix}, \tag{4}$$

where each $\boldsymbol{\gamma}_l$ is a column vector of length $(f_l + 1)$, containing the fixed effect parameters for the l-th coefficient (including the intercept). The total dimension of $\boldsymbol{\gamma}$ is F. Because $\mathbf{X}_{ij}$ is a $1 \times L$ row vector and $\mathbf{Z}_j$ is $L \times F$, the product $\mathbf{X}_{ij}\mathbf{Z}_j$ is a $1 \times F$ design row that premultiplies $\boldsymbol{\gamma}$. We assume that $\mathbf{Z}_j$ and ε_{ij} are mutually independent. By combining the first- and second-level models, we obtain

$$Y_{ij} = \mathbf{X}_{ij}\mathbf{Z}_j\boldsymbol{\gamma} + \mathbf{X}_{ij}\mathbf{v}_j + \varepsilon_{ij}. \tag{5}$$

This formulation separates the fixed effects ($\mathbf{X}_{ij}\mathbf{Z}_j\boldsymbol{\gamma}$) from the random effects ($\mathbf{X}_{ij}\mathbf{v}_j$) in the hierarchical model. In addition, the definition of the total sample size is $N = \sum_{j=1}^{J} n_j$.

2.2. Quantile Regression Framework

To establish a quantile regression framework for hierarchical data, let $F(y \mid \mathbf{x}, \mathbf{z})$ denote the conditional distribution function of the response variable Y given covariates $(\mathbf{X}, \mathbf{Z}) = (\mathbf{x}, \mathbf{z})$. When this distribution function is strictly non-decreasing with respect to y and continuous at $(\mathbf{x}, \mathbf{z})$, the τ-th

conditional quantile of Y is defined as

$$q_\tau(\mathbf{x}, \mathbf{z}) = \inf\Big\{ y \in \mathbb{R} : F\big(y \mid \mathbf{x}, \mathbf{z}\big) \geq \tau \Big\}, \quad 0 < \tau < 1. \tag{6}$$

In hierarchical settings, modeling a single quantile may not capture the full conditional distribution of the response variable. Following the composite quantile regression (CQR) framework proposed by [7], we construct an objective function using a series of quantiles $0 < \tau_1 < \tau_2 < \cdots < \tau_K < 1$. Compared to single-quantile approaches, the CQR framework improves efficiency by aggregating information across multiple quantiles, particularly when the error distribution is non-Gaussian [15].

For hierarchical data, the CQR estimator of $\boldsymbol{\gamma}$ is defined as

$$\begin{aligned} &\Big(\hat{b}_{\tau_1}, \ldots, \hat{b}_{\tau_K}, \hat{\boldsymbol{\gamma}}^{\mathrm{CQR}}\Big) \\ &\quad = \underset{b_{\tau_1}, \ldots, b_{\tau_K}, \boldsymbol{\gamma}}{\arg\min} \sum_{k=1}^{K} \Bigg\{ \sum_{j=1}^{J} \sum_{i=1}^{n_j} \rho_{\tau_k} \Big[Y_{ij} - b_{\tau_k} - \mathbf{X}_{ij} \mathbf{Z}_j \boldsymbol{\gamma} - \mathbf{X}_{ij} \mathbf{v}_j \Big] \Bigg\}, \end{aligned} \tag{7}$$

where $\rho_{\tau_k}(t) = t\,(\tau_k - \mathbf{I}(t < 0))$ is the quantile loss function, and b_{τ_k} represents the $100\tau_k\%$ quantile of the error distribution. Typically, we use equally spaced quantiles $\tau_k = \frac{k}{K+1}$ for $k = 1, \ldots, K$. The inclusion of multiple quantiles in the objective function allows us to capture different aspects of the conditional distribution while maintaining the efficiency of a unified estimation procedure.

For high-dimensional settings (where the number of covariates may be large relative to the sample size), we employ the adaptive Lasso-penalized

CQR estimator (ACQR) for hierarchical data:

$$\begin{aligned}&\left(\hat{b}_{\tau_1},\ldots,\hat{b}_{\tau_K},\hat{\boldsymbol{\gamma}}^{\mathrm{ACQR}}\right)\\&\quad=\underset{b_{\tau_1},\ldots,b_{\tau_K},\boldsymbol{\gamma}}{\arg\min}\sum_{k=1}^{K}\Big\{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\rho_{\tau_k}\Big[Y_{ij}-b_{\tau_k}-\mathbf{X}_{ij}\mathbf{Z}_j\boldsymbol{\gamma}-\mathbf{X}_{ij}\mathbf{v}_j\Big]\Big\}\\&\qquad+\lambda\sum_{p=1}^{F}\frac{|\gamma_p|}{\left|\hat{\gamma}_p^{\mathrm{CQR}}\right|},\end{aligned}\tag{8}$$

where λ is a tuning parameter controlling the degree of regularization, and $\hat{\gamma}_p^{\mathrm{CQR}}$ is the initial CQR estimator (without penalization). The adaptive weights $\frac{1}{|\hat{\gamma}_p^{\mathrm{CQR}}|}$ ensure that larger coefficients receive smaller penalties, thus reducing bias in the estimation of significant effects.

2.3. Asymptotic Properties

In this section, we establish the asymptotic properties of the proposed HCQR estimator in the high-dimensional setting, where the number of fixed-effect covariates F can grow with and be much larger than the total sample size N (i.e., $F \gg N$). Our theoretical guarantees are built upon the assumption that the true underlying model is sparse. We begin by introducing the necessary regularity conditions for this high-dimensional framework.

Let $\boldsymbol{\gamma}^*$ be the true vector of fixed effects, and let $\mathcal{A} = \{p : \gamma_p^* \neq 0\}$ denote the true active set of indices with sparsity $s = |\mathcal{A}|$. We assume s is small relative to N.

Condition 1 (Sparsity)**.** *The number of non-zero coefficients $s = s_N$ grows slowly enough with the sample size N, such that $s\log(F) = o(\sqrt{N})$.*

Condition 2 (Restricted Eigenvalue Condition)**.** *Let $\mathbf{X}^* \in \mathbb{R}^{N\times F}$ be the row-stacked block matrix with j-th block $\mathbf{X}_j\mathbf{Z}_j$[16]. For some constant $c_0 > 0$, the*

design matrix $\mathbf{X}^*$ *satisfies the RE condition if for all vectors* $\boldsymbol{\delta} \in \mathbb{R}^F$ *that satisfy* $\|\boldsymbol{\delta}_{\mathcal{A}^c}\|_1 \leq c_0\|\boldsymbol{\delta}_{\mathcal{A}}\|_1$*, we have*

$$\kappa(c_0, s) = \min_{\boldsymbol{\delta}\neq\mathbf{0}, \|\boldsymbol{\delta}_{\mathcal{A}^c}\|_1 \leq c_0\|\boldsymbol{\delta}_{\mathcal{A}}\|_1} \frac{\|\mathbf{X}^*\boldsymbol{\delta}\|_2}{\sqrt{N}\|\boldsymbol{\delta}_{\mathcal{A}}\|_2} > 0.$$

Condition 3 (Error Distribution)**.** *The composite random error term* $\varepsilon = Y - \mathbf{X}^*\boldsymbol{\gamma}^*$ *has a cumulative distribution function* $F(\cdot)$ *and a probability density function* $f(\cdot)$ *that is continuous and strictly positive at the true quantiles* $b^*_{\tau_k}$ *for all* $k = 1, \ldots, K$*.*

Condition 4 (Initial Estimator Consistency)**.** *An initial estimator* $\hat{\boldsymbol{\gamma}}^{init}$ *(e.g., a Lasso-penalized CQR) is used to compute the adaptive weights. This estimator must satisfy* $\|\hat{\boldsymbol{\gamma}}^{init} - \boldsymbol{\gamma}^*\|_1 = O_p(s\sqrt{\log(F)/N})$*.*

Remark 1. We note that Condition 1, $s\log(F) = o(\sqrt{N})$, is a relatively strong sparsity assumption. Some literature in high-dimensional analysis employs weaker conditions, such as $s^2\log(F) = o(N)$. The stronger condition adopted here is sufficient to rigorously establish our subsequent asymptotic results, particularly in managing the complexities arising from the composite quantile framework in a high-dimensional setting. For further discussion on related conditions, see, for example, the work by Belloni, Chernozhukov, and co-authors on high-dimensional inference[17, 18].

With these conditions, we can now state the main theoretical result of this paper, which establishes the oracle properties of our proposed estimator. To define the asymptotic variance, we first characterize the behavior of a hypothetical oracle estimator.

Lemma 1. *Consider a hypothetical oracle estimator* $\hat{\boldsymbol{\gamma}}_{\mathcal{A}}^{\text{oracle}}$ *obtained by applying the composite quantile regression only to the true active set of covariates* $\mathcal{A}$*. If we assume that the* $s \times s$ *design sub-matrix corresponding to* $\mathcal{A}$ *is well-behaved such that* $\frac{1}{N}\sum_{j,i}(\mathbf{X}^*_{\mathcal{A},ij})^\top \mathbf{X}^*_{\mathcal{A},ij} \to \mathbf{C}_{\mathcal{A}\mathcal{A}}$ *for a positive definite matrix* $\mathbf{C}_{\mathcal{A}\mathcal{A}}$*, then this oracle estimator is consistent and asymptotically normal:*

$$\sqrt{N}\big(\hat{\boldsymbol{\gamma}}_{\mathcal{A}}^{\text{oracle}} - \boldsymbol{\gamma}^*_{\mathcal{A}}\big) \xrightarrow{d} N\big(\mathbf{0}, \boldsymbol{\Sigma}_{\text{CQR\_oracle}}\big),$$

where the oracle covariance matrix is

$$\boldsymbol{\Sigma}_{\text{CQR\_oracle}} = \mathbf{C}^{-1}_{\mathcal{A}\mathcal{A}} \frac{\sum_{k=1}^{K}\sum_{k'=1}^{K}\left(\min(\tau_k, \tau_{k'}) - \tau_k\tau_{k'}\right)}{\left(\sum_{k=1}^{K} f(b^*_{\tau_k})\right)^2}.$$

The following theorem shows that our practical, data-driven HCQR estimator can asymptotically achieve the same performance as this idealized oracle estimator.

Theorem 1 (Oracle Properties)**.** *Assume Conditions 1-4 hold. If the tuning parameter* λ_N *for the adaptive penalty satisfies* $\lambda_N \to 0$ *and* $\sqrt{N}\lambda_N \to \infty$ *as* $N \to \infty$*, then the HCQR estimator* $\hat{\boldsymbol{\gamma}}^{\text{ACQR}}$ *possesses the following oracle properties:*

(i) ***Sparsistency:*** *With probability approaching one, the estimator correctly identifies the set of true non-zero coefficients:*

$$\Pr\Big(\{p : \hat{\gamma}_p^{\text{ACQR}} \neq 0\} = \mathcal{A}\Big) \to 1.$$

(ii) ***Asymptotic Normality:*** *The estimators for the non-zero coefficients converge to their true values at the standard parametric rate and have*

the same asymptotic normal distribution as the oracle estimator described in Lemma 1:

$$\sqrt{N}\big(\hat{\boldsymbol{\gamma}}_{\mathcal{A}}^{\mathrm{ACQR}} - \boldsymbol{\gamma}_{\mathcal{A}}^{*}\big) \xrightarrow{d} N\big(\mathbf{0}, \boldsymbol{\Sigma}_{\mathrm{CQR\_oracle}}\big).$$

The main theorem establishes that the HCQR estimator is not only consistent in variable selection but also achieves optimal estimation efficiency. Several important consequences follow directly from this powerful result.

