THE GRADUATE UNIVERSITY FOR
ADVANCED STUDIES, SOKENDAI

DOCTORAL THESIS

Sparse Modal Regression with Skew Noises

Author:
Kazuki Koyama

Supervisor:
Dr. Hironori Fujisawa

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy

in the

Department of Statistical Science
School of Multidisciplinary Sciences

March, 2025

The Graduate University for Advanced Studies, SOKENDAI

Abstract

School of Multidisciplinary Sciences

Department of Statistical Science

Doctor of Philosophy

Sparse Modal Regression with Skew Noises

by Kazuki Koyama

Sparse regression methods have been widely used in many fields for their statistical effectiveness and high interpretability. However, there are few sparse regression methods with skew noise, although statistical modeling using skewness is becoming more important, e.g., in the medical field. Azzalini's skew-normal distribution and its extensions are well-used for skew noise. Such skew regression methods have a severe problem with statistical interpretability because they model neither mean, median, nor mode. To overcome this problem, we propose a novel sparse regression method based on mode-invariant skew-normal noise. The regression model is easy to interpret in the proposed method because it always models a mode regardless of skewness. The proposed method is simple to implement and optimize, suggesting it is highly scalable to other machine learning methods. We also provide theoretical guarantees of the proposed method for the average excess risk and the estimation error. Numerical experiments on artificial and real-world data demonstrate that the proposed method performs significantly better and is more stable than other existing methods for various skew-noise data.

Acknowledgements

I would like to express my heartfelt gratitude to my supervisor, Dr. Hironori Fujisawa, for his unwavering support throughout my doctoral journey. His thoughtful guidance has greatly contributed to my growth, both academically and personally. I am certain that I would not be where I am today without his mentorship. Once again, I offer my deepest thanks for all his contributions.

I would like to extend my sincere gratitude to the members of my doctoral committee, Dr. Hideitsu Hino, Dr. Keisuke Yano, and Dr. Toshihiro Abe. Their constructive comments have greatly enhanced the quality of my research. I deeply appreciate their time, effort, and expertise in evaluating my work.

To all the students and alumni of the Fujisawa Lab, I am deeply thankful for their support and advice. Their encouragement and fellowship, not only in research but also in various aspects, have been invaluable. I sincerely hope for the continued success of their research and careers.

To my colleagues and superiors in my previous job, I am deeply grateful for the opportunity to pursue my doctoral studies. Although I had to resign due to personal circumstances, the values we cultivated together on social issues still greatly influence my current research. My experience with this wonderful team is an irreplaceable asset in my career.

Lastly, I must express my heartfelt appreciation to my family, especially my wife, who has always been there and stood by me. Although our circumstances dramatically changed during my doctoral journey, my wife's steadfast belief in me has been a guiding light. Her unwavering support and love have been my constant source of strength. I am profoundly grateful for everything she has done.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Starting with the Lasso (Tibshirani, 1996), an ℓ_1 regularization regression method is becoming popular for predictive modeling with relevant features. It has gained significant attention in many fields because of its ability to handle high-dimensional data efficiently and its exhaustive studies of methodological extensions and theoretical properties. (For details, e.g., see Bühlmann and Geer (2011); Hastie, Tibshirani, and Wainwright (2015).) As we have more data available and need more reasonable models, an ℓ_1 regularization regression method will likely continue to be more essential for high-dimensional data analysis.

The Lasso is based on the squared error; in other words, it focuses on a symmetric noise. In real-world data analysis, however, we need to treat skewness. For example, statistical modeling using skewness is indispensable in the medical field. Hossain and Beyene (2015) explored an application of the skew-normal distribution in the analysis of microRNA (miRNA) data, which often does not follow a normal distribution. Their findings, derived from both simulations and real miRNA dataset analyses, demonstrated that the skew-normal distribution could enhance the detection of differentially expressed miRNAs. This enhancement is particularly noticeable when the data is significantly skewed. Shafiei, Safarpoor, Jamalizadeh, and Tizhoosh (2020) introduced an automated stain normalization framework for histopathology images that uses a mixture of multivariate skew-normal distributions to capture both symmetric and non-symmetric observations. The method can model multi-modal and multiple correlated distributions, regardless of their non-symmetry, and has shown consistent and superior performance compared to existing methods. How to treat skewness is also a topic of interest in other fields, including financial, meteorological, and industrial data (Adcock, Eling, and Loperfido, 2015; Hao, Yang, and Berenguer, 2019; Simola, Dumusque, and Cisewski-Kehe, 2019).

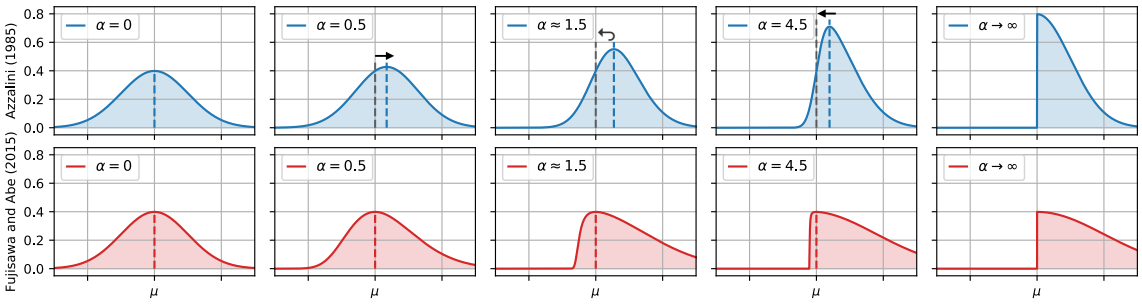

FIGURE 1.1: The difference between Azzalini's skew-normal distribution (upper row) and the mode-invariant skew-normal distribution (lower row) with the location parameter μ and the different skewness parameter α. The scale parameter is set to $\sigma = 1$. The blue and black dashed lines in the upper row represent the positions of the mode and μ in the former distribution, respectively. The mode of the former distribution depends complicatedly on the skewness parameter as well as the location and scale parameters. As α increases, the mode goes first to the right and then back to the left (to the original location). This behavior makes it difficult to interpret the former distribution. The red dashed line in the lower row represents the positions of both the mode and μ in the latter distribution. The mode of the latter distribution is invariant for any α. As $\alpha \to \infty$, both asymptotically approach half-normal distributions. The former is stretched vertically, but the latter is stretched horizontally.

Skew distributions have been studied extensively (Azzalini and Capitanio, 1999; Branco and Dey, 2001; Sahu, Dey, and Branco, 2003; Liseo and Loperfido, 2003; González-Farías, Dominguez-Molina, and Gupta, 2004; Arellano-Valle, Gómez, and Quintana, 2005; Arellano-Valle and Genton, 2005; Arellano-Valle and Azzalini, 2006). However, despite the intense need for skewness in real-world data analysis, there currently exist few relevant ℓ_1 regularization regression methods. The most significant study is Chen, Pourahmadi, and Maadooliat (2014), who assumed a regression model for the location parameter of the well-used Azzalini's skew-normal distribution (Azzalini, 1985) and then proposed an ℓ_1 regularization regression method with an efficient parameter estimation algorithm. Even if we move away from the ℓ_1 regularization, there exist few regularization methods; e.g., Cao, Wang, and Wu (2023) recently proposed the ridge regularization regression method for the mode of Azzalini's skew-normal distribution. This lack of attention may be because we would generally deal with skewed data using non-linear pre-processing with a power transform such as the Box–Cox transformation (Box and Cox, 1964) and Yeo–Johnson transformation (Yeo and Johnson, 2000). However, these non-linear transformations may not be suitable for a regression model with skewed noise because such transformations can provide symmetric distributions of outcomes but not symmetric noises due to asymmetric feature distributions. Furthermore, non-linear transformations can make it difficult to interpret a regression model since the non-linearity changes the meaning of an observation unit.

1.2 Our Contribution

In this thesis, we focus on a regression model whose noise is assumed to follow the mode-invariant skew-normal distribution introduced by Fujisawa and Abe (2015), and then we propose the regression estimator of the mode (location parameter) that is defined as the minimizer of the negative log-likelihood plus the ℓ_1 regularization, aiming to achieve both skewness modeling and statistical interpretability. The mode-invariant skew-normal distribution is a subclass of the Transformation-of-Scale (ToS) distribution (Jones, 2014; Jones, 2016) and is consistent with a normal distribution when not skewed. As the name implies, its mode is invariant regardless of skewness, where its location parameter coincides with its mode. Therefore, the proposed method always models a mode regardless of skewness, making it easy to interpret the statistical role of sparse features. This is equivalent to the reasonable assumption that a noiseless situation has the highest probability. Despite its helpful properties, there are few studies with the mode-invariant skew-normal distribution besides the basic maximum likelihood estimation (Fujisawa and Abe, 2015).

The proposed method differs significantly from Chen, Pourahmadi, and Maadooliat (2014)'s regression model for the location parameter of Azzalini's skew-normal distribution. Figure 1.1 illustrates different shapes between Azzalini's skew-normal distribution and the mode-invariant skew-normal distribution. Note that the location parameter of Azzalini's skew-normal distribution is neither mean, median, nor mode, which means that the role of Chen, Pourahmadi, and Maadooliat (2014)'s regression model is uncertain and difficult to interpret.

Our main contributions are as follows:

- Highlighting the attractiveness of the mode-invariant skew-normal distribution and assuming noise to follow it, we propose a novel ℓ_1 regularization regression method having both skewness modeling and statistical interpretability.
- Revealing partial convexity, we indicate that its optimization and implementation are relatively straightforward, although the proposed method involves complicated non-linear transformations.
- The proposed method is theoretically guaranteed in terms of its excess risk and estimation error by non-asymptotic analysis under mild assumptions usually adopted in the theoretical analysis of Lasso-type problems.

1.3 Outline

This thesis is organized as follows. In Chapter 2, we review sparse regression methods and skew distributions. In Chapter 3, the proposed method is introduced. In Chapter 4, the theoretical guarantees of the proposed method are provided by non-asymptotic analysis in terms of the excess risk and the estimation error. Such theoretical studies for skew regression models have not been shown so far. In Chapter 5, numerical experiments on artificial and real-world data are demonstrated. The results obtained in that study indicate that the proposed method can serve as a viable alternative to the Lasso. In Chapter 6, concluding remarks are given.

The main results of this thesis are based on the following journal paper:

- Kazuki Koyama, Takayuki Kawashima, and Hironori Fujisawa. "Sparse modal regression with mode-invariant skew noise". *Transactions on Machine Learning Research*, 2024.

Chapter 2

Preliminary

2.1 Sparse Regression

2.1.1 Problem Description

Let $\mathcal{X} \subset \mathbb{R}^P$ be the domain of input vector $\boldsymbol{X}$. Let $\mathcal{Y} \subset \mathbb{R}$ be the domain of output sample y. Suppose that we are given N independent and identically distributed (i.i.d.) samples $\mathcal{D}_N := \{(y_n, \boldsymbol{X}_n) \in \mathcal{Y} \times \mathcal{X}\}_{n=1}^N$.

We consider the following linear regression problem for an output $y \in \mathcal{Y}$ and a P-dimensional feature vector $\boldsymbol{X} \in \mathcal{X}$:

$$y = \boldsymbol{X}^\top \boldsymbol{\beta} + e, \tag{2.1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^P$ is a parameter vector and $e \in \mathbb{R}$ is a noise.

Let $\mathbf{y} := [y_1, \ldots, y_N]^\top$ and $\mathbf{X} := [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N]^\top$. In this thesis, unless otherwise specified, we assume that $\mathbf{X}$ is standardized such that $\frac{1}{N}\sum_{n=1}^N \mathbf{X}_{np}^2 = 1$ and that $\mathbf{y}$ and $\mathbf{X}$ are centralized such that $\sum_{n=1}^N y_n = 0$ and $\sum_{n=1}^N \mathbf{X}_{np} = 0$. This standardization for $\boldsymbol{X}$ ensures that each feature variable contributes equally to the output and prevents feature variables with larger scales from dominating the output. In addition, this centralization for y and $\boldsymbol{X}$ allows us to treat the intercept term as zero in regression models.

The most well-known approach to solving (2.1) is the least-squares method. In this method, we minimize the squared loss

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} := \operatorname*{argmin}_{\boldsymbol{\beta}} \frac{1}{2N} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2. \tag{2.2}$$

If $N > P$ and $\text{rank}(\mathbf{X}) = P$, we can analytically obtain the optimal solution of (2.2) as

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \tag{2.3}$$

Let $\boldsymbol{\beta}^0$ be the true parameter. Suppose that e follows the normal distribution and that the variance is known. In this case, the following excess risk of the squared-error estimator follows the chi-square distribution χ^2_P with P degrees of freedom (Montgomery, Peck, and Vining, 2021):

$$\frac{1}{\sigma^2} \|\mathbf{X}(\hat{\boldsymbol{\beta}}_{\text{OLS}} - \boldsymbol{\beta}^0)\|_2^2 \sim \chi^2_P. \tag{2.4}$$

Hence, we can obtain the following evaluation for (2.3):

$$\frac{1}{2N} \mathbb{E}\left[\|\mathbf{X}(\hat{\boldsymbol{\beta}}_{\text{OLS}} - \boldsymbol{\beta}^0)\|_2^2\right] = O\left(\frac{P}{N}\right). \tag{2.5}$$

Let $\mathcal{S}_0 := \{p : \beta_p^0 \neq 0\}$ and $\mathcal{S}_0^C := \{1, \ldots, P\} \backslash \mathcal{S}_0$. Suppose $s_0 < N$ with $s_0 := \#\mathcal{S}_0$. In contrast to the case where $N > P$, if $N < P$, we always have $\text{rank}(\mathbf{X}) \neq P$. Therefore, the optimal solution of (2.2) is not unique since $\mathbf{X}^\top \mathbf{X}$ is a singular matrix. If we completely know $\mathcal{S}_0$, we can employ a strategy to optimize (2.2) using the least-squares method with only the features corresponding to $\mathcal{S}_0$, such as

$$\hat{\boldsymbol{\beta}} := \begin{cases} \underset{\boldsymbol{\beta} \in \mathbb{R}^{s_0}}{\text{argmin}} \frac{1}{2N} \|\mathbf{y} - \mathbf{X}_{\mathcal{S}_0} \boldsymbol{\beta}\|_2^2 & (p \in \mathcal{S}_0) \\ \mathbf{0} & (p \in \mathcal{S}_0^C) \end{cases}. \tag{2.6}$$

In this case, following the same theory in this section, we have

$$\frac{1}{2N} \mathbb{E}\left[\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2\right] = O\left(\frac{s_0}{N}\right). \tag{2.7}$$

However, we typically can not know $\mathcal{S}_0$. Therefore, we need other techniques to estimate $\mathcal{S}_0$ from $\mathcal{D}_N$ directly beyond the least-squares method.

2.1.2 The Lasso

The Lasso (or the Least Absolute Shrinkage and Selection Operator) (Tibshirani, 1996) is a regression method to estimate a sparse model. Moreover, it helps in both variable selection and regularization and enhances the prediction accuracy

and interpretability of the statistical model. It is widely used in statistics and machine learning, in particular, for high-dimensional ($N < P$) data.

The Lasso adds an ℓ_1 regularization term to the squared loss. In response to (2.2), it optimizes

$$\hat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2N} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1, \tag{2.8}$$

where $\lambda \geq 0$ is a regularization parameter. In this way, the ℓ_1 regularization term can be applied to various regression models, including linear regression, logistic regression, and generalized linear models.

The Lasso offers several attractive properties. This section briefly describes some of them. In Sections 2.1.3 and 2.1.4, we will introduce theoretical guarantees for the excess risk and estimation error in detail.

Sparsity

The ℓ_1 regularization term in the Lasso has the effect of shrinking some of the estimated coefficients exactly to zero when λ is sufficiently large. This property makes the Lasso particularly useful for variable selection in high-dimensional settings.

Consider the mechanism for achieving a sparse optimal solution. We can replace the optimization problem (2.8) with the following equivalent optimization problem:

$$\hat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2N} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \quad \text{s.t. } \|\boldsymbol{\beta}\|_1 \leq t, \tag{2.9}$$

where t is some positive constant which depends on λ. The contours of the residual sum of squares are ellipsoids centered around the least squares solution. The ℓ_1 constraint forms a diamond shape in the coefficient space. Therefore, the optimal solution occurs at the intersection of the ellipsoid contours and the diamond-shaped constraint region. Due to the shape of the ℓ_1 norm constraint, the intersection often appears on the axes, which leads to some coefficients being exactly zero. Hence, the Lasso achieves sparsity (Bishop, 2006; Hastie, Tibshirani, and Friedman, 2009).

Interpretability

A simpler model with fewer variables is easier to interpret and understand. This is particularly important in fields where understanding the relationship between

variables is crucial. For example, in the medical field, identifying key biomarkers can lead to better diagnostic tools. In the financial field, understanding the impact of specific factors on economic outcomes can help inform policy decisions. In this sense, the ℓ_1 regularization term leads to high interpretability since it can produce sparse solutions.

Moreover, the ℓ_1 regularization term can help to handle multicollinearity. When feature variables are highly correlated, it tends to select one variable from a group of correlated variables and shrink the others to zero. This process not only reduces the complexity of the model but also enhances its interpretability by retaining only the most significant predictors.

In related studies, for example, the Independently Interpretable Lasso (IILasso) method (Takada, Suzuki, and Fujisawa, 2018; Takada, Suzuki, and Fujisawa, 2020) further improves interpretability by ensuring that each selected variable independently affects the objective variable. In addition, a critical review highlights the limitations and practical challenges of the Lasso when there are dependence structures among covariates (Freijeiro-González, Febrero-Bande, and González-Manteiga, 2022). This review provides insights into various derivatives and alternatives of the Lasso that can be more effective in different scenarios. These advancements in the Lasso methodologies underscore the importance of selecting appropriate regularization techniques to enhance model interpretability and performance.

Variable Selection

The *irrepresentable condition* is key to ensuring that the Lasso can consistently select the true model (Meinshausen and Yu, 2009; Bühlmann and Geer, 2011; Hastie, Tibshirani, and Wainwright, 2015). It involves the covariance structure of the predictors. Formally, the irrepresentable condition can be stated as follows:

$$\|\mathbf{X}_{\mathcal{S}_0^C}^\top \mathbf{X}_{\mathcal{S}_0} (\mathbf{X}_{\mathcal{S}_0}^\top \mathbf{X}_{\mathcal{S}_0})^{-1} \mathrm{sgn}(\boldsymbol{\beta}_{\mathcal{S}_0})\|_\infty \leq 1 - c_1, \tag{2.10}$$

where $c_1 \in (0, 1]$. This condition essentially means that the features not in the true model $\mathbf{X}_{\mathcal{S}_0^C}$ should not be too correlated with those in the true model $\mathbf{X}_{\mathcal{S}_0}$.

Another key is the *beta-min condition* (Leeb and Pötscher, 2003; Leeb and Pötscher, 2005; Bühlmann and Geer, 2011). It focuses on the magnitude of the true coefficients $\boldsymbol{\beta}$. The beta-min condition can be stated as follows:

$$\min_{j \in \mathcal{S}_0} |\beta_j^0| > c_2 \sqrt{\frac{\log P}{N}}, \tag{2.11}$$

for some positive constant c_2. This condition ensures that the true non-zero coefficients are large enough to be distinguished from zero, even in noise.

These conditions provide a framework for the consistency of the Lasso in variable selection. The irrepresentable condition ensures that the model structure is identifiable, while the beta-min condition ensures that the signal is strong enough to be detected. Both conditions must be satisfied so that the Lasso can reliably select the correct model (Zhao and Yu, 2006; Lahiri, 2021).

Prediction Accuracy

The Lasso improves prediction accuracy by balancing the trade-off between bias and variance through regularization. The regularization term in the objective function helps to reduce the variance of the model by shrinking some coefficients to zero. This can be understood through the following bias-variance trade-off (Hastie, Tibshirani, and Friedman, 2009; James, Witten, Hastie, and Tibshirani, 2013):

- **Bias**: Regularization introduces some bias into the model by shrinking the coefficients.
- **Variance**: By reducing the number of predictors (or setting some coefficients to zero), the Lasso reduces the model's complexity, which in turn reduces variance.

Moreover, the regularization term also helps to prevent overfitting by discouraging the model from becoming too complex. This leads to more robust models which better generalize to new data. The optimal value of λ is typically chosen through cross-validation.

2.1.3 Convex loss and ℓ_1 regularization

This section reviews the excess risk and the estimation error for general convex loss and ℓ_1 regularization. In particular, we focus on non-asymptotic analysis. As shown in Section 2.1.1, the order in (2.7) can be regarded as the excess risk of the Lasso under an ideal situation where we completely know $\mathcal{S}_0$. However, since such a situation usually does not occur, we need to consider the theoretical guarantees described here.

Let $f : \mathcal{X} \to \mathbb{R}$ be a regression function. Let $\ell_f(y, \boldsymbol{X}) : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$ be a loss function for f. For example, the squared loss corresponds to $\ell_f(y, \boldsymbol{X}) := (y - f(\boldsymbol{X}))^2$.

Let $\mathcal{F}$ be the set of all regression functions f such that the map $f \to \ell_f(y, \boldsymbol{X})$ is convex for any $(y, \boldsymbol{X})$. We define the true regression function f^0 as the minimizer of the theoretical risk

$$f^0 := \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{N} \sum_{(y_n, \boldsymbol{X}_n) \in \mathcal{D}_N} \mathbb{E}\left[\ell_f(y_n, \boldsymbol{X}_n)\right]. \tag{2.12}$$

Moreover, for any $f \in \mathcal{F}$, let the excess risk between f and f^0 be denoted by

$$\mathcal{E}(f \mid f^0) := \frac{1}{N} \sum_{(y_n, \boldsymbol{X}_n) \in \mathcal{D}_N} \left(\mathbb{E}\left[\ell_f(y_n, \boldsymbol{X}_n)\right] - \mathbb{E}\left[\ell_{f^0}(y_n, \boldsymbol{X}_n)\right]\right). \tag{2.13}$$

We focus on a linear regression function $f_{\boldsymbol{\beta}}(\boldsymbol{X}) := \boldsymbol{X}^\top \boldsymbol{\beta}$. The best linear approximation of the true regression function f^0 is given by

$$\boldsymbol{\beta}^0_{\text{linear}} := \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{N} \sum_{(y_n, \boldsymbol{X}_n) \in \mathcal{D}_N} \mathbb{E}\left[\ell_{f_{\boldsymbol{\beta}}}(y_n, \boldsymbol{X}_n)\right]. \tag{2.14}$$

The Lasso corresponds to solving the following optimization problem:

$$\hat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \left(\frac{1}{N} \sum_{(y_n, \boldsymbol{X}_n) \in \mathcal{D}_N} \ell_{f_{\boldsymbol{\beta}}}(y_n, \boldsymbol{X}_n) + \lambda \|\boldsymbol{\beta}\|_1 \right). \tag{2.15}$$

In this section, we are interested in the excess risk $\mathcal{E}(f_{\hat{\boldsymbol{\beta}}} \mid f^0)$ and the estimation error of $\hat{\boldsymbol{\beta}}$. To evaluate them, we need to introduce the *margin condition* and the *restricted eigenvalue condition*. We define these conditions sequentially.

Definition 1 (Margin Condition)**.** *Let $\mathcal{F}_{local} := \{\|f - f^0\|_\infty \leq \eta\} \subset \mathcal{F}$. We say that the margin condition holds with strictly convex function G, if for any $f \in \mathcal{F}_{local}$, we have*

$$\mathcal{E}(f \mid f^0) \geq G(\|f - f^0\|). \tag{2.16}$$

In particular, we say that the quadratic margin condition holds if $G(u) = cu^2$ with some positive constant c.

The margin condition evaluates the behavior of the theoretical risk near its minimizer (Bühlmann and Geer, 2011; Hastie, Tibshirani, and Wainwright, 2015). It means that the excess risk $\mathcal{E}(f \mid f^0)$ in the neighborhood $\mathcal{F}_{\text{local}}$ of f^0 is bounded below by a strictly convex function. This condition holds in many cases. For example, in the case of the squared loss with a fixed design, the quadratic margin condition holds.

Definition 2 (Restricted Eigenvalue Condition). *Let* $S_N := \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{X}_n \boldsymbol{X}_n^\top$. *We say that the* $(L, \mathcal{S})$*-restricted eigenvalue condition holds with* $L > 1$*, if*

$$\tilde{\kappa}^2(L, \mathcal{S}) := \inf_{\|\boldsymbol{\beta}_{\mathcal{S}^C}\|_1 \leq L\|\boldsymbol{\beta}_{\mathcal{S}}\|_1} \frac{\boldsymbol{\beta}^\top S_N \boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_2^2} > 0. \tag{2.17}$$

The restricted eigenvalue condition is crucial in analyzing the ℓ_1 penalized method, particularly in high-dimensional statistics. (Bickel, Ritov, and Tsybakov, 2009; Bühlmann and Geer, 2011; Hastie, Tibshirani, and Wainwright, 2015). For example, in the typical low-dimensional ($N > P$) setting, the least-squares loss is strongly convex since $\mathbf{X}^\top \mathbf{X}$ is positive definite and its smallest eigenvalue is not zero. However, in the high-dimensional ($N < P$) setting, $\mathbf{X}^\top \mathbf{X}$ is always rank-deficient. For this reason, we need to relax the strong convexity of the loss function.

If the parameter vector $\boldsymbol{\beta}^0_{\text{linear}}$ is sparse, it is known that for appropriate choices of λ, the error $\tilde{\boldsymbol{\beta}} := \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0_{\text{linear}}$ satisfies a cone constraint such that $\|\tilde{\boldsymbol{\beta}}_{\mathcal{S}^C}\|_1 \leq L\|\tilde{\boldsymbol{\beta}}_{\mathcal{S}}\|_1$ (Hastie, Tibshirani, and Wainwright, 2015). This condition ensures the imposition of strong convexity for this restricted set $\|\boldsymbol{\beta}_{\mathcal{S}^C}\|_1 \leq L\|\boldsymbol{\beta}_{\mathcal{S}}\|_1$.

We set $L = 3$ in this section, but this constant is quite arbitrary. We can replace it with any constant greater than one by adjusting other constants, which influences, for example, how to determine the lower bound for λ (Bühlmann and Geer, 2011).

Instead of the restricted eigenvalue condition, the following *compatibility condition* may be assumed (Geer and Bühlmann, 2009; Bühlmann and Geer, 2011). This condition is sufficient for the restricted eigenvalue condition.

Definition 3 (Compatibility Condition). *Let* $s := \#\mathcal{S}$. *We say that the* $(L, \mathcal{S})$*-compatibility condition holds with constant* $\tilde{\kappa}(L, \mathcal{S}) > 0$*, if for any* $\boldsymbol{\beta} \in \mathbb{R}^P$ *satisfying* $\|\boldsymbol{\beta}_{\mathcal{S}^C}\|_1 \leq L\|\boldsymbol{\beta}_{\mathcal{S}}\|_1$*, we have*

$$\|\boldsymbol{\beta}_{\mathcal{S}}\|_1^2 \leq \frac{s\boldsymbol{\beta}^\top S_N \boldsymbol{\beta}}{\tilde{\kappa}^2(L, \mathcal{S})}. \tag{2.18}$$

Based on the margin condition and the restricted eigenvalue condition, we have the following theorem, which leads to evaluating the excess risk and the estimation error (Bühlmann and Geer, 2011).