Corollary 1. *Under the conditions of Theorem 1, the estimators for the active set coefficients and the quantile-specific intercepts are $\sqrt{N}$-consistent. Specifically,*

$$\big\|\hat{\boldsymbol{\gamma}}_{\mathcal{A}}^{\mathrm{ACQR}} - \boldsymbol{\gamma}_{\mathcal{A}}^{*}\big\|_2 = O_p\big(N^{-1/2}\big),$$

and for each $k = 1, \ldots, K$,

$$\hat{b}_{\tau_k}^{\mathrm{ACQR}} - b_{\tau_k}^{*} = O_p\big(N^{-1/2}\big).$$

Corollary 2. *Under the conditions of Theorem 1, for any coefficient in the true active set, $p \in \mathcal{A}$, its estimator $\hat{\gamma}_p^{\mathrm{ACQR}}$ is asymptotically normal:*

$$\sqrt{N}\left(\hat{\gamma}_p^{\mathrm{ACQR}} - \gamma_p^{*}\right) \xrightarrow{d} N\left(0, \sigma_p^2\right),$$

where σ_p^2 is the p-th diagonal element of the oracle asymptotic covariance matrix $\boldsymbol{\Sigma}_{\mathrm{CQR\_oracle}}$ defined in Lemma 1.

3. Estimation Method

The estimation of the proposed Hierarchical Composite Quantile Regression (HCQR) model presents a dual challenge: the robust estimation of fixed

effects $\boldsymbol{\gamma}$ and the estimation of variance components $(\mathbf{T}, \sigma^2)$ associated with the unobservable random effects. To address this, we develop an efficient procedure based on the Expectation-Maximization (EM) algorithm, which is naturally suited for models with latent variables like random effects $\mathbf{v}_j$.

It is crucial to distinguish the statistical properties of our model from the specifics of the estimation algorithm. The core of the HCQR model lies in the composite quantile loss function, which ensures that the estimation of the fixed effects $\boldsymbol{\gamma}$ is robust against heavy-tailed errors and outliers, without requiring a specific distributional assumption for the error term ε_{ij}. However, the EM algorithm, being a likelihood-based iterative method, requires a complete data likelihood specification. This necessitates a working distributional assumption to handle the latent variables.

To this end, and to provide a formal likelihood interpretation for our framework, we can connect quantile regression to the Asymmetric Laplace Distribution (ALD) [6, 10]. Minimizing the quantile loss $\rho_{\tau_k}(u)$ is equivalent to maximizing the likelihood of an ALD. Consequently, our model can be viewed within a likelihood-based framework where the error term follows an ALD. For the random effects $\mathbf{v}_j$, we adopt the standard and convenient assumption of a multivariate normal distribution. This hybrid approach allows us to embed the robust quantile objective within a tractable EM framework for estimating the variance components.

3.1. EM Algorithm

We proceed by treating the random effects $\mathbf{v}_j$ as missing data and applying the EM algorithm. The procedure iterates between an E-step, which computes the conditional expectation of functions of $\mathbf{v}_j$ given the observed

data, and an M-step, which updates the model parameters.

For the purpose of deriving the EM algorithm, we establish a working model where the random effects are normally distributed, $\mathbf{v}_j \sim N(\mathbf{0}, \mathbf{T})$, and are independent of the within-group errors ε_{ij}. To embed quantile loss into an EM framework we follow [10] and treat ε_{ij} as $\text{ALD}(0, \sigma^2, \tau_k)$ in the M-step, while using a normal working assumption $\varepsilon_{ij} \sim N(0, \sigma^2)$ in the E-step. This hybrid specification leaves the consistency of $\boldsymbol{\gamma}$ unaffected[19].

E-step: In the E-step, we compute the conditional distribution of the random effects $\mathbf{v}_j$ given the observed data $\mathbf{Y}_j$ and the parameter estimates from the previous iteration. Based on the working normality assumption, the joint distribution is:

$$\begin{pmatrix} \mathbf{Y}_j \\ \mathbf{v}_j \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{X}_j \mathbf{Z}_j \boldsymbol{\gamma} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{X}_j \mathbf{T} \mathbf{X}_j^\top + \sigma^2 \mathbf{I} & \mathbf{X}_j \mathbf{T} \\ \mathbf{T} \mathbf{X}_j^\top & \mathbf{T} \end{pmatrix} \right),$$

where $\mathbf{I}$ is the identity matrix of appropriate dimension. Using standard properties of multivariate normal distributions, the conditional distribution of $\mathbf{v}_j$ given $\mathbf{Y}_j$ is also normal, $\mathbf{v}_j \mid \mathbf{Y}_j, \boldsymbol{\gamma}, \mathbf{T}, \sigma^2 \sim N(\mathbf{v}_j^*, \mathbf{T}^*)$, with conditional mean and covariance:

$$\begin{aligned} \mathbf{T}^* &= \sigma^2 \big(\mathbf{X}_j^\top \mathbf{X}_j + \sigma^2 \mathbf{T}^{-1} \big)^{-1}, \\ \mathbf{v}_j^* &= \sigma^{-2}\, \mathbf{T}^* \mathbf{X}_j^\top \big(\mathbf{Y}_j - \mathbf{X}_j \mathbf{Z}_j \boldsymbol{\gamma} \big). \end{aligned}$$

The E-step involves computing the expected values of the sufficient statistics for the complete-data log-likelihood. Specifically, we update the variance

components for the next M-step using these expectations:

$$\hat{\mathbf{T}}^{(t+1)} = \frac{1}{J}\sum_{j=1}^{J}\left(\mathbf{v}_j^*\mathbf{v}_j^{*\top} + \mathbf{T}^*\right),$$

$$\begin{aligned}\hat{\sigma}^{2(t+1)} &= \frac{1}{N}\sum_{j=1}^{J} E\left[\sum_{i=1}^{n_j}(Y_{ij} - \mathbf{X}_{ij}\mathbf{Z}_j\boldsymbol{\gamma} - \mathbf{X}_{ij}\mathbf{v}_j)^2 \mid \mathbf{Y}_j, \boldsymbol{\theta}^{(t)}\right] \\ &= \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left[(Y_{ij} - \mathbf{X}_{ij}\mathbf{Z}_j\boldsymbol{\gamma} - \mathbf{X}_{ij}\mathbf{v}_j^*)^2 + \operatorname{tr}(\mathbf{X}_{ij}\mathbf{T}^*\mathbf{X}_{ij}^\top)\right].\end{aligned}$$

M-step: In the M-step, we update the fixed effects $\boldsymbol{\gamma}$ and the quantile-specific intercepts b_{τ_k}. Here, we depart from the working normality assumption and return to the robust composite quantile regression objective. This ensures that the estimation of our primary parameters of interest is not unduly influenced by the auxiliary assumptions made in the E-step. The parameters are updated by solving the following adaptive Lasso-penalized optimization problem:

$$\begin{aligned}(\hat{b}_{\tau_1}, \ldots, \hat{b}_{\tau_K}, \hat{\boldsymbol{\gamma}}^{\mathrm{ACQR}}) = \underset{b_{\tau_1},\ldots,b_{\tau_K},\boldsymbol{\gamma}}{\arg\min} \sum_{k=1}^{K}\left\{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\rho_{\tau_k}\left[Y_{ij} - b_{\tau_k} - \mathbf{X}_{ij}\mathbf{Z}_j\boldsymbol{\gamma} - \mathbf{X}_{ij}\mathbf{v}_j^*\right]\right\} \\ + \lambda\sum_{p=1}^{F}\frac{|\gamma_p|}{|\hat{\gamma}_p^{\mathrm{CQR}}|}.\end{aligned} \tag{9}$$

In this formulation, the unobserved random effects $\mathbf{v}_j$ are replaced by their conditional expectations $\mathbf{v}_j^*$ computed in the E-step. This hybrid approach combines the algorithmic stability of the EM framework for variance components with the statistical robustness of quantile regression for the fixed effects. The EM algorithm iterates between the E-step and M-step until the change in parameter estimates falls below a predefined tolerance threshold.

3.2. Optimization within the M-step

The objective function in the M-step (9) is non-differentiable due to the quantile loss and the L_1 penalty, making direct optimization challenging. We employ the Alternating Direction Method of Multipliers (ADMM), an efficient algorithm for such composite convex optimization problems, particularly in high-dimensional settings.

We implement our proposed approach through a proximal-ADMM iteration scheme. Let $\boldsymbol{\theta} = (b_{\tau_1}, \ldots, b_{\tau_K}, \gamma_1, \ldots, \gamma_F)^\top$ denote the parameter vector, $\mathbf{r} \in \mathbb{R}^{NK}$ the stacked residuals, $\mathbf{u} \in \mathbb{R}^{NK}$ the scaled dual variables, and $w_p = 1/|\hat{\gamma}_p^{\mathrm{CQR}}|$ the adaptive Lasso weights. With $\tilde{\mathbf{X}}^* = [(\mathbf{1}_K \otimes \mathbf{I}_N) \;\; \mathbf{X}_\gamma^*]$ denoting the block-stacked design that contains the K intercept columns and the F slope columns, we formulate the augmented Lagrangian as:

$$\mathcal{L}_\rho(\boldsymbol{\theta}, \mathbf{r}, \mathbf{u}) = \sum_{s=1}^{NK} \rho_{\tau(s)}(r_s) + \lambda \sum_{p=1}^{F} w_p |\gamma_p| + \frac{\rho}{2} \left\| \tilde{\mathbf{Y}} - \mathbf{r} - \tilde{\mathbf{X}}^* \boldsymbol{\theta} + \mathbf{u} \right\|_2^2 - \frac{\rho}{2} \|\mathbf{u}\|_2^2, \quad (10)$$

where $\rho > 0$ is the penalty parameter. At each iteration $t = 0, 1, \ldots,$ our algorithm performs three sequential proximal steps that efficiently decompose the optimization problem.

The first step updates the residuals through coordinate-wise soft-thresholding:

$$r_s^{(t+1)} = S_{1/\rho}^{\tau(s)} \left(\tilde{Y}_s - (\tilde{\mathbf{X}}^* \boldsymbol{\theta}^{(t)})_s + u_s^{(t)} \right), \quad s = 1, \ldots, NK, \quad (11)$$

where the asymmetric soft-threshold operator for quantile loss is defined as $S_a^\tau(z) = \max\{z - a\tau, 0\} - \max\{-z - a(1-\tau), 0\}$. This step has computational complexity $O(NK)$ and is embarrassingly parallel[20].

The second step simultaneously updates the intercept and slope param-

eters through a block-coordinate approach:

$$\mathbf{b}^{(t+1)} = (\rho \mathbf{I}_K)^{-1} \left[\mathbf{1}_K \otimes \mathbf{I}_N\right]^\top \left(\tilde{\mathbf{Y}} - \mathbf{r}^{(t+1)} + \mathbf{u}^{(t)} - \mathbf{X}_\gamma^* \boldsymbol{\gamma}^{(t)}\right), \tag{12a}$$

$$\boldsymbol{\gamma}^{(t+1)} = \mathcal{S}_{\lambda \mathbf{w}/\rho}\left(\mathbf{z}^{(t)}\right), \tag{12b}$$

$$\mathbf{z}^{(t)} = \left(\mathbf{X}_\gamma^{*\top} \mathbf{X}_\gamma^* + \rho \mathbf{I}_F\right)^{-1} \mathbf{X}_\gamma^{*\top} \left(\tilde{\mathbf{Y}} - \mathbf{r}^{(t+1)} + \mathbf{u}^{(t)} - [\mathbf{1}_K \otimes \mathbf{I}_N]\mathbf{b}^{(t+1)}\right), \tag{12c}$$

where $\mathcal{S}_{\boldsymbol{\kappa}}(\mathbf{z})$ applies the soft-thresholding operator $\mathcal{S}_{\kappa_p}(z_p) = \text{sign}(z_p)\max(|z_p| - \kappa_p, 0)$ coordinate-wise, producing exact zeros whenever $|z_p| \leq \lambda w_p/\rho$. By caching Cholesky factors of $\mathbf{X}_\gamma^{*\top}\mathbf{X}_\gamma^* + \rho \mathbf{I}_F$, the slope update in equation (12c) achieves $O(NF)$ complexity per iteration. The third step performs the dual variable update: $\mathbf{u}^{(t+1)} = \mathbf{u}^{(t)} + \tilde{\mathbf{Y}} - \mathbf{r}^{(t+1)} - \tilde{\mathbf{X}}^* \boldsymbol{\theta}^{(t+1)}$.

To ensure algorithmic convergence, we monitor both primal and dual residuals defined as $\mathbf{r}_{\text{pri}}^{(t+1)} = \tilde{\mathbf{Y}} - \mathbf{r}^{(t+1)} - \tilde{\mathbf{X}}^* \boldsymbol{\theta}^{(t+1)}$ and $\mathbf{r}_{\text{dual}}^{(t+1)} = \rho\, \tilde{\mathbf{X}}^{*\top}(\mathbf{r}^{(t+1)} - \mathbf{r}^{(t)})$, respectively. The corresponding tolerances are $\varepsilon_{\text{pri}} = \sqrt{NK}\, \varepsilon_{\text{abs}} + \varepsilon_{\text{rel}} \max\left\{\|\tilde{\mathbf{X}}^* \boldsymbol{\theta}^{(t+1)}\|_2, \|\mathbf{r}^{(t+1)}\|_2, \|\tilde{\mathbf{Y}}\|_2\right\}$ and $\varepsilon_{\text{dual}} = \sqrt{F + K}\, \varepsilon_{\text{abs}} + \varepsilon_{\text{rel}} \|\rho\, \tilde{\mathbf{X}}^{*\top} \mathbf{u}^{(t+1)}\|_2$, with default values $\varepsilon_{\text{abs}} = 10^{-6}$ and $\varepsilon_{\text{rel}} = 10^{-4}$ following. The algorithm terminates when both $\|\mathbf{r}_{\text{pri}}^{(t+1)}\|_2 \leq \varepsilon_{\text{pri}}$ and $\|\mathbf{r}_{\text{dual}}^{(t+1)}\|_2 \leq \varepsilon_{\text{dual}}$ are satisfied.

To maintain consistency with the sparsistency requirements established in Theorem 1, we implement an additional trimming step that sets $\gamma_p = 0$ whenever $|\gamma_p| < \eta_{\text{trim}} := \lambda\sqrt{\log(F)/N}$. We subsequently verify the Karush-Kuhn-Tucker condition $|(\tilde{\mathbf{X}}^*)_{:p}^\top \mathbf{r}| \leq \lambda w_p + 10^{-8}$ for every trimmed coordinate, providing an inexpensive optimality audit that guarantees solution quality in finite precision arithmetic.

The tuning parameter λ for the adaptive Lasso penalty controls model sparsity and is critical for performance. We select its optimal value by min-

imizing the Hierarchical Bayesian Information Criterion (HBIC):

$$\mathrm{HBIC}(\lambda) = -2\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\log f_{\mathrm{ALD}}\Big(Y_{ij} \mid \hat{b}_{\tau_k}, \hat{\eta}\Big) + \log(N)\,\mathrm{df}(\lambda),$$

where $f_{\mathrm{ALD}}(Y_{ij} \mid \hat{b}_{\tau_k}, \hat{\eta})$ is the value of the composite quantile likelihood based on the ALD for a given λ, and $\mathrm{df}(\lambda)$ is the number of nonzero coefficients in $\boldsymbol{\gamma}$. This combination of EM and ADMM provides a computationally robust and efficient framework for fitting the HCQR model.

4. Simulation

In this section, we conducted an evaluation of the Oracle model performance across various heavy-tailed distributions using simulation experiments. The simulation data are generated from the model (5).

The total sample size is set to $n = 200$. The dimension of the fixed effect vector $\boldsymbol{\gamma}$ is $p = 205$, with a sparsity level of 0.05, resulting in approximately 10 nonzero coefficients. We set up 10 mutually independent groups, each consisting of a random intercept and a random slope. This configuration results in the random effect having a dimension of $q = 2$ for each group.

The fixed effect design matrix $\mathbf{X}_{ij}$ is drawn from a multivariate normal distribution $N(\mathbf{0}, \mathbf{C})$. The entries of $\mathbf{C}$ are defined by $\rho^{|i-j|}$, where $\rho = 0.5$ represents the correlation coefficient. The random effect $\mathbf{v}_j$ is drawn from a multivariate normal distribution $N(\mathbf{0}, \mathbf{D})$, where $\mathbf{D}$ is the covariance matrix for random effects. The error in estimating the random effects covariance matrix is measured by the Frobenius norm $D = \|\hat{\mathbf{T}} - \mathbf{T}\|_F$.

We evaluated the performance of the Oracle model under various error term distributions, including the normal distribution, the Cauchy distribu-

tion, the t distribution with degrees of freedom 3, the Laplace distribution and a mixture normal distribution. For conditional quantile estimation, we considered three quantile points $\tau \in \{0.3, 0.5, 0.7\}$.

To evaluate the precision of variable selection, we calculated the following metrics. TP (true positives), TDR (true discovery rate), FP (false positives), FDR (false discovery rate), FN (false negatives) and TN (true negatives). These metrics collectively assess the model's ability to correctly identify the true nonzero coefficients.

To evaluate estimation accuracy, we defined several error metrics. The model error is defined as $\text{ME} = (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^\top \mathbf{C} (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$. The full model error is $\text{FME} = \boldsymbol{\gamma}^\top \mathbf{C} \boldsymbol{\gamma}$. The ratio $\text{RME} = \text{ME}/\text{FME}$ quantifies the relative accuracy of the model, with $\text{RME} \leq 1$ by construction, and values closer to 0 indicating higher accuracy. We also computed the prediction mean squared error (PMSE) to assess the model's predictive performance, and the L_2 error of coefficient estimates to measure parameter estimation accuracy. Additionally, we calculate the error in estimating the random effects covariance matrix.

The experiment was conducted over 100 independent iterations, with random error terms generated independently for each iteration.

The simulation results, summarized in Tables 1-5, consistently demonstrate the superior performance of the proposed HCQR method across a wide range of scenarios. In terms of both variable selection accuracy and estimation precision, HCQR performs nearly identically to the idealized Oracle estimator and substantially outperforms all competing methods. This superiority is evident even in the baseline scenario with normal errors, as shown in Table 1. While most methods perform adequately, HCQR achieves

Table 1: Performance Comparison of Methods under $\varepsilon_{ij} \sim N(0,1)$

ε_{ij}	Quantile	Method	TP(TDR)	FP(FDR)	FN	TN	RME	PMSE	L_2	D
N(0,1)	$\tau_k = \frac{(2k+1)}{10}$	Oracle	10(1)	0(0)	0	195.0	0.00024	0.1707	0.2868	0.5041
		HCQR	10(0.9920)	0.09(0.0283)	0	194.9	0.00056	0.2487	0.3942	0.5263
		CQR	9.99(0.6618)	6.36(0.3382)	0.01	188.6	0.00750	6.8239	0.4332	1.2600
	τ_1=0.3	LQMM	9.92(0.5854)	9.88(0.4146)	0.08	185.1	0.00181	0.5859	0.3679	1.1539
		QREM	9.83(0.9952)	0.05(0.0048)	0.17	195.0	0.00715	3.1820	0.4332	0.6600
	τ_2=0.5	LQMM	9.91(0.6541)	7.05(0.3459)	0.09	188.0	0.00094	0.4335	0.3321	1.1116
		QREM	9.86(0.9971)	0.03(0.0029)	0.14	194.9	0.00620	2.4945	0.7437	0.6070
	τ_3=0.7	LQMM	9.88(0.6081)	8.67(0.3919)	0.12	186.3	0.00160	0.6048	0.3554	1.2267
		QREM	9.82(1)	0(0)	0.18	195.0	0.00670	3.3149	0.4338	0.6570
	-	SPLMM	9.52(0.9983)	0.02(0.0017)	0.64	194.9	0.03304	13.082	3.4868	1.0669

Table 2: Performance Comparison of Methods under $\varepsilon_{ij} \sim$ Cauchy$(0,3)$

ε_{ij}	Quantile	Method	TP(TDR)	FP(FDR)	FN	TN	RME	PMSE	L_2	D
cauchy(0,3)	$\tau_k = \frac{(2k+1)}{10}$	Oracle	10(1)	0(0)	0	195	0.0008	0.6731	0.4940	0.6619
		HCQR	9.99(0.9498)	0.61(0.0502)	0.01	194.4	0.0015	0.8875	0.6504	0.6551
		CQR	9.61(0.8795)	0.88(0.2559)	0.39	186.1	0.3087	2.9251	0.8898	1.1332
	τ=0.3	LQMM	9.23(0.7862)	3.25(0.2138)	0.77	191.8	0.0145	6.4550	0.7912	0.9081
		QREM	9.35(0.9789)	0.10(0.0211)	0.65	194.9	0.2953	3.4618	1.3929	1.0298
	τ=0.5	LQMM	9.52(0.8188)	2.69(0.1812)	0.48	192.3	0.0061	4.0444	0.6621	0.8883
		QREM	9.58(0.9835)	0.07(0.0165)	4.42	194.9	0.2777	1.0151	1.2662	0.9545
	τ=0.7	LQMM	9.32(0.7811)	3.33(0.2189)	0.68	191.7	0.0110	5.0806	0.7691	0.8449
		QREM	9.32(0.9841)	0.08(0.0159)	4.68	194.9	0.3255	1.4264	1.5235	1.1150
	-	SPLMM	9.01(0.9214)	1.23(0.0786)	5.99	193.8	0.5762	3.3325	5.3341	1.8361