Theorem 1. *Fix $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$. Let the empirical process of the loss function ℓ_{f_β} be denoted by*

$$\mathcal{V}_N(\boldsymbol{\beta}) := \frac{1}{N} \sum_{(y_n, \boldsymbol{X}_n) \in \mathcal{D}_N} \left(\ell_{f_\beta}(y_n, \boldsymbol{X}_n) - \mathbb{E}\left[\ell_{f_\beta}(y_n, \boldsymbol{X}_n) \right] \right). \tag{2.19}$$

Let $\mathcal{S}_0 := \{p : \beta^0_{linear,p} \neq 0\}$ and $s_0 := \#\mathcal{S}_0$. Suppose that the $(3, \mathcal{S}_0)$-restricted eigenvalue condition holds for $\mathcal{S}_0$. Suppose that the margin condition holds with strictly convex function G. Let H be the convex conjugate of G. Let

$$\epsilon_0 := \frac{3}{2} \mathcal{E}(f_{\beta^0_{linear}} \mid f^0) + H\left(\frac{4\lambda\sqrt{s_0}}{\tilde{\kappa}(3, \mathcal{S}_0)} \right). \tag{2.20}$$

Fix $\lambda_0 > 0$. Suppose $f_{\beta^0_{linear}} \in \mathcal{F}_{local}$. Suppose $f_\beta \in \mathcal{F}_{local}$ for any $\|\boldsymbol{\beta} - \boldsymbol{\beta}^0_{linear}\|_1 \leq \frac{\epsilon_0}{\lambda_0}$. Let

$$\mathcal{J} := \left\{ \sup_{\|\boldsymbol{\beta} - \boldsymbol{\beta}^0_{linear}\|_1 \leq \frac{\epsilon_0}{\lambda_0}} |\mathcal{V}_N(\boldsymbol{\beta}) - \mathcal{V}_N(\boldsymbol{\beta}^0_{linear})| \leq \epsilon_0 \right\}. \tag{2.21}$$

If the regularization parameter λ is chosen to satisfy $\lambda \geq 8\lambda_0$, then we have on $\mathcal{J}$,

$$\begin{aligned} \mathcal{E}(f_{\hat{\boldsymbol{\beta}}} \mid f^0) + \lambda \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0_{linear}\|_1 &\leq 6\mathcal{E}(f_{\beta^0_{linear}} \mid f^0) + 4H\left(\frac{4\lambda\sqrt{s_0}}{\tilde{\kappa}(3, \mathcal{S}_0)} \right) \\ &\left(= 4\epsilon_0 \right). \end{aligned} \tag{2.22}$$

In particular, if the quadratic margin condition holds with $G(u) = cu^2$, then we have

$$\mathcal{E}(f_{\hat{\boldsymbol{\beta}}} \mid f^0) + \lambda \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0_{linear}\|_1 \leq 6\mathcal{E}(f_{\beta^0_{linear}} \mid f^0) + \frac{16\lambda^2 s_0}{c\tilde{\kappa}^2(3, \mathcal{S}_0)}. \tag{2.23}$$

For practical purposes, we need to choose λ_0 so that the event $\mathcal{J}$ occurs with high probability. For this λ_0, we generally can take $\lambda_0 = O\left(\sqrt{\frac{\log P}{N}} \right)$. If we take $\lambda_0 = O\left(\sqrt{\frac{\log P}{N}} \right)$ and assume that the quadratic margin condition and the restricted eigenvalue condition hold, the estimation error $H\left(\frac{4\lambda\sqrt{s_0}}{\tilde{\kappa}(3, \mathcal{S}_0)} \right)$ behaves $O\left(\frac{s_0 \log P}{N} \right)$.

As an example of this case, we consider a typical situation corresponding to the ordinary Lasso. The ordinary Lasso has the squared loss with a fixed design, and the quadratic margin condition holds. Moreover, if the true regression function f^0 is linear, the approximation error $\mathcal{E}(f_{\beta^0_{\text{linear}}} \mid f^0)$ is zero since $f^0 = f_{\beta^0_{\text{linear}}}$. In this case, Theorem 1 gives a bound for both the excess risk $\mathcal{E}(f_{\hat{\boldsymbol{\beta}}} \mid f^0)$ and the

estimation error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0_{\text{linear}}\|_1$ as follows (Bühlmann and Geer, 2011).

Corollary 1. *Suppose the conditions of Theorem 1 with quadratic margin behavior. Suppose that the true regression function f^0 is linear. If we take $\lambda_0 = O\left(\sqrt{\frac{\log P}{N}}\right)$, then we have on $\mathcal{J}$,*

$$\mathcal{E}(f_{\hat{\boldsymbol{\beta}}} \mid f^0) = O_{\mathbb{P}}\left(\frac{s_0 \log P}{N}\right), \tag{2.24}$$

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0_{linear}\|_1 = O_{\mathbb{P}}\left(s_0\sqrt{\frac{\log P}{N}}\right). \tag{2.25}$$

From (2.7), we see that even if we completely knew $\mathcal{S}_0$, we could not take the order of the excess risk smaller than $O(\frac{s_0}{N})$. Corollary 1 shows that the excess risk (2.24) of the ordinary Lasso is only $O(\log P)$ worse than such an ideal situation.

2.1.4 Non-convex loss and ℓ_1 regularization

In this section, we extend convex loss treated in Section 2.1.3 to non-convex loss. In particular, we address ℓ_1 penalized likelihood problems. In such non-convex optimization problems, it is difficult to compute a global optimum. On the other hand, as shown in this section, the non-asymptotic theory is given for estimators defined by a global optimum.

We consider ℓ_1 penalized smooth likelihood problems, which are generally non-convex. The proposed method in this thesis belongs to this class, as described later. Suppose that the parameter space Ψ is a bounded subset of $\mathbb{R}^{P_\psi}$. Suppose that the true conditional probability density of y given $\boldsymbol{X}$ depends on $\boldsymbol{X}$ only through the true parameter function $\boldsymbol{\psi}^0(\boldsymbol{X}) \in \Psi$.

Let the probability density of y parametrized by $\boldsymbol{\psi} \in \Psi$ be denoted by $f_{\boldsymbol{\psi}}(y)$. With minus the log-likelihood as a loss function, we define the excess risk between $f_{\boldsymbol{\psi}}$ and $f_{\boldsymbol{\psi}^0}$ by

$$\begin{aligned}\mathcal{E}(\boldsymbol{\psi} \mid \boldsymbol{\psi}^0) :=& \mathbb{E}_{f_{\boldsymbol{\psi}^0}}\left[-\log f_{\boldsymbol{\psi}}(y)\right] - \mathbb{E}_{f_{\boldsymbol{\psi}^0}}\left[-\log f_{\boldsymbol{\psi}^0}(y)\right] \\ =& \mathbb{E}_{f_{\boldsymbol{\psi}^0}}\left[\log \frac{f_{\boldsymbol{\psi}^0}(y)}{f_{\boldsymbol{\psi}}(y)}\right].\end{aligned} \tag{2.26}$$

This excess risk is equivalent to the Kullback–Leibler divergence (Kullback and Leibler, 1951) between $f_{\boldsymbol{\psi}}$ and $f_{\boldsymbol{\psi}^0}$. For fixed $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$, we define the average

excess (Kullback–Leibler) risk between $f_{\boldsymbol{\psi}}$ and $f_{\boldsymbol{\psi}^0}$ by

$$\bar{\mathcal{E}}(\boldsymbol{\psi} \mid \boldsymbol{\psi}^0) := \frac{1}{N} \sum_{n=1}^{N} \mathcal{E}(\boldsymbol{\psi}(\boldsymbol{X}_n) \mid \boldsymbol{\psi}^0(\boldsymbol{X}_n)). \tag{2.27}$$

As shown in Section 2.1.3, the margin condition is key to evaluating the excess risk and the estimation error. In ℓ_1 penalized smooth likelihood problems, we can prove that the quadratic margin condition holds if the following three conditions hold (Bühlmann and Geer, 2011).

Condition 1. *Let $l_{\boldsymbol{\psi}}(y) := \log f_{\boldsymbol{\psi}}(y)$. There exists some function $G_3(y)$ such that*

$$\sup_{\boldsymbol{\psi} \in \Psi} \max_{(i_1,i_2,i_3) \in \{1,\dots,P_\psi\}^3} \left| \frac{\partial^3}{\partial \psi_1 \partial \psi_2 \partial \psi_3} l_{\boldsymbol{\psi}}(y) \right| \leq G_3(y), \tag{2.28}$$

$$\sup_{\boldsymbol{X} \in \mathcal{X}} \int G_3(y) f_{\boldsymbol{\psi}^0(\boldsymbol{X})}(y) dy < \infty. \tag{2.29}$$

Condition 2. *Let the Fisher information matrix of $f_{\boldsymbol{\psi}(\boldsymbol{X})}(y)$ be denoted by $\mathcal{I}(\boldsymbol{\psi}(\boldsymbol{X}))$. It holds that*

$$\inf_{\boldsymbol{X} \in \mathcal{X}} \Lambda_{\min}(\mathcal{I}(\boldsymbol{\psi}(\boldsymbol{X}))) > 0, \tag{2.30}$$

where $\Lambda_{\min}(A)$ is the smallest eigenvalue of A.

Condition 3. *For any $\epsilon > 0$, there exists some constant $\delta_\epsilon > 0$ such that*

$$\inf_{\boldsymbol{X} \in \mathcal{X}} \inf_{\|\boldsymbol{\psi} - \boldsymbol{\psi}^0(\boldsymbol{X})\|_2 > \epsilon} \mathcal{E}(\boldsymbol{\psi} \mid \boldsymbol{\psi}^0(\boldsymbol{X})) \geq \delta_\epsilon. \tag{2.31}$$

Theorem 2. *If Conditions 1, 2, and 3 hold, the quadratic margin condition holds such that*

$$\inf_{\boldsymbol{X} \in \mathcal{X}} \mathcal{E}(\boldsymbol{\psi} \mid \boldsymbol{\psi}^0(\boldsymbol{X})) \geq \frac{1}{c^2} \|\boldsymbol{\psi} - \boldsymbol{\psi}^0(\boldsymbol{X})\|_2^2, \tag{2.32}$$

with some positive constant c.

We now focus on the special case where the true conditional probability density of y given $\boldsymbol{X}$ linearly depends on $\boldsymbol{X}$. Let $\boldsymbol{\psi}^0(\boldsymbol{X})$ be composed of $\boldsymbol{\beta}^0 \in \mathbb{R}^P$ and $\boldsymbol{\xi}^0 \in \mathbb{R}^{P_\xi}$ satisfying $\boldsymbol{\psi}^0(\boldsymbol{X}) = [\boldsymbol{X}^\top \boldsymbol{\beta}^0, \boldsymbol{\xi}^{0\top}]^\top$. We assume that $\boldsymbol{\beta}^0$ is sparse. Then, we define a new parameter vector $\boldsymbol{\theta} := [\boldsymbol{\beta}^\top, \boldsymbol{\xi}^\top]^\top$ and assume that its parameter space Θ is a bounded subset of $\mathbb{R}^{P+P_\xi}$.

Letting $f_{\boldsymbol{\theta}}(y \mid \boldsymbol{X}) := f_{\boldsymbol{\psi}(\boldsymbol{X})}(y)$, we consider the following ℓ_1 penalized optimization problem:

$$\hat{\boldsymbol{\theta}} := \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} -\frac{1}{N}\sum_{n=1}^{N} \log f_{\boldsymbol{\theta}}(y_n \mid \boldsymbol{X}_n) + \lambda\|\boldsymbol{\beta}\|_1, \tag{2.33}$$

where $\lambda \geq 0$ is a regularization parameter. For this optimal solution $\hat{\boldsymbol{\theta}}$, we have the following theorem, which leads to evaluating the average excess risk and the estimation error (Bühlmann and Geer, 2011).

Theorem 3. *Suppose that Conditions 1, 2, and 3 hold. Let* $\mathcal{S}_0 := \{p : \beta_p^0 \neq 0\}$ *and* $s_0 := \#\mathcal{S}_0$. *Suppose that the* $(6, \mathcal{S}_0)$*-restricted eigenvalue condition holds. Fix* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$. *Let the average excess risk and the empirical process for* $\boldsymbol{\theta} \in \Theta$ *be denoted by*

$$\bar{\mathcal{E}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^0) := \frac{1}{N}\sum_{n=1}^{N} \mathcal{E}(\boldsymbol{\psi}(\boldsymbol{X}_n) \mid \boldsymbol{\psi}^0(\boldsymbol{X}_n)), \tag{2.34}$$

$$\mathcal{V}_N(\boldsymbol{\theta}) := \frac{1}{N}\sum_{n=1}^{N} \left(-\log f_{\boldsymbol{\theta}}(y_n \mid \boldsymbol{X}_n) - \mathbb{E}\left[-\log f_{\boldsymbol{\theta}}(y_n \mid \boldsymbol{X}_n)\right]\right). \tag{2.35}$$

Let $\|\boldsymbol{\theta}\|_* := \|\boldsymbol{\beta}\|_1 + \|\boldsymbol{\xi}\|_2$. *Fix* $\gamma \geq 1$ *and* $\lambda_0 > 0$. *Let*

$$\mathcal{J} := \left\{ \sup_{\boldsymbol{\theta}\in\Theta} \frac{\left|\mathcal{V}_N(\boldsymbol{\theta}) - \mathcal{V}_N(\boldsymbol{\theta}^0)\right|}{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\|_* \vee \lambda_0} \leq \gamma\lambda_0 \right\}. \tag{2.36}$$

If the regularization parameter λ *is chosen to satisfy* $\lambda \geq 2\gamma\lambda_0$, *then we have on* $\mathcal{J}$

$$\bar{\mathcal{E}}(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta}^0) + 2(\lambda - \gamma\lambda_0)\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 \leq 9(\lambda + \gamma\lambda_0)^2 c^2 \kappa^2 s_0, \tag{2.37}$$

where c *is the constant defined in Theorem 2,* $\kappa^{-1} := \tilde{\kappa}(6, \mathcal{S}_0)$.

We also need to choose λ_0 so that the event $\mathcal{J}$ occurs with high probability. For this λ_0, we generally can calculate $\lambda_0 = O\left(M_N \log N \sqrt{\frac{\log(N\vee P)}{N}}\right)$, where the rate for M_N depends on the model of interest. For example, for finite mixture regression models with ℓ_1 regularization, the rate is $M_N = O(\sqrt{\log N})$ (Städler, Bühlmann, and Geer, 2010), and for linear mixed effects models with ℓ_1 regularization, the rate is $M_N = O(\log N)$ (Schelldorfer, Bühlmann, and Geer, 2011).

If we take $\lambda_0 = O\left(M_N \log N \sqrt{\frac{\log(N\vee P)}{N}}\right)$, we can obtain convergence rates regarding the average excess risk $\bar{\mathcal{E}}(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta}^0)$ and the estimation error of $\hat{\boldsymbol{\beta}}$ from Theorem 3 (Bühlmann and Geer, 2011).

Corollary 2. *Suppose the conditions in Theorem 3. Suppose* $\lambda = 2\gamma\lambda_0$ *with* $\lambda = O\left(M_N \log N \sqrt{\frac{\log(N\vee P)}{N}}\right)$. *Then, we have*

$$\bar{\mathcal{E}}(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta}^0) = O_{\mathbb{P}}\left(\frac{s_0 M_N^2 (\log N)^2 \log(N \vee P)}{N}\right), \tag{2.38}$$

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 = O_{\mathbb{P}}\left(s_0 M_N \log N \sqrt{\frac{\log(N \vee P))}{N}}\right). \tag{2.39}$$

Corollary 2 shows that the average excess (Kullback–Leibler) risk achieves the optimal convergence rate up to the factor $O_{\mathbb{P}}\left(\frac{M_N^2 (\log N)^2 \log(N\vee P)}{N}\right)$ as if we knew which s_0 features were active. This order can be regarded as only $O(M_N^2 (\log N)^2)$ worse when compared to (2.24) for the ordinary Lasso.

2.2 Skew-Symmetric Distribution

In this section, we will review the *skew-symmetric* distributions. The term "skew-symmetric" may sound contradictory at first. This is based on the fact that these distributions are extensions of a simple, symmetric, unimodal distribution on $\mathbb{R}$ (e.g., the normal distribution). This concept was initially focused on by Azzalini (1985). However, as shown later, several alternative families of the skew-symmetric distributions have also been proposed.

Throughout this section, let g be a symmetric and unimodal probability density function on $\mathbb{R}$. Let G be its cumulative distribution function. In practical applications, we often consider this distribution as a location-scale family, defined as $g(x \mid \mu, \sigma) := \frac{1}{\sigma} g\left(\frac{x-\mu}{\sigma}\right)$. However, since this form is unnecessary for the review of this section, we simplify it by setting $\mu = 0$ and $\sigma = 1$. Then, we implicitly assume that $g(x)$ is symmetric and unimodal at $x = 0$.

For g, we usually use a simple two-parameter distribution consisting of location and scale, such as the normal distribution. On the other hand, we can employ a distribution with a third parameter to control tail-weight, such as the t distribution (Theodossiou, 1998).

We define a transformation function q (or r) to extend g to a skew-symmetric distribution as follows: let $q : \mathbb{R} \to \mathbb{R}$ be a monotonically increasing function; let $r : \mathbb{R} \to \mathbb{R}$ be the inverse function of q. The definition of q (or r) will be added as necessary in each subsection.

In the remaining subsections, we will define the families of the skew-symmetric distributions sequentially and then compare their strengths and weaknesses. Although many of these families are closely related, their statistical properties are greatly different. To highlight these relationships, we sometimes use differential functions, which may seem unnecessary. This is because the transformation function in one family can be interpreted as the Jacobian in another family.

2.2.1 Family I: Azzalini-type

The density of the Azzalini-type skew-symmetric distributions can be defined in the following general way (Azzalini and Capitanio, 2003; Wang, Boyer, and Genton, 2004).

Definition 4 (Family I: Azzalini-type)**.**

$$f_I(x) := 2g(x)q'(x), \tag{2.40}$$

where $q'(x) + q'(-x) = 1$.

For q' of this family, we often employ a cumulative distribution function of a symmetric distribution. This cumulative distribution function does not necessarily have to match G. Moreover, we often use the form $q'(\alpha x)$ for $\alpha \in \mathbb{R}$. This is because α can act as a skewness parameter for f_{I}. The ordinary Azzalini's skew normal (Azzalini, 1985) corresponds to $g(x) = \phi(x)$ and $q'(x) = \Phi(\alpha x)$, where ϕ and Φ are the probability density function and the cumulative distribution function of the standard normal distribution, respectively.

Formally, we can obtain the following stochastic representation (Azzalini, 2013): let $X_0 \sim g(x)$; let X_S be a random variable such that $X_S = 1$ with probability $q'(|X_0|)$ or $X_S = -1$ with probability $q'(-|X_0|)$; then $X_{\mathrm{I}} := X_S|X_0| \sim f_{\mathrm{I}}(x)$.

Family I and its multivariate generalizations have become the most popular in the literature on the skew-symmetric distribution, partly due to the comprehensive overviews provided by Azzalini and Capitanio (2003), Genton (2004), Azzalini (2005), Azzalini and Genton (2008), and Azzalini (2013).

In addition, one reason why this family is well-known is that it appears frequently in various simple modeling mechanisms. For example, densities f_{I} can be obtained as the marginal distribution of an untruncated variable when another (hidden) variable in a bivariate distribution is truncated. This leads to a weighted distribution, which arises from biased sampling mechanisms.

2.2.2 Family II: Transformation of Scale

Jones (2014) has proposed the following noteworthy framework for developing flexible families of skew-symmetric distributions, which is directly related to the proposed method in this thesis.

Definition 5 (Family II: Transformation of Scale)**.**

$$f_{\text{II}}(x) := 2g\left(r(x)\right), \tag{2.41}$$

where $q(x) - q(-x) = x$.

The form (2.41) is generally not a valid probability density function without the Jacobian. However, as long as the condition $q(x) - q(-x) = x$ holds, it is always a valid probability density function. We can unravel the reason from the linkage between f_{I} and f_{II}: $X_{\text{II}} \sim f_{\text{II}}$ can be derived from $X_{\text{I}} \sim f_{\text{I}}$ using the transformation $X_{\text{II}} = q(X_{\text{I}})$. That is why the derivative of the condition on q in Definition 5 is equivalent to the condition on q in Definition 4.

Family II is characterized by their densities of the form $g(r(x))$, and then $r(x)$ may be regarded as a transformation of scale for g. However, according to Jones (2014), "transformation of scale" refers to modifying the evaluation point x of a probability density function g rather than transforming the random variable associated with g. In that sense, Jones (2014) states that this approach is distinct from the more common method of transforming the random variable.

Family II has several attractive properties. The distribution $f_{\text{II}}(x)$ is unimodal if $g(x)$ is unimodal and its mode is explicitly taken at $x = q(0)$. This results from $f_{\text{II}}(x)$ maintaining the same values as $g(x)$ in the same order as x increases. The skewness properties of f_{II} can be interpreted in terms of the density-based skewness (Critchley and Jones, 2008) as opposed to the quantile-based skewness (Groeneveld and Meeden, 1984). For the distributions in this family, entropy remains invariant to the used transformation (Jones, 2014). Further details are given in Section 2.2.4.

As a subclass of Family II, Fujisawa and Abe (2015) has proposed the mode-invariant skew-normal distribution, which is assumed in the proposed method. This distribution has the attractive property that the mode is always invariant at $x = 0$ regardless of skewness. This property is achieved by satisfying $q(0) = 0$ in Definition 5. Specifically, they have introduced the following specific case for q (or r):

$$q_\alpha(x) := x + \rho_\alpha \frac{\sqrt{1 + \alpha^2 x^2} - 1}{\alpha}, \tag{2.42}$$

where $\alpha \in \mathbb{R}$ and $\rho_\alpha := 1 - \exp(-\alpha^2)$. The inverse function r with mode-invariance can be calculated as follows:

$$r_\alpha(x) := \begin{cases} \frac{1}{\alpha(1-\rho_\alpha^2)} \left(\alpha x + \rho_\alpha - \rho_\alpha \sqrt{(\alpha x + \rho_\alpha)^2 + (1-\rho_\alpha^2)} \right) & (\alpha \neq 0) \\ x & (\alpha = 0) \end{cases} . \tag{2.43}$$

The form (2.42) has other favorable properties besides mode-invariance. As shown in (2.43), it has a closed form for r, which makes it tractable both analytically and computationally. For $\alpha = 0$, since $r_0(x) = x$ corresponds to the identity transformation, $\tilde{f}_{\text{II}}(x) := g(r_\alpha(x))$ is equivalent to the base distribution $g(x)$. As α increases, the skewness increases monotonically. For $\alpha \to \infty$, $\tilde{f}_{\text{II}}(x)$ asymptotically approaches a half-normal distribution. The behavior for $\alpha < 0$ is symmetric at $\alpha = 0$ with it for $\alpha > 0$ (see Figure 1.1 for details). Again, for any α, the mode of $\tilde{f}_{\text{II}}(x)$ is invariant at $x = 0$.

2.2.3 Other Families

Family III: Two-piece

As an alternative approach to generating a skew distribution on $\mathbb{R}$, we can apply different scale factors to the half-lines $x < 0$ and $x > 0$. Such a distribution is called the two-piece distribution (or the split distribution), which has been revisited and parametrized in various forms over time. Hansen (1994) used this approach to introduce an asymmetric form of Student's t distribution, while Hinkley and Revankar (1977) independently developed a similar construction for an asymmetric Laplace distribution. Mudholkar and Hutson (2000) provides an overview and a variant of this concept. Here, we define the following form, introduced in Arellano-Valle, Gómez, and Quintana (2005).

Definition 6 (Family III: Two-piece)**.**

$$f_{\text{III}}(x) := \frac{2}{c_-(\alpha) + c_+(\alpha)} \left(g\left(\frac{x}{c_-(\alpha)}\right) \mathbb{1}\{x < 0\} + g\left(\frac{x}{c_+(\alpha)}\right) \mathbb{1}\{x \geq 0\} \right), \tag{2.44}$$

where $c_-(\alpha), c_+(\alpha) > 0$ with a skewness parameter α, and $\mathbb{1}\{\mathcal{A}\}$ is the indicator function, which means $\mathbb{1}\{\mathcal{A}\} = 1$ if $\mathcal{A}$ is true and $\mathbb{1}\{\mathcal{A}\} = 0$ if $\mathcal{A}$ is false.

Family III includes several subclasses which have been actively studied. For example, if setting $c_-(\alpha) = \frac{1}{\alpha}$ and $c_+(\alpha) = \alpha$ for $\alpha > 0$, (2.44) corresponds to the class in Fernández and Steel (1998); if setting $c_-(\alpha) = \frac{1-\alpha}{2}$ and $c_+(\alpha) = \frac{1+\alpha}{2}$ for $\alpha \in (-1, 1)$, it has been discussed in Jones (2014).

Jones (2014) has revealed that the latter example in the previous paragraph is included in Family Ⅱ. Indeed, it is also unimodal if g is unimodal and its mode is explicitly taken at $x = 0$. For this example, we can obtain the following transformation function:

$$r(x) = 2x \left(\frac{1}{1-\alpha} \mathbb{1}\{x < 0\} + \frac{1}{1+\alpha} \mathbb{1}\{x \geq 0\} \right). \tag{2.45}$$

Since Family Ⅱ includes many smooth transformation functions for x, such as (2.43), it is a serious drawback that (2.45) is generally non-smooth. However, it exhibits excellent tractability due to the simple nature of its transformation function. Moreover, the following stochastic representation, which Family Ⅲ has, is also a strength of (2.45): let $X_0 \sim g(x)$; let X_S be a random variable such that $X_S = c_-$ with probability $\frac{c_-}{c_-+c_+}$ or $X_S = c_+$ with probability $\frac{c_+}{c_-+c_+}$; then $X_{\text{Ⅲ}} := X_S |X_0| \sim f_{\text{Ⅲ}}(x)$.

Family Ⅳ: Transformation of Random Variable

Family Ⅱ is a valid probability density function without the Jacobian because of the condition $q(x) - q(-x) = x$. Conversely, by incorporating the Jacobian, we can remove this constraint.

Definition 7 (Family Ⅳ: Transformation of Random Variable)**.**

$$f_{\text{Ⅳ}}(x) := g\left(r(x)\right) / q'\left(r(x)\right). \tag{2.46}$$

The linkage between g and $f_{\text{Ⅳ}}$ is given by $X_{\text{Ⅳ}} := q(X) \sim f_{\text{Ⅳ}}(x)$ for $X \sim g(x)$, which is similar to the linkage between $f_{\text{Ⅰ}}$ and $f_{\text{Ⅱ}}$. Moreover, Family Ⅳ has a cumulative distribution function $G(r(x))$ and a quantile function $q(G^{-1}(x))$. For a comprehensive treatment of this transformation, we can refer to Ley and Paindaveine (2010).

Family Ⅴ: Probability Integral Transformation

If a random variable U follows the continuous uniform distribution on $[0, 1]$, then $X = G^{-1}(U) \sim g(x)$. This is called the (inverse) probability integral transformation. Using a different random variable V from U, we can obtain a skew distribution based on g. If a random variable V follows $q(x)$ on $[0, 1]$, $X = G^{-1}(V)$ follows the following distribution.

Definition 8 (Family V: Probability Integral Transformation)**.**

$$f_V(x) := g(x)q'(G(x)). \tag{2.47}$$

Family V has a cumulative distribution function $q(G(x))$ and a quantile function $G^{-1}(r(x))$. This shows that Family V essentially has the same analogy as Family IV with the roles of the transformation function and the initial random variable reversed: in Family IV, the initial random variable follows g, and the transformation function is any function; in Family V, the transformation function is G^{-1}, and the initial random variable is any random variable. For a comprehensive treatment of this transformation, we can refer to Ferreira and Steel (2006).

2.2.4 Comparison

We will compare the properties of the above five families, particularly focusing on the relationship between Family I and Family II. For a comprehensive comparison of all families, we can see Jones (2015).

Unimodality: Families II and III are unimodal if g is unimodal and its mode is explicitly given. As shown in Fujisawa and Abe (2015), these families include the mode-invariant skew-normal distribution, which means that the mode is always invariant regardless of skewness. On the other hand, Families I, IV, and V require case-by-case verification and typically lack explicit expressions for unimodality and their modes. For example, Azzalini (2013) demonstrates that $f_{\mathrm{I}}(x) := 2\phi(x)\Phi(x^3)$ becomes a multi-modal distribution. Although the ordinary Azzalini's skew-normal distribution (Azzalini, 1985) has unimodality, its mode has no analytic expression and depends on all the distribution parameters (see Figure 1.1).

Skewness: Family II can have a monotonic skewness parameter in the density-based sense (Critchley and Jones, 2008; Jones, 2014). In other words, if $q_1(x) \geq q_2(x)$ for any x with $f^1_{\mathrm{II}}(x) := 2g\left(r_1(x)\right)$ and $f^2_{\mathrm{II}}(x) := 2g\left(r_2(x)\right)$, f^1_{II} has a larger skewness in the density-based sense. Family I does not generally exhibit this type of skewness in both the quantile-based (Groeneveld and Meeden, 1984) and the density-based senses.

Tail-weight: In Families IV and V, we can control the tail-weight by choosing q (or r). In Family I, if q' is a cumulative distribution function or a survival function on $\mathbb{R}$, we can take different tail-weights. In Families II and III, we are always given the same tail-weight as g (Jones, 2015).

Physical Mechanism: Family I has strong generating mechanisms, especially in cases involving truncation and selection mechanisms, as described in Section 2.2.1. These mechanisms are actively studied and justified in many works (e.g., Genton (2004), Azzalini (2005), Marchenko and Genton (2012), and Azzalini (2013)), which makes Family I a potentially reasonable choice in many real-world situations. In this respect, Family I is significantly ahead of Family II. However, the mode-invariant skew-normal distribution (Fujisawa and Abe, 2015) may provide a new perspective on Family II. As in the proposed method, a regression model whose noise is assumed to follow this distribution can express a reasonable assumption that a noiseless situation has the highest probability. This is because we can utilize the fact that the regression estimator of the location parameter is completely equivalent to that of the mode. We can not achieve this type of modeling with Family I.

Interpretability: We use the term "interpretability" to refer to how easily the roles of these parameters can be understood. Specific parameters corresponding to particular features of distributions make it easier to understand these features. Families II and III have the attractive property that the mode is explicitly taken at $x = q(0)$. This leads to the mode-invariance in Fujisawa and Abe (2015) and allows us to model the mode with a single (location) parameter, as in the proposed method. However, Family I does not have such a property. Indeed, in the ordinary Azzalini's skew-normal distribution, the location parameter is neither mean, median, nor mode; these statistics (as well as variance and skewness) have complicated combinations of all the location, scale, and skewness parameters.

Chapter 3

Proposed Method

3.1 Skew Distribution

We use the mode-invariant skew-normal distribution (Fujisawa and Abe, 2015) with the parameter vector $\boldsymbol{\psi} := [\mu, \sigma, \alpha]^\top \in \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}$:

$$f(y \mid \boldsymbol{\psi}) := \frac{1}{\sigma}\phi\left(r_\alpha\left(\frac{y-\mu}{\sigma}\right)\right), \tag{3.1}$$

where ϕ is the probability density function of the standard normal distribution, r_α is a specific transformation function of scale, described later, and μ, σ, and α are the parameters of location, scale, and skewness, respectively. The function r_α must satisfy

$$r'_\alpha(u) > 0, \tag{3.2}$$
$$q'_\alpha(v) + q'_\alpha(-v) = 2, \tag{3.3}$$

where q_α is the inverse function of r_α. Note that the condition (3.2) ensures a monotone-increasing property of $r_\alpha(u)$. The model (3.1) is unimodal with the mode-invariant property $\text{Mode}[f(y \mid \boldsymbol{\psi})] = \mu$ for any σ and α.

Let $H_\alpha(v) := q_\alpha(v) - v$. While H_α (or r_α) can be freely designed if it satisfies the above conditions, this thesis employs the following specific function H_α because of its better analytical treatment:

$$H_\alpha(v) := \rho_\alpha \frac{\sqrt{1+\alpha^2 v^2} - 1}{\alpha}, \tag{3.4}$$

where $\rho_\alpha := 1 - \frac{1}{2}\exp(-\alpha^2)$.

Remark 1. *The above function ρ_α slightly differs from that proposed in Fujisawa and Abe (2015) because the Fisher information matrix is non-singular at $\alpha = 0$ when we use the above function but singular when we use the function proposed in Fujisawa and Abe*

(2015). The non-singularity is usually necessary to obtain some theoretical properties. A related proposition appears in Proposition 6.

3.2 Problem Description

We consider the linear regression model with the skew-unimodal noise, given by

$$y = \boldsymbol{X}^\top \boldsymbol{\beta} + e, \tag{3.5}$$

where $y \in \mathbb{R}$ is an output, $\boldsymbol{X} \in \mathbb{R}^P$ is a P-dimensional feature vector, $\boldsymbol{\beta} \in \mathbb{R}^P$ is a parameter vector, and e follows a skew-unimodal distribution $f(e \mid 0, \sigma, \alpha)$. Let $\boldsymbol{\psi}(\boldsymbol{X}) := [\boldsymbol{X}^\top \boldsymbol{\beta}, \sigma, \alpha]^\top$. The distribution of y can be expressed by $f(y \mid \boldsymbol{\psi}(\boldsymbol{X}))$, and the mode is $\boldsymbol{X}^\top \boldsymbol{\beta}$.

Let the log-likelihood function be denoted by $l(\boldsymbol{\theta} \mid \boldsymbol{X}, y) := \log f(y \mid \boldsymbol{\psi}(\boldsymbol{X}))$ with the parameter vector $\boldsymbol{\theta} := [\boldsymbol{\beta}^\top, \sigma, \alpha]^\top$. Let $\mathcal{D}_N := \{(y_n, \boldsymbol{X}_n)\}_{n=1}^N$ be the i.i.d. data. We define the loss function by

$$\begin{aligned} \ell(\boldsymbol{\theta} \mid \mathcal{D}_N) &:= -\frac{1}{N} \sum_{n=1}^N l(\boldsymbol{\theta} \mid \boldsymbol{X}_n, y_n) \\ &= -\frac{1}{N} \sum_{n=1}^N \log f(y_n \mid \boldsymbol{\psi}(\boldsymbol{X}_n)). \end{aligned} \tag{3.6}$$

Since we are interested in sparse regression, this thesis focuses on the Lasso-type problem

$$\min_{\boldsymbol{\theta}} \ \ell(\boldsymbol{\theta} \mid \mathcal{D}_N) + \lambda \|\boldsymbol{\beta}\|_1, \tag{3.7}$$

where $\lambda \geq 0$ is the regularization parameter.

3.3 Optimization

The loss function (3.6) is non-convex for $\boldsymbol{\theta}$ but convex for $\boldsymbol{\beta}$ when σ and α are fixed, as described in the following theorem.

Theorem 4. *The loss function $\ell(\boldsymbol{\theta} \mid \mathcal{D}_N)$ is convex with a Lipschitz continuous gradient for $\boldsymbol{\beta}$. In particular, if $S_N := \frac{1}{N} \sum_{n=1}^N \boldsymbol{X}_n \boldsymbol{X}_n^\top$ is positive definite, $\ell(\boldsymbol{\theta} \mid \mathcal{D}_N)$ is strongly convex for $\boldsymbol{\beta}$.*

The proof is given in Section 3.5.2. Using this theorem, we develop a problem-specific optimization method based on the block coordinate descent method. This

Algorithm 1 Optimization of the proposed method

Require: Hyper-parameter λ and initialized $\boldsymbol{\beta}$, σ, α

1: **while** until convergence **do**
2: **while** until convergence (of $\boldsymbol{\beta}$) **do**
3: **for** $n = 1, \ldots, N$ **do**
4: $z_n \leftarrow \frac{\alpha}{\sigma}\left(y_n - \boldsymbol{X}_n^\top \boldsymbol{\beta}\right) + \rho_\alpha$
5: $w_n \leftarrow \sqrt{z_n^2 + 1 - \rho_\alpha^2}$
6: $\tilde{y}_n \leftarrow \boldsymbol{X}_n^\top \boldsymbol{\beta} + \frac{\sigma}{2\alpha}\left(z_n - \frac{\rho_\alpha}{1+\rho_\alpha^2}\left(w_n + \frac{z_n^2}{w_n}\right)\right)$
7: **end for**
8: $\boldsymbol{\beta} \leftarrow \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1+\rho_\alpha^2}{N\sigma^2(1-\rho_\alpha^2)^2} \sum_{n=1}^{N}\left(\tilde{y}_n - \boldsymbol{X}_n^\top \boldsymbol{\beta}\right)^2 + \lambda\|\boldsymbol{\beta}\|_1$
9: **end while**
10: $\sigma \leftarrow \underset{\sigma:\,\sigma>0}{\operatorname{argmin}}\, \ell(\boldsymbol{\theta} \mid \mathcal{D}_N)$ with L-BFGS-B (or other valid algorithm)
11: $\alpha \leftarrow \underset{\alpha}{\operatorname{argmin}}\, \ell(\boldsymbol{\theta} \mid \mathcal{D}_N)$ with L-BFGS (or other valid algorithm)
12: **end while**

method allows us to separate the optimization problem into convex and non-convex and take advantage of the convenient property described in Theorem 4.

The update algorithm is summarized in Algorithm 1. We update $\boldsymbol{\beta}$, σ, and α alternately. First, we fix (σ, α) and optimize $\boldsymbol{\beta}$. We employ the Majorization-Minimization (MM) algorithm using the Taylor expansion with acceleration techniques (Jamshidian and Jennrich, 1997; Sun, Babu, and Palomar, 2016). The derivation of this MM algorithm is given in Section 3.5.3. As shown in line 8 of Algorithm 1, we can rewrite the $\boldsymbol{\beta}$ update as another Lasso-type problem and then use well-known software, e.g., the *sklearn.linear_model* package of Python. Next, we fix $(\boldsymbol{\beta}, \alpha)$ and optimize σ, and finally fix $(\boldsymbol{\beta}, \sigma)$ and optimize α. Although we need to consider the non-convexity for σ and α, their optimization problems are at most one variable. Hence, we can easily use reliable software, e.g. the *scipy.optimize* package of Python. In this thesis, we employ the L-BFGS algorithm (Liu and Nocedal, 1989), which is an iterative method for solving non-linear optimization problems. In particular, for the inequality constraint $\sigma > 0$, we can utilize the L-BFGS-B algorithm (Byrd, Lu, Nocedal, and Zhu, 1995; Zhu, Byrd, Lu, and Nocedal, 1997), which extends the L-BFGS algorithm to handle bounded constraints.

Let $\tilde{\ell}(\boldsymbol{\theta}) := \ell(\boldsymbol{\theta} \mid \mathcal{D}_N) + \lambda\|\boldsymbol{\beta}\|_1$. Let $\operatorname{dom} \tilde{\ell}$ be the effective domain of $\tilde{\ell}$. Following the definition of Tseng (2001), we say that $\boldsymbol{\theta} \in \operatorname{dom} \tilde{\ell}$ is a stationary point of $\tilde{\ell}$ if

for any $\boldsymbol{d} \in \mathbb{R}^{P+2}$,

$$\liminf_{a \downarrow 0} \frac{\tilde{\ell}(\boldsymbol{\theta} + a\boldsymbol{d}) - \tilde{\ell}(\boldsymbol{\theta})}{a} \geq 0. \tag{3.8}$$

Algorithm 1 guarantees convergence to a stationary point, as described in the following theorem.

Theorem 5. *Let $\{\boldsymbol{\theta}^{(t)}\}_{t=0,1,\ldots}$ be a sequence of $\boldsymbol{\theta}$ after updating σ in Algorithm 1. Then, every cluster point of $\{\boldsymbol{\theta}^{(t)}\}_{t=0,1,\ldots}$ is a stationary point of $\tilde{\ell}$.*

The proof is given in Section 3.5.4. In high-dimensional ($N < P$) settings, $S_N = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{X}_n \boldsymbol{X}_n^\top$ is typically not positive definite. This makes it difficult to accurately unravel the behavior and convergence of the algorithm. Furthermore, Algorithm 1 addresses a non-convex optimization problem, which is inherently more complex to analyze than standard convex optimization problems. Theorem 5 guarantees that all cluster points of $\{\boldsymbol{\theta}^{(t)}\}_{t=0,1,\ldots}$ in Algorithm 1 are locally convex in the sense of (3.8) without requiring any assumptions of S_N. Such theoretical results, which do not involve S_N, provide particularly valuable insights in high-dimensional settings.

Here, we describe how to set initial values in this thesis. The regression parameter $\boldsymbol{\beta}$ started from the estimated coefficients of the Lasso model for the same data. Using the residuals of its model, we initialized σ with their sample standard deviation and α with -1 or 1 corresponding to the sign of their sample skewness.

3.4 Related Work

3.4.1 Quantile Regression

Quantile regression has emerged as a powerful method in place of traditional mean regression. Since mean regression primarily focuses on the average effect, it may provide an inadequate representation of the data when assumptions of normality and homoscedasticity are violated. In contrast, quantile regression provides a comprehensive analysis of the conditional distribution of the response variable by estimating the conditional quantiles at each percentile. This allows us to explore the impact of predictors on various distribution segments, including the extremes. Several review papers offer valuable insights into quantile regression methodology and applications (Davino, Furno, and Vistocco, 2013; Huang, Zhang, Chen, and He, 2017; Koenker, 2017; Koenker, Chernozhukov, He, and Peng, 2017; Torossian, Picheny, Faivre, and Garivier, 2020; Sun and Lin, 2024).

Median regression is a specific form of quantile regression that focuses on estimating the 50th percentile (median). In descriptive statistics, measures of central tendency — such as the mean, median, and mode — serve as fundamental tools for summarizing large sets of observations into a single representative value. In unimodal symmetric distributions, the mean, median, and mode coincide. However, in unimodal skewed distributions, the mean shifts significantly towards the tail, the median shifts slightly in the same direction, and the mode shifts less than the mean and median. In other words, the median is typically located between the mean and mode. While median regression is more extensively studied and widely used than modal regression, in terms of a robust measure of central tendency, modal regression also deserves equal or greater attention.

3.4.2 Modal Regression

Modal regression offers a robust alternative, particularly when addressing skewed or heavy-tailed data. Unlike mean and median regression, modal regression seeks the "most likely" or most frequent value. This makes it less susceptible to outliers and can reveal trends obscured by skewed distributions. Moreover, it remains consistent even when applied to truncated data. For more comprehensive details, we can refer to (Yao and Li, 2014; Chen, 2018; Xiang and Yao, 2022). In particular, Example 1 in Yao and Li (2014) effectively demonstrates the difference between the modal and mean regression.

Modal regression can be categorized into three main types (Xiang and Yao, 2022):

Non-parametric modal regression: This approach offers a flexible estimation of the conditional mode without imposing any specific functional form on the relationship between the response and the covariates. It directly estimates the conditional density function non-parametrically, often using kernel density estimation, and identifies the mode as the point of maximum density. This approach can uncover complex relationships and patterns in the data that might be missed by traditional mean regression. However, this method can suffer from the "curse of dimensionality" when dealing with high-dimensional covariates. This leads to slower convergence rates, increased computational complexity, and a greater need for data (Kemp and Silva, 2012; Chen, Genovese, Tibshirani, and Wasserman, 2016; Chen, 2018; Feng, Fan, and Suykens, 2020).

Semi-parametric modal regression: This approach balances the flexibility of non-parametric methods and the efficiency of parametric methods. It achieves

this by imposing a certain structure on the conditional mode while allowing some components to be estimated non-parametrically. One example of a semi-parametric modal regression is the varying coefficient modal regression (Yao and Xiang, 2016). This model assumes that the conditional mode can be expressed as a linear combination of covariates, but the coefficients can vary smoothly with an index variable. For other examples, we can see Krief (2017) and Ota, Kato, and Hara (2019). Semi-parametric approach inherits both the advantages and disadvantages of parametric and non-parametric approaches, for better or worse.

Parametric modal regression: This approach assumes a specific functional form for the conditional mode, relying on distributions that can be conveniently parameterized by the mode, such as the gamma or beta distribution (Bourguignon, Leão, and Gallardo, 2020; Zhou and Huang, 2020; Zhou and Huang, 2022; Liu and Huang, 2024). This offers advantages in terms of efficiency and interpretability but carries the risk of model misspecification. Similar to mean regression, a link function connects the mode of the response to a linear predictor of covariates. This allows for modeling the relationship between the most frequent values of the response and the explanatory variables. The model's parameters, including the regression coefficients, are typically estimated using the maximum likelihood method. Standard errors and confidence intervals can be obtained based on the asymptotic properties of the maximum likelihood estimators.

The proposed method belongs to parametric modal regression. While Azzalini's skew-normal distribution can not accurately parametrize the mode, the proposed method can precisely formulate regression for the mode as the location parameter. As mentioned above, modal regression often employs the gamma and beta distributions. However, they may be insufficient to capture the skewness because, for example, they cannot represent the half-normal distribution. Therefore, the proposed method is a highly expressive parametric modal regression for skew distributions.

Furthermore, as discussed above, "flexibility" and "efficiency" are important in comparing parametric and non-parametric approaches. Regarding flexibility, the proposed method also carries the risk of model misspecification. However, as demonstrated later in Sections 5.1 and 5.2, the proposed method outperformed the comparative methods even in the case of model misspecification. Regarding efficiency, Theorem 4 shows that the proposed method is a convex function in terms of $\boldsymbol{\beta}$, and the non-convex parameters (σ and α) are only two-dimensional. Additionally, Theorem 5 guarantees that Algorithm 1 converges to a stationary

point. These two points make the proposed method quite manageable among parametric modal regressions.

3.4.3 Regularized Regression with Skew Distributions

The most famous skew distribution is Azzalini's skew-normal distribution (Azzalini, 1985), which is implemented in current standard scientific computing software. Many skew distributions were developed using Azzalini's basic idea (Azzalini and Capitanio, 1999; Branco and Dey, 2001; Sahu, Dey, and Branco, 2003; Liseo and Loperfido, 2003; González-Farías, Dominguez-Molina, and Gupta, 2004; Arellano-Valle, Gómez, and Quintana, 2005; Arellano-Valle and Genton, 2005; Arellano-Valle and Azzalini, 2006). The closest method to the proposed method is Chen, Pourahmadi, and Maadooliat (2014), which assumes a regression model for the location parameter of Azzalini's skew-normal distribution and applies an ℓ_1 regularization technique. However, the location parameter in these skew distributions is neither mean, median, nor mode; the mean, median, and mode have complicated combinations of all the location, scale, and skewness parameters. Even if the location is modeled by $\boldsymbol{X}^\top\boldsymbol{\beta}$ for Azzalini's skew-normal distribution, it is difficult to understand the role of features. The proposed modeling makes it easy to understand the role of features via the modeling of mode. The proposed modeling is suitable for an interpretable one in high-dimensional settings, but the modelings using Azzalini's skew-normal distribution are not.

3.5 Proofs

3.5.1 Upper Bounds on $|H_\alpha(v)|$ and $|r_\alpha(u)|$

Proposition 1. *We have* $|H_\alpha(v)| < \rho_\alpha|v|$, $|h_\alpha(v)| < \rho_\alpha$, $|h'_\alpha(v)| < \rho_\alpha|\alpha|$, *and* $|h''_\alpha(v)| < 3\rho_\alpha\alpha^2$. *From* $\rho_\alpha < 1$, *we also have* $|H_\alpha(v)| < |v|$, $|h_\alpha(v)| < 1$, $|h'_\alpha(v)| < |\alpha|$, *and* $|h''_\alpha(v)| < 3\alpha^2$.

Proof. It holds after simple calculation that

$$h_\alpha(v) = \frac{d}{dv}H_\alpha(v) = \frac{\rho_\alpha\alpha v}{\sqrt{1+\alpha^2v^2}}, \tag{3.9}$$

$$h'_\alpha(v) = \frac{d^2}{dv^2}H_\alpha(v) = \frac{\rho_\alpha\alpha}{(1+\alpha^2v^2)^{\frac{3}{2}}}, \tag{3.10}$$

$$h''_\alpha(v) = \frac{d^3}{dv^3}H_\alpha(v) = -\frac{3\rho_\alpha\alpha^3v}{(1+\alpha^2v^2)^{\frac{5}{2}}}. \tag{3.11}$$

We clearly see $|h_\alpha(v)| < \rho_\alpha$ and $|h'_\alpha(v)| < \rho_\alpha|\alpha|$. Since $H_\alpha(v) = \int_0^v h_\alpha(v)dv$, we have $|H_\alpha(v)| < \rho_\alpha|v|$. We also have

$$|h''_\alpha(v)| = 3\rho_\alpha\alpha^2\frac{|\alpha v|}{(1+\alpha^2v^2)^{\frac{1}{2}}}\frac{1}{(1+\alpha^2v^2)^2} \leq 3\rho_\alpha\alpha^2. \tag{3.12}$$

□

Proposition 2. *Suppose that Assumption 1 holds. Then, $r'_\alpha(u)$, $r''_\alpha(u)$, and $r'''_\alpha(u)$ are bounded and $|r_\alpha(u)|$ are bounded by $c|u|$ with some constant c.*

Proof. For $\alpha = 0$, since $r_\alpha(u) = u$, this proposition clearly holds. Hereafter, we consider the case $\alpha \neq 0$. By differentiating $u = q_\alpha(r_\alpha(u))$ with respect to u, we have $1 = q'_\alpha(r_\alpha(u))r'_\alpha(u)$ and then

$$r'_\alpha(u) = \frac{1}{q'_\alpha(r_\alpha(u))} = \frac{1}{1+h_\alpha(r_\alpha(u))}. \tag{3.13}$$

From Proposition 1, we have

$$0 < \frac{1}{1+\rho_\alpha} \leq r'_\alpha(u) \leq \frac{1}{1-\rho_\alpha}. \tag{3.14}$$

We can easily see that $1/2 < \rho_\alpha < c$ with some constant $c < 1$ because $\rho_\alpha = 1 - (1/2)\exp(-\alpha^2)$ and α is bounded from Assumption 1. This implies that $r'_\alpha(u)$ is bounded. From $r_\alpha(0) = 0$, we have $r_\alpha(u) = \int_0^u r'_\alpha(s)ds$ and then

$$|r_\alpha(u)| \leq \frac{1}{1-\rho_\alpha}|u| < c|u|. \tag{3.15}$$

By differentiating (3.13) with respect to u, we have

$$r''_\alpha(u) = -\frac{h'_\alpha(r_\alpha(u))r'_\alpha(u)}{\{1+h_\alpha(r_\alpha(u))\}^2} = -h'_\alpha(r_\alpha(u))\{r'_\alpha(u)\}^3. \tag{3.16}$$

By differentiating the above with respect to u, we have

$$r'''_\alpha(u) = -h''_\alpha(r_\alpha(u))\{r'_\alpha(u)\}^4 - 3h'_\alpha(r_\alpha(u))\{r'_\alpha(u)\}^2r''_\alpha(u). \tag{3.17}$$

From Proposition 1 and the boundedness of α, we can easily see $r''_\alpha(u)$ and $r'''_\alpha(u)$ are bounded. □

Proposition 3. *Fix $\alpha \neq 0$. For the definition in (3.4), the upper bounds of $\left|\frac{\partial}{\partial\alpha}H_\alpha(v)\right|$, $\left|\frac{\partial^2}{\partial\alpha^2}H_\alpha(v)\right|$, and $\left|\frac{\partial^3}{\partial\alpha^3}H_\alpha(v)\right|$ can be described by a linear function of $|v|$. They can also*

be described by a linear function of $|u|$. For $\alpha = 0$, the upper bounds of $\left|\frac{\partial}{\partial\alpha}H_\alpha(v)\right|$, $\left|\frac{\partial^2}{\partial\alpha^2}H_\alpha(v)\right|$, and $\left|\frac{\partial^3}{\partial\alpha^3}H_\alpha(v)\right|$ can be described by a quadratic function of $|v|$, a constant, and a quartic function of $|v|$, respectively. They can also be described by a quadratic function of $|u|$, a constant, and a quartic function of $|u|$, respectively.

Proof. Analytically, for fixed $\alpha \neq 0$,

$$\begin{aligned}\frac{\partial}{\partial\alpha}H_\alpha(v) &= \left(\frac{\alpha\exp(-\alpha^2)}{\rho_\alpha} + \frac{1}{\alpha\sqrt{1+\alpha^2v^2}}\right)H_\alpha(v)\\ &\leq \left(\frac{\alpha\exp(-\alpha^2)}{\rho_\alpha} + \frac{1}{\alpha}\right)H_\alpha(v),\end{aligned} \tag{3.18}$$

$$\begin{aligned}\frac{\partial^2}{\partial\alpha^2}H_\alpha(v) &= \left(\frac{(\rho_\alpha - 2\alpha^2)\exp(-\alpha^2)}{\rho_\alpha^2} - \frac{v^2}{(1+\alpha^2v^2)^{\frac{3}{2}}} - \frac{1}{\alpha^2\sqrt{1+\alpha^2v^2}}\right)H_\alpha(v)\\ &\quad + \left(\frac{\alpha\exp(-\alpha^2)}{\rho_\alpha} + \frac{1}{\alpha\sqrt{1+\alpha^2v^2}}\right)^2 H_\alpha(v)\\ &\leq \left(\frac{|\rho_\alpha - 2\alpha^2|\exp(-\alpha^2)}{\rho_\alpha^2} + \frac{2\sqrt{3}}{9\alpha^2} + \frac{1}{\alpha^2}\right)H_\alpha(v)\\ &\quad + \left(\frac{\alpha\exp(-\alpha^2)}{\rho_\alpha} + \frac{1}{\alpha}\right)^2 H_\alpha(v),\end{aligned} \tag{3.19}$$

$$\begin{aligned}\frac{\partial^3}{\partial\alpha^3}H_\alpha(v) &= \Bigg(-\frac{2\alpha(\rho_\alpha - 2\alpha)\exp(-\alpha^2)}{\rho_\alpha^2} - \frac{2\alpha(\rho_\alpha - 2\alpha^2)\exp(-2\alpha^2)}{\rho_\alpha^3}\\ &\qquad - \frac{2\alpha(1+\rho_\alpha)\exp(-\alpha^2)}{\rho_\alpha^2}\Bigg)H_\alpha(v)\\ &\quad + \left(\frac{3\alpha v^4}{(1+\alpha^2v^2)^{\frac{5}{2}}} + \frac{v^2}{\alpha(1+\alpha^2v^2)^{\frac{3}{2}}} + \frac{2}{\alpha^3\sqrt{1+\alpha^2v^2}}\right)H_\alpha(v)\\ &\quad + 3\left(\frac{(\rho_\alpha - 2\alpha^2)\exp(-\alpha^2)}{\rho_\alpha^2} - \frac{v^2}{(1+\alpha^2v^2)^{\frac{3}{2}}} - \frac{1}{\alpha^2\sqrt{1+\alpha^2v^2}}\right)\\ &\qquad \times \left(\frac{\alpha\exp(-\alpha^2)}{\rho_\alpha} + \frac{1}{\alpha\sqrt{1+\alpha^2v^2}}\right)H_\alpha(v)\\ &\quad + \left(\frac{\alpha\exp(-\alpha^2)}{\rho_\alpha} + \frac{1}{\alpha\sqrt{1+\alpha^2v^2}}\right)^3 H_\alpha(v)\\ &\leq \Bigg(\frac{2\alpha|\rho_\alpha - 2\alpha|\exp(-\alpha^2)}{\rho_\alpha^2} + \frac{2\alpha|\rho_\alpha - 2\alpha^2|\exp(-2\alpha^2)}{\rho_\alpha^3}\\ &\qquad + \frac{2\alpha(1+\rho_\alpha)\exp(-\alpha^2)}{\rho_\alpha^2}\Bigg)H_\alpha(v)\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{\alpha^3}\left(\frac{48\sqrt{5}}{125}+\frac{2\sqrt{3}}{9}+2\right)H_\alpha(v)\\
&+3\left(\frac{|\rho_\alpha-2\alpha^2|\exp(-2\alpha^2)}{\rho_\alpha^2}+\frac{2\sqrt{3}}{9\alpha^2}+\frac{1}{\alpha^2}\right)\\
&\quad\times\left(\frac{\alpha\exp(-2\alpha^2)}{\rho_\alpha}+\frac{1}{\alpha}\right)H_\alpha(v)\\
&+\left(\frac{\alpha\exp(-\alpha^2)}{\rho_\alpha}+\frac{1}{\alpha}\right)^3 H_\alpha(v). \qquad (3.20)
\end{aligned}
$$

Therefore, since $|H_\alpha(v)| < \rho_\alpha|v|$ from Proposition 1, it holds that the upper bounds of $\left|\frac{\partial}{\partial\alpha}H_\alpha(v)\right|$, $\left|\frac{\partial^2}{\partial\alpha^2}H_\alpha(v)\right|$, and $\left|\frac{\partial^3}{\partial\alpha^3}H_\alpha(v)\right|$ can be described by a linear function of $|v|$. Furthermore, from (3.15), their upper bounds can also be described by a linear function of $|u|$.