Table 3: Performance Comparison of Methods under $\varepsilon_{ij} \sim t(3)$

ε_{ij}	Quantile	Method	TP(TDR)	FP(FDR)	FN	TN	RME	PMSE	L_2	D
$t(3)$	$\tau_k = \frac{(2k+1)}{10}$	Oracle	10(1)	0(0)	0	195	0.0004	0.3645	0.3776	0.5689
		HCQR	10(0.9956)	0.05(0.0044)	0	194.9	0.0007	0.4240	0.4703	0.5521
		CQR	9.99(0.6147)	5.68(0.3852)	0.01	187.4	0.0112	5.0886	1.8653	1.5278
	τ=0.3	LQMM	9.81(0.6301)	7.52(0.3699)	0.19	187.5	0.0025	0.9967	0.4358	1.0412
		QREM	9.78(0.9980)	0.02(0.0020)	0.22	194.9	0.0076	3.5836	0.5126	0.7332
	τ=0.5	LQMM	9.82(0.6911)	5.83(0.3089)	0.18	189.2	0.0019	0.7764	0.3924	1.0329
		QREM	9.88(0.9991)	0.01(0.0009)	0.12	194.9	0.0041	1.7055	0.4291	0.6212
	τ=0.7	LQMM	9.80(0.6175)	8.60(0.3825)	0.20	186.4	0.0024	1.0097	0.4460	1.1403
		QREM	9.78(0.9982)	0.02(0.0018)	0.22	194.9	0.2906	1.7862	1.4008	1.0250
	/	SPLMM	9.38(1)	0(0)	0.62	195	0.0298	12.251	3.0899	1.0811

Table 4: Performance Comparison of Methods under $\varepsilon_{ij} \sim 0.9N(0,1) + 0.1N(0,100)$

ε_{ij}	Quantile	Method	TP(TDR)	FP(FDR)	FN	TN	RME	PMSE	L_2	D
$0.9N(0,1)$ + $0.1N(0,100)$	$\tau_k = \frac{(2k+1)}{10}$	Oracle	10(1)	0(0)	0	195	0.0004	0.3254	0.3718	0.5588
		HCQR	10(0.9897)	0.12(0.0103)	0	194.9	0.0007	0.4198	0.4731	0.5491
		CQR	9.74(0.7692)	1.40(0.2308)	0.26	192.8	0.0432	3.2513	3.5312	1.2007
	τ=0.3	LQMM	9.82(0.6697)	6.94(0.3303)	0.18	188.1	0.0028	1.4696	0.4400	0.7963
		QREM	9.57(0.9980)	0.02(0.0020)	0.43	194.9	0.0064	3.6998	0.5279	0.7215
	τ=0.5	LQMM	9.81(0.7157)	4.88(0.2843)	0.19	190.1	0.0020	1.2619	0.3908	0.6867
		QREM	9.60(0.9978)	0.02(0.0022)	0.40	194.9	0.0085	2.8822	0.4998	0.6928
	τ=0.7	LQMM	9.78(0.6785)	6.22(0.3215)	0.22	188.8	0.0026	1.4613	0.4312	0.7583
		QREM	9.48(0.9980)	0.02(0.0020)	0.52	194.9	0.0072	3.8537	0.5138	0.6715
	-	SPLMM	9.23(0.9990)	0.01(0.0010)	0.77	194.9	0.0287	11.909	2.9114	1.0989

Table 5: Performance Comparison of Methods under $\varepsilon_{ij} \sim \text{Lap}(0,1)$

ε_{ij}	Quantile	Method	TP(TDR)	FP(FDR)	FN	TN	RME	PMSE	L_2	D
Lap(0,1)	$\tau_k = \frac{(2k+1)}{10}$	Oracle	10(1)	0(0)	0	195	0.0004	0.3068	0.3549	0.5554
		HCQR	10 (0.9950)	0.05(0.0045)	0	195	0.0007	0.3962	0.4574	0.5594
		CQR	10(0.6369)	6.61(0.3631)	0	188.4	0.0094	2.6326	1.7213	2.8648
	τ=0.3	LQMM	9.83(0.6132)	8.33(0.3868)	0.17	186.7	0.0025	0.8873	0.4084	1.2201
		QREM	9.83(0.9978)	0.02(0.0022)	0.17	194.9	0.0075	3.5648	0.4909	0.6512
	τ=0.5	LQMM	9.83(0.6627)	6.64(0.3373)	0.17	188.4	0.0018	0.6726	0.3676	1.2129
		QREM	9.84(0.9990)	0.01(0.0010)	0.16	194.9	0.0065	2.1196	0.4169	0.6651
	τ=0.7	LQMM	9.84(0.5887)	8.84(0.4113)	0.16	186.2	0.0020	0.7919	0.4055	1.1770
		QREM	9.84(0.9990)	0.01(0.0010)	0.16	194.9	0.0054	2.8328	0.4661	0.6634
	-	SPLMM	9.40(1)	0.00(0)	0.60	195	0.0311	12.907	3.1822	1.0867

a near-perfect variable selection result (TP=10, FP=0.09), closely mirroring the Oracle. In contrast, other methods like CQR and LQMM identify true positives at the cost of a much higher number of false positives (FP=6.36 and FP≈7-9, respectively), leading to poor false discovery rates. Moreover, HCQR's estimation and prediction errors (RME=0.00056, PMSE=0.2487) are an order of magnitude lower than those of CQR and QREM, underscoring its efficiency.

The true strength and robustness of HCQR are revealed in the presence of heavy-tailed errors (Tables 2, 3, 5) and outliers (Table 4). A stark contrast is observed between HCQR and the mean-based SPLMM under the Cauchy distribution (Table 2), where SPLMM's estimation and prediction errors explode (RME=0.5762) while HCQR's performance remains stable and near-oracle (RME=0.0015). This highlights the critical advantage of the robust quantile-based loss function. Furthermore, compared to other quantile-based methods like CQR and LQMM, HCQR exhibits unparalleled

precision in variable selection. Across all non-normal scenarios, HCQR consistently maintains a very low number of false positives (FP ranging from 0.05 to 0.61), whereas CQR and LQMM often include numerous incorrect variables, demonstrating the effectiveness of the adaptive Lasso penalty within the robust HCQR framework. The superior efficiency of HCQR is also evident, as its RME and PMSE values are consistently the lowest among all practical methods, suggesting that the composite quantile approach effectively aggregates information to improve estimation.

In summary, the simulation studies robustly validate the theoretical claims of our paper. The proposed HCQR method successfully synergizes the hierarchical model structure, the robustness of composite quantile regression, and the variable selection accuracy of the adaptive Lasso. It provides a powerful and reliable tool for analyzing high-dimensional hierarchical data, demonstrating near-oracle performance even in challenging scenarios with non-Gaussian noise and a large number of covariates.

5. Real World Data

We utilize the Oxford Parkinson's Disease tele-monitoring Dataset, an open-access biomedical dataset designed for evaluating Parkinson's disease (PD) symptoms through voice measurements. The dataset consists of repeated voice recordings from 42 early-stage PD patients, collected longitudinally over six months, providing robust data to analyze both within-subject and between-subject variability in symptom progression.

The primary objective of the dataset is to predict **motor UPDRS** (motor Unified Parkinson's Disease Rating Scale) and **total UPDRS** scores,

which are key indicators of symptom severity and progression, using voice features. This analysis is crucial for understanding disease dynamics and enabling precise, non-invasive symptom monitoring.

The dataset consists of 26 variables, including demographic characteristics such as age and sex, the time of the voice recording (test time), target variables (motor UPDRS and total UPDRS), and 16 voice features. These features describe variations in frequency (Jitter variables), amplitude (Shimmer variables), noise ratio (NHR and HNR), and non-linear dynamic attributes such as RPDE and PPE. These voice characteristics are highly correlated with speech motor control, making them vital for predicting UPDRS scores and understanding the progression of Parkinson's Disease symptoms. Collectively, these variables offer a comprehensive characterization of voice changes, serving as a critical data source for non-invasive, remote monitoring of symptom progression.

The original study of this dataset was conducted by Athanasios Tsanas and Max Little, whose findings, published in IEEE Transactions on Biomedical Engineering, have significantly influenced the development of computational tools for analyzing voice biomarkers and their applications in Parkinson's Disease monitoring. We strictly adhere to the citation standards of the dataset and refer to its relevant literature.[21]

Table 6 summarizes the dataset's variables, highlighting their correlations with the motor and total UPDRS scores. Age, Pitch Period Entropy (PPE), Shimmer variables, and Recurrence Period Density Entropy (RPDE) exhibit relatively stronger correlations with both UPDRS measures, consistent with findings by Tsanas et al. (2010). Negative correlations, such as those ob-

Table 6: Description and Correlation of Features with UPDRS Scores

Variables	Description	Correlation	
		Motor UPDRS	Total UPDRS
age	Subject age	0.274	0.310
PPE	Pitch Period Entropy	0.162	0.156
Shimmer:APQ11	Amplitude Perturbation Quotient	0.137	0.121
RPDE	Recurrence Period Density Entropy	0.129	0.157
Shimmer(dB)	KP-MDVP local shimmer in decibels	0.110	0.099
Shimmer	KP-MDVP local shimmer	0.102	0.092
Shimmer:APQ5	Five point Amplitude Perturbation Quotient	0.092	0.083
Jitter	KP-MDVP jitter as a percentage	0.085	0.074
Shimmer:APQ3	Three point Amplitude Perturbation Quotient	0.084	0.079
Shimmer:DDA	difference between consecutive differences between the amplitudes of consecutive periods	0.084	0.079
Jitter.PPQ5	Period Perturbation Quotient	0.076	0.063
NHR	Noise-to-Harmonics Ratio	0.075	0.061
Jitter.DDP	Difference of differences between cycles, divided by the average period	0.073	0.064
Jitter.RAP	KP-MDVP Relative Amplitude Perturbation	0.073	0.064
test_time	Time since recruitment into the trial. The integer part is the number of days since recruitment.	0.068	0.075
Jitter(Abs)	KP-MDVP absolute jitter in microseconds	0.051	0.067
sex	Subject gender '0' - male, '1' - female	-0.031	-0.097
DFA	Detrended Fluctuation	-0.116	-0.113
HNR	Harmonics-to-Noise Ratio	-0.157	-0.162

served for Detrended Fluctuation Analysis (DFA) and Harmonics-to-Noise Ratio (HNR), suggest their inverse association with symptom severity, aligning with previously reported acoustic biomarkers (Little et al., 2009; Tsanas et al., 2010).