Next, we consider the case $\alpha \to 0$. Let $g(x) := \sqrt{1+x^2}-1$. Letting the m-th order derivative of $g(x)$ be denoted by $g^{(m)}(x)$, we have

$$
\begin{aligned}
g^{(1)}(x) &= \frac{x}{\sqrt{1+x^2}}, &(3.21)\\
g^{(2)}(x) &= \frac{1}{(1+x^2)^{\frac{3}{2}}}, &(3.22)\\
g^{(3)}(x) &= -\frac{3x}{(1+x^2)^{\frac{5}{2}}}, &(3.23)
\end{aligned}
$$

Then, it holds after simple calculations that

$$
\begin{aligned}
\lim_{\alpha\to0}\frac{g(\alpha v)}{\alpha} &= 0, &(3.24)\\
\lim_{\alpha\to0}\frac{\partial}{\partial\alpha}\frac{g(\alpha v)}{\alpha} &= \lim_{\alpha\to0}\left(-\frac{g(\alpha v)}{\alpha^2}+\frac{v\cdot g^{(1)}(\alpha v)}{\alpha}\right)\\
&= \frac{1}{2}v^2, &(3.25)\\
\lim_{\alpha\to0}\frac{\partial^2}{\partial\alpha^2}\frac{g(\alpha v)}{\alpha} &= \lim_{\alpha\to0}\left(\frac{2g(\alpha v)}{\alpha^3}-\frac{2v\cdot g^{(1)}(\alpha v)}{\alpha^2}+\frac{v^2\cdot g^{(2)}(\alpha v)}{\alpha}\right)\\
&= 0, &(3.26)\\
\lim_{\alpha\to0}\frac{\partial^3}{\partial\alpha^3}\frac{g(\alpha v)}{\alpha} &= \lim_{\alpha\to0}\left(-\frac{6g(\alpha v)}{\alpha^4}+\frac{6v\cdot g^{(1)}(\alpha v)}{\alpha^3}-\frac{3v^2\cdot g^{(2)}(\alpha v)}{\alpha^2}+\frac{v^3\cdot g^{(3)}(\alpha v)}{\alpha}\right)\\
&= -\frac{3}{4}v^4. &(3.27)
\end{aligned}
$$

It also holds that

$$\begin{aligned}\lim_{\alpha\to 0}\rho_\alpha &= \lim_{\alpha\to 0}\left(1-\frac{1}{2}\exp(-\alpha^2)\right)\\ &= \frac{1}{2}, \qquad (3.28)\\ \lim_{\alpha\to 0}\frac{\partial}{\partial\alpha}\rho_\alpha &= \lim_{\alpha\to 0}\alpha\exp(-\alpha^2)\\ &= 0, \qquad (3.29)\\ \lim_{\alpha\to 0}\frac{\partial^2}{\partial\alpha^2}\rho_\alpha &= \lim_{\alpha\to 0}(1-2\alpha^2)\exp(-\alpha^2)\\ &= 1, \qquad (3.30)\\ \lim_{\alpha\to 0}\frac{\partial^3}{\partial\alpha^3}\rho_\alpha &= \lim_{\alpha\to 0}2\alpha(2\alpha^2-3)\exp(-\alpha^2)\\ &= 0. \qquad (3.31)\end{aligned}$$

Note that $H_\alpha(v)=\rho_\alpha\frac{g(\alpha v)}{\alpha}$. Using from (3.24) through (3.31), we have

$$\begin{aligned}\lim_{\alpha\to 0}\frac{\partial}{\partial\alpha}H_\alpha(v) &= \lim_{\alpha\to 0}\left(\frac{\partial}{\partial\alpha}\rho_\alpha\cdot\frac{g(\alpha v)}{\alpha}+\rho_\alpha\cdot\frac{\partial}{\partial\alpha}\frac{g(\alpha v)}{\alpha}\right)\\ &= \frac{1}{4}v^2, \qquad (3.32)\\ \lim_{\alpha\to 0}\frac{\partial^2}{\partial\alpha^2}H_\alpha(v) &= \lim_{\alpha\to 0}\left(\frac{\partial^2}{\partial\alpha^2}\rho_\alpha\cdot\frac{g(\alpha v)}{\alpha}+2\cdot\frac{\partial}{\partial\alpha}\rho_\alpha\cdot\frac{\partial}{\partial\alpha}\frac{g(\alpha v)}{\alpha}+\rho_\alpha\cdot\frac{\partial^2}{\partial\alpha^2}\frac{g(\alpha v)}{\alpha}\right)\\ &= 0, \qquad (3.33)\\ \lim_{\alpha\to 0}\frac{\partial^3}{\partial\alpha^3}H_\alpha(v) &= \lim_{\alpha\to 0}\left(\frac{\partial^3}{\partial\alpha^3}\rho_\alpha\cdot\frac{g(\alpha v)}{\alpha}+3\cdot\frac{\partial^2}{\partial\alpha^2}\rho_\alpha\cdot\frac{\partial}{\partial\alpha}\frac{g(\alpha v)}{\alpha}\right.\\ &\qquad \left.+3\cdot\frac{\partial}{\partial\alpha}\rho_\alpha\cdot\frac{\partial^2}{\partial\alpha^2}\frac{g(\alpha v)}{\alpha}+\rho_\alpha\cdot\frac{\partial^3}{\partial\alpha^3}\frac{g(\alpha v)}{\alpha}\right)\\ &= \frac{3}{2}v^2-\frac{3}{8}v^4. \qquad (3.34)\end{aligned}$$

Therefore, the upper bounds of $\left|\frac{\partial}{\partial\alpha}H_\alpha(v)\right|$, $\left|\frac{\partial^2}{\partial\alpha^2}H_\alpha(v)\right|$, and $\left|\frac{\partial^3}{\partial\alpha^3}H_\alpha(v)\right|$ as $\alpha\to 0$ can be described by a quadratic function of $|v|$, a constant, and a quartic function of $|v|$, respectively. From (3.15), their upper bounds can also be described by a quadratic function of $|u|$, a constant, and a quartic function of $|u|$, respectively. □

Proposition 4. *Fix $\alpha\neq 0$. For the definition in (3.4), the upper bounds of $\left|\frac{\partial}{\partial\alpha}r_\alpha(u)\right|$, $\left|\frac{\partial^2}{\partial\alpha^2}r_\alpha(u)\right|$, and $\left|\frac{\partial^3}{\partial\alpha^3}r_\alpha(u)\right|$ can be described by a linear function of $|u|$, a quadratic function of $|u|$, and a cubic function of $|u|$, respectively. For $\alpha=0$, the upper bounds of*

$\left|\frac{\partial}{\partial\alpha}r_\alpha(u)\right|$, $\left|\frac{\partial^2}{\partial\alpha^2}r_\alpha(u)\right|$, and $\left|\frac{\partial^3}{\partial\alpha^3}r_\alpha(u)\right|$ can be described by a quadratic function of $|u|$, a quartic function of $|u|$, and a sextic function of $|u|$, respectively.

Proof. Using the Euler's chain rule,

$$\frac{\partial}{\partial\alpha}r_\alpha(u) = -r'_\alpha(u)\frac{\partial}{\partial\alpha}H_\alpha(v), \tag{3.35}$$

$$\frac{\partial^2}{\partial\alpha^2}r_\alpha(u) = r''_\alpha(u)\left(\frac{\partial}{\partial\alpha}H_\alpha(v)\right)^2 - r'_\alpha(u)\frac{\partial^2}{\partial\alpha^2}H_\alpha(v), \tag{3.36}$$

$$\frac{\partial^3}{\partial\alpha^3}r_\alpha(u) = -r'''_\alpha(u)\left(\frac{\partial}{\partial\alpha}H_\alpha(v)\right)^3 + 3r''_\alpha(u)\frac{\partial}{\partial\alpha}H_\alpha(v)\cdot\frac{\partial^2}{\partial\alpha^2}H_\alpha(v) - r'_\alpha(u)\frac{\partial^3}{\partial\alpha^3}H_\alpha(v). \tag{3.37}$$

Therefore, from Proposition 2 and Proposition 3, the proof is complete. □

3.5.2 Proof of Theorem 4

Let $f_*(u \mid \alpha) := \phi(r_\alpha(u))$. Let $l_*(\alpha \mid u) := \log f_*(u \mid \alpha)$. For $\alpha = 0$, since the optimization problem can be regarded as an ordinary least squares method, this theorem clearly holds. Therefore, in the following, we consider the case $\alpha \neq 0$. From $u = (y-\mu)/\sigma$ and $\mu = \boldsymbol{X}^\top\boldsymbol{\beta}$, since

$$\begin{aligned}\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\ell(\boldsymbol{\theta} \mid \mathcal{D}_N) &= -\frac{1}{N}\sum_{n=1}^N \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}l(\boldsymbol{\beta},\sigma,\alpha \mid \boldsymbol{X}_n, y_n)\\ &= -\frac{1}{N\sigma^2}\sum_{n=1}^N \boldsymbol{X}_n\boldsymbol{X}_n^\top\frac{\partial^2}{\partial u_n^2}l_*(\alpha \mid u_n),\end{aligned} \tag{3.38}$$

it is sufficient for Theorem 4 to verify the convexity of $l_*(\alpha \mid u_n)$ for u_n. Letting $v_n := r_\alpha(u_n)$, we have

$$\begin{aligned}-\frac{\partial^2}{\partial u_n^2}l_*(\alpha \mid u_n) &= \frac{1}{2}\frac{\partial^2}{\partial u_n^2}r_\alpha^2(u_n)\\ &= \left(r'_\alpha(u_n)\right)^2 + r_\alpha(u_n)r''_\alpha(u_n)\\ &= \frac{1+h_\alpha(v_n)-v_nh'_\alpha(v_n)}{\left(1+h_\alpha(v_n)\right)^3},\end{aligned} \tag{3.39}$$

where the last equality holds from the calculation in the proof of Proposition 2.

Since $|h_\alpha(v_n)| < \rho_\alpha < 1$ from Proposition 1, the denominator of (3.39) is bounded by

$$0 < (1-\rho_\alpha)^3 < \left(1+h_\alpha(v_n)\right)^3 < (1+\rho_\alpha)^3 < 8. \tag{3.40}$$

The numerator of (3.39) is obviously 1 for $v_n = 0$. Consider the case $v_n \neq 0$. The numerator of (3.39) is expressed as

$$\begin{aligned} 1 + h_\alpha(v_n) - v_n h'_\alpha(v_n) &= 1 - v_n^2 \left(\frac{h_\alpha(v_n)}{v_n} \right)' \\ &= 1 - \rho_\alpha \left(\frac{\alpha v_n}{\sqrt{1 + \alpha^2 v_n^2}} \right)^3, \end{aligned} \tag{3.41}$$

and then

$$0 < 1 - \rho_\alpha < 1 + h_\alpha(v_n) - v_n h'_\alpha(v_n) < 1 + \rho_\alpha < 2. \tag{3.42}$$

Combining (3.39), (3.40), and (3.42), we have

$$0 < \frac{1 - \rho_\alpha}{(1 + \rho_\alpha)^3} < \frac{1 + h_\alpha(v_n) - v_n h'_\alpha(v_n)}{(1 + h_\alpha(v_n))^3} < \frac{1 + \rho_\alpha}{(1 - \rho_\alpha)^3} < \infty. \tag{3.43}$$

Thus, there exist some constants $C_L := (1 - \rho_\alpha)/(1 + \rho_\alpha)^3 > 0$ and $C_U := (1 + \rho_\alpha)/(1 - \rho_\alpha)^3 > 0$, satisfying

$$\frac{C_L}{N\sigma^2} \sum_{n=1}^{N} \boldsymbol{X}_n \boldsymbol{X}_n^\top \prec \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \ell(\boldsymbol{\theta} \mid \mathcal{D}_N) \prec \frac{C_U}{N\sigma^2} \sum_{n=1}^{N} \boldsymbol{X}_n \boldsymbol{X}_n^\top. \tag{3.44}$$

From Assumption 1, α is bounded and then we see that C_L is bounded above zero and C_U is bounded below. In particular, if S_N is positive definite, the minimum eigenvalue of $\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \ell(\boldsymbol{\theta} \mid \mathcal{D}_N)$ is positive from (3.44). Hence, $\ell(\boldsymbol{\theta} \mid \mathcal{D}_N)$ is strongly convex for $\boldsymbol{\beta}$. The proof is complete.

3.5.3 Derivation of MM Algorithm

Consider fixing (σ, α) and optimizing $\boldsymbol{\beta}$. According to Fujisawa and Abe (2015), the function r_α corresponding to (3.4) is

$$r_\alpha(u) = \begin{cases} \frac{1}{\alpha(1-\rho_\alpha^2)} \left(\alpha u + \rho_\alpha - \rho_\alpha \sqrt{(\alpha u + \rho_\alpha)^2 + (1 - \rho_\alpha^2)} \right) & (\alpha \neq 0) \\ u & (\alpha = 0) \end{cases}. \tag{3.45}$$

Let $\ell_F(\boldsymbol{\beta}) := \ell(\boldsymbol{\theta} \mid \mathcal{D}_N) + \lambda \|\boldsymbol{\beta}\|_1$ with fixed (σ, α). For $\alpha \neq 0$, the optimize problem (3.7) is rewritten as

$$\arg\min_{\boldsymbol{\beta}} \ell_F(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} -\frac{1}{N} \sum_{n=1}^{N} \log f(y_n \mid \boldsymbol{X}_n^\top \boldsymbol{\beta}, \sigma, \alpha) + \lambda \|\boldsymbol{\beta}\|_1$$

$$
\begin{aligned}
&= \arg\min_{\boldsymbol{\beta}} \frac{1}{2N} \sum_{n=1}^{N} \left(\frac{z_n - \rho_\alpha w_n}{\alpha(1-\rho_\alpha^2)} \right)^2 + \lambda \|\boldsymbol{\beta}\|_1 \\
&= \arg\min_{\boldsymbol{\beta}} \frac{1}{2N} \sum_{n=1}^{N} \frac{z_n^2 - 2\rho_\alpha z_n w_n + \rho_\alpha^2 (z_n^2 + 1 - \rho_\alpha^2)}{\alpha^2 (1-\rho_\alpha^2)^2} + \lambda \|\boldsymbol{\beta}\|_1 \\
&= \arg\min_{\boldsymbol{\beta}} \frac{1}{2N} \sum_{n=1}^{N} \frac{(1+\rho_\alpha^2) z_n^2 - 2\rho_\alpha z_n w_n}{\alpha^2 (1-\rho_\alpha^2)^2} + \lambda \|\boldsymbol{\beta}\|_1 \\
&= \arg\min_{\boldsymbol{\beta}} \frac{1+\rho_\alpha^2}{2N\alpha^2(1-\rho_\alpha^2)^2} \sum_{n=1}^{N} \left(z_n^2 - \frac{2\rho_\alpha}{1+\rho_\alpha^2} z_n w_n \right) + \lambda \|\boldsymbol{\beta}\|_1,
\end{aligned}
\tag{3.46}
$$

where $z_n(\boldsymbol{\beta}) := \frac{\alpha}{\sigma}(y_n - \boldsymbol{X}_n^\top \boldsymbol{\beta}) + \rho_\alpha$ and $w_n(\boldsymbol{\beta}) := \sqrt{z_n^2 + 1 - \rho_\alpha^2}$. Thus, there exists some constant C_F, satisfying

$$
\ell_F(\boldsymbol{\beta}) = \frac{1+\rho_\alpha^2}{2N\alpha^2(1-\rho_\alpha^2)^2} \sum_{n=1}^{N} \left(z_n^2 - \frac{2\rho_\alpha}{1+\rho_\alpha^2} z_n w_n \right) + \lambda \|\boldsymbol{\beta}\|_1 + C_F. \tag{3.47}
$$

Note that $\frac{\partial}{\partial \boldsymbol{\beta}} z_n = -\frac{\alpha}{\sigma} \boldsymbol{X}_n$ and $\frac{\partial}{\partial \boldsymbol{\beta}} w_n = -\frac{\alpha}{\sigma} \frac{z_n}{w_n} \boldsymbol{X}_n$. Letting $\hat{l}_n(\boldsymbol{\beta}) := z_n^2 - \frac{2\rho_\alpha}{1+\rho_\alpha^2} z_n w_n$, we have

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}} \hat{l}_n(\boldsymbol{\beta}) &= -\frac{2\alpha}{\sigma} z_n \boldsymbol{X}_n + \frac{2\alpha\rho_\alpha}{\sigma(1+\rho_\alpha^2)} \left(w_n + \frac{z_n^2}{w_n} \right) \boldsymbol{X}_n \\
&= -\frac{2\alpha}{\sigma} z_n \boldsymbol{X}_n + \frac{2\alpha\rho_\alpha}{\sigma(1+\rho_\alpha^2)} \left(2w_n - \frac{1-\rho_\alpha^2}{w_n} \right) \boldsymbol{X}_n,
\end{aligned}
\tag{3.48}
$$

$$
\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \hat{l}_n(\boldsymbol{\beta}) = \frac{2\alpha^2}{\sigma^2} \boldsymbol{X}_n \boldsymbol{X}_n^\top - \frac{2\alpha^2\rho_\alpha}{\sigma^2(1+\rho_\alpha^2)} \left(2 + \frac{1-\rho_\alpha^2}{w_n^2} \right) \frac{z_n}{w_n} \boldsymbol{X}_n \boldsymbol{X}_n^\top. \tag{3.49}
$$

Since $\left(2 + \frac{1-\rho_\alpha^2}{w_n^2}\right) \frac{z_n}{w_n} \to \pm 2$ as $z_n \to \pm\infty$ and

$$
\frac{\partial}{\partial z_n} \left(2 + \frac{1-\rho_\alpha^2}{w_n^2} \right) \frac{z_n}{w_n} = \frac{3(1-\rho_\alpha^2)^2}{w_n^5} > 0, \tag{3.50}
$$

we have $\left| \left(2 + \frac{1-\rho_\alpha^2}{w_n^2}\right) \frac{z_n}{w_n} \right| < 2$. Then, using $0 \le \frac{\rho_\alpha}{1+\rho_\alpha^2} < \frac{1}{2}$ from $0 \le \rho_\alpha < 1$, we have from (3.49),

$$
\boldsymbol{O} \prec \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \hat{l}_n(\boldsymbol{\beta}) \prec \frac{4\alpha^2}{\sigma^2} \boldsymbol{X}_n \boldsymbol{X}_n^\top. \tag{3.51}
$$

From (3.47) and (3.51), we can construct a surrogate function $\ell_G(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)})$ of $\ell_F(\boldsymbol{\beta})$ near $\boldsymbol{\beta}^{(t)}$ based on the second-order Taylor expansion as follows (see Lemma 12

of Sun, Babu, and Palomar (2016) for details):

$$\begin{aligned}\ell_F(\boldsymbol{\beta}) \leq & \frac{1+\rho_\alpha^2}{2N\alpha^2(1-\rho_\alpha^2)^2}\sum_{n=1}^{N}\left(\hat{l}_n(\boldsymbol{\beta}^{(t)}) + \left.\frac{\partial \hat{l}_n}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}}(\boldsymbol{\beta}-\boldsymbol{\beta}^{(t)})\right. \\ & \left. + \frac{2\alpha^2}{\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\beta}^{(t)})^\top \boldsymbol{X}_n\boldsymbol{X}_n^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^{(t)})\right) + \lambda\|\boldsymbol{\beta}\|_1 + C_F \\ =: & \ \ell_G(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}). \end{aligned} \tag{3.52}$$

Therefore, from (3.48) and (3.52), we can update $\boldsymbol{\beta}^{(t)}$ at the t-th iteration as follows:

$$\begin{aligned}\boldsymbol{\beta}^{(t+1)} &= \arg\min_{\boldsymbol{\beta}} \ell_G(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)}) \\ &= \arg\min_{\boldsymbol{\beta}} \frac{1+\rho_\alpha^2}{2N\alpha^2(1-\rho_\alpha^2)^2}\sum_{n=1}^{N}\left(\left.\frac{\partial \hat{l}_n}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}}(\boldsymbol{\beta}-\boldsymbol{\beta}^{(t)})\right. \\ &\qquad \left. + \frac{2\alpha^2}{\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\beta}^{(t)})^\top \boldsymbol{X}_n\boldsymbol{X}_n^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^{(t)})\right) + \lambda\|\boldsymbol{\beta}\|_1 \\ &= \arg\min_{\boldsymbol{\beta}} \frac{1+\rho_\alpha^2}{N\sigma^2(1-\rho_\alpha^2)^2}\sum_{n=1}^{N}\left(\tilde{y}_n - \boldsymbol{X}_n^\top\boldsymbol{\beta}\right)^2 + \lambda\|\boldsymbol{\beta}\|_1, \end{aligned} \tag{3.53}$$

where $\tilde{y}_n := \boldsymbol{X}_n^\top\boldsymbol{\beta}^{(t)} + \frac{\sigma}{2\alpha}\left(z_n^{(t)} - \frac{\rho_\alpha}{1+\rho_\alpha^2}\left(w_n^{(t)} + \frac{z_n^{(t)2}}{w_n^{(t)}}\right)\right)$ with $z_n^{(t)} := z_n(\boldsymbol{\beta}^{(t)})$ and $w_n^{(t)} := w_n(\boldsymbol{\beta}^{(t)})$.

For $\alpha = 0$, the optimize problem (3.7) is rewritten as the ordinary Lasso, i.e.,

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2N\sigma^2}\sum_{n=1}^{N}\left(y_n - \boldsymbol{X}_n^\top\boldsymbol{\beta}\right)^2 + \lambda\|\boldsymbol{\beta}\|_1. \tag{3.54}$$

3.5.4 Proof of Theorem 5

First, we fix (σ, α) and consider the convergence of the MM algorithm for $\boldsymbol{\beta}$. We will prove that the surrogate function $\ell_G(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)})$ in Section 3.5.3 is a first-order surrogate of $\ell_F(\boldsymbol{\beta})$ near $\boldsymbol{\beta}^{(t)}$ based on Definition 2.1 of Mairal (2013).

From (3.52), $\ell_G(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)})$ is obviously a majorant function of $\ell_F(\boldsymbol{\beta})$. From (3.44) and (3.51), we have

$$\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\ell_F(\boldsymbol{\beta}) \prec \frac{C_U}{N\sigma^2}\sum_{n=1}^{N}\boldsymbol{X}_n\boldsymbol{X}_n^\top \prec \underbrace{\frac{C_U}{\sigma^2}\Lambda_{\max}(S_N)}_{=:L_F}\boldsymbol{I}, \tag{3.55}$$

$$\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\ell_G(\boldsymbol{\beta}\mid\boldsymbol{\beta}^{(t)}) \prec \frac{2(1+\rho_\alpha^2)}{N\sigma^2(1-\rho_\alpha^2)^2}\sum_{n=1}^N \boldsymbol{X}_n\boldsymbol{X}_n^\top \prec \underbrace{\frac{2(1+\rho_\alpha^2)}{\sigma^2(1-\rho_\alpha^2)^2}\Lambda_{\max}(S_N)}_{=:L_G}\boldsymbol{I}, \quad (3.56)$$

where $\Lambda_{\max}(S_N)$ is the largest eigenvalue of $S_N := \frac{1}{N}\sum_{n=1}^N \boldsymbol{X}_n\boldsymbol{X}_n^\top$. Hence, $\frac{\partial}{\partial\boldsymbol{\beta}}\ell_F(\boldsymbol{\beta})$ and $\frac{\partial}{\partial\boldsymbol{\beta}}\ell_G(\boldsymbol{\beta}\mid\boldsymbol{\beta}^{(t)})$ are L_F-Lipschitz continuous and L_G-Lipschitz continuous, respectively. Let $\ell_H(\boldsymbol{\beta}\mid\boldsymbol{\beta}^{(t)}) := \ell_G(\boldsymbol{\beta}\mid\boldsymbol{\beta}^{(t)}) - \ell_F(\boldsymbol{\beta})$. Then, $\frac{\partial}{\partial\boldsymbol{\beta}}\ell_H(\boldsymbol{\beta}\mid\boldsymbol{\beta}^{(t)})$ is also (L_F+L_G)-Lipschitz continuous because for any $\boldsymbol{\beta}_\mathbf{x}$ and $\boldsymbol{\beta}_\mathbf{y}$,

$$\begin{aligned}
&\left\|\frac{\partial}{\partial\boldsymbol{\beta}}\ell_H(\boldsymbol{\beta}_\mathbf{x}\mid\boldsymbol{\beta}^{(t)}) - \frac{\partial}{\partial\boldsymbol{\beta}}\ell_H(\boldsymbol{\beta}_\mathbf{y}\mid\boldsymbol{\beta}^{(t)})\right\| \\
&\le \left\|\frac{\partial}{\partial\boldsymbol{\beta}}\ell_G(\boldsymbol{\beta}_\mathbf{x}\mid\boldsymbol{\beta}^{(t)}) - \frac{\partial}{\partial\boldsymbol{\beta}}\ell_G(\boldsymbol{\beta}_\mathbf{y}\mid\boldsymbol{\beta}^{(t)})\right\| + \left\|\frac{\partial}{\partial\boldsymbol{\beta}}\ell_F(\boldsymbol{\beta}_\mathbf{x}) - \frac{\partial}{\partial\boldsymbol{\beta}}\ell_F(\boldsymbol{\beta}_\mathbf{y})\right\| \\
&\le L_G\|\boldsymbol{\beta}_\mathbf{x}-\boldsymbol{\beta}_\mathbf{y}\| + L_F\|\boldsymbol{\beta}_\mathbf{x}-\boldsymbol{\beta}_\mathbf{y}\| \\
&= (L_F+L_G)\|\boldsymbol{\beta}_\mathbf{x}-\boldsymbol{\beta}_\mathbf{y}\|. \qquad (3.57)
\end{aligned}$$

Moreover, we have $\ell_H(\boldsymbol{\beta}^{(t)}\mid\boldsymbol{\beta}^{(t)}) = 0$ and $\frac{\partial}{\partial\boldsymbol{\beta}}\ell_H(\boldsymbol{\beta}^{(t)}\mid\boldsymbol{\beta}^{(t)}) = 0$.

Therefore, $\ell_G(\boldsymbol{\beta}\mid\boldsymbol{\beta}^{(t)})$ is a first-order surrogate of $\ell_F(\boldsymbol{\beta})$ near $\boldsymbol{\beta}^{(t)}$. From Proposition 2.2 of Mairal (2013), we have for any $t \ge 1$ and some positive constant R,

$$\tilde{\ell}(\boldsymbol{\beta}^{(t)},\sigma,\alpha) - \tilde{\ell}(\boldsymbol{\beta}^*,\sigma,\alpha) \le \frac{2(L_F+L_G)R^2}{t+2}, \quad (3.58)$$

where $\tilde{\ell}(\boldsymbol{\theta}) := \ell(\boldsymbol{\theta}\mid\mathcal{D}_N) + \lambda\|\boldsymbol{\beta}\|_1$ and $\boldsymbol{\beta}^*$ is a minimizer of $\tilde{\ell}$ with fixed (σ,α). Hence, it holds that $\boldsymbol{\beta}^{(t)} \to \boldsymbol{\beta}^*$ as $t\to\infty$.

Next, we regard Algorithm 1 as the block coordinate descent (BCD) method for the three blocks $(\boldsymbol{\beta},\sigma,\alpha)$ and consider its convergence based on Theorem 4.1 (c) of Tseng (2001). When using the cyclic rule (see Section 2 of Tseng (2001)) to update the BCD method for B blocks, Theorem 4.1 (c) of Tseng (2001) states the following.

Theorem 4.1 (c) of Tseng (2001). *Assume that the level set $\Theta^0 := \{\boldsymbol{\theta}\mid\tilde{\ell}(\boldsymbol{\theta}\mid\mathcal{D}_N) \le \tilde{\ell}(\boldsymbol{\theta}^{(0)}\mid\mathcal{D}_N)\}$ is compact and that $\tilde{\ell}$ is continuous on Θ^0. If $\tilde{\ell}(\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_B)$ has at most one minimum in $\boldsymbol{\theta}_b$ for $b=2,\dots,B-1$, then every cluster point $\boldsymbol{\theta}^\dagger$ of $\{\boldsymbol{\theta}^{(t)}\}_{t:=(B-1) \bmod B}$ is a coordinatewise minimum point of $\tilde{\ell}$. In addition, if $\tilde{\ell}$ is regular at $\boldsymbol{\theta}^\dagger$, then $\boldsymbol{\theta}^\dagger$ is a stationary point of $\tilde{\ell}$.*

We assume that there always exists $(y_n,\boldsymbol{X}_n)\in\mathcal{D}_N$ such that $y_n - \boldsymbol{X}_n^\top\boldsymbol{\beta} > 0$ after updating $\boldsymbol{\beta}$ by the MM algorithm, and also, there always exists $(y_n,\boldsymbol{X}_n)\in\mathcal{D}_N$ such that $y_n - \boldsymbol{X}_n^\top\boldsymbol{\beta} < 0$. Under this assumption, it holds that $\tilde{\ell}(\boldsymbol{\theta}\mid\mathcal{D}_N)\to\infty$ as $\sigma\downarrow 0$, $\sigma\to\infty$, $\alpha\to-\infty$, or $\alpha\to\infty$.

Let dom ℓ and dom $\tilde{\ell}$ be the effective domain of ℓ and $\tilde{\ell}$, respectively. We can see that $\tilde{\ell}(\boldsymbol{\theta} \mid \mathcal{D}_N)$ is continuous in dom $\tilde{\ell}$ and the level set $\boldsymbol{\Theta}^0$ is compact. From Lemma 3.1 of Tseng (2001), $\tilde{\ell}$ is regular at any $\boldsymbol{\theta} \in$ dom $\tilde{\ell}$ since dom ℓ is open and ℓ is Gâteaux-differentiable on dom ℓ. Therefore, from Theorem 4.1 (c) of Tseng (2001), if $\tilde{\ell}(\boldsymbol{\theta} \mid \mathcal{D}_N)$ has at most one minimum in σ, then Theorem 5 holds. We will prove that this necessary condition holds.

Let $\chi(u) := u r_\alpha(u) r'_\alpha(u)$. We have

$$\begin{aligned}\chi'(u) &= r_\alpha(u) r'_\alpha(u) + u \left(r'_\alpha(u)\right)^2 + u r_\alpha(u) r''_\alpha(u) \\ &= r_\alpha(u) r'_\alpha(u) + u \left(\left(r'_\alpha(u)\right)^2 + r_\alpha(u) r''_\alpha(u) \right). \end{aligned} \tag{3.59}$$

Since $r'_\alpha(u) > 0$ from (3.14) and $r_\alpha(0) = 0$ from (3.45), it holds that

$$\begin{aligned}\mathrm{sgn}(r_\alpha(u) r'_\alpha(u)) &= \mathrm{sgn}(r(u)) \\ &= \mathrm{sgn}(u). \end{aligned} \tag{3.60}$$

Since $\left(r'_\alpha(u)\right)^2 + r_\alpha(u) r''_\alpha(u) > 0$ from (3.39) and (3.43), it holds that

$$\mathrm{sgn}\left(u \left(\left(r'_\alpha(u)\right)^2 + r_\alpha(u) r''_\alpha(u) \right) \right) = \mathrm{sgn}(u). \tag{3.61}$$

Hence, we have

$$\mathrm{sgn}(\chi'(u)) = \mathrm{sgn}(u). \tag{3.62}$$

Let $u_n := \frac{y_n - \boldsymbol{X}_n^\top \boldsymbol{\beta}}{\sigma}$. For $y_n - \boldsymbol{X}_n^\top \boldsymbol{\beta} > 0$, $u_n \in (0, \infty)$ decreases monotonically as σ increases. Since $\chi'(u_n) > 0$ for $u_n \in (0, \infty)$ from (3.62), $\chi(u_n)$ is a strictly decreasing function for σ. For $y_n - \boldsymbol{X}_n^\top \boldsymbol{\beta} < 0$, $u_n \in (-\infty, 0)$ increases monotonically as σ increases. Since $\chi'(u_n) < 0$ for $u_n \in (-\infty, 0)$ from (3.62), $\chi(u_n)$ is also a strictly decreasing function for σ. For $y_n - \boldsymbol{X}_n^\top \boldsymbol{\beta} = 0$, $\chi(u_n) = \chi(0) = 0$.

Thus, $\frac{\partial}{\partial \sigma} \ell(\boldsymbol{\theta} \mid \mathcal{D}_N)$ is a strictly increasing function for σ, because

$$\begin{aligned}\frac{\partial}{\partial \sigma} \ell(\boldsymbol{\theta} \mid \mathcal{D}_N) &= -\frac{1}{N} \sum_{n=1}^{N} \frac{\partial}{\partial \sigma} l(\boldsymbol{\theta} \mid \boldsymbol{X}_n, y_n) \\ &= \frac{1}{N\sigma} \sum_{n=1}^{N} \left(1 - u_n r_\alpha(u_n) r'_\alpha(u_n)\right) \\ &= \frac{1}{N\sigma} \left(N - \sum_{n=1}^{N} \chi(u_n) \right). \end{aligned} \tag{3.63}$$

Moreover, there exists only one $\sigma(=: \tilde{\sigma})$ satisfying $\sum_{n=1}^{N} \chi(u_n) = N$, because

$$\lim_{\sigma \downarrow 0} \sum_{n=1}^{N} \chi(u_n) = \underbrace{\lim_{u_n \to -\infty} \sum_{u_n < 0} \chi(u_n)}_{\to \infty} + \underbrace{\lim_{u_n \to \infty} \sum_{u_n > 0} \chi(u_n)}_{\to \infty}$$
$$= \infty, \tag{3.64}$$

$$\lim_{\sigma \to \infty} \sum_{n=1}^{N} \chi(u_n) = \underbrace{\lim_{u_n \uparrow 0} \sum_{u_n < 0} \chi(u_n)}_{\to 0} + \underbrace{\lim_{u_n \downarrow 0} \sum_{u_n > 0} \chi(u_n)}_{\to 0}$$
$$= 0. \tag{3.65}$$

$\tilde{\sigma}$ is also the only solution satisfying $\frac{\partial}{\partial \sigma} \ell(\boldsymbol{\theta} \mid \mathcal{D}_N) = 0$ and the only minimum of $\ell(\boldsymbol{\theta} \mid \mathcal{D}_N)$. The proof is complete.

Note that from Theorem 4, if S_N is positive definite, $\ell(\boldsymbol{\theta} \mid \mathcal{D}_N)$ is strongly convex for $\boldsymbol{\beta}$ and has only one minimum in $\boldsymbol{\beta}$. In this case, σ can be optimized first by swapping lines 2 to 9 and line 10 in Algorithm 1, and then every cluster point in a sequence of $\boldsymbol{\theta}$ after updating $\boldsymbol{\beta}$ is a stationary point.

Chapter 4

Theoretical Results

This section provides theoretical guarantees for the proposed method by non-asymptotic analysis. First, we present some basic properties of the first, second, and third derivatives of $\ell(\boldsymbol{\psi} \mid y) = \log f(y \mid \boldsymbol{\psi})$. Next, using them, we show that the excess risk has a quadratic margin. Finally, we show the upper bounds of the average excess risk and the estimation error, and then we verify the favorable properties of the proposed method. All proofs of propositions and theorems in this section are given in Section 4.6.

4.1 Basic Properties of Likelihood

We present some basic properties of the first, second, and third derivatives of $\ell(\boldsymbol{\psi} \mid y)$. These properties are the basis of non-asymptotic analysis.

Suppose that $\mathcal{X} = \{\boldsymbol{X}\}$ is included in a bounded space of $\mathbb{R}^P$. We assume that the parameter spaces $\Psi = \{\boldsymbol{\psi}\}$ and $\Theta = \{\boldsymbol{\theta}\}$ are bounded in the following sense. Hereafter, we suppose that the following assumption is satisfied.

Assumption 1. *For some positive constant K, we suppose*

$$\Psi \subset \tilde{\Psi} = \{\boldsymbol{\psi} : |\mu| \leq K,\ 1/K \leq |\sigma| \leq K,\ |\alpha| \leq K\}, \tag{4.1}$$

$$\Theta \subset \tilde{\Theta} = \{\boldsymbol{\theta} : \boldsymbol{\psi}(\boldsymbol{X}) \in \tilde{\Psi} \textit{ for any } \boldsymbol{X} \in \mathcal{X}\}. \tag{4.2}$$

The true parameters $\boldsymbol{\psi}^0$ and $\boldsymbol{\theta}^0$ are included in Ψ and Θ, respectively.

Let the score function (first derivative) of $l(\boldsymbol{\psi} \mid y) := \log f(y \mid \boldsymbol{\psi})$ be denoted by

$$s_{\boldsymbol{\psi}}(y) := \frac{\partial}{\partial \boldsymbol{\psi}} l(\boldsymbol{\psi} \mid y). \tag{4.3}$$

Proposition 5. *We have*

$$\|s_{\boldsymbol{\psi}}(y)\|_\infty \leq C_{1,2}|y|^2 + C_{1,1}|y| + C_{1,0}, \tag{4.4}$$

where $C_{1,2}, C_{1,1}, C_{1,0}$ *are some positive constants.*

The score functions of regression parameters μ and σ are bounded by the first and second-order polynomials of $|y|$. This is the same property as the normal linear regression model. We may expect that the score function of skew parameter α is bounded by a third-order polynomial of $|y|$ because the skewness usually depends on the third moment. However, it is bounded by a second-order polynomial of $|y|$. Proposition 5 implies that the proposed model is easy to treat α as well as σ.

Let the Fisher information matrix of $f_{\boldsymbol{\psi}}(y) := f(y \mid \boldsymbol{\psi})$ be denoted by

$$\mathcal{I}(\boldsymbol{\psi}) := \mathbb{E}_{f_\psi}\left[-\frac{\partial^2}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\top} l(\boldsymbol{\psi} \mid y)\right]. \tag{4.5}$$

Proposition 6. *Let* $\boldsymbol{\psi}^0(\boldsymbol{X}) := [\boldsymbol{X}^\top\boldsymbol{\beta}^0, \sigma^0, \alpha^0]^\top$. *The Fisher information matrix* $\mathcal{I}(\boldsymbol{\psi}^0(\boldsymbol{X}))$ *is uniformly positive definite for any* $\boldsymbol{X}$, *in other words,* $\inf_{\boldsymbol{X}} \Lambda_{\min}(\mathcal{I}(\boldsymbol{\psi}^0(\boldsymbol{X}))) > 0$, *where* $\Lambda_{\min}(A)$ *is the smallest eigenvalue of* A.

If the model is simple, we can easily prove a similar proposition to Proposition 6 because the Fisher information matrix is easily calculated. However, the model (3.1) is not simple, and it is not easy to verify Proposition 6, as seen in a devised proof of Section 4.6.1. We found that the constant factor ρ_α proposed by Fujisawa and Abe (2015) does not satisfy Proposition 6 at $\alpha = 0$. This is why a constant factor ρ_α is slightly modified, as described in Remark 1, and then we can prove Proposition 6.

Proposition 7. *There exists some function* $G_3(y)$ *such that*

$$\sup_{(i_1,i_2,i_3)\in\{1,2,3\}^3}\left|\frac{\partial^3}{\partial\psi_{i_1}\partial\psi_{i_2}\partial\psi_{i_3}} l(\boldsymbol{\psi} \mid y)\right| \leq G_3(y),$$
$$C_3 := \sup_{\boldsymbol{X}} \int G_3(y) f(y \mid \boldsymbol{\psi}^0(\boldsymbol{X}))dy < \infty. \tag{4.6}$$

4.2 Excess Risk and Margin

Let the Kullback–Leibler divergence between f_{ψ^0} and f_{ψ}, which is called the *excess risk*, be denoted by

$$\mathcal{E}(\psi \mid \psi^0) := \mathbb{E}_{f_{\psi^0}}\left[\log \frac{f_{\psi^0}(y)}{f_{\psi}(y)}\right]. \tag{4.7}$$

Proposition 8. *For any $\epsilon > 0$, there exists some constant $\delta_\epsilon > 0$ such that*

$$\inf_{X} \inf_{\|\psi - \psi^0(X)\|_2 > \epsilon} \mathcal{E}(\psi \mid \psi^0(X)) \geq \delta_\epsilon. \tag{4.8}$$

Proposition 8 shows the *identifiability condition*, which is usually necessary to estimate the true parameter by a statistical method correctly. The identifiability condition is often assumed implicitly, but it is precisely verified here because the model (3.1) is complicated due to skew noise.

Theorem 6. *There exists some constant $C > 0$ such that*

$$\inf_{X} \frac{\mathcal{E}\left(\psi \mid \psi^0(X)\right)}{\|\psi - \psi^0(X)\|_2^2} \geq \frac{1}{C^2}. \tag{4.9}$$

Theorem 6 says that the proposed method has a *quadratic margin*, implying the error $\mathcal{E}\left(\psi \mid \psi^0(X)\right)$ presents a sharper behavior around $\psi^0(X)$ than the well-used squared error $\|\psi - \psi^0(X)\|_2^2$ up to a constant factor.

4.3 Convergence Rate

The following is a well-used condition in non-asymptotic analysis for the ℓ_1 penalized method, which is called the *restricted eigenvalue condition*.

Assumption 2. *Let $\mathcal{S}_0 := \{p : \beta_p^0 \neq 0\}$ and $\mathcal{S}_0^C := \{1, \ldots, P\} \backslash \mathcal{S}_0$. With $L > 1$, define the $(L, \mathcal{S}_0)$-restricted eigenvalue constant by*

$$\tilde{\kappa}^2(L, \mathcal{S}_0) := \inf_{\|\beta_{\mathcal{S}_0^C}\|_1 \leq L\|\beta_{\mathcal{S}_0}\|_1} \frac{\beta^\top S_N \beta}{\|\beta\|_2^2}. \tag{4.10}$$

We set $L = 6$ and assume $0 < \tilde{\kappa}(L, \mathcal{S}_0) \leq 1$.

The squared loss of the simple linear regression model for a low-dimensional case is strongly convex because the Hessian matrix of the squared loss concerning the regression parameter vector $\boldsymbol{\beta}$ is S_N and it is positive definite. This property provides the strength of the least squared estimator. However, for the high-dimensional case $N < P$, S_N may not be of full rank; in other words, its smallest eigenvalue may be zero, and then the squared loss is not strongly convex. The non-asymptotic analysis reveals that if the smallest eigenvalue of S_N under the *restricted condition* $\|\boldsymbol{\beta}_{\mathcal{S}_0^{\mathbb{C}}}\|_1 \leq L\|\boldsymbol{\beta}_{\mathcal{S}_0}\|_1$ is positive, such as Assumption 2, we can show theoretical guarantees of the ℓ_1 regularized estimator (Bühlmann and Geer, 2011; Hastie, Tibshirani, and Wainwright, 2015).

In Assumption 2, we set $L = 6$ to simplify later notations. We can replace it with any constant greater than one by adjusting how to choose the regularization parameter. Based on this assumption, we have the following important theorem.

Theorem 7. *Suppose that Assumption 2 is satisfied. Fix $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$. Let the average excess risk and the empirical process of the loss function (3.6) be denoted by*

$$\bar{\mathcal{E}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^0) := \frac{1}{N}\sum_{n=1}^{N} \mathcal{E}\left(\boldsymbol{\psi}(\boldsymbol{X}_n) \mid \boldsymbol{\psi}^0(\boldsymbol{X}_n)\right), \tag{4.11}$$

$$\mathcal{V}_N(\boldsymbol{\theta}) := \frac{1}{N}\sum_{n=1}^{N} \left(l\left(\boldsymbol{\psi}(\boldsymbol{X}_n) \mid y_n\right) - \mathbb{E}\left[l\left(\boldsymbol{\psi}(\boldsymbol{X}_n) \mid y\right)\right]\right). \tag{4.12}$$

Let $\|\boldsymbol{\theta}\|_* := \|\boldsymbol{\beta}\|_1 + \|[\sigma, \alpha]^\top\|_2$. Fix $\gamma \geq 1$ and $\lambda_0 > 0$. Let

$$\mathcal{J} := \left\{ \frac{\left|\mathcal{V}_N(\hat{\boldsymbol{\theta}}) - \mathcal{V}_N(\boldsymbol{\theta}^0)\right|}{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\|_* \vee \lambda_0} \leq \gamma\lambda_0 \right\}. \tag{4.13}$$

If the regularization parameter λ is chosen to satisfy $\lambda \geq 2\gamma\lambda_0$, then we have on $\mathcal{J}$

$$\bar{\mathcal{E}}(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta}^0) + 2(\lambda - \gamma\lambda_0)\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 \leq 9(\lambda + \gamma\lambda_0)^2 C^2 \kappa^2 s_0, \tag{4.14}$$

where C is the constant defined in Theorem 6, $\kappa^{-1} := \tilde{\kappa}(L, \mathcal{S}_0)$, and $s_0 := \#\mathcal{S}_0$.

Suppose

$$\lambda_0 = O\left((\log N)^2 \sqrt{\frac{\log(N \vee P)}{N}}\right). \tag{4.15}$$

This order is necessary to prove Theorem 8 and is relevant to the following two corollaries. From Theorem 7, we can obtain convergence rates regarding the average excess risk $\bar{\mathcal{E}}(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta}^0)$ and the estimation error of $\hat{\boldsymbol{\beta}}$.

Corollary 3. *Suppose* $\lambda = 2\gamma\lambda_0$ *with* (4.15) *and the same assumptions as in Theorem 7. Then, we have*

$$\bar{\mathcal{E}}(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta}^0) = O_{\mathbb{P}}\left(\frac{s_0(\log N)^4 \log(N \vee P)}{N}\right), \tag{4.16}$$

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 = O_{\mathbb{P}}\left(s_0(\log N)^2\sqrt{\frac{\log(N \vee P))}{N}}\right). \tag{4.17}$$

The orders of the average excess risk (4.16) and the estimation error (4.17) are $(\log N)^4$ and $(\log N)^2$ larger than those of the ordinary Lasso, respectively. These are the same as in some well-known complicated models, such as the linear mixed effects models with ℓ_1 regularization (Schelldorfer, Bühlmann, and Geer, 2011). A similar result is also seen in Corollary 4.

4.4 Feature Selection

When the non-zero coefficients are sufficiently large, which is the so-called beta-min condition (Bühlmann, 2013), we can show that the estimated set of non-zero coefficients $\hat{\mathcal{S}} := \{p : \hat{\beta}_p \neq 0\}$ includes the true set of non-zero coefficients $\mathcal{S}_0$ with high probability.

Corollary 4. *Suppose* $\lambda = 2\gamma\lambda_0$ *with* (4.15) *and the same assumptions as in Theorem 7. Suppose the non-zero coefficients of* $\boldsymbol{\beta}_0$ *are sufficiently large so that*

$$\min_{p \in \mathcal{S}_0} |\beta_p^0| \gg O\left(s_0(\log N)^2\sqrt{\frac{\log(N \vee P))}{N}}\right). \tag{4.18}$$

Then, with high probability, it holds $\hat{\mathcal{S}} \supseteq \mathcal{S}_0$.

As in Corollary 3, the order in (4.18) is $(\log N)^2$ larger when compared to the ordinary Lasso.

4.5 Reasonable Success Probability

The statements after Section 4.3 hold when the event $\mathcal{J}$ in (4.13) occurs. Finally, we show that the event $\mathcal{J}$ occurs with high probability.

Theorem 8. *Suppose* $\lambda = 2\gamma\lambda_0$ *with* (4.15) *and the same assumptions as in Theorem 7. The event* $\mathcal{J}$ *occurs with probability at least* $1 - \delta$*, where with some positive constants*

$C_{4,1}, C_{4,2} > 0$,

$$\delta := C_{4,1} \exp\left(-\frac{\gamma^2 (\log N)^2 \log(N \vee P)}{C_{4,2}^2}\right) + \frac{1}{N}. \tag{4.19}$$

4.6 Proofs

4.6.1 Proof of Propositions

Preliminary Propositions

Proposition 9. *The M-th order moment of $U \sim f_*(u \mid \alpha)$ is finite. The M-th order moment of $Y \sim f(y \mid \boldsymbol{X}^\top \boldsymbol{\beta}, \sigma, \alpha)$ is also finite.*

Proof. From Proposition 1, it holds that

$$\begin{aligned}
\mathbb{E}_{U \sim f_*}\left[U^m\right] &= \int u^m \phi\left(r_\alpha(u)\right) du \\
&= \int (v + H_\alpha(v))^m \phi(v)(1 + h_\alpha(v)) dv \\
&\leq \int |v + H_\alpha(v)|^m |1 + h_\alpha(v)| \phi(v) dv \\
&\leq \int (|v| + |H_\alpha(v)|)^m (1 + |h_\alpha(v)|) \phi(v) dv \\
&\leq \int (|v| + \rho_\alpha |v|)^m (1 + \rho_\alpha) \phi(v) dv \\
&= (1 + \rho_\alpha)^{m+1} \int |v|^m \phi(v) dv \\
&= (1 + \rho_\alpha)^{m+1} \mathbb{E}_{V \sim \phi}\left[|V|^m\right] < \infty.
\end{aligned} \tag{4.20}$$

Furthermore, since $y = \sigma u + \mu$, it obviously holds that

$$\begin{aligned}
\mathbb{E}_{Y \sim f}\left[Y^m\right] &= \mathbb{E}_{U \sim f_*}\left[(\sigma U + \boldsymbol{X}^\top \boldsymbol{\beta})^m\right] \\
&= \sum_{k=0}^{m} \binom{m}{k} \sigma^k \left(\boldsymbol{X}^\top \boldsymbol{\beta}\right)^{m-k} \mathbb{E}_{U \sim f_*}\left[U^k\right] < \infty.
\end{aligned} \tag{4.21}$$

□

Proposition 10. *Suppose that Assumption 1 is satisfied. Then, the third-order derivative of the log-likelihood function $l(\boldsymbol{\theta} \mid \boldsymbol{X}, y)$, more precisely, $\left|\frac{\partial^3}{\partial\theta_{i_1}\partial\theta_{i_2}\partial\theta_{i_3}} l(\boldsymbol{\theta} \mid \boldsymbol{X}, y)\right|$ for any $(i_1, i_2, i_3) \in \{1, \ldots, P+2\}^3$ are bounded by a polynomial of $|y|$.*

Proof. The scores of the log-likelihood function $l(\boldsymbol{\theta} \mid \boldsymbol{X}, y)$ is

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta} \mid \boldsymbol{X}, y) &= \frac{\partial}{\partial u}\frac{\partial u}{\partial \mu}\frac{\partial \mu}{\partial \boldsymbol{\beta}} l_*(\alpha \mid u) \\
&= \frac{1}{\sigma} r_\alpha(u) r'_\alpha(u) \boldsymbol{X}, \qquad (4.22) \\
\frac{\partial}{\partial \sigma} l(\boldsymbol{\theta} \mid \boldsymbol{X}, y) &= \frac{\partial}{\partial u}\frac{\partial u}{\partial \sigma} l_*(\alpha \mid u) + \frac{\partial}{\partial \sigma}\log\frac{1}{\sigma} \\
&= \frac{u}{\sigma} r_\alpha(u) r'_\alpha(u) - \frac{1}{\sigma}, \qquad (4.23) \\
\frac{\partial}{\partial \alpha} l(\boldsymbol{\theta} \mid \boldsymbol{X}, y) &= \frac{\partial}{\partial \alpha} l_*(\alpha \mid u) \\
&= -r_\alpha(u)\frac{\partial}{\partial \alpha} r_\alpha(u). \qquad (4.24)
\end{aligned}$$

The second-order derivatives of (4.22), (4.23), and (4.24) with respect to $\boldsymbol{\theta}$ consist of

$$\begin{aligned}
&u,\ r_\alpha(u),\ r'_\alpha(u),\ r''_\alpha(u),\ r'''_\alpha(u), \\
&\frac{\partial}{\partial \alpha} r_\alpha(u),\ \frac{\partial^2}{\partial \alpha^2} r_\alpha(u),\ \frac{\partial^3}{\partial \alpha^3} r_\alpha(u),\ \frac{\partial}{\partial \alpha} r'_\alpha(u),\ \frac{\partial}{\partial \alpha} r''_\alpha(u),\ \frac{\partial^2}{\partial \alpha^2} r'_\alpha(u), \qquad (4.25)
\end{aligned}$$

as terms involving u. From Proposition 2 and Proposition 4, the second to eighth functions of (4.25) are bounded by some sextic function of u. Consider the remaining last three functions of (4.25). We have

$$\begin{aligned}
\frac{\partial}{\partial \alpha} r'_\alpha(u) &= -r''_\alpha(u)\frac{\partial}{\partial \alpha} H_\alpha(v), \qquad (4.26) \\
\frac{\partial}{\partial \alpha} r''_\alpha(u) &= -r'''_\alpha(u)\frac{\partial}{\partial \alpha} H_\alpha(v), \qquad (4.27) \\
\frac{\partial^2}{\partial \alpha^2} r'_\alpha(u) &= r'''_\alpha(u)\left(\frac{\partial}{\partial \alpha} H_\alpha(v)\right)^2 - r''_\alpha(u)\frac{\partial^2}{\partial \alpha^2} H_\alpha(v). \qquad (4.28)
\end{aligned}$$

From Proposition 2 and Proposition 4, the absolutes of the above three functions are bounded by some quintic function of $|u|$. Since $u = (y - \mu)/\sigma$ and Assumption 1, the proof is complete. □

Proposition 11. *Let $\xi(q_\alpha(v)) := \eta(v)$. Then, we have*

$$\mathbb{E}_{U \sim f_*(u|\alpha)}[\xi(U)] = \mathbb{E}_{V \sim \phi(v)}[\eta(V)(1 + h_\alpha(V))]. \qquad (4.29)$$

Proof. On the basis of integration by substitution,

$$\mathbb{E}_{U \sim f_0(u|\alpha)}[\xi(U)] = \int \xi(u)\phi(r_\alpha(u))du$$

$$
\begin{aligned}
&= \int \eta(v)\phi(v)(1+h_\alpha(v))dv \\
&= \mathbb{E}_{V\sim\phi(v)}[\eta(V)(1+h_\alpha(V))]. \qquad (4.30)
\end{aligned}
$$

□

For a condition $\mathcal{A}$, let $\mathbb{1}\{\mathcal{A}\}$ be the indicator function, which means $\mathbb{1}\{\mathcal{A}\} = 1$ if $\mathcal{A}$ is true and $\mathbb{1}\{\mathcal{A}\} = 0$ if $\mathcal{A}$ is false.