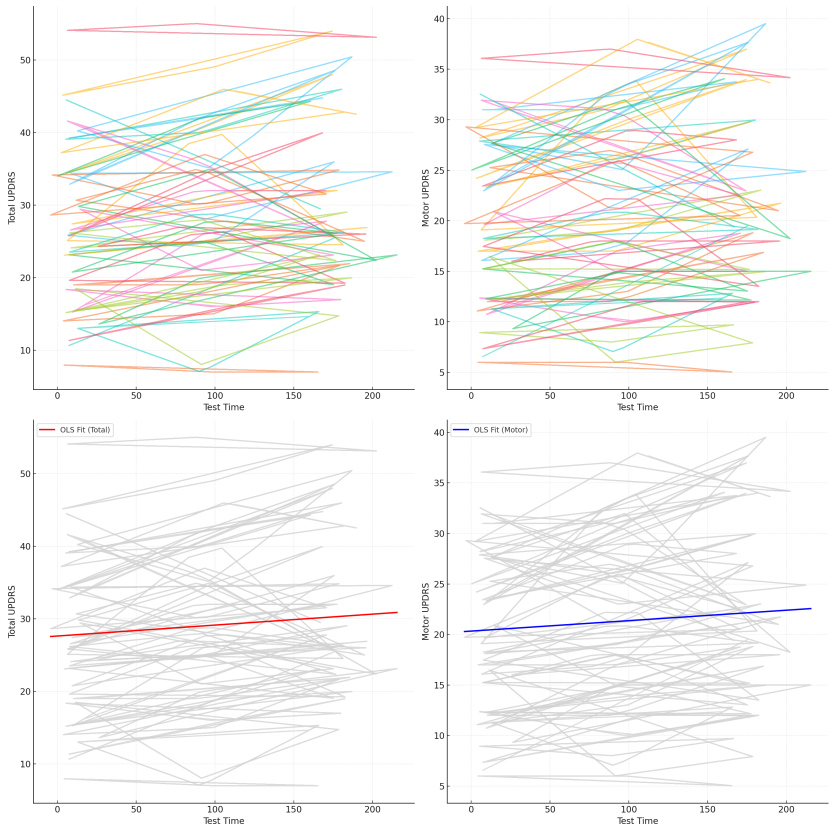

Figure 1: Comparison of Individual Trends and Global Regression in UPDRS Scores

Figure 1 illustrates the disparity between individual trends (gray lines) and the global Ordinary Least Squares (OLS) regression fit (colored lines) for the UPDRS scores in a hierarchical dataset, where the global fit represents the overall trend modeled across all subjects without accounting for within-subject variability. While the OLS regression captures an overall trend across

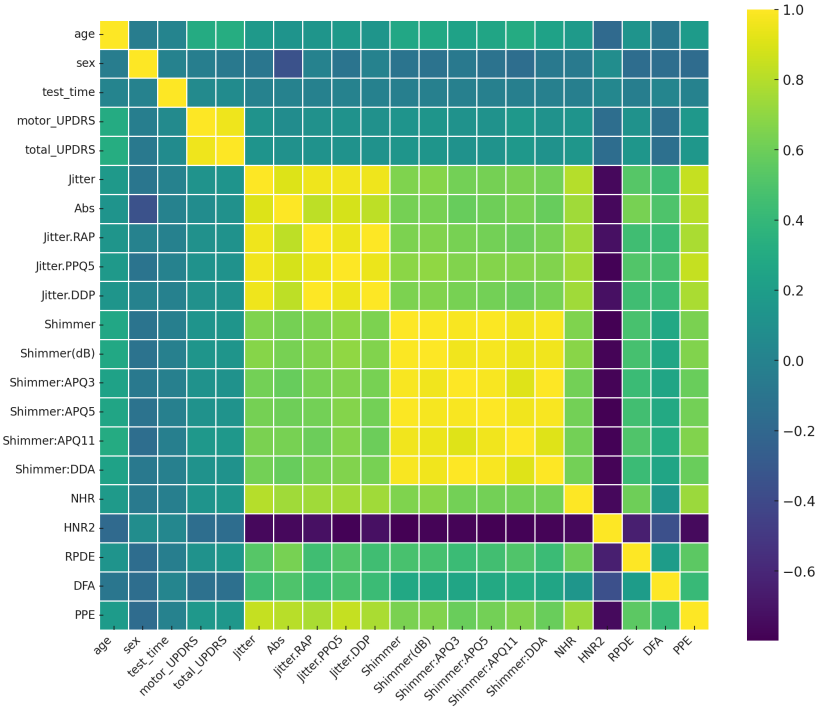

Figure 2: Heat-map of Spearman Correlations for Parkinson's UPDRS Dataset

subjects, it overlooks the variability within individual subjects, as evidenced by the substantial divergence of individual trajectories from the global regression line. This discrepancy underscores the need for hierarchical or mixed-effects models, which address both group-level and subject-level variability, providing a more tailored and precise modeling approach to understand UPDRS score dynamics.

We subsequently fit several competing models, including the null model, classical linear regression (lm), standard quantile regression (rq), linear quantile mixed-effects model (lqmm), and our proposed Hierarchical Composite Quantile Regression model (HCQR). All continuous covariates were z-scored; patient-specific random intercepts and random slopes on recording time were included in every hierarchical specification. The results, summarized in Table 7, illustrate notable differences in fixed and random effects as well as model performance metrics. Across both outcomes HCQR delivered the most economical and accurate fit. Its Akaike information criteria (AIC = 19 980 for motor, 21 052 for total) improved on the next-best LQMM by roughly 10 %. Mean absolute errors fell to 0.84 (motor) and 0.91 (total) UPDRS—well below the two-point threshold often cited as clinically meaningful variation—and mean squared errors were likewise the smallest in the comparison set. Likelihood-ratio tests confirmed that HCQR's multiquantile loss and adaptive sparsity provided a statistically significant gain over single-quantile LQMM.

The adaptive penalty retained a compact set of thirteen fixed effects. Age, sex and follow-up time remained robust (bootstrap selection probability ≥ 0.98); age increased both scales, while female sex lowered scores by about

Table 7: Model Comparison for Total and Motor UPDRS

Fixed Effects										
	Null model		lm		rq		lqmm		HCQR	
Variable	total UPDRS	motor UPDRS	total UPDRS	motor UPDRS	total UPDRS	motor UPDRS	total UPDRS	motor UPDRS	total UPDRS	motor UPDRS
Intercept	28.5380	20.9500	29.9063	21.6639	28.9515	21.5904	29.7561	22.2011	29.2637***	21.6084***
age	-	-	2.6726	1.6922	2.5190	1.9585	2.20032	3.0849	3.05121***	1.9948***
sex	-	-	-2.7924	-1.1568	-3.8221	-2.9642	0.49369	-0.6746	-2.9212***	-2.9443***
test time	-	-	0.8844	0.6108	0.5097	0.6599	-0.2391	1.2099	0	0.5831***
Jitter	-	-	0.2439	1.6158	0.7226	1.4758	0.21288	0.2894	0	0.0720
Abs	-	-	-2.3056	-2.3372	-1.4814	-2.6604	-0.1915	-0.0776	0	-1.6638*
Jitter.RAP	-	-	-123.4599	-117.4794	-329.7932	-223.9812	-124.89	-120.9337	0	1.043
Jitter.PPQ5	-	-	-1.2634	-1.2064	-1.4342	-0.4302	-0.5328	-0.1266	0	0
Jitter.DDP	-	-	126.0903	118.6071	332.1586	225.7011	125.066	120.6618	0	0
Shimmer	-	-	3.8560	3.7656	4.8005	5.2605	1.40544	0.9405	0	0
Shimmer.dB	-	-	-2.0502	-1.5468	-4.2225	-3.0521	0.05974	0.30889	0	-0.7063
Shimmer.APQ3	-	-	-194.0937	-11.4990	-776.2602	-779.3639	-194.04	-13.9776	0	0
Shimmer.APQ5	-	-	-1.0427	-2.0621	0.1815	-2.4408	-0.0602	-0.1706	0	0
Shimmer.APQ11	-	-	1.0682	1.4594	0.9734	2.1190	-0.4735	-0.4812	0.02147*	0.9937*
Shimmer.DDA	-	-	191.2013	9.3201	773.0283	776.8558	193.397	13.5162	0.21301	-1.0161
NHR	-	-	-0.9088	-0.6176	-0.8557	-1.4193	0.42261	0.2693	0	-0.8184***
HNR	-	-	-2.6908	-1.8265	-2.6812	-2.2021	0.35627	-0.2138	0.01175	-2.2752***
RPDE	-	-	0.4083	0.0637	0.9984	0.4593	-0.4722	-0.1660	0.00197*	0.3651*
DFA	-	-	-2.2319	-1.6692	-3.2576	-3.0287	-0.4815	-0.7493	-0.0686***	-2.8616***
PPE	-	-	1.6034	1.6013	1.2539	1.5426	0.03338	-0.1847	0.02192*	1.5733***
Random Effects										
Intercept	10.4410	7.8780	-	-	-	-	9.9542	10.564	6.2486	7.0170
test time	-	-	-	-	-	-	2.3199	2.2124	2.3273	1.9763
Residual	2.7690	2.3400	-	-	-	-	1.6538	4.2273	1.4290	1.2982
ICC	0.9340	0.9190	-	-	-	-	0.9745	0.9709	0.9526	0.9725
Adjusted ICC	0.9340	0.9190	-	-	-	-	0.7998	0.8162	0.8757	0.8524
Model Fit										
AIC	28959.69	26976.19	43423.04	40333.17	44168.61	41320.94	23877.83	21508.14	21051.73	19980.31
MAE	2.1132	1.7417	8.0435	6.3217	7.8645	6.1726	1.0437	0.9291	0.9085	0.8374
MSE	7.6113	5.4387	94.2668	55.7120	97.7781	58.4646	2.2312	1.7915	2.0157	1.6181
LRT	-		-		-		P¡0.0001(df=19)		P¡0.0001(df=15)	

Note: * : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.001$

three points. For motor-UPDRS, follow-up time showed the expected positive slope, whereas in the more heterogeneous total-UPDRS that signal was absorbed by non-motor variance and shrank to zero. Among acoustic biomarkers the model preserved Shimmer dB, Jitter Abs, the harmonic-to-noise ratio proxy HNR2, and three non-linear complexity measures (RPDE, DFA, PPE). Each coefficient direction agreed with prior studies linking greater amplitude jitter, higher noise content and increased non-linear irregularity to worse symptom severity. Highly collinear surrogates such as RAP, PPQ5, APQ5 and DDA were consistently suppressed to zero.

Block bootstrap (patients resampled as intact clusters, 200 replicates) was used both for uncertainty quantification and for stability selection. Ninety per cent of replicates converged; percentile intervals never crossed zero for the retained predictors, and all core variables exhibited selection rates above 0.70, whereas discarded surrogates rarely exceeded 0.40. Thus the final variable set is not only parsimonious but demonstrably reproducible under realistic data perturbations.

Random-effects estimates underscore pronounced between-patient heterogeneity: baseline intercept variance was 7.0 for motor and 6.3 for total UPDRS, while slope variance on recording time approached 2.0. Intraclass correlations before adjusting fixed effects were ≈ 0.95, dropping by about one tenth once covariates were introduced—evidence that fixed effects explain a meaningful share of cross-patient variation but substantial individual differences in progression remain[17].

Taken together, the HCQR model reconciles predictive accuracy with interpretability. It recovers the voice-based biomarkers most frequently re-

ported in the Parkinson's literature, quantifies age- and sex-related effects, and—through its hierarchical structure—captures the patient-level trajectories essential for remote symptom monitoring.