Proposition 12. *For $V \sim \mathcal{N}(0,1)$ and a sufficiently large positive constant M, it holds that*

$$
\mathbb{E}\left[|V|^m \mathbb{1}\{|V| > M\}\right] \leq \begin{cases} \exp\left(-\frac{M^2}{2}\right) & (m=0) \\ \left(M^{m-1} + O\left(M^{m-2}\right)\right)\exp\left(-\frac{M^2}{2}\right) & (m=1,2,3,4) \end{cases} \qquad (4.31)
$$

Proof. For $m = 0$,

$$
\begin{aligned}
\mathbb{E}\left[\mathbb{1}\{|V| > M\}\right] &= 2\int_M^\infty \phi(v)dv \\
&= 2\int_0^\infty \phi(v+M)dv \\
&\leq 2\exp\left(-\frac{M^2}{2}\right)\int_0^\infty \phi(v)dv \\
&= \exp\left(-\frac{M^2}{2}\right). \qquad (4.32)
\end{aligned}
$$

For $m = 1,2,3,4$,

$$
\begin{aligned}
\mathbb{E}\left[|V|\,\mathbb{1}\{|V| > M\}\right] &= 2\int_M^\infty v\phi(v)dv \\
&= \sqrt{\frac{2}{\pi}}\exp\left(-\frac{M^2}{2}\right) \\
&\leq \exp\left(-\frac{M^2}{2}\right), \qquad (4.33)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\left[|V|^2\,\mathbb{1}\{|V| > M\}\right] &= 2\int_M^\infty v^2\phi(v)dv \\
&= \sqrt{\frac{2}{\pi}}M\exp\left(-\frac{M^2}{2}\right) + \mathbb{E}\left[\mathbb{1}\{|V| > M\}\right] \\
&\leq \sqrt{\frac{2}{\pi}}M\exp\left(-\frac{M^2}{2}\right) + \exp\left(-\frac{M^2}{2}\right)
\end{aligned}
$$

$$\leq (M+1)\exp\left(-\frac{M^2}{2}\right), \tag{4.34}$$

$$\begin{aligned}
\mathbb{E}\left[|V|^3 \mathbb{1}\{|V|>M\}\right] &= 2\int_M^\infty v^3\phi(v)dv \\
&= \sqrt{\frac{2}{\pi}}M^2\exp\left(-\frac{M^2}{2}\right) + 2\mathbb{E}\left[|V|\mathbb{1}\{|V|>M\}\right] \\
&\leq \sqrt{\frac{2}{\pi}}M^2\exp\left(-\frac{M^2}{2}\right) + 2\exp\left(-\frac{M^2}{2}\right) \\
&\leq (M^2+2)\exp\left(-\frac{M^2}{2}\right),
\end{aligned} \tag{4.35}$$

$$\begin{aligned}
\mathbb{E}\left[|V|^4 \mathbb{1}\{|V|>M\}\right] &= 2\int_M^\infty v^4\phi(v)dv \\
&= \sqrt{\frac{2}{\pi}}M^3\exp\left(-\frac{M^2}{2}\right) + 3\mathbb{E}\left[|V|^2\mathbb{1}\{|V|>M\}\right] \\
&\leq \sqrt{\frac{2}{\pi}}M^3\exp\left(-\frac{M^2}{2}\right) + 3(M+1)\exp\left(-\frac{M^2}{2}\right) \\
&\leq (M^3+3M+3)\exp\left(-\frac{M^2}{2}\right).
\end{aligned} \tag{4.36}$$

□

Proposition 13. *For $U \sim f_*(u \mid \alpha)$ and a sufficiently large positive constant M, it holds that*

$$\mathbb{E}\left[|U|^m \mathbb{1}\{|U|>M\}\right] \leq \begin{cases} 2\exp\left(-\frac{M^2}{8}\right) & (m=0) \\ 4\left(M^{m-1}+O\left(M^{m-2}\right)\right)\exp\left(-\frac{M^2}{8}\right) & (m=1,2,3,4) \end{cases} \tag{4.37}$$

Proof. From Proposition 1, we have $|q_\alpha(v)| \leq |v| + |H_\alpha(v)| < 2|v|$ and $|1+h_\alpha(v)| < 1+|h_\alpha(v)| < 2$. In addition to this, using Proposition 11 and Proposition 12, it holds that

$$\begin{aligned}
&\mathbb{E}\left[|U|^m \mathbb{1}\{|U|>M\}\right] \\
&= \mathbb{E}\left[|q_\alpha(V)|^m (1+h_\alpha(V))\mathbb{1}\{|q_\alpha(V)|>M\}\right] \\
&\leq 2^{m+1}\mathbb{E}\left[|V|^m \mathbb{1}\left\{|V|>\frac{M}{2}\right\}\right]
\end{aligned}$$

$$
\begin{aligned}
&= \begin{cases} 2\exp\left(-\frac{M^2}{8}\right) & (m=0) \\ 2^{m+1}\left(\left(\frac{M}{2}\right)^{m-1} + O\left(\left(\frac{M}{2}\right)^{m-2}\right)\right)\exp\left(-\frac{M^2}{8}\right) & (m=1,2,3,4) \end{cases} \\
&= \begin{cases} 2\exp\left(-\frac{M^2}{8}\right) & (m=0) \\ 4\left(M^{m-1} + O\left(M^{m-2}\right)\right)\exp\left(-\frac{M^2}{8}\right) & (m=1,2,3,4) \end{cases}
\end{aligned} \tag{4.38}
$$

□

Proposition 14. *For $Y \sim f(y \mid \zeta)$ and a sufficiently large positive constant M, it holds that*

$$
\begin{aligned}
&\mathbb{E}\left[|Y|^m \mathbb{1}\{|Y| > M\}\right] \\
&\leq \begin{cases} 2\exp\left(-\frac{1}{8}\left(\frac{M-K}{K}\right)^2\right) & (m=0) \\ 4K\left(M^{m-1} + O\left(M^{m-2}\right)\right)\exp\left(-\frac{1}{8}\left(\frac{M-K}{K}\right)^2\right) & (m=1,2,3,4) \end{cases}
\end{aligned} \tag{4.39}
$$

where K is based on the parameter space (4.1).

Proof. Note that $|\mu| \leq K$ and $\sigma \leq K$ from (4.1). Using Proposition 13, it holds that

$$
\begin{aligned}
&\mathbb{E}\left[|Y|^m \mathbb{1}\{|Y| > M\}\right] \\
&= \mathbb{E}\left[|\mu + \sigma U|^m \mathbb{1}\{|\mu + \sigma U| > M\}\right] \\
&\leq \mathbb{E}\left[(K + K|U|)^m \mathbb{1}\{K + K|U| > M\}\right] \\
&= K^m \sum_{i=0}^{m} {}_m\mathrm{C}_i \mathbb{E}\left[|U|^i \mathbb{1}\left\{|U| > \frac{M-K}{K}\right\}\right] \\
&\leq 2K^m \exp\left(-\frac{1}{8}\left(\frac{M-K}{K}\right)^2\right) \\
&\quad + 4K^m \sum_{i=1}^{m} {}_m\mathrm{C}_i \left(\left(\frac{M-K}{K}\right)^{i-1} + O\left(\left(\frac{M-K}{K}\right)^{i-2}\right)\right)\exp\left(-\frac{1}{8}\left(\frac{M-K}{K}\right)^2\right) \\
&= \begin{cases} 2\exp\left(-\frac{1}{8}\left(\frac{M-K}{K}\right)^2\right) & (m=0) \\ 4K\left(M^{m-1} + O\left(M^{m-2}\right)\right)\left(-\frac{1}{8}\left(\frac{M-K}{K}\right)^2\right) & (m=1,2,3,4) \end{cases}
\end{aligned} \tag{4.40}
$$

□

Proposition 15. *Let the upper bound of (4.4) be denoted by $G_5(y)$. Let*

$$
G(y \mid M) := G_5(y)\mathbb{1}\{G_5(y) > M\}. \tag{4.41}
$$

Then, for $Y \sim f(y \mid \boldsymbol{\zeta})$ and a sufficiently large positive constant M, there exist some positive constants $C_{5,1}$, $C_{5,2}$, and $C_{5,3}$ such that

$$\mathbb{E}\left[G(Y \mid M)\right] \leq C_{5,1} M^{\frac{1}{2}} \exp(-C_{5,3} M), \tag{4.42}$$

$$\mathbb{E}\left[G^2(Y \mid M)\right] \leq C_{5,2} M^{\frac{3}{2}} \exp(-C_{5,3} M). \tag{4.43}$$

Proof. From Proposition 5, we know

$$G_5(y) = C_{1,2}|y|^2 + C_{1,1}|y| + C_{1,0} > M, \tag{4.44}$$

and then we have

$$|y| > \frac{-C_{1,1} + \sqrt{C_{1,1}^2 + 4C_{1,2}(M - C_{1,0})}}{2C_{1,2}} \geq \sqrt{C_{5,4} M} \tag{4.45}$$

where $C_{5,4}$ is some positive constant. From Proposition 14,

$$\begin{aligned}
\mathbb{E}\left[G(Y \mid M)\right] &\leq \mathbb{E}\left[\left(C_{1,2}|Y|^2 + C_{1,1}|Y| + C_{1,0}\right) \mathbb{1}\left\{|Y| > \sqrt{C_{5,4} M}\right\}\right] \\
&\leq 4C_{1,2} K \left(\sqrt{C_{5,4} M} + O(1)\right) \exp\left(-\frac{1}{8}\left(\frac{\sqrt{C_{5,4} M} - K}{K}\right)^2\right) \\
&\leq C_{5,1} M^{\frac{1}{2}} \exp(-C_{5,3} M),
\end{aligned} \tag{4.46}$$

where $C_{5,1}$ and $C_{5,2}$ are some positive constants.

$$\begin{aligned}
\mathbb{E}\left[G^2(Y \mid M)\right] &\leq \mathbb{E}\left[\left(C_{1,2}|y|^2 + C_{1,1}|y| + C_{1,0}\right)^2 \mathbb{1}\left\{|Y| > \sqrt{C_{5,4} M}\right\}\right] \\
&\leq 4C_{1,2}^2 K \left(\left(\sqrt{C_{5,4} M}\right)^3 + O(M)\right) \exp\left(-\frac{1}{8}\left(\frac{\sqrt{C_{5,4} M} - K}{K}\right)^2\right) \\
&\leq C_{5,2} M^{\frac{3}{2}} \exp(-C_{5,3} M),
\end{aligned} \tag{4.47}$$

where $C_{5,2}$ is some positive constant. □

Proposition 16. *Let $G(y)$ be the function defined by (4.41). Let*

$$F(y \mid M) := G(y \mid M) + \mathbb{E}\left[G(Y \mid M) \mid \boldsymbol{X}\right]. \tag{4.48}$$

Let $M_N := \frac{3}{C_{5,3}} \log N$ *for* $N \geq \exp\left(\frac{27C_{5,2}^2}{C_{5,3}^3}\right) \vee \left(\frac{3C_{5,1}^2}{C_{5,3}}\right)^{\frac{1}{4}} \vee e$. *Then, it holds that*

$$\mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N} F(Y_n \mid M_N) > \frac{2\log N}{N}\right) \leq \frac{1}{N}. \tag{4.49}$$

Proof. From $N \geq \exp\left(\frac{27C_{5,2}^2}{C_{5,3}^3}\right) \vee \left(\frac{3C_{5,1}^2}{C_{5,3}}\right)^{\frac{1}{4}} \vee e$, we have

$$\frac{3^{\frac{3}{2}}C_{5,2}}{C_{5,3}^{\frac{3}{2}}} \leq \sqrt{\log N} \leq 2\sqrt{\log N} - \frac{\sqrt{3}C_{5,1}}{\sqrt{C_{5,3}}N^2}. \tag{4.50}$$

Therefore, using Proposition 15, it holds

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N} F(Y_n \mid M_N) > \frac{2\log N}{N}\right) \\
&\leq \mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N} \left(G(Y_n \mid M_N) + \mathbb{E}\left[G(Y \mid M_N \mid \boldsymbol{X}_n)\right]\right) > \frac{2\log N}{N}\right) \\
&= \mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N} G(Y_n \mid M_N) > \frac{2\log N}{N} - \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}\left[G(Y \mid M_N) \mid \boldsymbol{X}_n\right]\right) \\
&\leq \mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N} G(Y_n \mid M_N) > \frac{2\log N}{N} - C_{5,1}M_N^{\frac{1}{2}}\exp(-C_{5,3}M_N)\right) \\
&= \mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N} G(Y_n \mid M_N) > \frac{\sqrt{\log N}}{N}\left(2\sqrt{\log N} - \frac{\sqrt{3}C_{5,1}}{\sqrt{C_{5,3}}N^2}\right)\right) \\
&\leq \mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N} G(Y_n \mid M_N) > \frac{\log N}{N}\right) \\
&\leq \mathbb{E}\left[G^2(Y \mid M_N)\right]\left(\frac{N}{\log N}\right)^2 \\
&\leq C_{5,2}M_N^{\frac{3}{2}}\exp(-C_{5,3}M_N)\left(\frac{N}{\log N}\right)^2 \\
&= \frac{3^{\frac{3}{2}}C_{5,2}}{C_{5,3}^{\frac{3}{2}}}\frac{1}{N\sqrt{\log N}} \\
&\leq \frac{1}{N}
\end{aligned} \tag{4.51}$$

□

Proof of Proposition 5

Proof. Note that $u = (y - \boldsymbol{X}^\top \boldsymbol{\beta})/\sigma$. This proposition is derived immediately from Assumption 1, (4.22), (4.23), (4.24), Proposition 2, and Proposition 4. □

Proof of Proposition 6

Proof. For simplicity of notation, let $\omega(v) := \sigma \frac{\partial}{\partial \alpha} H_\alpha(v)$. From the scores (4.22), (4.23), and (4.24), the Fisher information matrix can be described as follows by using Proposition 11:

$$
\begin{aligned}
&\mathbb{E}\left[-\frac{\partial^2}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^\top} l(\boldsymbol{\psi} \mid Y)\right] \\
&= \frac{1}{\sigma^2} \mathbb{E}_{V \sim \phi(v)} \left[\frac{1}{1 + h_\alpha(V)} \begin{bmatrix} V^2 & V^2 H_\alpha & -V^2 \omega \\ V^2 H_\alpha & V^2 H_\alpha^2 + V^4 & -V^2 H_\alpha \omega \\ -V^2 \omega & -V^2 H_\alpha \omega & V^2 \omega^2 \end{bmatrix} \right] - \frac{1}{\sigma^2} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\
&\succeq \frac{1}{2\sigma^2} \mathbb{E}_{V \sim \phi(v)} \begin{bmatrix} V^2 & V^2 H_\alpha & -V^2 \omega \\ V^2 H_\alpha & V^2 H_\alpha^2 + 3 & -V^2 H_\alpha \omega \\ -V^2 \omega & -V^2 H_\alpha \omega & V^2 \omega^2 \end{bmatrix} - \frac{1}{\sigma^2} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\
&= \frac{1}{2\sigma^2} \mathbb{E}_{V \sim \phi(v)} \begin{bmatrix} V^2 & V^2 H_\alpha & -V^2 \omega \\ V^2 H_\alpha & V^2 H_\alpha^2 + 1 & -V^2 H_\alpha \omega \\ -V^2 \omega & -V^2 H_\alpha \omega & V^2 \omega^2 \end{bmatrix} \\
&= \frac{1}{2\sigma^2} \mathbb{E}_{V \sim v^2 \phi(v)} \begin{bmatrix} 1 & H_\alpha & -\omega \\ H_\alpha & H_\alpha^2 + 1 & -H_\alpha \omega \\ -\omega & -H_\alpha \omega & \omega^2 \end{bmatrix}. \qquad (4.52)
\end{aligned}
$$

Hereafter, since $v^2 \phi(v)$ is a probability density function on $v \in \mathbb{R}$, the random variable V is regarded as $V \sim v^2 \phi(v)$ for notational simplicity. Note that matrix (4.52) is clearly positive semidefinite, we only need to check that its determinant is positive. For $\alpha \to 0$, we have

$$
\begin{aligned}
\lim_{\alpha \to 0} H_\alpha(v) &= \lim_{\alpha \to 0} \rho_\alpha \frac{\alpha v^2}{\sqrt{1 + \alpha^2 v^2} + 1} \\
&= 0, \qquad (4.53) \\
\lim_{\alpha \to 0} \omega(v) &= \sigma \lim_{\alpha \to 0} \left(\exp(-\alpha^2) \left(\sqrt{1 + \alpha^2 v^2} - 1 \right) + \rho_\alpha \frac{v^2}{\sqrt{1 + \alpha^2 v^2} \left(\sqrt{1 + \alpha^2 v^2} + 1 \right)} \right) \\
&= \frac{\sigma v^2}{4}, \qquad (4.54)
\end{aligned}
$$

and then we see that the determinant of (4.52) is positive because

$$\begin{aligned}
&\det\left(\mathbb{E}_{V\sim v^2\phi(v)}\begin{bmatrix}1 & H_\alpha & -\omega\\ H_\alpha & H_\alpha^2+1 & -H_\alpha\omega\\ -\omega & -H_\alpha\omega & \omega^2\end{bmatrix}\right)\\
&=\begin{vmatrix}1 & 0 & -\frac{\sigma}{4}\mathbb{E}_{V\sim v^2\phi(v)}\left[V^2\right]\\ 0 & 1 & 0\\ -\frac{\sigma}{4}\mathbb{E}_{V\sim v^2\phi(v)}\left[V^2\right] & 0 & \frac{\sigma^2}{16}\mathbb{E}_{V\sim v^2\phi(v)}\left[V^4\right]\end{vmatrix}\\
&=\frac{\sigma^2}{16}\left(\mathbb{E}_{V\sim v^2\phi(v)}\left[V^4\right]-\mathbb{E}^2_{V\sim v^2\phi(v)}\left[V^2\right]\right)\\
&=\frac{3}{4}\sigma^2\\
&>0.
\end{aligned}\tag{4.55}$$

Consider the case $\alpha \neq 0$. We have

$$\begin{aligned}
&\det\left(\mathbb{E}_{V\sim v^2\phi(v)}\begin{bmatrix}1 & H_\alpha & -\omega\\ H_\alpha & H_\alpha^2+1 & -H_\alpha\omega\\ -\omega & -H_\alpha\omega & \omega^2\end{bmatrix}\right)\\
&=\begin{vmatrix}1 & \mathbb{E}\left[H_\alpha\right] & -\mathbb{E}\left[\omega\right]\\ \mathbb{E}\left[H_\alpha\right] & \mathbb{E}\left[H_\alpha^2\right]+1 & -\mathbb{E}\left[H_\alpha\omega\right]\\ -\mathbb{E}\left[\omega\right] & -\mathbb{E}\left[H_\alpha\omega\right] & \mathbb{E}\left[\omega^2\right]\end{vmatrix}\\
&=\left(\mathbb{E}\left[H_\alpha^2\right]\mathbb{E}\left[\omega^2\right]-\mathbb{E}\left[H_\alpha^2\right]\mathbb{E}^2\left[\omega\right]-\mathbb{E}^2\left[H_\alpha\right]\mathbb{E}\left[\omega^2\right]\right)\\
&\quad+2\mathbb{E}\left[H_\alpha\omega\right]\mathbb{E}\left[H_\alpha\right]\mathbb{E}\left[\omega\right]-\mathbb{E}^2\left[H_\alpha\omega\right]+\left(\mathbb{E}\left[\omega^2\right]-\mathbb{E}^2\left[\omega\right]\right)\\
&>-\mathbb{E}^2\left[H_\alpha\right]\mathbb{E}^2\left[\omega\right]+2\mathbb{E}\left[H_\alpha\omega\right]\mathbb{E}\left[H_\alpha\right]\mathbb{E}\left[\omega\right]-\mathbb{E}^2\left[H_\alpha\omega\right]+\mathbb{V}\left[\omega\right]\\
&=-\left(\mathbb{E}\left[H_\alpha\omega\right]-\mathbb{E}\left[H_\alpha\right]\mathbb{E}\left[\omega\right]\right)^2+\mathbb{V}\left[\omega\right]\\
&=\mathbb{V}\left[\omega\right]-\mathrm{Cov}^2\left[H_\alpha,\omega\right],
\end{aligned}\tag{4.56}$$

where the inequality is based on $\mathbb{V}\left[H_\alpha\right]\mathbb{V}\left[\omega\right] = \left(\mathbb{E}\left[H_\alpha^2\right]-\mathbb{E}^2\left[H_\alpha\right]\right)\left(\mathbb{E}\left[\omega^2\right]-\mathbb{E}^2\left[\omega\right]\right) > 0$. To examine the sign of (4.56), we evaluate $\mathbb{V}\left[H_\alpha\right]$ concretely, starting with (3.4), i.e.,

$$\begin{aligned}
\mathbb{V}_{V\sim v^2\phi(v)}\left[H_\alpha\right] &= \rho_\alpha^2\mathbb{V}_{V\sim v^2\phi(v)}\left[\sqrt{V^2+\frac{1}{\alpha^2}}\right]\\
&=\rho_\alpha^2\left(\mathbb{E}_{V\sim v^2\phi(v)}\left[V^2+\frac{1}{\alpha^2}\right]-\mathbb{E}^2_{V\sim v^2\phi(v)}\left[\sqrt{V^2+\frac{1}{\alpha^2}}\right]\right)\\
&=\rho_\alpha^2\left(3+\frac{1}{\alpha^2}-\mathbb{E}^2_{V\sim v^2\phi(v)}\left[\sqrt{V^2+\frac{1}{\alpha^2}}\right]\right).
\end{aligned}\tag{4.57}$$

Here, we focus on the behavior of $\mathbb{E}_{V\sim v^2\phi(v)}\left[\sqrt{V^2+\frac{1}{\alpha^2}}\right]$ with respect to α. We have

$$\begin{aligned}
&\frac{\partial}{\partial\alpha}\frac{1}{\sqrt{2+\frac{1}{\alpha^2}}}\mathbb{E}_{V\sim v^2\phi(v)}\left[\sqrt{V^2+\frac{1}{\alpha^2}}\right]\\
&=\frac{\partial}{\partial\alpha}\int_{\mathbb{R}}\sqrt{\frac{1+\alpha^2v^2}{1+2\alpha^2}}v^2\phi(v)dv\\
&=\int_{\mathbb{R}}\frac{\partial}{\partial\alpha}\sqrt{\frac{1+\alpha^2v^2}{1+2\alpha^2}}v^2\phi(v)dv\\
&=\frac{\alpha}{(1+2\alpha^2)^{\frac{3}{2}}}\int_{\mathbb{R}}\left(\frac{v^4}{\sqrt{1+\alpha^2v^2}}-\frac{2v^2}{\sqrt{1+\alpha^2v^2}}\right)\phi(v)dv\\
&=\frac{\alpha}{(1+2\alpha^2)^{\frac{3}{2}}}\int_{\mathbb{R}}\left(\frac{v^2}{\sqrt{1+\alpha^2v^2}}-\frac{\alpha^2v^4}{(1+\alpha^2v^2)^{\frac{3}{2}}}\right)\phi(v)dv\\
&=\frac{\alpha}{(1+2\alpha^2)^{\frac{3}{2}}}\int_{\mathbb{R}}\frac{v^2}{(1+\alpha^2v^2)^{\frac{3}{2}}}\phi(v)dv,
\end{aligned}\tag{4.58}$$

where the fourth equality is derived by

$$\begin{aligned}
&\int_{\mathbb{R}}\frac{v^4}{\sqrt{1+\alpha^2v^2}}\phi(v)dv\\
&=-\int_{\mathbb{R}}\frac{v^3}{\sqrt{1+\alpha^2v^2}}\left(\phi(v)\right)'dv\\
&=-2\underbrace{\lim_{R\to\infty}\frac{R^3\phi(R)}{\sqrt{1+\alpha^2R^2}}}_{=0}+\int_{\mathbb{R}}\left(\frac{3v^2}{\sqrt{1+\alpha^2v^2}}-\frac{\alpha^2v^4}{(1+\alpha^2v^2)^{\frac{3}{2}}}\right)\phi(v)dv.
\end{aligned}\tag{4.59}$$

Hence, $\frac{\partial}{\partial\alpha}\frac{1}{2+\frac{1}{\alpha^2}}\mathbb{E}^2_{V\sim v^2\phi(v)}\left[\sqrt{V^2+\frac{1}{\alpha^2}}\right]$ is monotonically increasing function of α^2. Moreover,

$$\begin{aligned}
\frac{1}{2+\frac{1}{\alpha^2}}\mathbb{E}^2_{V\sim v^2\phi(v)}\left[\sqrt{V^2+\frac{1}{\alpha^2}}\right]&>\lim_{\alpha^2\to0}\frac{1}{2+\frac{1}{\alpha^2}}\mathbb{E}^2_{V\sim v^2\phi(v)}\left[\sqrt{V^2+\frac{1}{\alpha^2}}\right]\\
&=1.
\end{aligned}\tag{4.60}$$

Thus, we have

$$\mathbb{E}^2_{V\sim v^2\phi(v)}\left[\sqrt{V^2+\frac{1}{\alpha^2}}\right]>2+\frac{1}{\alpha^2},\tag{4.61}$$

and $\mathbb{V}_{V\sim v^2\phi(v)}[H_\alpha] < \rho_\alpha^2 < 1$ from (4.57). For $\alpha \neq 0$, $H_\alpha(v)$ and $\omega(v)$ aren't obviously constant functions. Then, $\mathbb{V}_{V\sim v^2\phi(v)}[\omega(V)] < \infty$ because the upper bound of $\left|\frac{\partial}{\partial\alpha}H_\alpha(v)\right|$ is a linear function of $|u|$ from Proposition 3. Therefore, since $0 < \mathbb{V}[H_\alpha] < 1$ and $0 < \mathbb{V}[\omega] < \infty$, it holds that

$$\mathrm{Cov}^2[H_\alpha, \omega] < \mathbb{V}[H_\alpha]\,\mathbb{V}[\omega] < \mathbb{V}[\omega]. \tag{4.62}$$

From (4.56) and (4.62), we see that $\mathcal{I}(\boldsymbol{\psi}) = \mathbb{E}\left[\frac{\partial^2}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\top} l(\boldsymbol{\psi} \mid Y)\right]$ is positive definite for any $\boldsymbol{\psi}$. Consider the case $\boldsymbol{\psi} = \boldsymbol{\psi}^0(\boldsymbol{X}) = (\boldsymbol{X}^\top\beta^0, \sigma^0, \alpha^0)$. Since $\boldsymbol{\psi}^0(\boldsymbol{X})$ belongs to the bounded set Ψ, we have $\inf_X \Lambda_{\min}(\mathcal{I}(\boldsymbol{\psi}^0(\boldsymbol{X}))) > 0$. □

Proof of Proposition 7

Proof. From Proposition 10, with some natural number M and some positive constant $C_{3,m}$ for $m = 0, 1, \ldots, M$, we have

$$\max_{(i_1,i_2,i_3)\in\{1,2,3\}^3} \left|\frac{\partial^3}{\partial\psi_{i_1}\partial\psi_{i_2}\partial\psi_{i_3}} l(\boldsymbol{\theta} \mid \boldsymbol{X}, y)\right| \leq \sum_{m=0}^{M} C_{3,m}|y|^m. \tag{4.63}$$

Since the m-th order moment of $Y \sim f(y \mid \boldsymbol{X}^\top\boldsymbol{\beta}, \sigma, \alpha)$ is finite from Proposition 9, the proof is complete. □

Proof of Proposition 8

Proof. Applying the Taylor expansion around the point $\boldsymbol{\psi}^0$ for (4.7), there exists some parameter vector $\tilde{\boldsymbol{\psi}} \in \{t\boldsymbol{\psi}^0 + (1-t)\boldsymbol{\psi} \in \mathbb{R}^3 \mid t \in [0,1]\}$ such that

$$\begin{aligned}\mathcal{E}(\boldsymbol{\psi} \mid \boldsymbol{\psi}^0) = \mathcal{E}(\boldsymbol{\psi}^0 \mid \boldsymbol{\psi}^0) &+ \left.\frac{\partial\mathcal{E}}{\partial\boldsymbol{\psi}^\top}\right|_{\boldsymbol{\psi}=\boldsymbol{\psi}^0}(\boldsymbol{\psi} - \boldsymbol{\psi}^0) \\ &+ \frac{1}{2}(\boldsymbol{\psi} - \boldsymbol{\psi}^0)^\top \left.\frac{\partial^2\mathcal{E}}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\top}\right|_{\boldsymbol{\psi}=\tilde{\boldsymbol{\psi}}}(\boldsymbol{\psi} - \boldsymbol{\psi}^0).\end{aligned} \tag{4.64}$$

Excess Risk (4.7) satisfies $\mathcal{E}(\boldsymbol{\psi}^0 \mid \boldsymbol{\psi}^0) = 0$ and $\frac{\partial\mathcal{E}}{\partial\boldsymbol{\psi}}(\boldsymbol{\psi}^0 \mid \boldsymbol{\psi}^0) = 0$ by its definition. Because $\mathcal{I}(\tilde{\boldsymbol{\psi}})$ is positive definite from Proposition 6, its smallest eigenvalue $\Lambda_{\min}(\mathcal{I}(\tilde{\boldsymbol{\psi}}))$ is positive. Thus, the quadratic term of (4.64) is evaluated as

$$(\boldsymbol{\psi} - \boldsymbol{\psi}^0)^\top \left.\frac{\partial^2\mathcal{E}}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\top}\right|_{\boldsymbol{\psi}=\tilde{\boldsymbol{\psi}}}(\boldsymbol{\psi} - \boldsymbol{\psi}^0) \geq \Lambda_{\min}(\mathcal{I}(\tilde{\boldsymbol{\psi}}))\|\boldsymbol{\psi} - \boldsymbol{\psi}^0\|_2^2. \tag{4.65}$$

Therefore, letting $C_2 := \inf_{X\in\mathcal{X}} \Lambda_{\min}(\mathcal{I}(\tilde{\boldsymbol{\psi}}(\boldsymbol{X}))) > 0$, based on

$$\begin{aligned}\inf_{X\in\mathcal{X}} \inf_{\boldsymbol{\psi}\in\Theta:\|\boldsymbol{\psi}-\boldsymbol{\psi}^0\|_2>\epsilon} \mathcal{E}(\boldsymbol{\psi} \mid \boldsymbol{\psi}^0) &\geq \inf_{X\in\mathcal{X}} \inf_{\boldsymbol{\psi}\in\Theta:\|\boldsymbol{\psi}-\boldsymbol{\psi}^0\|_2>\epsilon} \frac{1}{2}\Lambda_{\min}(\mathcal{I}(\tilde{\boldsymbol{\psi}}(\boldsymbol{X})))\|\boldsymbol{\psi}-\boldsymbol{\psi}^0\|_2^2 \\ &> \frac{\epsilon^2}{2} \inf_{X\in\mathcal{X}} \Lambda_{\min}(\mathcal{I}(\tilde{\boldsymbol{\psi}}(\boldsymbol{X}))) \\ &> \frac{\epsilon^2 C_2}{2}, \end{aligned} \tag{4.66}$$

we can take $\delta_\epsilon = \frac{\epsilon^2 C_2}{2} > 0$ as δ_ϵ satisfying Proposition 8. □

4.6.2 Proof of Theorem 6

If Condition 1, 2, and 3 in Section 9.4.1.1 of Bühlmann and Geer (2011) holds, the proposition holds immediately from Lemma 9.1 of Bühlmann and Geer (2011). These Condition 1, 2, and 3 correspond to Proposition 7, 6, and 8, which were already shown, respectively. The proof is complete.