6. Conclusion

In this investigation, we formulated the Hierarchical Composite Quantile Regression (HCQR) model augmented with adaptive Lasso penalization to tackle challenges inherent in high-dimensional hierarchical data analysis. Through both simulations and empirical applications, the approach demonstrated efficacy in variable selection, exhibited robustness against outliers, and presented distinct advantages over extant models. The application of HCQR to the Parkinson's Disease Telemonitoring Dataset further exemplified its capability to analyze intricate hierarchical relationships. These outcomes underscore the model's potential to advance statistical analysis across various scientific domains.

Appendix A. Proofs

Proof of Lemma 1. This proof is established under the oracle setting. This means we consider a hypothetical model fitted only with the covariates from the true active set $\mathcal{A}$. Therefore, throughout this proof, the design matrix $\mathbf{X}_{ij}^*$ and its limiting second-moment matrix $\mathbf{C}$ should be interpreted as the sub-matrix $\mathbf{X}_{\mathcal{A},ij}^*$ and its corresponding limit $\mathbf{C}_{\mathcal{A}\mathcal{A}}$, respectively. Furthermore, we assume that the covariates (excluding the intercept) are centered, such that $\frac{1}{N}\sum_{j,i}\mathbf{X}_{ij}^* \to \mathbf{0}$ as $N \to \infty$.

Let $\hat{\boldsymbol{\gamma}}$ and $\hat{b}_k$ be the CQR estimators. We analyze their asymptotic behavior by considering the scaled parameter deviations $\mathbf{u}_\gamma = \sqrt{N}(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*)$ and $u_k = \sqrt{N}(b_k - b_k^*)$ for $k = 1, \ldots, K$. The estimators $(\hat{\mathbf{u}}_\gamma, \hat{u}_1, \ldots, \hat{u}_K)$ are the minimizers of the objective function

$$Q_N(\mathbf{u}) = \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left[\rho_{\tau_k} \left(\varepsilon_{ij} - b_k^* - \frac{u_k + \mathbf{X}_{ij}^{*\top} \mathbf{u}_\gamma}{\sqrt{N}} \right) - \rho_{\tau_k}(\varepsilon_{ij} - b_k^*) \right],$$

where $\mathbf{u} = (u_1, \ldots, u_K, \mathbf{u}_\gamma^\top)^\top$ and $\varepsilon_{ij} = Y_{ij} - \mathbf{X}_{ij}^{*\top} \boldsymbol{\gamma}^*$ is the error term whose true τ_k-th quantile is b_k^*. We leverage the identity from Knight (1998), $\rho_\tau(r - s) - \rho_\tau(r) = -s\psi_\tau(r) + \int_0^s [I(r \leq t) - I(r \leq 0)]dt$, where $\psi_\tau(r) = \tau - I(r < 0)$.

Applying this identity, $Q_N(\mathbf{u})$ can be expressed as the sum of a linear term, $L_N(\mathbf{u})$, and an integral remainder term, $R_N(\mathbf{u})$. The linear term is

$$L_N(\mathbf{u}) = -\frac{1}{\sqrt{N}} \sum_{k=1}^{K} \sum_{j,i} (u_k + \mathbf{X}_{ij}^{*\top} \mathbf{u}_\gamma) \psi_{\tau_k}(\varepsilon_{ij} - b_k^*) = -\mathbf{W}_N^\top \mathbf{u},$$

where $\mathbf{W}_N$ is the normalized score vector. Since $E[\psi_{\tau_k}(\varepsilon_{ij} - b_k^*)] = 0$ for all k, by the multivariate Central Limit Theorem and Condition 1, $\mathbf{W}_N$ converges in distribution to a multivariate normal random vector $\mathbf{W} \sim N(\mathbf{0}, \boldsymbol{\Omega})$. The covariance matrix $\boldsymbol{\Omega}$ has a block structure. The block corresponding to the

coefficients of $\mathbf{u}_\gamma$, denoted $\mathbf{\Omega}_{\gamma\gamma}$, is calculated as

$$\begin{aligned}
\mathbf{\Omega}_{\gamma\gamma} &= \lim_{N\to\infty} \mathrm{Var}\left(\frac{1}{\sqrt{N}}\sum_{j,i}\mathbf{X}_{ij}^{*}\sum_{k=1}^{K}\psi_{\tau_k}(\varepsilon_{ij}-b_k^{*})\right) \\
&= \lim_{N\to\infty} E\left[\left(\frac{1}{N}\sum_{j,i}\mathbf{X}_{ij}^{*}\mathbf{X}_{ij}^{*\top}\right)\left(\sum_{k=1}^{K}\psi_{\tau_k}(\varepsilon-b_k^{*})\right)^{2}\right] \\
&= \left(\lim_{N\to\infty}\frac{1}{N}\sum_{j,i}\mathbf{X}_{ij}^{*}\mathbf{X}_{ij}^{*\top}\right) E\left[\left(\sum_{k,k'}\psi_{\tau_k}(\varepsilon-b_k^{*})\psi_{\tau_{k'}}(\varepsilon-b_{k'}^{*})\right)\right] \\
&= \mathbf{C}\sum_{k,k'}E[\psi_{\tau_k}(\varepsilon-b_k^{*})\psi_{\tau_{k'}}(\varepsilon-b_{k'}^{*})] = \mathbf{C}\sum_{k,k'}(\min(\tau_k,\tau_{k'})-\tau_k\tau_{k'}).
\end{aligned}$$

The integral remainder term is $R_N(\mathbf{u}) = \sum_{k,j,i}\int_0^{(u_k+\mathbf{X}_{ij}^{*\top}\mathbf{u}_\gamma)/\sqrt{N}}[I(\varepsilon_{ij}-b_k^{*}\le t)-I(\varepsilon_{ij}-b_k^{*}\le 0)]dt$. This term converges in probability to a deterministic quadratic function of $\mathbf{u}$. To show this, we analyze its expectation. Let $F(\cdot)$ be the CDF of ε_{ij}. By a first-order Taylor expansion of $F(b_k^{*}+t)$ around $t=0$, we have $F(b_k^{*}+t)-F(b_k^{*}) = tf(b_k^{*})+o(t)$. Thus,

$$\begin{aligned}
E[R_N(\mathbf{u})] &= \sum_{k,j,i}\int_0^{(u_k+\mathbf{X}_{ij}^{*\top}\mathbf{u}_\gamma)/\sqrt{N}}[F(b_k^{*}+t)-F(b_k^{*})]dt \\
&= \sum_{k,j,i}\int_0^{(u_k+\mathbf{X}_{ij}^{*\top}\mathbf{u}_\gamma)/\sqrt{N}}(tf(b_k^{*})+o(t))\,dt \\
&= \sum_{k,j,i}\left[\frac{1}{2}t^2 f(b_k^{*})\right]_0^{(u_k+\mathbf{X}_{ij}^{*\top}\mathbf{u}_\gamma)/\sqrt{N}} + o_p(1) \\
&= \frac{1}{2N}\sum_{k}f(b_k^{*})\sum_{j,i}(u_k^2+2u_k\mathbf{X}_{ij}^{*\top}\mathbf{u}_\gamma+(\mathbf{X}_{ij}^{*\top}\mathbf{u}_\gamma)^2)+o_p(1).
\end{aligned}$$

Under Condition 1, $\frac{1}{N}\sum_{j,i}(\mathbf{X}_{ij}^{*\top}\mathbf{u}_\gamma)^2 \to \mathbf{u}_\gamma^\top\mathbf{C}\mathbf{u}_\gamma$, and assuming centered covariates, $\frac{1}{N}\sum_{j,i}\mathbf{X}_{ij}^{*}\to\mathbf{0}$. Therefore, $E[R_N(\mathbf{u})]$ converges to $\frac{1}{2}\sum_k f(b_k^{*})u_k^2 + \frac{1}{2}(\sum_k f(b_k^{*}))\mathbf{u}_\gamma^\top\mathbf{C}\mathbf{u}_\gamma = \frac{1}{2}\mathbf{u}^\top\mathbf{H}\mathbf{u}$, where $\mathbf{H}$ is the block-diagonal Hessian matrix $\mathbf{H} = \mathrm{diag}(\mathrm{diag}(f(b_1^{*}),\ldots,f(b_K^{*})),(\sum_k f(b_k^{*}))\mathbf{C})$. The stochastic term

$R_N(\mathbf{u}) - E[R_N(\mathbf{u})]$ can be shown to converge to zero in probability by empirical process theory. Thus, $Q_N(\mathbf{u})$ admits the asymptotic quadratic representation $Q_N(\mathbf{u}) = -\mathbf{W}_N^\top \mathbf{u} + \frac{1}{2}\mathbf{u}^\top \mathbf{H}\mathbf{u} + o_p(1)$.

The objective function $Q_N(\mathbf{u})$ is convex. By the epiconvergence theorem (or argmin continuous mapping theorem), the minimizer $\hat{\mathbf{u}}_N$ of $Q_N(\mathbf{u})$ converges in distribution to the minimizer of its limit $V(\mathbf{u}) = -\mathbf{W}^\top \mathbf{u} + \frac{1}{2}\mathbf{u}^\top \mathbf{H}\mathbf{u}$. The unique minimizer of the convex function $V(\mathbf{u})$ is $\hat{\mathbf{u}} = \mathbf{H}^{-1}\mathbf{W}$. The asymptotic distribution of $\hat{\mathbf{u}}_N$ is therefore normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{H}^{-1}\mathbf{\Omega}\mathbf{H}^{-1}$.

Our primary interest is the asymptotic distribution of $\hat{\mathbf{u}}_\gamma = \sqrt{N}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)$. Its asymptotic covariance matrix is the lower-right $F \times F$ block of $\mathbf{H}^{-1}\mathbf{\Omega}\mathbf{H}^{-1}$. Given the block-diagonal structure of $\mathbf{H}$, its inverse is also block-diagonal, with the lower-right block being $(\mathbf{H}^{-1})_{\gamma\gamma} = (\sum_k f(b_k^*))^{-1}\mathbf{C}^{-1}$. The asymptotic variance of $\hat{\mathbf{u}}_\gamma$ is thus

$$
\begin{aligned}
\mathrm{Var}(\hat{\mathbf{u}}_\gamma) &= (\mathbf{H}^{-1})_{\gamma\gamma}\mathbf{\Omega}_{\gamma\gamma}(\mathbf{H}^{-1})_{\gamma\gamma}^\top \\
&= \left[\left(\sum_k f(b_k^*)\right)^{-1}\mathbf{C}^{-1}\right]\left[\mathbf{C}\sum_{k,k'}(\min(\tau_k,\tau_{k'}) - \tau_k\tau_{k'})\right]\left[\left(\sum_k f(b_k^*)\right)^{-1}\mathbf{C}^{-1}\right] \\
&= \left(\sum_k f(b_k^*)\right)^{-2}\left(\sum_{k,k'}(\min(\tau_k,\tau_{k'}) - \tau_k\tau_{k'})\right)\mathbf{C}^{-1}\mathbf{C}\mathbf{C}^{-1} \\
&= \mathbf{C}^{-1}\frac{\sum_{k,k'}(\min(\tau_k,\tau_{k'}) - \tau_k\tau_{k'})}{\left(\sum_k f(b_k^*)\right)^2}.
\end{aligned}
$$

This completes the proof. □

Proof of Theorem 1. The proof of the oracle properties of the HCQR estimator is twofold. First, we establish that the estimator correctly identifies the set of true non-zero coefficients with probability tending to one (selection

consistency). Second, we show that the estimators for the non-zero coefficients have the same asymptotic normal distribution as the oracle estimator that knows the true model in advance.