4.6.3 Proof of Theorem 7

If Condition 1, 2, and 3 in Section 9.4.1.1 and Condition 4 in Section 9.4.2 of Bühlmann and Geer (2011) holds, the proposition holds immediately from Theorem 9.1 of Bühlmann and Geer (2011). These Condition 1, 2, and 3 correspond to Proposition 7, 6, and 8, which were already shown, respectively. Condition 4 correspond to Assumption 2. The proof is complete.

4.6.4 Proof of Theorem 8

Theorem 8 is similar to Corollary 9.1 of Bühlmann and Geer (2011). The difference is a model. The former model is a regression model with the mode-invariant skew-normal noise, but the latter model is a finite mixture model. The proof of Corollary 9.1 of Bühlmann and Geer (2011) is based on two parts; one is a model-specific part, and the other is not related to a model. The model-specific part consists of Proposition 9.4 and Lemma 9.3 of Bühlmann and Geer (2011). We prove these propositions for the former model. Using them, we can apply the same proof technique as in Bühlmann and Geer (2011) to prove Theorem 8.

Proposition 9.4 and Lemma 9.3 of Bühlmann and Geer (2011) show how to obtain an upper bound of the score function and how to set M_N. For our sparse regression model with the mode-invariant skew-normal noise, the upper bound of the score function is obtained in Proposition 5, which corresponds to Proposition 9.4 of Bühlmann and Geer (2011). In this case, following the approach described

in Bühlmann and Geer (2011), we set $M_N = O(\log N)$, and then we can derive Proposition 16, which corresponds to Lemma 9.3 of Bühlmann and Geer (2011). The proof is complete.

Chapter 5

Experimental Results

This section demonstrates the performance of the proposed method with synthetic and real-world datasets.

5.1 Skew-Noises with Mode Zero

The simulation model is $y = \boldsymbol{X}^\top \boldsymbol{\beta}^0 + e$. The feature vector $\boldsymbol{X}_n \in \mathbb{R}^{P=100}$ was generated randomly with each element following the continuous uniform distribution on $[-1, 1] \in \mathbb{R}$. The true parameter was set to $\boldsymbol{\beta}^0 = [1.0, -0.9, 0.8, \ldots, -0.1, 0, \ldots, 0]^\top \in \mathbb{R}^{100}$ with 10% non-zero coefficients of varying magnitudes and signs. Let the random variable following (3.1) be denoted by $\mathcal{SN}(\mu, \sigma, \alpha)$. Let the log-normal variable be denoted by $\ln \mathcal{N}(\mu, \sigma)$. The noise e was generated from four types of distribution: $\mathcal{SN}(0, 1, 1)$ (**Data 1**), $\mathcal{SN}(0, 1, 2)$ (**Data 2**), $\ln \mathcal{N}(0, 0.5) - \exp(-0.5^2)$ (**Data 3**), and $\ln \mathcal{N}(0, 1) - \exp(-1^2)$ (**Data 4**). Data 3 and Data 4 are the log-normal variable adjusted to mode zero. These are misspecified models. The noise distributions are depicted in the first column of Figure 5.1.

In Algorithm 1, we employed the *scipy.optimize.minimize(method='L-BFGS-B')* function of Python to optimize σ (line 10) and α (line 11). Specifically, the function was executed with $\sigma > 0$ for σ and without any bounds for α. All other settings of the function were used at their default values.

We compared the proposed method with the ordinary Lasso (Tibshirani, 1996) ("Lasso"), the ordinary Lasso applied to the Yeo–Johnson transformed output ("Lasso+YJ"), and another skew-noise Lasso (Chen, Pourahmadi, and Maadooliat, 2014) ("Chen+") that assumes a regression model for the location parameter of Azzalini's skew-normal distribution. The tuning power parameter in the Yeo–Johnson transformation was determined by maximum likelihood estimation. The sample size was set to $N = 500$. We conducted 50 experiments with different random seeds. For each trial, the regularization coefficient λ was adjusted by 5-fold cross-validation based on the log-likelihood loss, in which the

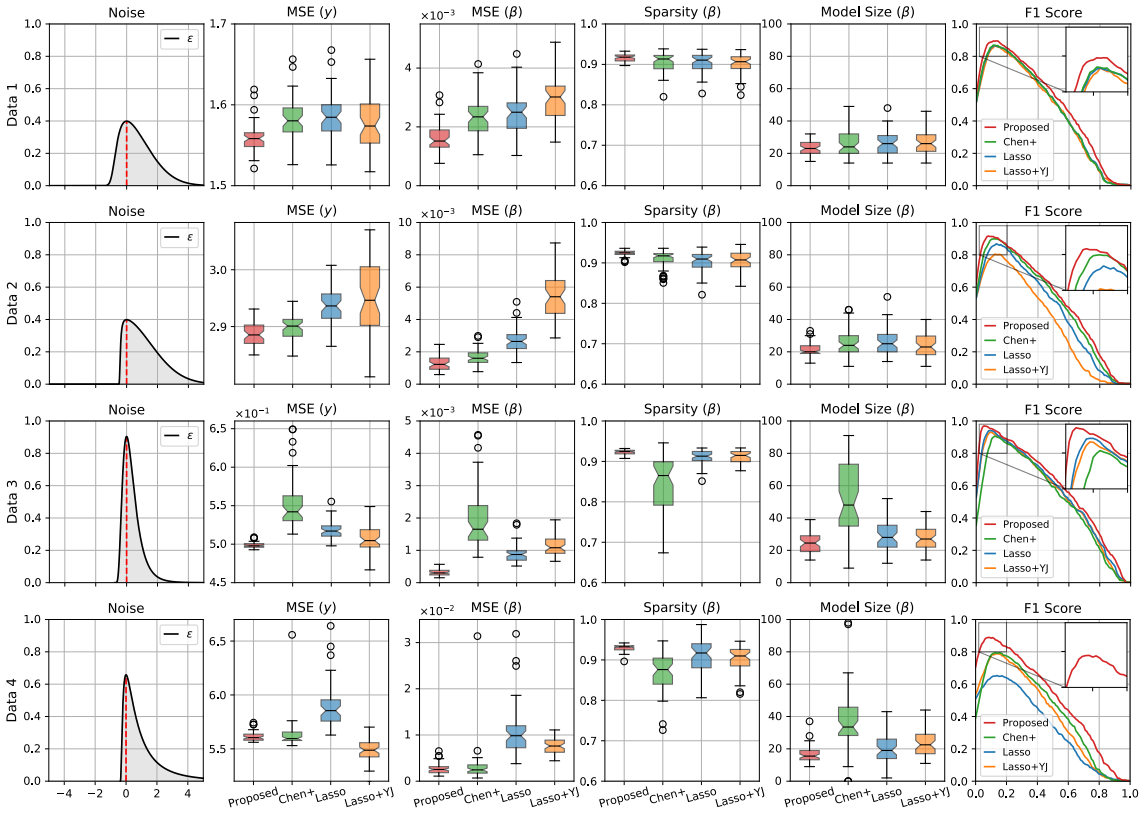

FIGURE 5.1: Comparison of four methods with four different types of simulation noises in terms of five evaluation measures. From top to bottom, each row results from Data 1, Data 2, Data 3, and Data 4, respectively. The first column plots the probability density function of the noise. The second through fifth columns are box plots of some scores. The last column is the mean F1 score for each threshold (horizontal axis). In all figures, the proposed method, Chen+ (Chen, Pourahmadi, and Maadooliat, 2014), Lasso, and Lasso with the Yeo–Johnson transformation are in red, green, blue, and yellow, respectively. All experiments were conducted for 50 runs with different random seeds.

numbers of training and validation data for each trial were set to 400 and 100, respectively.

Five evaluation measures are shown in the second through six columns of Figure 5.1. The second column is the Mean Squared Error (MSE) for the prediction: $\mathrm{MSE}(\hat{y}) := \sum_{n:\mathrm{test}}(\hat{y}_n - y_n^0)^2/N$, where the MSE was estimated from 2,000 test data generated under the same conditions. The third column is the MSE of the estimated coefficients: $\mathrm{MSE}(\hat{\boldsymbol{\beta}}) := \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2^2/P$. Since "Lasso+YJ" is based on the Yeo–Johnson transformed output, the resulting estimator $\hat{\boldsymbol{\beta}}$ does not generally converge to $\boldsymbol{\beta}^0$. Therefore, $\mathrm{MSE}(\hat{\boldsymbol{\beta}})$ is expected to be inferior to other methods. The fourth column is a measure for sparsity by Hurley and Rickard (2009) as the most robust measure founded on a weighted sum of estimated all coefficients for evaluating sparsity. The measure range is $[0, 1] \in \mathbb{R}$; as a coefficient vector is sparser, its value is higher (see Section 5.4.1 for details). The fifth column is the

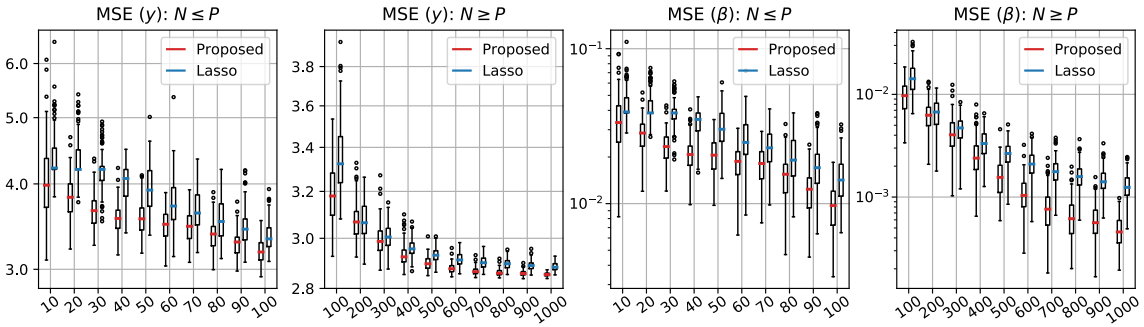

FIGURE 5.2: Box plots of MSEs for Data 2 with $P = 100$ fixed and various sample sizes N. The vertical axis is a logarithmic scale. The horizontal axis is the sample size N. All experiments were conducted for 100 runs with different random seeds.

size of the estimated model, which is defined as the number of non-zero coefficients: Model Size$(\hat{\boldsymbol{\beta}}) := \#\{\hat{\beta}_p \neq 0 : 1 \leq p \leq P\}$. The last column is the F1 score (Van Rijsbergen, 2004) defined by

$$\text{F1 Score} := 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{5.1}$$

where Precision and Recall are defined by $\text{TP}/(\text{TP} + \text{FP})$ and $\text{TP}/(\text{TP} + \text{FN})$, respectively, and TP, FP, and FN mean the true positive, false positive, and false negative, respectively. The positivity (negativity) of the estimated coefficient is defined by $|\hat{\beta}_p|/\|\hat{\boldsymbol{\beta}}\|_\infty > \delta_T$ for a threshold $\delta_T \in [0, 1]$ (horizontal axis). As the feature selection is more successful at each threshold, the F1 score is larger. This means that a better method presents an upper curve.

As a result of Figure 5.1, the proposed method outperformed the comparative methods in all the cases, indicating better prediction and feature selection with a smaller number of features. We see that Chen+ is inferior to Lasso in Data 3. This might be because Azzalini's skew-normal distribution could not adequately model such skewed noise.

Figure 5.2 shows box plots of MSEs for 100 trials regarding Data 2 with $P = 100$ fixed and various sample sizes N. The regularization coefficient λ was tuned by 5-fold cross-validation with 80% training and 20% validation data. The Lasso is a benchmark against the proposed method. The proposed method presented stable behaviors for $N < P$ as well as $N > P$.

In Section 5.4.2, we also show the results with a linear regression model without regularization, assuming Azzalini's skew-normal distribution, skew-t distribution, and skew-Cauchy distribution for noise (Arellano-Valle and Azzalini, 2013; Azzalini and Arellano-Valle, 2013; Azzalini and Salehi, 2020), on Data 1 through Data 4.

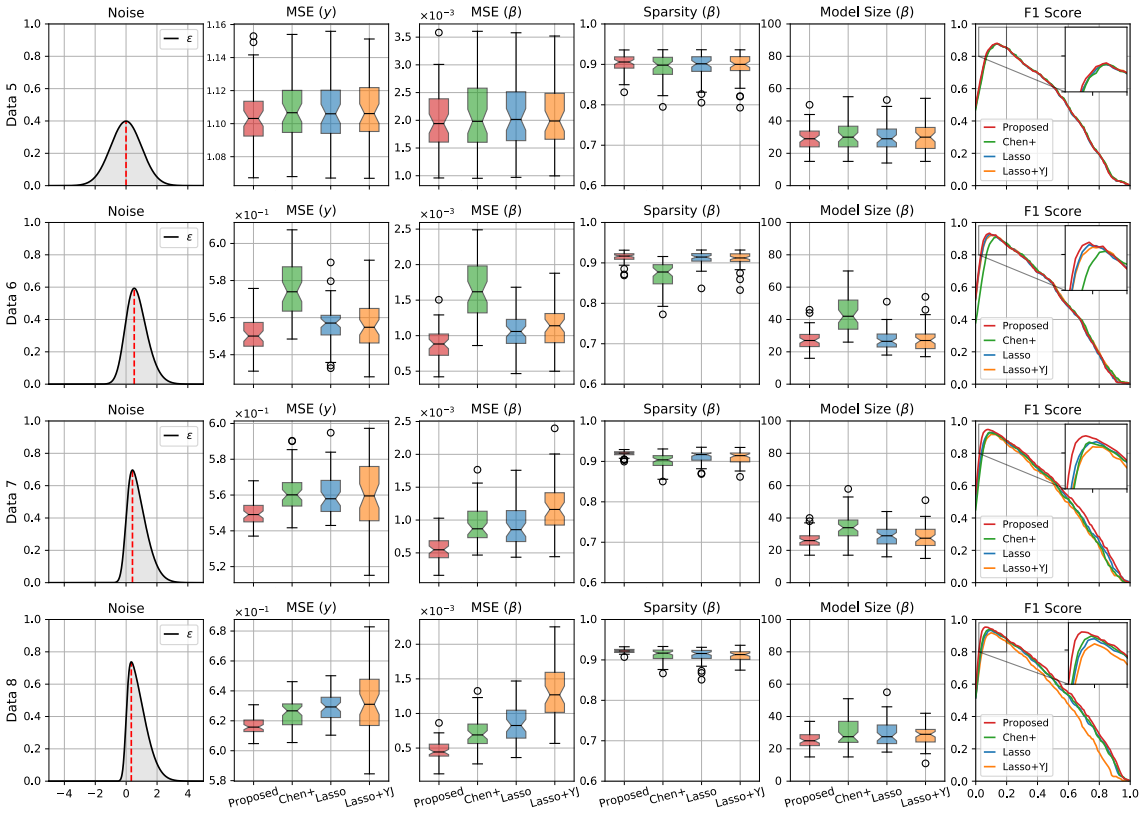

FIGURE 5.3: Comparison of four methods with four different types of simulation noises in terms of five evaluation measures. From top to bottom, each row results from Data 5, Data 6, Data 7, and Data 8, respectively. The first column plots the probability density function of the noise. The second through fifth columns are box plots of some scores. The last column is the mean F1 score for each threshold (horizontal axis). In all figures, the proposed method, Chen+ (Chen, Pourahmadi, and Maadooliat, 2014), Lasso, and Lasso with the Yeo–Johnson transformation are in red, green, blue, and yellow, respectively. All experiments were conducted for 50 runs with different random seeds.

5.2 Normal and Azzalini's Skew-Noises

We explore both normal and Azzalini's skew-normal noises under the same conditions as in Section 5.1. Let the random variable following Azzalini's skew-normal be denoted by $\mathcal{SN}_{\mathrm{A}}(\mu, \sigma, \alpha)$. The noise e was generated from additional four types of distribution: $\mathcal{N}(0,1)$ (**Data 5**), $\mathcal{SN}_{\mathrm{A}}(0,1,2)$ (**Data 6**), $\mathcal{SN}_{\mathrm{A}}(0,1,4)$ (**Data 7**), and $\mathcal{SN}_{\mathrm{A}}(0,1,6)$ (**Data 8**). Notably, Azzalini's skew-normal noise is not mode-invariant, i.e., its location parameter does not coincide with its mode, which significantly violates the assumptions of the proposed method.

The results are shown in Figure 5.3. In Data 5, the four methods presented similar behaviors, as expected. In Data 6, Data 7, and Data 8, the proposed method outperformed the other methods with the difference becoming more significant as α increased. Surprisingly, Chen+ presented similar results to Lasso, although Chen+ was slightly worse. This would be because the objective function of Chen+

TABLE 5.1: Results of PDGFR ($N = 79$, $P = 320$).

	$\mathbf{MSE}(\hat{y})$	$\mathbf{Sparsity}(\hat{\beta})$	$\mathbf{Size}(\hat{\beta})$
Proposed	$\mathbf{7.00 \times 10^{-1}}$ (3.16×10^{-1})	$\mathbf{9.90 \times 10^{-1}}$ (3.16×10^{-3})	$\mathbf{5.30 \times 10^{0}}$ (1.84×10^{0})
Chen+	10.9×10^{-1} (2.54×10^{-1})	8.18×10^{-1} (3.72×10^{-1})	8.17×10^{0} (4.82×10^{0})
Lasso	7.28×10^{-1} (4.79×10^{-1})	9.85×10^{-1} (9.82×10^{-3})	8.90×10^{0} (5.57×10^{0})
Lasso + YJ	12.7×10^{-1} (4.22×10^{-1})	9.85×10^{-1} (9.82×10^{-3})	8.93×10^{0} (5.57×10^{0})

TABLE 5.2: Results of MTP ($N = 274$, $P = 1142$).

	$\mathbf{MSE}(\hat{y})$	$\mathbf{Sparsity}(\hat{\beta})$	$\mathbf{Size}(\hat{\beta})$
Proposed	$\mathbf{8.43 \times 10^{-1}}$ (1.65×10^{-1})	$\mathbf{9.91 \times 10^{-1}}$ (2.05×10^{-3})	$\mathbf{2.09 \times 10^{1}}$ (5.45×10^{0})
Chen+	10.8×10^{-1} (1.88×10^{-1})	9.72×10^{-1} (4.58×10^{-3})	5.93×10^{1} (7.47×10^{0})
Lasso	23.4×10^{-1} (84.9×10^{-1})	9.82×10^{-1} (6.66×10^{-3})	4.25×10^{1} (13.9×10^{0})
Lasso + YJ	11.6×10^{-1} (3.84×10^{-1})	9.82×10^{-1} (6.66×10^{-3})	4.25×10^{1} (13.9×10^{0})

is hard to optimize due to troublesome non-convexity with complicated combinations of all parameters. Similarly, for Data 3 (skew data), Chen+ was worse than Lasso. In comparison, the proposed method can present stable behaviors because the objective function is convex for many variables $\boldsymbol{\beta} \in \mathbb{R}^P$, although it is non-convex for only two variables $\sigma, \alpha \in \mathbb{R}$, as shown in Theorem 1.

5.3 Real-World Medical Data

We applied the proposed method to the following two medical datasets: **PDGFR** (Platelet Derived Growth Factor Receptor) consists of $N = 79$ samples and $P = 320$ features, where the outcome is the ability to inhibit PDGFR phosphorylation (Guha and Jurs, 2004). **MTP** (MelTing Point) includes $N = 274$ samples and $P = 1142$ features, where the outcome is the melting point of drug-like compounds (Karthikeyan, Glen, and Bender, 2005). These datasets have many features compared to their sample size and may hold skewed noise, as described later. We compared the proposed method with "Chen+" (Chen, Pourahmadi, and Maadooliat, 2014), "Lasso" (Tibshirani, 1996), and "Lasso+YJ", as in the previous section.

We used 80% of the samples for training and the remaining 20% for testing. Then, we tuned each regularization coefficient λ with 5-fold cross-validation using 20% of the training samples (i.e., 16% of all samples) as validation data. These samples were generated randomly, and 30 trials were conducted with different random seeds. We calculated $\mathrm{MSE}(\hat{y})$, Sparsity of $\hat{\boldsymbol{\beta}}$, and Model $\mathrm{Size}(\hat{\boldsymbol{\beta}})$ and evaluated their performance.

Table 5.1 and Table 5.2 show the results of PDGFR and MTP, respectively, with the means and standard deviations (in parentheses) of the evaluation measures.

For both datasets, the proposed method simultaneously achieved the smallest prediction error and the smallest model size. In particular, for MTP, the proposed method reduced the model size to less than half when compared to the other methods. Regarding the residuals of the test data after training with the Lasso, the mean of the normalized skewness of the residuals was -0.81 for PDGFR and 0.75 for MTP.

In Section 5.4.3, we also applied the proposed method to more easily interpretable data, specifically the Engineering Graduate Salary (EGS) prediction data (Aggarwal, Srikant, and Nisar, 2016). EGS has fewer features than the two datasets above, and the meanings of all features are specifically provided. These properties allow us to consider the validity of the estimated active features. The result shows that the proposed method outperformed the comparative methods and could select more reasonable features.

5.4 Additional Experimental Details and Results

5.4.1 Definition of Sparsity

In this chapter, we used the measure for sparsity (denoted by "Sparsity"), as introduced by Hurley and Rickard (2009). In this section, we briefly review its definition.

Given a vector $\boldsymbol{\beta} \in \mathbb{R}^P$, we sort the coefficients from the smallest to largest, i.e., $\beta_{(1)} \leq \ldots \leq \beta_{(P)}$, where (p) is the p-th new index after sorting. Then, the "Sparsity" of $\boldsymbol{\beta}$ is defined by

$$\text{Sparsity}(\boldsymbol{\beta}) = 1 - \frac{2}{\|\boldsymbol{\beta}\|_1} \sum_{p=1}^{P} |\beta_{(p)}| \left(\frac{P - p + \frac{1}{2}}{P} \right). \tag{5.2}$$

This measure is based on a weighted sum of all coefficients, evaluating how important a particular coefficient is to an overall sparsity. We have $\text{Sparsity}(\boldsymbol{\beta}) = 0$ for the least sparse case where all coefficients have the same value, and $\text{Sparsity}(\boldsymbol{\beta}) = 1$ for the most sparse case where there exist few non-zero coefficients under the situation $P \to \infty$. Therefore, the measure can evaluate sparsity of $\boldsymbol{\beta}$.