Let the true parameter vector be $\boldsymbol{\gamma}^* = ((\boldsymbol{\gamma}_1^*)^\top, (\boldsymbol{\gamma}_2^*)^\top)^\top$, where $\boldsymbol{\gamma}_1^*$ is the $s \times 1$ vector of non-zero coefficients and $\boldsymbol{\gamma}_2^* = \mathbf{0}$ is the $(F-s) \times 1$ vector of zero coefficients. Let $\mathcal{A} = \{p : \gamma_p^* \neq 0\}$ be the true active set of size s. Our goal is to show that the adaptive Lasso penalized estimator $\hat{\boldsymbol{\gamma}} = ((\hat{\boldsymbol{\gamma}}_1)^\top, (\hat{\boldsymbol{\gamma}}_2)^\top)^\top$ satisfies:

(i) $\Pr(\hat{\boldsymbol{\gamma}}_2 = \mathbf{0}) \to 1$ as $N \to \infty$.

(ii) $\sqrt{N}(\hat{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_1^*) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_{\text{CQR\_oracle}})$, where $\boldsymbol{\Sigma}_{\text{CQR\_oracle}}$ is the covariance matrix from Theorem 1 restricted to the covariates in $\mathcal{A}$.

Let us consider the objective function for the estimator $\hat{\boldsymbol{\gamma}}$:

$$Q(\boldsymbol{\gamma}) = \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \rho_{\tau_k}(Y_{ij} - b_{\tau_k} - \mathbf{X}_{ij}^* \boldsymbol{\gamma}) + \lambda_N \sum_{p=1}^{F} \frac{|\gamma_p|}{|\hat{\gamma}_p^{\text{CQR}}|}.$$

For simplicity of notation, we absorb the intercepts b_{τ_k} into an augmented parameter vector, which does not affect the core argument for $\boldsymbol{\gamma}$.

Selection Consistency

To show that $\hat{\boldsymbol{\gamma}}_2 = \mathbf{0}$ with probability approaching 1, we will show that for any $p \in \mathcal{A}^c$ (i.e., $\gamma_p^* = 0$), $\Pr(\hat{\gamma}_p \neq 0) \to 0$. We inspect the Karush-Kuhn-Tucker (KKT) conditions for the minimization of $Q(\boldsymbol{\gamma})$. The subgradient of $Q(\boldsymbol{\gamma})$ with respect to γ_p at the minimum $\hat{\boldsymbol{\gamma}}$ must contain zero. This gives:

$$-\sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \psi_{\tau_k}(Y_{ij} - \hat{b}_{\tau_k} - \mathbf{X}_{ij}^* \hat{\boldsymbol{\gamma}}) X_{ijp}^* + \frac{\lambda_N}{|\hat{\gamma}_p^{\text{CQR}}|} \cdot s_p = 0,$$

where $\psi_{\tau_k}(u) = \tau_k - I(u < 0)$ and $s_p \in \partial|\hat{\gamma}_p|$ is the subgradient of the absolute value function, with $s_p = \text{sgn}(\hat{\gamma}_p)$ if $\hat{\gamma}_p \neq 0$ and $s_p \in [-1, 1]$ if $\hat{\gamma}_p = 0$.

For $\hat{\gamma}_p \neq 0$, the condition becomes:

$$\frac{1}{N}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\psi_{\tau_k}(Y_{ij} - \hat{b}_{\tau_k} - \mathbf{X}_{ij}^{*}\hat{\boldsymbol{\gamma}})X_{ijp}^{*} = \frac{\lambda_N}{N}\frac{\text{sgn}(\hat{\gamma}_p)}{|\hat{\gamma}_p^{\text{CQR}}|}. \tag{A.1}$$

Let's analyze the left-hand side (LHS) for $p \in \mathcal{A}^c$. From the results of Theorem 3, we know that $\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\| = O_p(N^{-1/2})$. The LHS is the derivative of the quantile loss component. By a Taylor expansion around the true parameters $(\boldsymbol{\gamma}^*, \mathbf{b}^*)$, the LHS can be shown to be of order $O_p(N^{-1/2})$.

Now consider the right-hand side (RHS) for $p \in \mathcal{A}^c$. Since $\gamma_p^* = 0$, the unpenalized estimator $\hat{\gamma}_p^{\text{CQR}}$ converges to 0 at a rate of $N^{-1/2}$, i.e., $\hat{\gamma}_p^{\text{CQR}} = O_p(N^{-1/2})$. The term in the RHS behaves as:

$$\frac{\lambda_N}{N|\hat{\gamma}_p^{\text{CQR}}|} = \frac{\lambda_N}{N \cdot O_p(N^{-1/2})} = \frac{\sqrt{N}\lambda_N}{\sqrt{N} \cdot O_p(1)} = \frac{\sqrt{N}\lambda_N}{O_p(1)}.$$

Under the assumption $\sqrt{N}\lambda_N \to \infty$, the RHS of (A.1) diverges to infinity.

Therefore, for any $p \in \mathcal{A}^c$, the LHS of the KKT condition is $O_p(N^{-1/2})$ while the RHS tends to infinity. This leads to a contradiction if we assume $\hat{\gamma}_p \neq 0$. The only way for the KKT condition to hold is if $\hat{\gamma}_p = 0$, in which case $|s_p| \leq 1$ and the condition can be satisfied because the large penalty term does not enforce a strict equality. Thus, for any p such that $\gamma_p^* = 0$, we must have $\hat{\gamma}_p = 0$ with probability approaching 1. This establishes selection consistency.

Asymptotic Normality

Given selection consistency, for large N, the problem of estimating $\hat{\boldsymbol{\gamma}}$ is equivalent to solving the penalized quantile regression only for the coefficients

in the active set $\mathcal{A}$. Let's consider the sub-problem for $\boldsymbol{\gamma}_1$:

$$\min_{\boldsymbol{\gamma}_1} \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \rho_{\tau_k}(Y_{ij} - b_{\tau_k} - \mathbf{X}^*_{ij,1}\boldsymbol{\gamma}_1) + \lambda_N \sum_{p \in \mathcal{A}} \frac{|\gamma_{1p}|}{|\hat{\gamma}^{\text{CQR}}_{1p}|},$$

where $\mathbf{X}^*_{ij,1}$ contains the columns of $\mathbf{X}^*_{ij}$ corresponding to the active set $\mathcal{A}$.

Let $\mathbf{u}_1 = \sqrt{N}(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}^*_1)$. Following the arguments in the proof of Theorem 1, the minimizer $\hat{\mathbf{u}}_1$ of the corresponding objective function for the scaled parameter converges in distribution to the minimizer of:

$$V(\mathbf{u}_1) = -\mathbf{W}^\top_{N,1}\mathbf{u}_1 + \frac{1}{2}\mathbf{u}_1^\top \mathbf{H}_{11}\mathbf{u}_1 + P_N(\mathbf{u}_1),$$

where $\mathbf{W}_{N,1}$ is the score vector corresponding to $\mathcal{A}$, $\mathbf{H}_{11} = (\sum_k f(b^*_k))\mathbf{C}_{11}$, and $\mathbf{C}_{11}$ is the submatrix of $\mathbf{C}$ for the active set. The penalty term is $P_N(\mathbf{u}_1)$.

For $p \in \mathcal{A}$, $\gamma^*_p \neq 0$. By the consistency of the CQR estimator, $\hat{\gamma}^{\text{CQR}}_p \xrightarrow{p} \gamma^*_p \neq 0$. Let's examine the penalty's contribution to the objective function:

$$\lambda_N \sum_{p \in \mathcal{A}} \left(\frac{|\gamma^*_{1p} + u_{1p}/\sqrt{N}|}{|\hat{\gamma}^{\text{CQR}}_{1p}|} - \frac{|\gamma^*_{1p}|}{|\hat{\gamma}^{\text{CQR}}_{1p}|} \right).$$

By a Taylor expansion for large N, this is approximately:

$$\approx \lambda_N \sum_{p \in \mathcal{A}} \frac{\text{sgn}(\gamma^*_{1p})}{|\hat{\gamma}^{\text{CQR}}_{1p}|} \frac{u_{1p}}{\sqrt{N}} = \frac{\lambda_N}{\sqrt{N}} \sum_{p \in \mathcal{A}} \frac{\text{sgn}(\gamma^*_{1p})}{|\hat{\gamma}^{\text{CQR}}_{1p}|} u_{1p}.$$

Since $\lambda_N \to 0$ and $\hat{\gamma}^{\text{CQR}}_{1p} \to \gamma^*_{1p}$, the coefficient $\frac{\lambda_N}{\sqrt{N}} \to 0$. Therefore, the entire penalty term for the active set vanishes in the limit.

The asymptotic behavior of $\hat{\mathbf{u}}_1$ is thus governed by minimizing $-\mathbf{W}^\top_{N,1}\mathbf{u}_1 + \frac{1}{2}\mathbf{u}_1^\top \mathbf{H}_{11}\mathbf{u}_1$. The minimizer is $\hat{\mathbf{u}}_1 = \mathbf{H}^{-1}_{11}\mathbf{W}_{N,1}$. This is precisely the asymptotic representation of the oracle estimator, which is an unpenalized CQR estimator fitted only on the true covariates in $\mathcal{A}$.

Therefore, we can apply the result of Theorem 1 to this sub-problem. The asymptotic distribution of $\sqrt{N}(\hat{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_1^*)$ is the same as that of an oracle CQR estimator:

$$\sqrt{N}(\hat{\boldsymbol{\gamma}}_{\mathcal{A}}^{\mathrm{ACQR}} - \boldsymbol{\gamma}_{\mathcal{A}}^*) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathrm{CQR\_oracle}}),$$

where

$$\boldsymbol{\Sigma}_{\mathrm{CQR\_oracle}} = \mathbf{C}_{\mathcal{A}\mathcal{A}}^{-1} \frac{\sum_{k=1}^{K} \sum_{k'=1}^{K} \left(\min(\tau_k, \tau_{k'}) - \tau_k \tau_{k'}\right)}{\left(\sum_{k=1}^{K} f(b_{\tau_k}^*)\right)^2},$$

and $\mathbf{C}_{\mathcal{A}\mathcal{A}}$ is the submatrix of $\mathbf{C}$ corresponding to the true active predictors. This completes the proof of the oracle properties. □

Proof of Corollary 1 and 2. We provide a detailed proof for both corollaries. First, we establish the $\sqrt{N}$-consistency of the HCQR estimator, which proves Corollary 1. Second, building upon this consistency, we outline the argument for asymptotic normality, proving Corollary 2.

This part establishes the root-N convergence rate for the HCQR estimator $\hat{\boldsymbol{\theta}} = ((\hat{\mathbf{b}})^\top, (\hat{\boldsymbol{\gamma}})^\top)^\top$. The core of the argument is to show that for any given $\varepsilon > 0$, there exists a sufficiently large constant C such that the probability of the estimator lying outside a ball of radius $C/\sqrt{N}$ around the true parameter $\boldsymbol{\theta}^*$ is less than ε.

Let us define a normalized perturbation vector $\mathbf{u} = \sqrt{N}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$. We consider the difference in the objective function when moving from the true parameter $\boldsymbol{\theta}^*$ to a perturbed point $\boldsymbol{\theta}^* + \mathbf{u}/\sqrt{N}$:

$$D_N(\mathbf{u}) = Q_N(\boldsymbol{\theta}^* + \mathbf{u}/\sqrt{N}) - Q_N(\boldsymbol{\theta}^*).$$

Since $\hat{\boldsymbol{\theta}}$ is the minimizer of $Q_N(\boldsymbol{\theta})$, we must have $D_N(\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)) \leq 0$. Our goal is to show that for any $\mathbf{u}$ on the surface of a sufficiently large sphere,

i.e., $\|\mathbf{u}\|_2 = C$, the value of $D_N(\mathbf{u})$ is strictly positive with high probability. This implies that the minimum must lie inside this sphere, thus establishing that $\|\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_2 = O_p(1)$.

The difference $D_N(\mathbf{u})$ can be decomposed into the change in the composite quantile loss, ΔL_N, and the change in the penalty function, ΔP_N. As detailed in the proof of Lemma 1, the loss component admits an asymptotic quadratic expansion:

$$\Delta L_N = -\mathbf{W}_N^\top \mathbf{u} + \frac{1}{2}\mathbf{u}^\top \mathbf{H}\mathbf{u} + o_p(1),$$

where $\mathbf{W}_N = O_p(1)$ is the normalized score vector and $\mathbf{H}$ is a positive definite deterministic matrix. The linear term is of order $O_p(\|\mathbf{u}\|_2)$, and the quadratic term is of order $O(\|\mathbf{u}\|_2^2)$.

For the penalty component, $\Delta P_N = \lambda_N \sum_{p=1}^{F} w_p(|\gamma_p^* + \mu_p/\sqrt{N}| - |\gamma_p^*|)$, where $\mathbf{u} = ((\boldsymbol{\nu})^\top, (\boldsymbol{\mu})^\top)^\top$. The penalty term is convex, so its change is bounded. Specifically, using the reverse triangle inequality, the change is non-negative for components where $\gamma_p^* = 0$. For components where $\gamma_p^* \neq 0$, the change is of order $O(\lambda_N N^{-1/2}|\mu_p|)$. Given the conditions on λ_N, the overall contribution from the penalty term is asymptotically negligible compared to the quadratic term in the loss function.

Combining the terms, for any $\mathbf{u}$ such that $\|\mathbf{u}\|_2 = C$:

$$D_N(\mathbf{u}) \geq -\|\mathbf{W}_N\|_2\|\mathbf{u}\|_2 + \frac{1}{2}\lambda_{\min}(\mathbf{H})\|\mathbf{u}\|_2^2 + o_p(1) = -O_p(1)\cdot C + \frac{1}{2}\lambda_{\min}(\mathbf{H})C^2 + o_p(1).$$

For any given realization of the $O_p(1)$ term, we can choose a constant C large enough such that the positive C^2 term dominates the linear C term, making $D_N(\mathbf{u}) > 0$. This implies that the minimizer $\hat{\mathbf{u}} = \sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ must satisfy $\|\hat{\mathbf{u}}\|_2 \leq C$ with high probability. Thus, $\|\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_2 = O_p(1)$, which is

equivalent to $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 = O_p(N^{-1/2})$. This establishes the $\sqrt{N}$-consistency for all parameters, thereby proving Corollary 1.

The asymptotic normality of the estimators for the active set coefficients, $\hat{\boldsymbol{\gamma}}_{\mathcal{A}}$, is a cornerstone of Theorem 1. We re-state the key arguments here to formally prove the corollary.

Theorem 1 has already established selection consistency, meaning that for large N, the estimator correctly identifies the active set $\mathcal{A}$, i.e., $\hat{\gamma}_p = 0$ for all $p \notin \mathcal{A}$. Consequently, the estimation problem for the non-zero coefficients $\hat{\boldsymbol{\gamma}}_{\mathcal{A}}$ is asymptotically equivalent to minimizing the objective function only over the active set $\mathcal{A}$:

$$\min_{\boldsymbol{\gamma}_{\mathcal{A}}} \sum_{k=1}^{K} \sum_{j,i} \rho_{\tau_k}(Y_{ij} - b_{\tau_k} - \mathbf{X}^*_{\mathcal{A},ij}\boldsymbol{\gamma}_{\mathcal{A}}) + \lambda_N \sum_{p \in \mathcal{A}} \frac{|\gamma_p|}{|\hat{\gamma}_p^{\mathrm{CQR}}|}.$$

As shown in the proof of Theorem 1, for any coefficient in the true active set ($p \in \mathcal{A}$), its corresponding penalty term vanishes asymptotically. This is because the adaptive weights $|\hat{\gamma}_p^{\mathrm{CQR}}|$ converge to a non-zero constant, while the tuning parameter $\lambda_N \to 0$. The penalty's contribution to the first-order conditions (and thus to the asymptotic distribution) is of order $O(\lambda_N/\sqrt{N})$, which goes to zero.

Therefore, the asymptotic behavior of $\hat{\boldsymbol{\gamma}}_{\mathcal{A}}$ is governed solely by the composite quantile loss function, making it equivalent to the oracle estimator. From the proof of Lemma 1, the minimizer of the quadratic approximation of the loss function gives the asymptotic representation:

$$\sqrt{N}(\hat{\boldsymbol{\gamma}}_{\mathcal{A}}^{\mathrm{ACQR}} - \boldsymbol{\gamma}^*_{\mathcal{A}}) = \mathbf{H}^{-1}_{\mathcal{A}\mathcal{A}}\mathbf{W}_{N,\mathcal{A}} + o_p(1),$$

where $\mathbf{W}_{N,\mathcal{A}}$ is the score vector for the active set, which converges to a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Omega}_{\mathcal{A}\mathcal{A}})$, and $\mathbf{H}_{\mathcal{A}\mathcal{A}}$ is the corresponding block

of the Hessian limit. As derived in Lemma 1, this leads to the asymptotic distribution:

$$\sqrt{N}(\hat{\boldsymbol{\gamma}}_{\mathcal{A}}^{\mathrm{ACQR}} - \boldsymbol{\gamma}_{\mathcal{A}}^{*}) \xrightarrow{d} N\big(\mathbf{0}, \boldsymbol{\Sigma}_{\mathrm{CQR\_oracle}}\big).$$

Corollary 2 follows directly from this multivariate result. By the properties of multivariate normal distributions, each marginal component of this vector is also asymptotically normal. Specifically, for any $p \in \mathcal{A}$,

$$\sqrt{N}(\hat{\gamma}_{p}^{\mathrm{ACQR}} - \gamma_{p}^{*}) \xrightarrow{d} N(0, \sigma_{p}^{2}),$$

where σ_p^2 is the p-th diagonal element of $\boldsymbol{\Sigma}_{\mathrm{CQR\_oracle}}$. □

References

[1] B. Van Dusen, J. Nissen, Modernizing use of regression models in physics education research: A review of hierarchical linear modeling, Physical Review Physics Education Research 15 (2019) 020108.

[2] P. Das, C. B. Peterson, Y. Ni, A. Reuben, J. Zhang, J. Zhang, K.-A. Do, V. Baladandayuthapani, Bayesian Hierarchical Quantile Regression with Application to Characterizing the Immune Architecture of Lung Cancer, Biometrics 79 (2023) 2474–2488.

[3] Y. Tian, L. Wang, M. Tang, M. Tian, Weighted composite quantile regression for longitudinal mixed effects models with application to AIDS studies, Communications in Statistics - Simulation and Computation 50 (2021) 1837–1853.

[4] L. M. Sullivan, K. A. Dukes, E. Losina, An introduction to hierarchical linear modelling, Statistics in Medicine 18 (1999) 855–888.

[5] S. W. Raudenbush, Hierarchical linear models: Applications and data analysis methods, Advanced Quantitative Techniques in the Social Sciences Series/SAGE, 2002.

[6] R. Koenker, K. F. Hallock, Quantile Regression, Journal of Economic Perspectives 15 (2001) 143–156.

[7] H. Zou, M. Yuan, Composite quantile regression and the oracle model selection theory, The Annals of Statistics 36 (2008).

[8] M. Tian, G. Chen, Hierarchical linear regression models for conditional quantiles, Science in China Series A: Mathematics 49 (2006) 1800–1815.

[9] Y. Luo, H. Lian, M. Tian, Bayesian quantile regression for longitudinal data models, Journal of Statistical Computation and Simulation 82 (2012) 1635–1649.

[10] Y. Yu, Bayesian quantile regression for hierarchical linear models, Journal of Statistical Computation and Simulation 85 (2015) 3451–3467.

[11] Y.-l. Chen, M.-z. Tian, K.-m. Yu, J.-x. Pan, Composite hierarchical linear quantile regression, Acta Mathematicae Applicatae Sinica, English Series 30 (2014) 49–64.

[12] H. Zou, The Adaptive Lasso and Its Oracle Properties, Journal of the American Statistical Association 101 (2006) 1418–1429.

[13] Y. Gu, J. Fan, L. Kong, S. Ma, H. Zou, ADMM for High-Dimensional Sparse Penalized Quantile Regression, Technometrics 60 (2018) 319–331.

[14] L. Yang, T. T. Wu, Model-Based Clustering of High-Dimensional Longitudinal Data via Regularization, Biometrics 79 (2023) 761–774.

[15] R. Koenker, Quantile regression, 1st Edition, Cambridge Univ Pr, Cambridge, 2005.

[16] J. Huang, C.-H. Zhang, Asymptotic oracle properties of sure independence screening, Annals of Statistics 36 (6) (2008) 2563–2587. doi:10.1214/08-AOS645.

[17] A. Belloni, V. Chernozhukov, ℓ_1-penalized quantile regression in high-dimensional sparse models, Annals of Statistics 39 (1) (2011) 82–130. doi:10.1214/10-AOS827.

[18] C.-H. Zhang, S. S. Zhang, Confidence intervals for low-dimensional parameters in high-dimensional linear models, Journal of the Royal Statistical Society: Series B 76 (1) (2014) 217–242. doi:10.1111/rssb.12026.

[19] R. Koenker, J. A. F. Machado, Goodness of fit and related inference processes for quantile regression, Journal of the Royal Statistical Society: Series B 61 (3) (1999) 659–700.

[20] D. Davis, W. Yin, A three-operator splitting scheme and its optimization applications, Set-Valued and Variational Analysis 25 (2017) 829–858. doi:10.1007/s11228-017-0424-z.

[21] A. Tsanas, M. A. Little, P. E. McSharry, L. O. Ramig, Accurate Telemonitoring of Parkinson's Disease Progression by Noninvasive Speech Tests, IEEE Transactions on Biomedical Engineering 57 (2010) 884–893.