5.4.2 Comparison with Other Methods

Adding to Section 5.1, we also compared the proposed model with a linear regression model without regularization, assuming Azzalini's skew-normal distribution, skew-t distribution, and skew-Cauchy distribution for noise (Arellano-Valle

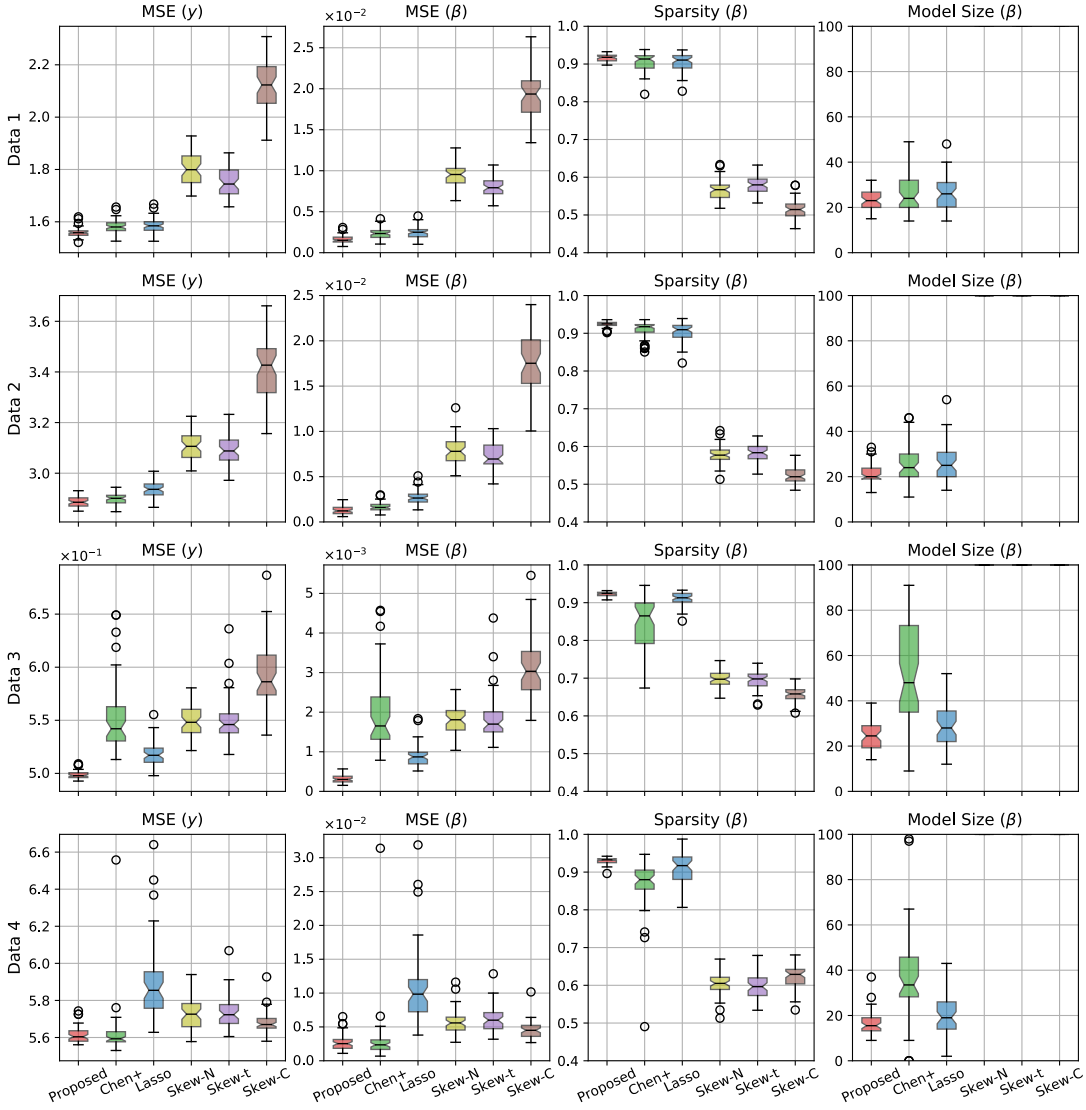

FIGURE 5.4: Comparison of six methods with four different types of simulation noises in terms of four evaluation measures. From top to bottom, each row results from Data 1, Data 2, Data 3, and Data 4, respectively. All experiments were conducted for 50 runs with different random seeds.

and Azzalini, 2013; Azzalini and Arellano-Valle, 2013; Azzalini and Salehi, 2020). These methods are denoted by "skew-N," "skew-t," and "skew-C," respectively. The experimental setting and data are the same as in Section 5.1.

Four evaluation measures, i.e., MSE($\hat{y}$), MSE($\hat{\boldsymbol{\beta}}$), Sparsity($\hat{\boldsymbol{\beta}}$), and Model Size($\hat{\boldsymbol{\beta}}$), are shown in Figure 5.4. The definitions of these measures are the same as in Section 5.1. As a result of Figure 5.4, the proposed method outperformed the comparative methods in all cases, indicating better prediction with a smaller number of features. Note that the model size for "skew-N," "skew-t," and "skew-C" is always 100 due to no regularization term, and the results of "Proposed," "Chen+," and "Lasso" are the same as in Figure 5.1.

TABLE 5.3: Results of EGS ($N = 3998$, $P = 25$).

	$\mathbf{MSE}(\hat{y})$	$\mathbf{Sparsity}(\hat{\beta})$	$\mathbf{Size}(\hat{\beta})$
Proposed	$\mathbf{2.11 \times 10^1}$ (3.14×10^1)	$\mathbf{4.62 \times 10^{-1}}$ (2.84×10^{-2})	$\mathbf{2.31 \times 10^1}$ (1.05×10^0)
Chen+	4.97×10^1 (7.59×10^1)	1.16×10^{-1} (16.8×10^{-2})	8.33×10^1 (12.0×10^0)
Lasso	4.89×10^1 (7.57×10^1)	3.78×10^{-1} (3.72×10^{-2})	2.44×10^1 (0.68×10^0)
Lasso + YJ	2.64×10^1 (3.86×10^1)	3.78×10^{-1} (3.72×10^{-2})	2.44×10^1 (0.68×10^0)

5.4.3 Real-World Financial Data

We applied the proposed method to the Engineering Graduate Salary (**EGS**) prediction data (Aggarwal, Srikant, and Nisar, 2016), which provides engineering graduates' employment outcomes with standardized assessment scores. EGS has fewer features than the two medical datasets in Section 5.3, and the meanings of all features are specifically given. These properties allow us to consider the validity of the estimated active features. For simplicity, we used only numerical type features (see Table 5.4 for details). The data used here consisted of $N = 3998$ and $P = 25$. Other experimental conditions were the same as in Section 5.3.

Table 5.3 shows the result of EGS with the means and standard deviations (in parentheses) of the evaluation measures. The proposed method simultaneously achieved the smallest prediction error and the smallest model size. Regarding the residuals of the test data after training with the Lasso, the mean of the normalized skewness of the residuals was 2.24.

We also obtained interesting results. Figure 5.5 shows box plots of the estimated coefficients for each method. The top 10 coefficients are picked out based on the averaged absolute values over 30 trials. For example, a graduate's quantitative ability ("Quant") ranked first for the proposed method, while a graduation year ("12graduation") ranked first for Lasso and Lasso+YJ. In the proposed method, "12graduation" ranked fourth with about half an effect. It may sound strange that the most relevant factor for salary is the graduation year. As the box plots indicate, the coefficient of this feature was sometimes zero. Moreover, an English score ("English") and a college tier ("CollegeTier") ranked second and third in the proposed method. From these points, the proposed method could select more reasonable features.

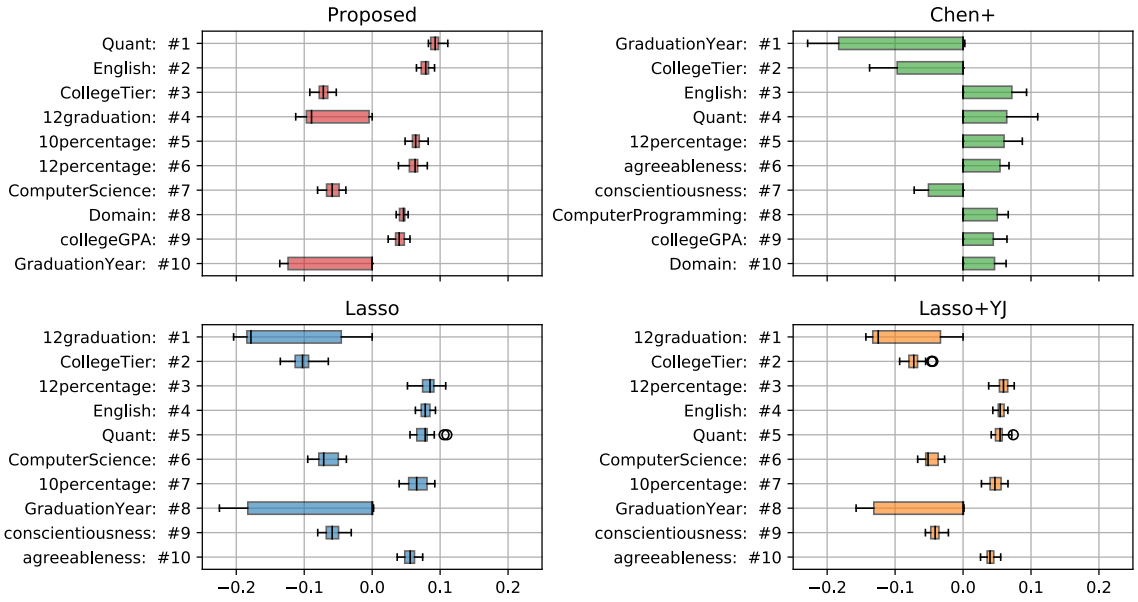

FIGURE 5.5: Box plots of the estimated coefficients for EGS. In the vertical axis, the top 10 coefficients are picked out based on the averaged absolute values over 30 runs with different random seeds. Details of each feature are described in Table 5.4.

TABLE 5.4: Description of features in EGS data (see Table 2 in Aggarwal, Srikant, and Nisar (2016) for details).

Feature	Description
10percentage	Overall marks obtained in grade 10 examinations
12graduation	Year of graduation (senior year high school)
12percentage	Overall marks obtained in grade 12 examinations
CollegeID	Unique ID identifying the college which the candidate attended
CollegeTier	Tier of college (computed from the average AMCAT scores)
CollegeGPA	Aggregate GPA at graduation
CollegeCityID	A unique ID to identify the city in which the college is located
CollegeCityTier	The tier of the city in which the college is located
GraduationYear	Year of graduation (Bachelor's degree)
English	Score in AMCAT's English section
Logical	Score in AMCAT's Logical ability section
Quant	Score in AMCAT's Quantitative ability section
Domain	Score in AMCAT's Domain module
ComputerProgramming	Score in AMCAT's Computer programming section
ElectronicsAndSemicon	Score in AMCAT's Electronics & Semiconductor Engineering section
ComputerScience	Score in AMCAT's Computer Science section
MechanicalEngg	Score in AMCAT's Mechanical Engineering section
ElectricalEngg	Score in AMCAT's Electrical Engineering section
TelecomEngg	Score in AMCAT's Telecommunication Engineering section
CivilEngg	Score in AMCAT's Civil Engineering section
conscientiousness	Score in one of the sections of AMCAT's personality test
agreeableness	Score in one of the sections of AMCAT's personality test
extraversion	Score in one of the sections of AMCAT's personality test
nueroticism	Score in one of the sections of AMCAT's personality test
openess_to_experience	Score in one of the sections of AMCAT's personality test

Chapter 6

Conclusion

In this thesis, we have proposed a novel sparse regression model whose noise is assumed to follow the mode-invariant skew-normal distribution. The proposed method always models a mode regardless of skewness and is highly adaptable and interpretable to skew noise. We have also shown that the proposed method is straightforward to implement and optimize, and it has theoretical guarantees for the average excess risk and the estimation error by non-asymptotic analysis. The results of numerical experiments demonstrated that the proposed method achieved both skewness modeling and statistical interpretability simultaneously, even for high-dimensional data, highlighting its significant effectiveness.

As a next challenge, we focus on extending to the *mode-invariant skew-t* distribution. To our knowledge, no study has explicitly dealt with this distribution besides its basic idea being mentioned in Fujisawa and Abe (2015). The first problem would be to prove the positive definiteness of the Fisher information matrix, whose proof could be complicated, such as in the proof of Proposition 6. Its optimization algorithm could also be problematic. However, these may be overcome in terms of t distribution being a scale mixture of normal distribution. It is also interesting to extend from univariate to multivariate noise for some machine learning algorithms. Fortunately, a multivariate mode-invariant skew-normal distribution has already been proposed by Abe and Fujisawa (2019), suggesting that we could quickly tackle and formulate models.

References

Abe, Toshihiro, and Hironori Fujisawa. "Multivariate skew distributions with mode-invariance through the transformation of scale". *Japanese Journal of Statistics and Data Science*, vol. 2, 2019, pp. 529–44.

Adcock, Christopher, Martin Eling, and Nicola Loperfido. "Skewed distributions in finance and actuarial science: a review". *The European Journal of Finance*, vol. 21, nos. 13–14, 2015, pp. 1253–81.

Aggarwal, Varun, Shashank Srikant, and Harsh Nisar. "AMEO 2015: A dataset comprising AMCAT test scores, biodata details and employment outcomes of job seekers". *Proceedings of the 3rd IKDD Conference on Data Science, 2016*. 2016, pp. 1–2.

Arellano-Valle, Reinaldo B., and Adelchi Azzalini. "On the unification of families of skew-normal distributions". *Scandinavian Journal of Statistics*, vol. 33, no. 3, 2006, pp. 561–74.

———. "The centred parameterization and related quantities of the skew-*t* distribution". *Journal of Multivariate Analysis*, vol. 113, 2013, pp. 73–90.

Arellano-Valle, Reinaldo B., and Marc G. Genton. "On fundamental skew distributions". *Journal of Multivariate Analysis*, vol. 96, no. 1, 2005, pp. 93–116.

Arellano-Valle, Reinaldo B., Héctor W. Gómez, and Fernando A. Quintana. "Statistical inference for a general class of asymmetric distributions". *Journal of Statistical Planning and Inference*, vol. 128, no. 2, 2005, pp. 427–43.

Azzalini, Adelchi. "A class of distributions which includes the normal ones". *Scandinavian Journal of Statistics*, vol. 12, no. 2, 1985, pp. 171–78.

———. *The skew-normal and related families*. Cambridge University Press, 2013.

———. "The skew-normal distribution and related multivariate families". *Scandinavian Journal of Statistics*, vol. 32, no. 2, 2005, pp. 159–88.

Azzalini, Adelchi, and Reinaldo B. Arellano-Valle. "Maximum penalized likelihood estimation for skew-normal and skew-*t* distributions". *Journal of Statistical Planning and Inference*, vol. 143, no. 2, 2013, pp. 419–33.

Azzalini, Adelchi, and Antonella Capitanio. "Distributions generated by perturbation of symmetry with emphasis on a multivariate skew *t*-distribution". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 65, no. 2, 2003, pp. 367–89.

———. "Statistical applications of the multivariate skew normal distribution". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, 1999, pp. 579–602.

Azzalini, Adelchi, and Marc G. Genton. "Robust likelihood methods based on the skew-*t* and related distributions". *International Statistical Review*, vol. 76, no. 1, 2008, pp. 106–29.

Azzalini, Adelchi, and Mahdi Salehi. "Some computational aspects of maximum likelihood estimation of the skew-*t* distribution". *Computational and Methodological Statistics and Biostatistics: Contemporary Essays in Advancement*, 2020, pp. 3–28.

Bickel, Peter J., Ya'acov Ritov, and Alexandre B. Tsybakov. "Simultaneous analysis of Lasso and Dantzig selector". *The Annals of Statistics*, vol. 37, no. 4, 2009, pp. 1705–32.

Bishop, Christopher M. *Pattern recognition and machine learning*. Springer, 2006.

Bourguignon, Marcelo, Jeremias Leão, and Diego I. Gallardo. "Parametric modal regression with varying precision". *Biometrical Journal*, vol. 62, no. 1, 2020, pp. 202–20.

Box, George E. P., and David R. Cox. "An analysis of transformations". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 26, no. 2, 1964, pp. 211–43.

Branco, Márcia D., and Dipak K. Dey. "A general class of multivariate skew-elliptical distributions". *Journal of Multivariate Analysis*, vol. 79, no. 1, 2001, pp. 99–113.

Bühlmann, Peter. "Statistical significance in high-dimensional linear models". *Bernoulli*, vol. 19, no. 4, 2013, pp. 1212–42.

Bühlmann, Peter, and Sara van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer, 2011.

Byrd, Richard H., Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. "A limited memory algorithm for bound constrained optimization". *SIAM Journal on scientific computing*, vol. 16, no. 5, 1995, pp. 1190–208.

Cao, Xingyun, Danlu Wang, and Liucang Wu. "Performance of ridge estimator in skew-normal mode regression model". *Communications in Statistics - Simulation and Computation*, vol. 52, no. 3, 2023, pp. 1164–77.

Chen, Lianfu, Mohsen Pourahmadi, and Mehdi Maadooliat. "Regularized multivariate regression models with skew-*t* error distributions". *Journal of Statistical Planning and Inference*, vol. 149, 2014, pp. 125–39.

Chen, Yen-Chi. "Modal regression using kernel density estimation: a review". *WIREs Computational Statistics*, vol. 10, no. 4, 2018, e1431.

Chen, Yen-Chi, Christopher R. Genovese, Ryan J. Tibshirani, and Larry Wasserman. "Nonparametric modal regression". *The Annals of Statistics*, vol. 44, no. 2, 2016, pp. 489–514.

Critchley, Frank, and M. C. Jones. "Asymmetry and gradient asymmetry functions: Density-based skewness and kurtosis". *Scandinavian Journal of Statistics*, vol. 35, no. 3, 2008, pp. 415–37.

Davino, Cristina, Marilena Furno, and Domenico Vistocco. *Quantile regression: theory and applications*. John Wiley & Sons, 2013.

Feng, Yunlong, Jun Fan, and Johan A. K. Suykens. "A statistical learning approach to modal regression". *Journal of Machine Learning Research*, vol. 21, no. 2, 2020, pp. 1–35.

Fernández, Carmen, and Mark F. J. Steel. "On Bayesian modeling of fat tails and skewness". *Journal of the American Statistical Association*, vol. 93, no. 441, 1998, pp. 359–71.

Ferreira, José T. A. S., and Mark F. J. Steel. "A constructive representation of univariate skewed distributions". *Journal of the American Statistical Association*, vol. 101, no. 474, 2006, pp. 823–29.

Freijeiro-González, Laura, Manuel Febrero-Bande, and Wenceslao González-Manteiga. "A critical review of LASSO and its derivatives for variable selection under dependence among covariates". *International Statistical Review*, vol. 90, no. 1, 2022, pp. 118–45.

Fujisawa, Hironori, and Toshihiro Abe. "A family of skew distributions with mode-invariance through transformation of scale". *Statistical Methodology*, vol. 25, 2015, pp. 89–98.

Geer, Sara van de, and Peter Bühlmann. "On the conditions used to prove oracle results for the Lasso". *Electronic Journal of Statistics*, vol. 3, 2009, pp. 1360–92.

Genton, Marc G. *Skew-elliptical distributions and their applications: a journey beyond normality*. CRC Press, 2004.

González-Farías, Graciela, J. Armando Dominguez-Molina, and Arjun Gupta. "The closed skew-normal distribution". *Skew-elliptical distributions and their applications: a journey beyond normality*, 2004, pp. 25–42.

Groeneveld, Richard A., and Glen Meeden. "Measuring skewness and kurtosis". *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 33, no. 4, 1984, pp. 391–99.

Guha, Rajarshi, and Peter C. Jurs. "Development of linear, ensemble, and nonlinear models for the prediction and interpretation of the biological activity of a set of PDGFR inhibitors". *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 6, 2004, pp. 2179–89.

Hansen, Bruce E. "Autoregressive conditional density estimation". *International Economic Review*, vol. 35, no. 3, 1994, pp. 705–30.

Hao, Songhua, Jun Yang, and Christophe Berenguer. "Degradation analysis based on an extended inverse Gaussian process model with skew-normal random effects and measurement errors". *Reliability Engineering & System Safety*, vol. 189, 2019, pp. 261–70.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009.

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

Hinkley, David V., and Nagesh S. Revankar. "Estimation of the Pareto law from underreported data: A further analysis". *Journal of Econometrics*, vol. 5, no. 1, 1977, pp. 1–11.

Hossain, Ahmed, and Joseph Beyene. "Application of skew-normal distribution for detecting differential expression to microRNA data". *Journal of Applied Statistics*, vol. 42, no. 3, 2015, pp. 477–91.

Huang, Qi, Hanze Zhang, Jiaqing Chen, and Mengying He. "Quantile regression models and their applications: A review". *Journal of Biometrics & Biostatistics*, vol. 8, no. 3, 2017, pp. 1–6.

Hurley, Niall, and Scott Rickard. "Comparing measures of sparsity". *IEEE Transactions on Information Theory*, vol. 55, no. 10, 2009, pp. 4723–41.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. Springer, 2013.

Jamshidian, Mortaza, and Robert I. Jennrich. "Acceleration of the EM algorithm by using quasi-Newton methods". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 59, no. 3, 1997, pp. 569–87.

Jones, M. C. "Generating distributions by transformation of scale". *Statistica Sinica*, vol. 24, 2014, pp. 749–71.

———. "On bivariate transformation of scale distributions". *Communications in Statistics - Theory and Methods*, vol. 45, no. 3, 2016, pp. 577–88.

———. "On families of distributions with shape parameters". *International Statistical Review*, vol. 83, no. 2, 2015, pp. 175–92.

Karthikeyan, Muthukumarasamy, Robert C. Glen, and Andreas Bender. "General melting point prediction based on a diverse compound data set and artificial neural networks". *Journal of Chemical Information and Modeling*, vol. 45, no. 3, 2005, pp. 581–90.

Kemp, Gordon C. R., and J. M. C. Santos Silva. "Regression towards the mode". *Journal of Econometrics*, vol. 170, no. 1, 2012, pp. 92–101.

Koenker, Roger. "Quantile regression: 40 years on". *Annual Review of Economics*, vol. 9, 2017, pp. 155–76.

Koenker, Roger, Victor Chernozhukov, Xuming He, and Limin Peng. *Handbook of quantile regression*. CRC Press, 2017.

Koyama, Kazuki, Takayuki Kawashima, and Hironori Fujisawa. "Sparse modal regression with mode-invariant skew noise". *Transactions on Machine Learning Research*, 2024. https://openreview.net/forum?id=63r6M1JkXm.

Krief, Jerome M. "Semi-linear mode regression". *The Econometrics Journal*, vol. 20, no. 2, 2017, pp. 149–67.

Kullback, Solomon, and Richard A. Leibler. "On information and sufficiency". *The Annals of Mathematical Statistics*, vol. 22, no. 1, 1951, pp. 79–86.

Lahiri, Soumendra N. "Necessary and sufficient conditions for variable selection consistency of the LASSO in high dimensions". *The Annals of Statistics*, vol. 49, no. 2, 2021, pp. 820–44.

Leeb, Hannes, and Benedikt M. Pötscher. "Model selection and inference: Facts and fiction". *Econometric Theory*, vol. 21, no. 1, 2005, pp. 21–59.

———. "The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations". *Econometric Theory*, vol. 19, no. 1, 2003, pp. 100–142.

Ley, Christophe, and Davy Paindaveine. "Multivariate skewing mechanisms: a unified perspective based on the transformation approach". *Statistics & Probability Letters*, vol. 80, nos. 23–24, 2010, pp. 1685–94.

Liseo, Brunero, and Nicola Loperfido. "A Bayesian interpretation of the multivariate skew-normal distribution". *Statistics & Probability Letters*, vol. 61, no. 4, 2003, pp. 395–401.

Liu, Dong C., and Jorge Nocedal. "On the limited memory BFGS method for large scale optimization". *Mathematical Programming*, vol. 45, no. 1, 1989, pp. 503–28.

Liu, Qingyang, and Xianzheng Huang. "Parametric modal regression with error in covariates". *Biometrical Journal*, vol. 66, no. 1, 2024, e2200348.

Mairal, Julien. "Optimization with first-order surrogate functions". *Proceedings of the 30th International Conference on Machine Learning*. PMLR, 2013, pp. 783–91.

Marchenko, Yulia V., and Marc G. Genton. "A Heckman selection-*t* model". *Journal of the American Statistical Association*, vol. 107, no. 497, 2012, pp. 304–17.

Meinshausen, Nicolai, and Bin Yu. "Lasso-type recovery of sparse representations for high-dimensional data". *The Annals of Statistics*, vol. 37, no. 1, 2009, pp. 246–70.

Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

Mudholkar, Govind S., and Alan D. Hutson. "The epsilon–skew–normal distribution for analyzing near-normal data". *Journal of Statistical Planning and Inference*, vol. 83, no. 2, 2000, pp. 291–309.

Ota, Hirofumi, Kengo Kato, and Satoshi Hara. "Quantile regression approach to conditional mode estimation". *Electronic Journal of Statistics*, vol. 13, no. 2, 2019, pp. 3120–60.

Sahu, Sujit K., Dipak K. Dey, and Márcia D. Branco. "A new class of multivariate skew distributions with applications to Bayesian regression models". *Canadian Journal of Statistics*, vol. 31, no. 2, 2003, pp. 129–50.

Schelldorfer, Jürg, Peter Bühlmann, and Sara van de Geer. "Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization". *Scandinavian Journal of Statistics*, vol. 38, no. 2, 2011, pp. 197–214.

Shafiei, Sobhan, Amir Safarpoor, Ahad Jamalizadeh, and Hamid R. Tizhoosh. "Class-agnostic weighted normalization of staining in histopathology images using a spatially constrained mixture model". *IEEE Transactions on Medical Imaging*, vol. 39, no. 11, 2020, pp. 3355–66.

Simola, Umberto, Xavier Dumusque, and Jessi Cisewski-Kehe. "Measuring precise radial velocities and cross-correlation function line-profile variations using a Skew Normal density". *Astronomy & Astrophysics*, vol. 622, 2019, A131.

Städler, Nicolas, Peter Bühlmann, and Sara van de Geer. "ℓ_1-penalization for mixture regression models". *Test*, vol. 19, 2010, pp. 209–56.

Sun, Ying, Prabhu Babu, and Daniel P. Palomar. "Majorization-minimization algorithms in signal processing, communications, and machine learning". *IEEE Transactions on Signal Processing*, vol. 65, no. 3, 2016, pp. 794–816.

Sun, Ying, and Fuming Lin. "Introduction and some recent advances in Lp quantile regression". *Journal of Applied Mathematics and Physics*, vol. 12, no. 11, 2024, pp. 3827–41.

Takada, Masaaki, Taiji Suzuki, and Hironori Fujisawa. "Independently interpretable Lasso: A new regularizer for sparse regression with uncorrelated variables". *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. PMLR, 2018, pp. 454–63.

———. "Independently interpretable Lasso for generalized linear models". *Neural Computation*, vol. 32, no. 6, 2020, pp. 1168–221.

Theodossiou, Panayiotis. "Financial data and the skewed generalized *t* distribution". *Management Science*, vol. 44, no. 12, 1998, pp. 1650–61.

Tibshirani, Robert. "Regression shrinkage and selection via the Lasso". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 58, no. 1, 1996, pp. 267–88.

Torossian, Léonard, Victor Picheny, Robert Faivre, and Aurélien Garivier. "A review on quantile regression for stochastic computer experiments". *Reliability Engineering & System Safety*, vol. 201, 2020, p. 106858.

Tseng, Paul. "Convergence of a block coordinate descent method for nondifferentiable minimization". *Journal of Optimization Theory and Applications*, vol. 109, 2001, pp. 475–94.

Van Rijsbergen, Cornelis Joost. *The geometry of information retrieval*. Cambridge University Press, 2004.

Wang, Jiuzhou, Joseph Boyer, and Marc G. Genton. "A skew-symmetric representation of multivariate distributions". *Statistica Sinica*, vol. 14, no. 4, 2004, pp. 1259–70.

Xiang, Sijia, and Weixin Yao. "Modal regression for skewed, truncated, or contaminated data with outliers". *Advances and Innovations in Statistics and Data Science*, ed. by Wenqing He, Liqun Wang, Jiahua Chen, and Chunfang Devon Lin, Springer, 2022, pp. 257–73.

Yao, Weixin, and Longhai Li. "A new regression model: modal linear regression". *Scandinavian Journal of Statistics*, vol. 41, no. 3, 2014, pp. 656–71.

Yao, Weixin, and Sijia Xiang. "Nonparametric and varying coefficient modal regression". *arXiv preprint arXiv:1602.06609*, 2016.

Yeo, In-Kwon, and Richard A. Johnson. "A new family of power transformations to improve normality or symmetry". *Biometrika*, vol. 87, no. 4, 2000, pp. 954–59.

Zhao, Peng, and Bin Yu. "On model selection consistency of Lasso". *Journal of Machine Learning Research*, vol. 7, no. 90, 2006, pp. 2541–63.

Zhou, Haiming, and Xianzheng Huang. "Bayesian beta regression for bounded responses with unknown supports". *Computational Statistics & Data Analysis*, vol. 167, 2022, p. 107345.

———. "Parametric mode regression for bounded responses". *Biometrical Journal*, vol. 62, no. 7, 2020, pp. 1791–809.

Zhu, Ciyou, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization". *ACM Transactions on Mathematical Software (TOMS)*, vol. 23, no. 4, 1997, pp. 550–60.